frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fbioe.2020.00701

A Deep Learning Framework to Predict Tumor Tissue-of-Origin Based on Copy Number Alteration

Ying Liang [1†], Haifeng Wang [2†], Jialiang Yang [3†], Xiong Li [4], Chan Dai [3], Peng Shao [1], Geng Tian [3], Bo Wang [3] and Yinglong Wang [1]*

[1] College of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang, China, [2] Department of Urology, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China, [3] Geneis (Beijing) Co. Ltd., Beijing, China, [4] School of Software, East China Jiaotong University, Nanchang, China

Cancer of unknown primary site (CUPS) is a type of metastatic tumor for which the sites of tumor origin cannot be determined. Precise diagnosis of the tissue origin for metastatic CUPS is crucial for developing treatment schemes to improve patient prognosis. Recently, there have been many studies using various cancer biomarkers to predict the tissue-of-origin (TOO) of CUPS. However, only a very few of them use copy number alteration (CNA) to trance TOO. In this paper, a two-step computational framework called CNA_origin is introduced to predict the tissue-of-origin of a tumor from its gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a convolution neural network (CNN). Based on real datasets released from the public database, CNA_origin had an overall accuracy of 83.81% on 10-fold cross-validation and 79% on independent datasets for predicting tumor origin, which improved the accuracy by 7.75 and 9.72% compared with the method published in a previous paper. Our results suggested that the autoencoder model can extract key characteristics of CNA and that the CNN classifier model developed in this study can predict the origin of tumors robustly and effectively. CNA_origin was written in Python and can be downloaded from https://github.com/YingLianghnu/CNA_origin.

Keywords: tumor, tissue-of-origin, copy number alteration, autoencoder, convolution neural network

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Guohua Huang,
Shaoyang University, China
Khanh N. Q. Le,
Taipei Medical University, Taiwan
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China

****Correspondence:***
Ying Liang
aliang1229@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *29 March 2020*
Accepted: *04 June 2020*
Published: *05 August 2020*

Citation:
Liang Y, Wang H, Yang J, Li X, Dai C, Shao P, Tian G, Wang B and Wang Y (2020) A Deep Learning Framework to Predict Tumor Tissue-of-Origin Based on Copy Number Alteration. Front. Bioeng. Biotechnol. 8:701. doi: 10.3389/fbioe.2020.00701

1. INTRODUCTION

Cancer metastasis is the process in which tumor cells fall off from the primary site, enter the circulatory system, transfer to other parts of the body, and continue to grow. In about 3–5% of metastatic tumors, the sites of origin cannot be found, and this is known as cancer of unknown primary site (CUPS). Patients diagnosed with CUPS are treated with broad-spectrum anticancer drugs and have a low median survival time of 9–12 months. Precise diagnosis of the tissue of origin for metastatic CUP is essential for deciding on the treatment scheme to improve the patient's prognosis (Chen et al., 2017). Clinical, imaging and pathological examination are used to detect the tissue of origin, but these approaches can only determine the tissue of origin in about 50–80% of CUP patients.

Recently, a large number of studies have tried to use cancer biomarkers to predict the primary tumor site for CUPs so as to provide much-needed guidelines for timely patient care and cancer therapy (Liang et al., 2016; Grewal et al., 2019; Wang et al., 2019; Zheng et al., 2019). The gene expression patterns in tumors have high specificity, and so these the most widely used biomarkers for tumor classification (Bloom et al., 2004; Tothill et al., 2005; Staub et al., 2010; Wu et al., 2010; Handorf et al., 2013; Xu et al., 2016; Wang et al., 2018; Li et al., 2019). For example, Li used the within-sample relative gene expression orderings of gene pairs within individual samples to identify a prediction signature (Li et al., 2019). Wang proposed a general framework to identify a subset of genes for each tumor subtype and presented a corresponding classification model for distinguishing different tumor subtypes (Wang et al., 2018). Xu established a comprehensive database integrating microarray- and sequencing-based gene expression profiles of 16,674 tumor samples covering 22 common human tumor types to discriminate the origins of tumor tissue, which will be an additional useful tool for determining the tumor origin (Xu et al., 2016).

DNA methylation and miRNA regulate the expression of genes involved in numerous biological processes (Rosenfeld et al., 2008; Rosenwald et al., 2010; Ferracin et al., 2011; Mueller et al., 2011; Søkilde et al., 2014). Tang developed a user-friendly webserver to predict tumor origin by identifying highly tissue-specific CpG sites and miRNA expression (Tang et al., 2017). Bae tried to discover tissue-specific methylation markers and predicted the tissue-of-origin in CUPS (Bae et al., 2018). Yang proposed an inverse space sparse representation model to distinguish tumor origins considering the characteristics of gene-based tumor data (Yang et al., 2019). Visual imagery is one of the main methods used by pathologists to assess the stage, type, and subtype of tumors (Shi et al., 2016; Coudray et al., 2018; Mohsen et al., 2018). Coudray employed visual inspection of histopathology slides to classify lung adenocarcinoma, lung squamous cell carcinoma, and normal lung tissue, which achieved performance comparable to that of pathologists (Coudray et al., 2018). Ultrasound imaging can also be used for tumor detection and diagnosis with a deep polynomial network algorithm (Shi et al., 2016).

