Computational and Structural Biotechnology Journal 19 (2021) 5008–5018

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

Machine learning and deep learning methods that use omics data for metastasis prediction

Somayah Albaradei [a,b,1], Maha Thafar [a,c], Asim Alsaedi [d,e], Christophe Van Neste [a], Takashi Gojobori [a,f], Magbubah Essack [a,*,1], Xin Gao [a,*]

[a] Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[b] King Abdulaziz University, Faculty of Computing and Information Technology, Jeddah, Saudi Arabia
[c] Taif University, Collage of Computers and Information Technology, Taif, Saudi Arabia
[d] King Saud bin Abdulaziz University for Health Sciences, Jeddah, Saudi Arabia
[e] King Abdulaziz Medical City, Jeddah, Saudi Arabia
[f] Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 January 2021
Received in revised form 16 August 2021
Accepted 2 September 2021
Available online 4 September 2021

Keywords:
Cancer
Metastasis
Machine learning
Deep learning
Artificial intelligence

ABSTRACT

Knowing metastasis is the primary cause of cancer-related deaths, incentivized research directed towards unraveling the complex cellular processes that drive the metastasis. Advancement in technology and specifically the advent of high-throughput sequencing provides knowledge of such processes. This knowledge led to the development of therapeutic and clinical applications, and is now being used to predict the onset of metastasis to improve diagnostics and disease therapies. In this regard, predicting metastasis onset has also been explored using artificial intelligence approaches that are machine learning, and more recently, deep learning-based. This review summarizes the different machine learning and deep learning-based metastasis prediction methods developed to date. We also detail the different types of molecular data used to build the models and the critical signatures derived from the different methods. We further highlight the challenges associated with using machine learning and deep learning methods, and provide suggestions to improve the predictive performance of such methods.

Contents

Abbreviations: Acc, Accuracy; AE, autoencoder; ANN, Artificial Neural Network; AUC, area under the curve; BC, Betweenness centrality; BH, Benjamini-Hochberg; BioGRID, Biological General Repository for Interaction Datasets; CCP, compound covariate predictor; CEA, Carcinoembryonic antigen; CNN, convolution neural networks; CV, cross-validation; DBN, deep belief network; DDBN, discriminative deep belief network; DEGs, differentially expressed genes; DIP, Database of Interacting Proteins; DNN, Deep neural network; DT, Decision Tree; EMT, epithelial-mesenchymal transition; GA, Genetic Algorithm; GANs, generative adversarial networks; GEO, Gene Expression Omnibus; HCC, hepatocellular carcinoma; HPRD, Human Protein Reference Database; FC, fully connected; k-CV, k-fold cross validation; KNN, K-nearest neighbor; LIMMA, linear models for microarray data; LOOCV, Leave-one-out cross-validation; LR, Logistic Regression; L-SVM, linear SVM; MCCV, Monte Carlo cross-validation; MLP, multilayer perceptron; mRMR, minimum redundancy maximum relevance; NPV, negative predictive value; PCA, Principal component analysis; PPI, protein-protein interaction; PPV, positive predictive value; RC, ridge classifier; RF, Random Forest; RFE, recursive feature elimination; RMA, robust multi-array average; RNN, recurrent neural networks; Se, sensitivity; SGD, stochastic gradient descent; SMOTE, synthetic minority over-sampling technique; Sp, specificity; SVM, Support Vector Machine; TCGA, The Cancer Genome Atlas.

* Corresponding authors.
E-mail addresses: magbubah.essack@kaust.edu.sa (M. Essack), xin.gao@kaust.edu.sa (X. Gao).
[1] Shared first author.

https://doi.org/10.1016/j.csbj.2021.09.001

1. Metastasis

It is well known that once cancer has metastasized, treatment becomes much more challenging [1]. The reason being, primary cancer and metastasis have different properties that make the latter highly aggressive [2], and the drugs chosen to target the primary tumors do not necessarily target the metastases (secondary tumors). Thus, research effort has been directed towards predicting metastasis and therapeutic approaches that prevent metastasis.

Metastasis occurs through a series of cellular events such as the tumor cells detaching from the primary site, and then these detached cells are captured and transported throughout the bloodstream via blood flow to invade and colonize at a secondary site eventually [2–5]. However, the primary tumor site consists of a heterogeneous and genetically diverse tumor population, and not all tumor cells are capable of metastasizing [6]. According to the "*seed and soil*" metastasis hypothesis, both the seed (cancer cell) as well as the soil (secondary site) determine site selection. Only those cells that develop all the favorable traits for the metastatic cascade metastasize successfully [2,7]. These favorable traits include developing resistance against anoikis (induction of apoptosis upon detachment) during epithelial-mesenchymal transition (EMT) [2,8], the migratory capability of the tumor cells increasing through the formation of invadopodia [7], the tumor cells exhibiting more epithelial type behavior to be able to settle at the secondary site [9,10]. The tumor cells produce factors that modify the tumor microenvironment at secondary sites, and the secondary site must have the corresponding receptors to facilitate metastasis. For example, cancer cells expressing CXCR4 metastasize to tissue expressing the sole ligand for CXCR4, chemokine CXCL12 [11,12]. The mutual coordination determines the success of metastasis [13,14].

Understanding these mechanisms led to several therapeutic approaches to inhibit metastasis [2], including blocking EMT and reversing anoikis-resistance using quinazoline-based anoikis inducers and PPARγ, TRKB, and SRC inhibitors [15]. The second approach to inhibit metastasis blocks cell motility via cadherin, integrins, selectin, and CD44 [16], using N-cadherin inhibitors, integrin antagonists, selectin inhibitors, and CD44 antagonists. The cell motility is also blocked when interfering with invadopodia formation using growth factor inhibitor. The third approach deals with the interference of the interaction between tumors, the seed, and the tumor microenvironments, the soil. Inhibitors of VEGF, FGF, PDGF, and EGFR signaling also target EMT and, thus, tumor invasiveness [17]. In this regard, tumor models have been used to show that inhibitors of SRC, TWIST1, and TKS5 are promising agents to inhibit EMT [18]. Other mechanisms include interference with inflammation, inhibition of integrin signaling, interfering with the hypoxia process, remodeling of extracellular matrix (ECM), and inhibition of cancer-associated fibroblasts signaling [2]. Ideally, targeting both seed and soil should produce the synergistic effect needed to prevent metastasis, but a cure is still not in sight. Also, these approaches may have been successful in mouse models or *in vitro* cell cultures, but none of them were, as yet, successfully used to treat metastases in the clinic setting.

Consequently, early detection of metastasis-prone tumors and residual tumors metastases are also being pursued to improve clinical decision-making concerning treatment plans. In this regard, identifying the molecular features uniquely associated with a metastasized tumor is being pursued through various technologies [19]. Researchers have also developed several computational methods to predict metastasis and the key features associated with it. These *in silico* methods were developed using diverse features. Some methods incorporated clinicopathological-based features [20–22], other methods incorporated image-based features [23–25] or text-based features [26], while the more recent methods incorporate omics-based data as features. This review summarizes the *in silico* metastasis-related prediction methods that incorporate omics data as features. We highlight the lessons learned from developing these methods and further discuss different challenges in this field and how we believe deep learning (DL) can build a metastasis prediction model with generalizability and high prediction accuracy (Acc).

2. Computational modeling used to predict metastasis

Metastasis prediction, like any other prediction task, involves multiple steps. Generally, the overall metastasis prediction workflow starts with 1/ defining the problem statement, i.e., the goal of classification, 2/ omics data type selection, i.e., mRNA, microRNA, DNA methylation, etc., 3/ data preprocessing, 4/ feature selection, and 5/ model building (Fig. 1). Some studies have also incorporated downstream analysis and annotation methods such as pathway and gene ontology enrichment analysis in the workflow, after model building. This review focuses on classifiers that can distinguish metastatic and non-metastatic samples, functioning as metastasis predictors.

2.1. Data types that provide features used in metastasis prediction models

Most metastasis prediction models try to solve a binary classification problem that classifies samples as metastatic or non-metastatic. The models' discriminative power depends on the features used, as features that add more to distinguishing between samples provide higher prediction accuracy [27–30]. However, these features that distinguish between samples are not always apparent. Therefore, feature selection plays an integral part in building classification models.

Some studies used conventional parameters such as histomorphology, immunohistochemistry-based features and others used

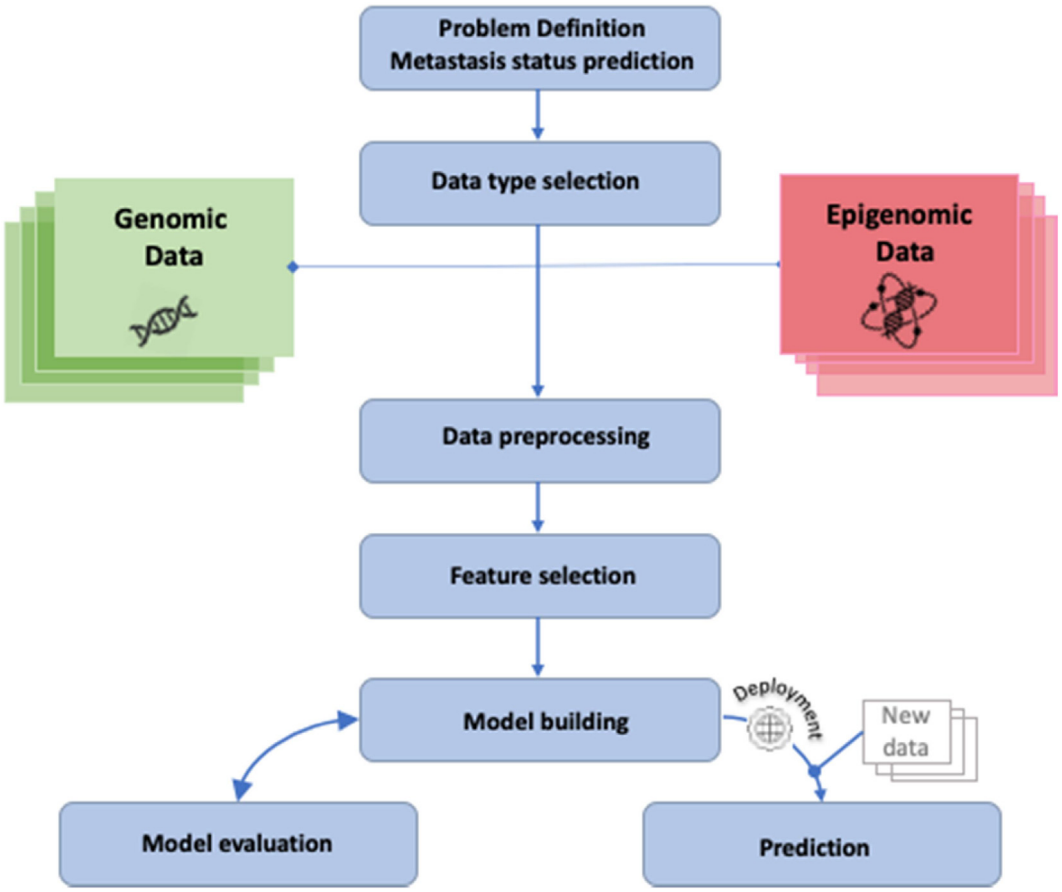

Fig. 1. General metastasis prediction workflow. First: problem definition, second: omics data type selection (e.g., mRNA, microRNA, DNA methylation, etc.), third: data preprocessing (quality measures, impute missing values, etc.), fourth: feature selection, fifth: model building by choosing machine learning (ML)/deep learning (DL) approaches, sixth: model evaluation by fine-tuning the model until it becomes satisfactory to use in implementation, and seventh: model deployment as a software program for clinical practice, or as a tool to be used in clinical research.

clinical, radiological parameters as features, but omics-based features (mRNA and microRNA expression data, and methylation information) appear to be more popular for inclusion in building such models. This predilection towards omics-based information is expected, as such high-throughput approaches can deliver a huge amount of data at tremendously reduced costs. Nevertheless, what may come as a surprise is that, thus far, most studies only used a single type of omics data to build the models except for a recent study published by Bhalla and colleagues in 2019 [28] that leveraged data from more than one omics type.

2.2. Data preprocessing and feature selection

Data processing is an essential step to improve the overall quality and performance of the model. It involves data cleaning, such as dealing with missing values or extreme observations such as erroneous or outliers. It also involves data transformation such as changing the data type (discretization), e.g., transferring categorical data into numerical data, or changing the range of the data value (normalization), e.g., quantile normalization or z-score normalization. There is no "one-size-fits-all" rule for selecting the normalization or transformation method. It depends on deep understanding and data exploration [31]. The data transformation step is usually followed by feature selection to reduce the dimensionality of the data and remove the uninformative features if there is a large number of input features in the dataset. Methods generally used for feature selection include filter methods, wrapper methods, or embedded methods [32]. In filter methods, features are selected independently from the prediction model. For Example, betweenness centrality (BC) has been commonly used in the metastasis prediction task [33–36] as a filter-based method to rank genes based on their importance in biological networks such as protein–protein interaction (PPI) networks. Then, an increment of N top-ranked genes is fed to the classifier until no increase in performance is acquired. Other statistical methods are also employed to filter out features that are not differential among classification groups. For example, Metri and colleagues [35] employed linear models for microarray data (LIMMA), a Bioconductor package to perform differential gene expression analysis, and selected only significantly (p-value <= 0.05) differentially expressed genes (fold change of ±2). Similarly, Wu and colleagues [37] performed the Mann-Whitney *U* test with Benjamini-Hochberg (BH) multiple-testing correction to select only differentially methylated probes for downstream analysis. Principal component analysis (PCA) and WEKA-FCBF were also used as filter-based feature selection methods [28]. The student's *t*-test has also been utilized to compute the mean value of each of the N features for each class and retained only those features wherein the difference between means is statistically significant [29].

In the wrapper methods, features are selected based on the resulting performance of the prediction model. For example, the underlying idea of the genetic algorithm (GA), used as a feature selection step [37], is to generate several random possible solutions, which represent different genes (features), and then combine the best solutions in an iterative process. There are also many wrapper methods designed for applications where the number of

Table 1
Machine learning (ML) models developed for metastasis prediction that incorporate omics data as features.

Problem Statement	Study	Features used	ML methods	Validation technique	Accuracy measures
We analyzed the expression profiles of HCC samples without or with intra-hepatic metastases	Ye et al. [41]	mRNA expression (Microarray)	CCP	LOOCV	Acc (0.85)
Metastasis prediction in breast cancer.	Burton et al. [40]	mRNA expression (Microarray)	RF, LR, SVM, ANN, Voting	Internal on the same microarray platform 10-fold CV	Acc (Voting: ~0.87)
				External test set from the same microarray platform	Acc (Voting: 0.73)
				External test set from different microarray platforms	Acc (RF: 0.61)
Prediction of osteosarcoma metastasis.	He et al. [33]	mRNA expression (Microarray)	SVM	Independent test sets	Average Acc (~0.98), Se (1.0), Sp (~0.97), PPV (~0.95), NPV (1.0), AUC (~0.98)
Predict cutaneous melanoma metastasis.	Wei et al. [34]	mRNA expression (Microarray and RNAseq)	SVM	Independent test sets	Average Acc (~0.97), Se (1.0), Sp (~0.94), PPV (~0.97), NPV (1.0), AUC (~0.97)
Prediction of metastatic melanoma.	Metri et al. [35]	mRNA expression (Microarray and RNAseq)	Adaboost and RF	CV	Acc (~0.87)
Predict breast cancer metastasis.	Tuo et al. [36]	mRNA expression (Microarray and RNAseq)	SVM	Independent test sets	Average Acc (~0.94)
Predict lymph node metastasis in endometrial cancer	Ahsen et al. [38]	microRNA expression (Microarray)	weighted SVM	Independent test set	Acc (0.85)
Prediction of brain metastasis in lung adenocarcinoma	Zhao et al. [39]	microRNA expression (Microarray)	RF	Independent test set	Acc (0.91)
Predict lymph node metastasis in stomach cancer	Wu et al.[37]	DNA methylation (Microarray)	RF	MCCV	Average AUC (~0.78)
Classify metastatic and primary tumors of skin cutaneous melanoma	Bhalla et al.[28]	mRNA expression (RNAseq)	DT, KNN, RF, LR, RC, SVM	Independent test set	SVM: Se (0.89), Sp (0.90), Acc (0.89), MCC (0.73), AUC (0.95).
		microRNA expression (RNAseq)			SVM: Se (0. 90), Sp (0.79), Acc (0.88), MCC (0.66), AUC (0.89).
		DNA methylation (Microarray)			LR: Se (0.78), Sp (0.71), Acc (0.77), MCC (0.44), AUC (0.85).
		Ensemble of three (mRNA, microRNA, DNA methylation)			SVM: Se (0.86), Sp (0.95), Acc (0.88), MCC (0.71), AUC (0.93)
Cross-cancer metastasis signature in the microRNA-mRNA.	Lee et al.[29]	mRNA expression (RNAseq)	LASSO, RF, SVM	MCCV	RF: AUC (0.74)
		microRNA expression (RNAseq)			RF: AUC (0.64)

Shortcuts in the fourth column are: Compound covariate predictor (CCP), random forest (RF), logistic regression (LR), support vector machine (SVM), artificial neural network (ANN), decision tree (DT), K-nearest neighbor (KNN), ridge classifier (RC), and linear model trained with L_1-norm regularization (LASSO). Shortcut in the fifth column are: Leave-one-out cross-validation (LOOCV), cross-validation (CV), Matthews correlation coefficient (MCC), and Monte Carlo cross-validation (MCCV).

samples is smaller than the number of measured features per sample, such as recursive feature elimination (RFE), support vector machines with L_1-norm regularization (SVM-L_1), and lone star [28,38].

In the embedded methods, the feature selection process is integrated as part of the learning algorithm and occurs during the prediction process (not as a preprocessing step). The qualities of both the filter and wrapper methods are combined in the embedded methods. For example, random forest (RF) capitalizes on its own variable selection process by simultaneously performing feature selection and prediction. The RF has been used by [39] to select the most discriminative microRNAs features and predict metastasis for lung adenocarcinoma.

2.3. In silico models developed for metastasis prediction

Most of the metastasis prediction models developed so far employ a binary classification approach, where the two classes included in the study are usually metastatic and non-metastatic samples. To solve this classification problem, these models were developed using different ML methods and various input features. Here, we summarize the studies that used omics-based features (Table 1) for such metastasis-related binary classification problems and also focused on a class of ML known as DL, that we anticipate will improve predictive performance. These studies were collated based on a Google scholar search with the query, (metasta*[Title/Abstract] AND (machine learning OR deep learning), restricted to 2000–2019 and we retrieved just over 250 articles, of which only 14 articles were relevant to our focus. Specifically, we focused on works that use omics data, we exclude works on images, clinicopathologic, radiomics, or any other data types. Since our focus is only on the work that predict whether the cancer is in its primary (non-metastatic) or metastatic state, we also exclude works that predict the origin of metastatic cancer or to predict the clinical metastasis time.

2.3.1. ML models that incorporate omics data as features

The first and most popular omics data type used in metastasis prediction models is mRNA expression data [33–36,40]. However, in the last three years, we have seen DNA methylation [37] and microRNA expression [38,39] data used as features in these metastasis prediction models.

A few research groups [38,39,41] carried out experiments to obtain the omics data. However, for most studies, the omics data were obtained from Gene Expression Omnibus (GEO) or The Cancer Genome Atlas (TCGA). When dealing with such datasets, the general workflow is to process the data through standard bioinformatics pipelines for data filtering, pre-processing, and normalization. Then, differential gene expression is determined using robust statistical methods; this allows for identifying critical genes associated with differentiating the metastatic and non-metastatic samples. The most differentially expressed genes can then be used by the models to differentiate the metastatic samples from the non-metastatic ones.

Since specific microRNAs as well as alterations in DNA methylation levels are recognized hallmarks of human cancers [42–44] both microRNA and DNA methylation profiles were also recently examined in the context of metastasis prediction. Now, similar to mRNAs, microRNAs, and DNA methylation are emerging as powerful predictors of metastatic cancer.

All the models use validation techniques to estimate the generalization accuracy of a model on unseen data. The validation techniques are either part of the CV (including MCCV and LOOCV) category or are independent (external) test sets. Both methods use a test set (i.e., data to validate the model's performance). On the one hand, CV is a technique that involves partitioning the original dataset into a training set used to train the model and a test set used to validate the analysis. It uses multiple train-test splits, and the estimation of accuracy is averaged over all splits to get the total effectiveness of the model. On the other hand, using an independent test set means testing a model on totally different data than it was trained on, to provide an unbiased learning performance estimate. Many works [29,35,37,41] use CV as a validation method. It allows the model to train on multiple train-test splits, thus indicating how well the model will perform on unseen data. However, reporting the model's performance on a completely independent test set is recommended to provide an unambiguous statistical validation [28,33,34,36,38–40]. In fact, some scientific journals provided new guidelines demanding external cohort validation [45–47]. Also, ML/DL models with higher generalizability are suitable for clinical setting use [45].

Additionally, because omics data typically present a high-dimensional imbalanced characteristic, performance metrics are fundamental in evaluating the quality of the learning methods and learned models. The literature defines many different evaluation metric categories. These are the threshold metrics that quantify the classification prediction error (e.g., Acc, PPV, NPV, Sp, Sn, MCC, and F-score), the ranking methods, and metrics that evaluate the classifiers based on how effective they are at separating classes (e.g., AUC). In a nutshell, classification accuracy is almost universally inappropriate for imbalanced classification. The reason is, a high accuracy (or low error) is achievable by a no-skill model that only predicts the majority class. For imbalanced classification, the F-score is a popular metric. Also, we can combine Sensitivity and Specificity into a single score that balances both concerns, called the geometric mean or G-Mean. For more detail, see [48].

2.3.2. ML models that incorporate mRNA data as features

Ye and colleagues [41], in 2003, were the first to use omics data, specifically mRNA expression data, for metastasis prediction using compound covariate predictor (CCP). They demonstrated that the metastasis predictor could distinguish metastasis-free primary hepatocellular carcinoma (HCC) from primary HCC that has associated metastatic lesions, but it could not distinguish between the primary HCC and its associated metastatic lesions because the gene expression signatures were too similar. Their results suggest that one of the lead genes in this signature, osteopontin, is a diagnostic marker and a potential therapeutic target for metastatic HCC. They proposed a robust signature model that correctly classified 85% of the validation sample set.

Since then, ML methods have been developed for metastasis prediction using specific protein(s), histomorphological, and clinicopathological data, but they did not include omics data [49,50]. Thus, these methods were biased in terms of the methods not being developed using complete expression profiles. Nonetheless, these studies point out characteristics to improve prediction performance, including 1/ increasing the data that can be used as features to distinguish between samples, 2/ establishing "rules" that define how the classifier must use the features, and 3/ considering various ML methods as they impact the prediction accuracy.

Thus, in 2012, Burton and colleagues [40] compared the performance of eight classification methods to predict metastasis in breast cancer using microarray gene expression data. Classifiers built include RF, logistic regression (LR), support vector machine (SVM) with different kernels, artificial neural network (ANN), and a voting-based classifier. The voting-based classifier combines the results of the seven classifiers into one voting decision. The voting-based model has the advantage of reducing the variance between the different classification models and significantly increased the model performance when validating the models on the test dataset from the same platform. However, the RF model, a voting-like model, proved to be the most robust model when validating a test dataset from the different platforms as shown in Table 1.

Since then, most classification models developed for metastasis prediction by many groups were SVM-based. These works employed microarray-based gene expression data from GEO or TCGA to predict metastasis of different cancers such as osteosarcoma [33], cutaneous melanoma [34], and breast cancer [36]. They employed the SVM-based classifier to classify the samples as metastatic or non-metastatic and trained and validated the model's performance on independent datasets. They processed the gene expression data using the standard procedure of missing value treatment, background correction, and normalization. They first used statistical methods to identify the differentially expressed genes (DEGs). Next, they constructed PPI networks using some of the available databases such as the Human Protein Reference Database (HPRD), the Database of Interacting Proteins (DIP) or the Biological General Repository for Interaction Datasets (BioGRID). They used the networks to calculate BC, which they used to rank genes because high BC indicates the gene is an essential intermediary in the regulatory network. The top-ranking genes were included in the SVM classifier as features. They reported the overall average Acc of these models on test datasets (shown in Table 1).

Several top-ranking genes in the metastatic osteosarcoma samples [33] were associated with apoptosis regulation, cell proliferation, actin cytoskeleton processes, and cancer pathways such as the TGF-β signaling pathway. Several cancer-related pathways were also significantly over-represented by the top-ranking genes in the cutaneous melanoma study [34], including focal adhesion, regulation of actin cytoskeleton, apoptosis, and cell proliferation. The study results from [36] suggested that CDK2, CDKN1A, E2F1, and MYC are potential feature genes in metastatic breast cancer. They overcame the limitation of data-driven algorithms, which do not consider any biological information of the component genes as input, by implementing a condition-specific biological network approach to select genes that distinguish metastatic melanoma from primary melanoma.

During the same timeframe when the SVM-based approaches were developed, Metri and colleagues [35] also developed a method to differentiate metastatic melanoma from primary melanoma but instead used Adaboost around RF. Their dataset consisted of a training set and two independent test sets. The study results suggest six genes (KRT16, KIT, ALDH1A1, SPRR3, HSP90AB1, and TMEM45B) that function in metastatic melanoma processes, such as phospholipid metabolic process, protein-lipid complex assembly, regulation of inflammatory response, negative regulation of protein kinase activity, and regulation of the innate immune response. Although Metri and colleagues [35] considered more biological information in their feature selection step and showed promising results using Adaboost around RF, SVM showed superior results when applied to various feature selection methods. This shows that SVM is less sensitive to input parameter choices and reflects the ability of SVM-based approaches to distinguish between classes in complex datasets.

In the Supplementary Material we further provide the sample sizes used in all the studies, as well as the total number of features used in each study and the number of features used by each ML model to distinguish between samples. For these ML models that use mRNA data as features (ranging from 62,977 to 13,733 features), the number of features used in the model is ∼1.66% to 0.04% of the total number of features used to develop the models.

2.3.3. ML models that incorporate microRNA and DNA methylation multi-omics data as features

Similar to mRNAs expression profile, more and more expression profiles for microRNA and DNA methylation are also becoming available [51,52]. Thus, a few studies have also used microRNA and DNA methylation data for metastasis prediction in the last five years (Table1). Microarray-based microRNA expression profiles were used to identify crucial metastasis-related microRNAs to predict metastatic lymph node risk in endometrial cancer [38] and identify brain metastasis-related microRNAs in lung adenocarcinoma patients [39].

Ahsen and colleagues [38] designed a unique feature selection algorithm called lone star (specifically developed to identify a small number of discriminative features when the number of features is more than the number of samples) [38]. The application of the lone star algorithm resulted in a set of 18 discriminative microRNAs. Using these 18 microRNAs as input to a weighted SVM classifier achieved high Acc for independent test datasets (Acc 0.85; see Table 1). Functional annotation of these microRNAs revealed about 23 cancer-associated genes targeted by these microRNAs.

Zhao and colleagues [39] used the standard statistical feature selection algorithms in the feature selection step. They identified eight significantly differentially expressed microRNAs between BM+ (patients with brain metastasis) and BM- (patients with non-brain metastasis) groups. An RF supervised classification algorithm was then employed for building a predictive model that pinpointed three relevant microRNAs (miR-210, miR-214, and miR-15a) as powerful predictors of brain metastasis. This classifier achieved an overall Acc of 0.91. Functional annotation revealed that these three microRNAs target about 2914 protein-coding genes. Gene ontology and pathway enrichment study of these protein-coding genes revealed that these microRNAs target genes mostly associated with metabolic and mitotic cell cycle processes, and are mainly involved in cancer metastasis-related signaling pathways.

To use methylation-based molecular markers to predict lymph node metastasis in stomach cancer, Wu and colleagues [37] proposed an RF classifier using methylation data from TCGA. They performed three preprocessing steps. First, they carried out differential methylation analysis to extract significantly differentiating probes between metastatic and non-metastatic samples. Next, the feature selection technique, minimum redundancy maximum relevance (mRMR), was applied to remove redundant features. In the final step, they implemented a genetic algorithm-based method to extract the most relevant probes. A total of 12 methylation probes were finally used as features and fed to an RF model to classify lymph node metastasis in stomach cancer. These probes are known to be associated with lymph node metastasis-related genes such as HOXD1, NMT1, and SEMA3E. The data was randomly split 100 times for training and testing to build the model. The model achieved an average area under the curve (AUC) of ∼0.78.

Thus far, we have described several studies that use only a single type of omics data but, a few recent studies also showed the use of more than one omics (multi-omics) data as input features. Bhalla and colleagues [28] used multi-omics data (i.e., mRNA expression, microRNA expression, and methylation data from TCGA) with ML to classify metastatic and primary skin cutaneous melanoma tumors. They compared performance using different feature selection techniques and different ML methods (decision tree (DT), K-nearest neighbor (KNN), RF, LR, ridge classifier (RC), SVM) for individual omics data and multi-omics data. The best performing model using mRNA data was SVM, with 17 features selected using SVM-L_1. The best performing model using microRNA data was SVM, with 32 features selected using WEKA-FCBF. The best performing model using DNA methylation data was LR, with 38 features selected using WEKA-FCBF. They also developed an ensemble learning model wherein mRNA, microRNA, and DNA methylation were the input features for an SVM. Although the ensemble model achieved an Acc comparable with using mRNA data only, it notably achieved the highest specificity (Sp of 0.95) among all models (see Table 1). This study identified *CASP7, S100A7, C7, KRT14, MMP3, LOC642587,* hsa-mir-203b, and hsa-mir-205 as the genes and microRNAs that are potential critical genomic features contributing to the oncogenesis of melanoma. They further suggested *CDK14, ESM1, NFATC3, ZNF827*, *C7orf4,* and *ZSWIM7* as novel putative markers for SKCM metastasis.

In another study, Lee and colleagues [29] performed a cross-cancer integrated analysis using mRNA and microRNA expression data and developed classifiers to differentiate metastatic samples and primary tumor samples across 11 cancer types. Top 64, 128, and 256 features were selected using the student's *t*-test. Three models were then trained (a linear model trained with L_1-norm regularization (LASSO), RF, and SVM) and evaluated across 100 Monte Carlo cross-validation (MCCV). Using only the mRNA expression data, the RF model with 256 features achieved the highest AUC when predicting metastasis (AUC 0.74). Similarly, using only the microRNA expression, the RF model with 256 features achieved the highest AUC when predicting metastasis (AUC 0.64). Then, they integrated the mRNA and microRNA data to model how well the metastasis-associated microRNA can predict the metastasis-associated mRNA. Several microRNAs previously linked to cancer metastasis or progression were identified as critical genomic features, including miR-301b, miR-423, and miR-1296. However, they did not develop or provide an ensemble model using both the mRNA and microRNA expression data as features to show how this alternative approach may impact the AUC.

In summary, the choice of the ML method varied from study to study. More comparable studies are needed, i.e., studies that use more than one classifier and provide multiple accuracy measures, but for now, it seems SVM is the better performing classifier, followed by RF. The current omics-based methods implementing a system biology's approach for predicting metastasis is a critical approach since microRNA and DNA methylation levels affect gene expression, and therefore should be considered when developing a model. However, such analysis requires a considerable amount of

data, and feature selection becomes a complicated process that hampers model generalizability.

3. Exploring DL-based metastasis prediction models

Improving the models in terms of predictive performance and model generalizability requires a different approach. DL could be this approach as it directly captures the nonlinear and complex relationships of high dimensional or noisy biological data [53–56], i.e., DL incorporates an automatic feature selection process. Moreover, the DL approaches applied to different cancer-related prediction tasks performed better or on par with other ML approaches that require a feature selection step [57]. Nonetheless, DL approaches developed to improve the metastasis prediction task are still few up-to-date.

3.1. DL architecture, training, and models

DL is a class of ML that consists of an input layer, multiple hidden (processing) layers, and an output layer. DL is a representation-learning method that progressively learns multiple input data [58]. DL is composed of nonlinear modules that transform the previous level's representation into a higher and more abstract representation level. With multiple nonlinear layers, the DL model can learn complex functions that map the input to the desired output. One compelling advantage of DL methods is that the levels of representation (features) are learned from the raw data using a general-purpose learning procedure and are not subjected to the handcrafted feature engineering limitations.

Training a DL network is an optimization problem [58]. The objective is to minimize the difference between the predicted output values and the real or actual values by minimizing the error defined by an appropriate loss function. Briefly, an input is forward propagated until it reaches the output layer, then the error is determined by comparing the predicted and actual output values using the loss function. The DL network is trained by calculating the gradient of the loss function and then using a backpropagation algorithm to propagate the error backward, from the output layer to the input layer. This allows the gradients for the weights to be computed and then adjusted using gradient methods, such as stochastic gradient descent (SGD) or other variants such as the Adam algorithm. The weights being updated after each iteration allows the desired learning to be achieved.

There are three commonly used DL models: deep neural network (DNN), convolution neural networks (CNN), and recurrent neural networks (RNN). The DNN represents a fully connected neural network such as multilayer perceptron (MLP), autoencoder (AE), and deep belief network (DBN). The architecture of MLP (Fig. 2A) consists of input, multiple hidden layers, and output layers. Using the backpropagation technique, MLP continuously adjusts the weights between two neurons to correctly establish a network between the output and input layers [59].

The architecture of AE (Fig. 2B) is composed of an encoder layer, bottleneck (code), and decoder layers [60]. Encoder layers capture the essential features from the input and produce the coding layer. On the other hand, the decoder layers use the coding layer to reconstruct the input data. The most discriminative features automatically learned from the raw data are in the coding layer. Another variant of AE called denoising autoencoder (DAE) corrupts the input by adding some noise to force the hidden layers to learn more generalized, meaningful, and essential features and prevent them from merely learning the identity function.

DBN (Fig. 2C) is a generative model consisting of stacks of restricted Boltzmann machines (RBM). An RBM is composed of a visible layer and a hidden layer with undirected connections between them. RBM is also a generative model that establishes the correct relationship between the visible and hidden layers to efficiently extract the essential features from the original data [61]. Thus, DBN is an unsupervised approach that learns to probabilistically reconstruct the inputs, while the hidden layer acts as feature detectors [62].

CNN (Fig. 2D), in general, consists of multiple convolutions (CONV) and pooling (POOL) followed by fully connected (FC) layers [63]. The filters used in the CONV layers create feature maps that indicate the detected features' locations and strength in the input. Those filters are automatically learned from the training dataset to distinguish between the output classes accurately. Local conjunctions of features from the previous layer are also detected during the CONV layers. The CONV layer is usually followed by a POOL layer that merges semantically similar features into one, thereby reducing the dimensionality of representations and creating invari-

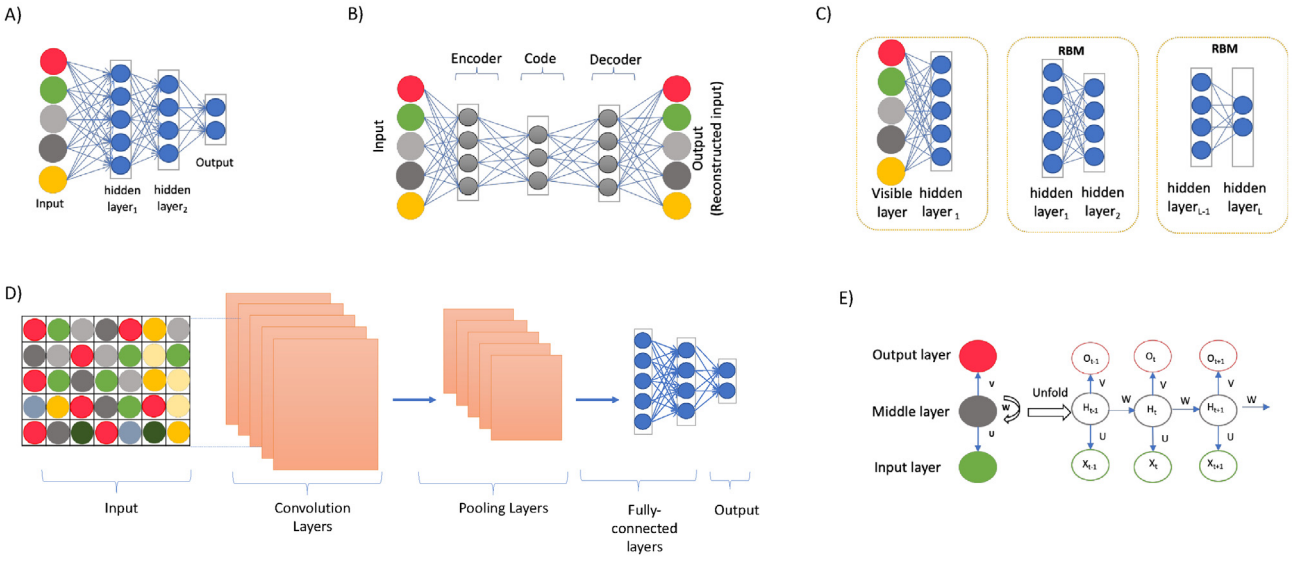

Fig. 2. Commonly used DL architectures. A) A multilayer perceptron (MLP) structure that contains multi hidden layers. B) A typical autoencoder (AE) structure. C) The basic structure of the deep belief network (DBN) network (stack of restricted Boltzmann machines (RBM)). D) A simple convolution neural networks (CNN) model structure. (E) The structure of recurrent neural networks (RNN) and the structure after unfolding by time.

Table 2
Studies that have included deep learning (DL) in several forms in their metastasis prediction workflows.

Problem Statement	Study	Features used	DL approach	Accuracy	ML approach	Accuracy
Identify whether there are lymph node metastases in colorectal cancer patients	Karabulut & Ibrikci [66]	mRNA expression	DDBN	0.91	SVM RF KNN	0.83 0.88 0.79
Predict the occurrence of metastatic events in breast cancer patients	Chereda et al. [67]	mRNA expression	GraphCNN	0.76	RF KNN	0.88 0.79
Predict if the cancer is in its metastatic state in colorectal cancer patients	Albaradei et al. [68]	DNA methylation	CNN	0.96	–	–

Abbreviations: Discriminative deep belief network (DDBN), graph convolution neural networks (graph CNN) architecture and convolution neural networks (CNN) architectures.

ance to distortions in the input data. The FC layers are the last layers in CNN that perform the final output. Recently, CNN has been generalized to work on arbitrarily structured graphs as many relevant real-world datasets come in the form of graphs or networks. Graph CNN is a promising extension of CNN in non-Euclidean domains such as graphs [64].

The architecture of RNN (Fig. 2E) allows it to recognize patterns in sequential data. RNN has loops that allow information to be carried across nodes while reading a sequential input. The critical element is the hidden state "memory", which captures information calculated in the previous steps [65].

3.2. DL-based metastasis prediction models

Recently, Karabulut *et al.* [66], Chereda *et al.* [67], and Albaradei *et al.* [68] proposed DL models to predict the metastasis using omics data (see Table 2, Fig. 3).

In 2017, Karabulut and colleagues [66] proposed a discriminative deep belief network (DDBN) to demonstrate the DL approach's ability to produce a powerful decision support model using gene expression data. This study included gene expression profiles of patient samples with recurrent and non-recurrent laryngeal cancer, bladder urothelial cells with and without cancer, as well as colorectal cancer with and without lymph node metastasis from the BioGPS portal [69]. In light of this review, we will only focus on the colorectal cancer dataset. The proposed DDBN is a DBN with a discriminative fine-tuning step at the end layer to provide a supervised approach. The fine-tuning procedure implemented using a backpropagation algorithm. They implemented two preprocessing steps 1/ selecting important features using information gain technique, and 2/ oversampling the minority class using synthetic minority over-sampling technique (SMOTE). With 10-folds CV, the averaged Acc for predicting the metastasis of the colorectal cancer samples show that the DL method DDBN outperform several ML methods such as SVM, RF, and KNN (Table 2).

Subsequently, Chereda and colleagues [67] developed a DL model that applied the graph CNN technique by exploiting the PPI graph as prior knowledge for predicting breast cancer's metastasis. The study included gene expression data from 10 GEO microarray datasets. They preprocessed each dataset using the robust multi-array average (RMA) probe summary algorithm[70], and then, combined the datasets based on the HG-UI33A array probes and applied quantile normalization. HPRD was used to construct the PPI network, and gene expression values were mapped

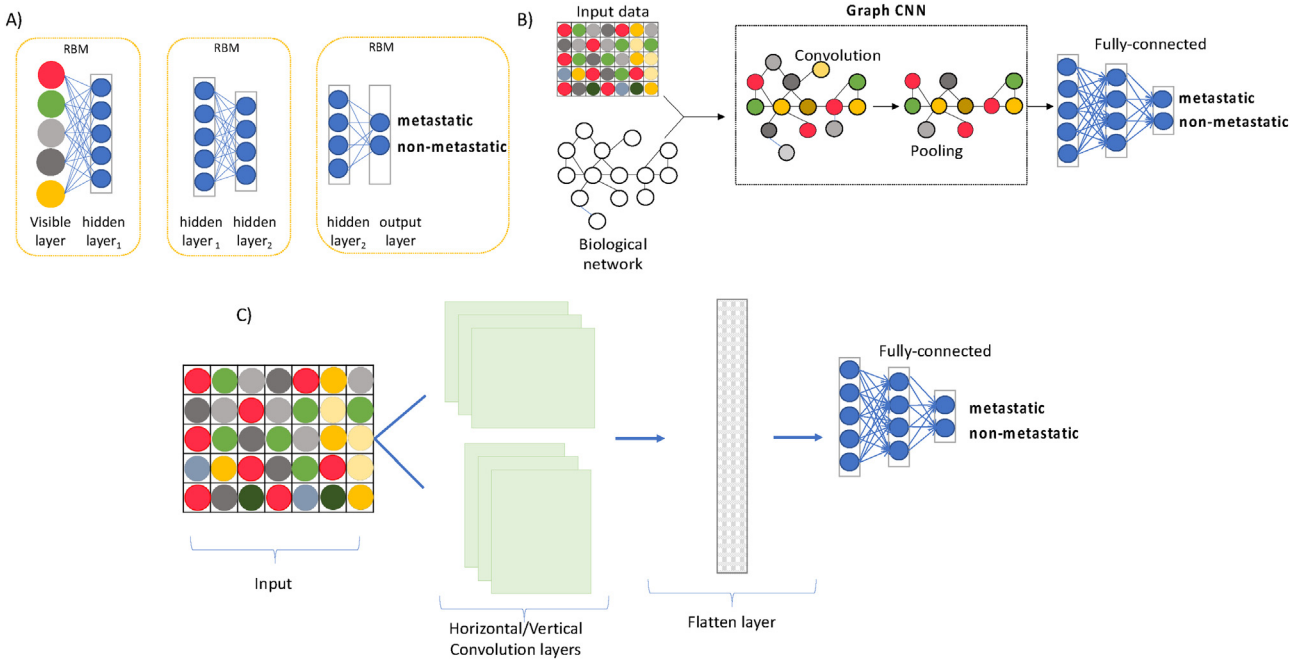

Fig. 3. The deep learning architectures used to predict metastasis. A) Discriminative deep belief network (DDBN) architecture used by (Karabulut & Ibrikci, 2017 [66]). B) Graph convolution neural networks (graph CNN) architecture used by (Chereda et al., 2019 [67]). C) Convolution neural networks (CNN) architecture used by (Albaradei et al., 2019 [68]).

to the vertices in the network, where edges reflect interactions, regulation, and signal flow. The resulting network had 207 connected components from which only the most significant components with 6888 vertices were used in the graph CNN. The graph CNN consists of two CONV layers with 32 filters, one POOL layer, and two FC layers. To show the strength of this proposed DL model, they compared its performance with several ML models. Using 10-fold CV, the Acc suggested the DL model outperform several ML models such as lasso LR, and RF (Table 2). Also, to show the strength of the graph CNN model, among other DL models, they further compared it with MLP, which showed less Acc (Acc of 0.74 (±1.84)).

In 2019 Albaradei and colleagues [68] also proposed a DL model called Deep2Met to predict metastatic colorectal cancer but used DNA methylation data. The study included 90 metastatic and 211 non-metastatic colorectal cancer DNA methylation profiles from TCGA. As a preprocessing step, probes with null values were eliminated, and the remaining beta values were represented as a 2D matrix. The proposed Deep2Met model consists of six parallel CONV layers with horizontal and vertical filters, followed by the ReLU layer, then a FC layer followed by a softmax layer to classify cases into metastatic or non-metastatic categories. Deep2Met model achieved AUC and an average F-scores of 0.97 and 0.95, respectively.

These initial studies demonstrate that DL metastasis prediction models usually outperform traditional ML models. However, none of the current studies have attempted to use DL with high dimensional multi-omics data as yet, the outcome of which would be interesting to know.

4. Conclusion: limitations and future directions

In this review, we summarized work that applied ML with omics data as features to predict metastasis. Most of the studies used gene expression omics data, but the few studies that use microRNA and DNA methylation data showed high prediction accuracy. These results suggest these data types have features essential to the metastasis process and, consequently, the prediction task. However, there is a lack of consistency in the discovered biomarkers, such as in the three ML-based breast cancer studies [36,40], which may be a consequence of the differences in profiling technologies, data processing steps, as well as genetic variation associated with specific patient populations. Thus, extracting the relevant features is a critical but difficult task that needs reproducibility [71,72]. Fortunately, modern DL cracks the code for training stability, generalization, and scale on big data. Consequently, DL is expected to better deal with high dimensional multi-omics data, and it can extract features directly from the raw data [73–75]. Thus, recent studies have explored using DL methods to overcome or limit ML-related challenges.

Initial studies demonstrate DL metastasis prediction models usually outperform traditional ML models. They used different DL architectures to build the metastasis prediction models, such as simple DBN and CNNs. However, none have tested the use of AE [60]. AE is a powerful DL technique that learns how to compress and encode data efficiently then learns how to reconstruct the data back from the reduced encoded representation. AE reduces data dimensions by ignoring the noise as well as extracting useful and essential biological features. AE has been successfully used to analyze high-dimensional gene expression data [76,77], and to integrate heterogeneous data [78,79]. Also, it produced promising results when identifying multi-omics features linked to the differential survival of HCC patients [80]. Besides AE, exploring other DL methods such as DNN and RNN [59] may also yield promising results. Furthermore, understanding the various models' advantages should facilitate combining them to build more powerful models [74].

Even with the advantage of DL-based metastasis predictions, it faces several challenges. For example, DL needs more data than traditional ML models to avoid overfitting and propose a more generic prediction model. However, using some new techniques such as zero-shot learning [81] and few-shot learning [82] limits this obstacle to some extent. Other options include using data augmentation techniques to increase data volume/size [83] and generative adversarial networks (GANs) [84] that generate artificial data based on the original dataset [85]. Also, in terms of the imbalance in available omics data, methods such as resampling and cost-sensitive learning [86] can be used. Nonetheless, more extensive and more diverse training sets are generally necessary to obtain models that can generalize well to a broader range.

Another critical challenge is the complexity of metastasis events itself, as several interrelated biological events drive metastasis, and many factors are involved. That is, omics features such as mRNA expression, microRNA expression, and DNA methylation influence each other; they are not independent variables. Thus, when these features are considered independent features when building classifiers, crucial interdependence information that should increase prediction accuracy might be lost. Therefore, it is challenging but essential to encode the interactome behavior while building the model. Another challenge is working with instances where the number of features is extensively larger than the available number of samples for model training and validation. Nonetheless, DL can handle a large number of features; thus, we believe that the application of DL-based approaches in combination with multi-omics data is the future direction in this field. Lastly, since there are few datasets available in the public domain, especially metastasis-related ones, more experiments need to be conducted across different cancer types to create a pan-cancer metastasis database, which can support metastasis prediction research.

To make the DL model practically useful, one has to show the resulting overall model's biology-related meaning. A simple way to interpret a CNN model is to visualize the filters as they might give a sense of what local features the DL has detected. Another way to determine feature importance in a DL model is to perturb inputs and observe its impact on the DL output. For example, given the gene expression profile of a sample, some genes can be systematically varied while the rest of the genes remain fixed, then the changes in the output of DL can be monitored. Several perturbation-based methods have been proposed to make DL models more interpretable, such as saliency maps [87]. Despite the simplicity of these methods, they are computationally expensive, as, for each perturbed input, a separate forward propagation through the DL is required to compute the output. Thus, a backpropagation-based method has been proposed to interpret DL models efficiently. In such a method, a backward propagation from the output layer is performed using gradients to calculate the feature's importance in the input layer [88]. Although DL models are still far from being clinically applicable, they are promising artificial intelligence tools that could support precision treatments in clinics.

The safe and timely translation of AI research into clinically validated and appropriately regulated systems can benefit everyone [89]. But rigorous research and tackling challenges and biases inherent in ML/DL models are required before we can realize the substantial impact of these models' unique contributions to clinical decision-making. Also, we need uniformity in guidelines [89] for required randomized control trials evaluating the performance of new ML/DL models in real-life settings. Nevertheless, working

through these challenges, we will overcome existing barriers, and ML/DL models may eventually meet their expectations to integrate into clinical decision-making and transform the data-driven evolution of precision medicine.

Funding

The research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST) through the Awards Nos. BAS/1/1059-01-01, BAS/1/1624-01-01, FCC/1/1976-20-01, and FCC/1/1976-26-01.

CRediT authorship contribution statement

Somayah Albaradei: Conceptualization, Writing - original draft, Writing - review & editing. **Maha Thafar:** Writing - original draft. **Asim Alsaedi:** Conceptualization, Writing - original draft. **Christophe Van Neste:** Data curation. **Takashi Gojobori:** Writing - review & editing. **Magbubah Essack:** Conceptualization, Writing - original draft, Writing - review & editing, Data curation. **Xin Gao:** Conceptualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.09.001.

References

[1] Dillekås H, Rogers MS, Straume O. Are 90% of deaths from cancer caused by metastases? Cancer Med 2019;8:5574–6.
[2] Guan X. Cancer metastases: challenges and opportunities. Acta Pharmac Sinica B 2015;5:402–18.
[3] Krakhmal NV, Zavyalova MV, Denisov EV, et al. Cancer invasion: patterns and mechanisms. Acta Naturae 2015;7:17–28.
[4] Seyfried T. Cancer as a metabolic disease: on the origin, management, and prevention of cancer. John Wiley & Sons; 2012.
[5] Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell 2011;147:275–92.
[6] Eccles SA, Welch DR. Metastasis: recent discoveries and novel treatment strategies. Lancet 2007;369:1742–57.
[7] Huang B, Jolly MK, Lu M, et al. Modeling the transitions between collective and solitary migration phenotypes in cancer metastasis. Sci Rep 2015;5:17379.
[8] Yeung KT, Yang J. Epithelial-mesenchymal transition in tumor metastasis. Mol Oncol 2017;11:28–39.
[9] Banyard J, Bielenberg DR. The role of EMT and MET in cancer dissemination. Connect Tissue Res 2015;56:403–13.
[10] Yao D, Dai C, Peng S. Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation. Mol Cancer Res 2011;9:1608–20.
[11] Kang Y, Siegel PM, Shu W, et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell 2003;3:537–49.
[12] Wendel C, Hemping-Bovenkerk A, Krasnyanska J, et al. CXCR4/CXCL12 participate in extravasation of metastasizing breast cancer cells within the liver in a rat model. PLoS ONE 2012;7:e30046.
[13] Fidler IJ, Kripke ML. The challenge of targeting metastasis. Cancer Metastasis Rev 2015;34:635–41.
[14] Steeg PS. Targeting metastasis. Nat Rev Cancer 2016;16:201–18.
[15] Sakamoto S, Kyprianou N. Targeting anoikis resistance in prostate cancer metastasis. Mol Aspects Med 2010;31:205–14.
[16] Wells A, Grahovac J, Wheeler S, et al. Targeting tumor cell motility as a strategy against invasion and metastasis. Trends Pharmacol Sci 2013;34:283–9.
[17] Fang H, Declerck YA. Targeting the tumor microenvironment: from understanding pathways to effective clinical trials. Cancer Res 2013;73:4965–77.
[18] Paz H, Pathak N, Yang J. Invading one step at a time: the role of invadopodia in tumor metastasis. Oncogene 2014;33:4193–202.
[19] Griffith OL, Gray JW. 'Omic approaches to preventing or managing metastatic breast cancer. Breast Cancer Res 2011;13.
[20] Takada M, Sugimoto M, Masuda N, et al. Prediction of postoperative disease-free survival and brain metastasis for HER2-positive breast cancer patients treated with neoadjuvant chemotherapy plus trastuzumab using a machine learning algorithm. Breast Cancer Res Treat 2018;172:611–8.
[21] Qiu C, Jiang L, Cao Y, et al. Factors associated with de novo metastatic disease in invasive breast cancer: comparison of artificial neural network and logistic regression models. Transl Cancer Res 2019;8:77–86.
[22] Tapak L, Shirmohammadi-Khorram N, Amini P, et al. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Global Health 2019;7:293–9.
[23] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.
[24] Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
[25] Roth HR, Lu L, Liu J, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans Med Imaging 2016;35:1170–81.
[26] Yala A, Barzilay R, Salama L, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 2017;161:203–11.
[27] Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS ONE 2011;6:e28210.
[28] Bhalla S, Kaur H, Dhall A, et al. Prediction and analysis of skin cancer progression using genomics profiles of patients. Sci Rep 2019;9:15790.
[29] Lee SC, Quinn A, Nguyen T, et al. A cross-cancer metastasis signature in the microRNA–mRNA axis of paired tissue samples. Mol Biol Rep 2019;46:5919–30.
[30] Albaradei S, Napolitano F, Thafar MA, et al. MetaCancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. Comput Struct Biotechnol J 2021;19:4404–11.
[31] Xie Y, Jeong KS, Pan W, et al. A case study on choosing normalization methods and test statistics for two-channel microarray data. Comp Funct Genomics 2004;5:432–44.
[32] Stańczyk U. Feature evaluation by filter, wrapper, and embedded approaches. Feature Select Data Pattern Recogn 2015:29–44.
[33] He Y, Ma J, Ye X. A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy. Int J Mol Med 2017;40:1357–64.
[34] Wei D. A multigene support vector machine predictor for metastasis of cutaneous melanoma. Mol Med Rep 2018;17:2907–14.
[35] Metri R, Mohan A, Nsengimana J, et al. Identification of a gene signature for discriminating metastatic from primary melanoma using a molecular interaction network approach. Sci Rep 2017;7:17314.
[36] Tuo Y, An N, Zhang M. Feature genes in metastatic breast cancer identified by MetaDE and SVM classifier methods. Mol Med Rep 2018;17:4281–90.
[37] Wu J, Xiao Y, Xia C, et al. Identification of biomarkers for predicting lymph node metastasis of stomach cancer using clinical DNA methylation data. Dis Markers 2017;2017:5745724.
[38] Ahsen ME, Boren TP, Singh NK, et al. Sparse feature selection for classification and prediction of metastasis in endometrial cancer. BMC Genomics 2017;18:233.
[39] Zhao S, Yu J, Wang L. Machine learning based prediction of brain metastasis of patients with IIIA-N2 lung adenocarcinoma by a three-miRNA signature. Transl Oncol 2018;11:157–67.
[40] Burton M, Thomassen M, Tan Q, et al. Gene expression profiles for predicting metastasis in breast cancer: a cross-study comparison of classification methods. ScientificWorldJournal 2012;2012:380495.
[41] Ye Q-H, Qin L-X, Forgues M, et al. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9:416–23.
[42] Jansson MD, Lund AH. MicroRNA and cancer. Mol Oncol 2012;6:590–610.
[43] Pfeifer GP. Defining driver DNA methylation changes in human cancer. Int J Mol Sci 2018;19:1166.
[44] Cui M, Wang H, Yao X, et al. Circulating microRNAs in cancer: potential and challenge. Front Genet 2019;10:626.
[45] Adlung L, Cohen Y, Mor U, et al. Machine learning in clinical decision making. Med 2021.
[46] Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. Radiol Soc North Am 2020.
[47] Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323.
[48] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recogn Lett 2009;30:27–38.
[49] Chen J-Q, Zhan W-H, He Y-L, et al. Prediction of lymph node metastasis with binary logistic regression in gastric carcinoma. Zhonghua Wei Chang Wai Ke Za Zhi 2005;8:436–9.
[50] Mitra AP, Almal AA, George B, et al. The use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. BMC Cancer 2006;6:159.
[51] Kim J, Yao F, Xiao Z, et al. MicroRNAs and metastasis: small RNAs play big roles. Cancer Metastasis Rev 2018;37:5–15.
[52] Okugawa Y, Grady WM, Goel A. Epigenetic alterations in colorectal cancer: emerging biomarkers. Gastroenterology 2015;149:1204–1225.e1212.

[53] Raza K. Analysis of microarray data using artificial intelligence based techniques. Biotechnology 2019:865–88.
[54] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69.
[55] Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nat Biotechnol 2018;36:829–38.
[56] Park Y, Kellis M. Deep learning for regulatory genomics. Nat Biotechnol 2015;33:825.
[57] Fakoor R, Ladhak F, Nazi A, et al. Using deep learning to enhance cancer diagnosis and classification. Proceedings of the international conference on machine learning. researchgate.net, 2013.
[58] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[59] Svozil D, Kvasnicka V, Jí Pospichal. Introduction to multi-layer feed-forward neural networks. Chemometrics Intelligent Lab Syst 1997;39:43–62.
[60] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504–7.
[61] Hinton G. Boltzmann Machines, Encyclopedia of Machine Learning and Data Mining 2014:1–7.
[62] Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput 2006;18:1527–54.
[63] LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1:541–51.
[64] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Lee DD, Sugiyama M, Luxburg UV, editors. Advances in Neural Information Processing Systems 29. Curran Associates Inc; 2016. p. 3844–52.
[65] Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1989;1:270–80.
[66] Karabulut EM, Ibrikci T. Discriminative deep belief networks for microarray based cancer classification 2017.
[67] Chereda H, Bleckmann A, Kramer F, et al. Utilizing molecular network information via graph convolutional neural networks to predict metastatic event in breast cancer. Stud Health Technol Inform 2019;267:181–6.
[68] Albaradei S, Thafar M, Van Neste C et al. Metastatic State of Colorectal Cancer can be Accurately Predicted with Methylome. Proceedings of the 2019 6th International Conference on Bioinformatics Research and Applications. Association for Computing Machinery, 2019, 125–130.
[69] Wu C, Orozco C, Boyer J, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 2009;10:1–8.
[70] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
[71] Drier Y, Domany E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? PLoS ONE 2011;6:e17795.
[72] Veytsman B, Cui T, Baranova A. Practical detection of biological age: why it is not a trivial task. Biomarkers of Human Aging Springer 2019:7–21.
[73] Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Briefings Bioinf 2020.
[74] Zhang Z, Zhao Y, Liao X, et al. Deep learning in omics: a survey and guideline. Brief Functional Genomics 2019;18:41–57.
[75] Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. Nat Genet 2019;51:12–8.
[76] Chen L, Cai C, Chen V, et al. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinf 2016;17 (Suppl 1):9.
[77] Khalili M, Alavi MH, Khodakarim S et al. Prediction of the thromboembolic syndrome: an application of artificial neural networks in gene expression data analysis 2016.
[78] Miotto R, Li L, Kidd BA, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094.
[79] Chen Q, Song X, Yamada H et al. Learning deep representation from big and heterogeneous data for traffic accident inference. Thirtieth AAAI Conference on Artificial Intelligence. aaai.org, 2016.
[80] Chaudhary K, Poirion OB, Lu L et al. Deep Learning based multi-omics integration robustly predicts survival in liver cancer.
[81] Palatucci M, Pomerleau D, Hinton GE, et al. Zero-shot learning with semantic output codes. In: Bengio Y, Schuurmans D, Lafferty JD, editors. Advances in Neural Information Processing Systems 22. Curran Associates Inc; 2009. p. 1410–8.
[82] Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 2006;28:594–611.
[83] Mikolajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018.
[84] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, editors. Advances in Neural Information Processing Systems 27. Curran Associates Inc; 2014. p. 2672–80.
[85] Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol 2020;38:143–5.
[86] Kukar M, Kononenko I. Cost-sensitive learning with neural networks. ECAI pdfssemanticscholarorg 1998:445–9.
[87] Ribeiro MT, Singh S, Guestrin C. "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. dl.acm.org. p. 1135–44.
[88] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of the 34th International Conference on Machine Learning-Volume 70. dl.acm.org, 2017, 3145-3153.
[89] Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:1–9.