As yet, few studies have investigated the roles of genome variants on tissue-of-origin in CUPS. Genome variants include mutation, small insertion, and deletion (INEDL) and copy number alteration (CNA). CNA is amplification and deletion of genomic sequences ranging from kilobases (Kb) to megabases (Mb) in size, which covers 360 Mb and encompasses hundreds of genes, disease loci, and functional elements (Redon et al., 2006). As the main genetic marker of the genome, CNA can affect the gene function through gene dose, gene breakage, gene fusion, and position effects and is closely related to the occurrence and development of tumor (Poduri et al., 2013). CNA also plays an increasingly important role in targeted therapy, personalized treatment, and prognosis judgment for tumors. Marquard developed a tool named TumorTracer by using publicly available somatic mutation data to train random forest classifiers and thus to identify the tissue of origin. This was demonstrated to be accurate enough to aid in the clinical diagnosis of cancers with unknown primary origin (Marquard et al., 2015). Zhang conducted a comprehensive genome-wide analysis of CNAs from six cancer types and selected 19 discriminative genes for tumor classification, but their overall prediction accuracy was about 75% (Zhang et al., 2016). In the current study, a computational method called CNA_origin is proposed to predict the tissue of origin with the information of gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a convolution neural network (CNN). This predictor successfully learned the inherent information of gene copy number and exhibited superior performance to classical algorithms for the same benchmark datasets.

TABLE 1 | Number of samples per tissue for CNA profiles.

Primary site	Histology	CNA datasets
Breast	BRCA (Breast invasive carcinoma)	847
Colorectal	COADREAD (Colorectal adenocarcinoma)	575
Brain	GBM (Glioblastoma multiforme)	563
Kidney	KIRC (Kidney renal clear cell carcinoma)	490
Ovarian	OV (Ovarian serous cystadenocarcinoma)	562
Uterine	UCEC (Uterine Corpus Endometrial Carcinoma)	443

2. MATERIALS AND METHODS

2.1. Datasets

The copy number signal was produced by Affymetrix SNP 6.0 arrays for the set of samples in the cancer genome atlas (TCGA) study, as generated with the Firehose analysis pipeline. The preprocessing analysis of the dataset was performed with GISTIC (Beroukhim et al., 2007). These datasets were from primary solid tumor samples released by MSKCC in 2013 that could be downloaded from http://cbio.mskcc.org/cancergenomics/pancan_tcga/. The datasets with a sample size greater than 400 were selected. The details of all tissue samples, including tumor status, histopathology details, and sample sizes, are summarized in **Table 1**.

Each sample had 24,174 genes with discrete copy number values denoted as "–2," "–1," "0," "1," "2," where "–2" was homozygous deletion, "–1" was heterozygous loss, "0" was diploid, "1" was one copy gain and "2" was high-level amplification or multiple-copy gain (Ciriello et al., 2013). The CNA values were scaled to [–1, 1] with Equation (1).

$$x' = \frac{x}{|x|_{max}} \tag{1}$$

where x was the CNA value of the gene, $|x|_{max}$ was the maximum absolute value of CNA among samples, and x' was the value after correction.

2.2. Feature Extraction

Each sample had 24,174 gene-level CNA values. High dimensionality and small sample sizes have seriously obscured the intrinsic nature of CNA data. In this paper, CNA_origin applied a stacked autoencoder (SAE) to extract the features of CNA values, which converted the high-dimensional data into low-dimensional codes by training a multilayer neural network with small central layers to reconstruct high-dimensional input vectors (Hinton and Salakhutdinov, 2006). The SAE consisted of an adaptive multilayer "encoder" network and an asymmetric "decoder" network, and high-dimensional abstraction whilst maintaining the key information was achieved for feature reduction with the help of hidden nodes in the code layer, as illustrated in **Figure 1A**.

In the encoder network, the 24,174 gene-level CNA values used as inputs were mapped to the latent representation of next layer using Equation (2).

$$X^{[i]} = f(W_i X^{[i-1]} + b_i) \tag{2}$$

where $f(x) = max\,(0, x)$ was ReLU activation function, b_i was the bias of layer i, and W_i was the weight between layer i-1 and i. In the decoder network, the code layer was used to reconstruct the input by a reverse mapping using Equation (3).

$$X^{[i]} = f({W_i}' X^{[i-1]} + {b_i}') \tag{3}$$

where ${W_i}' = W_i^T$. The tanh activation function $f\,(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ was added to predict the final value, and the dimensionality of the final output layer was the same as that of the input layer. To determine the optimized parameters of W and b, layer-by-layer pretraining was used to minimize the error between the input X and output X'. The middle features were extracted through hidden nodes in the code layer.

CNA_origin was implemented in Python 3.7.3 using Keras (2.24) with the backend of TensorFlow (1.14.0). For the feature extraction of gene CNA, the neuron numbers in symmetrical hidden layers were set at 4,096, 1,024, 256, 100, 256, 1,024, and 4,096, respectively. The middlemost 100 neurons represented the extracted features, as it was found that features with more than 100 dimensions were not helpful to improve the classifier performance. The initial learning rate was set to 0.01, batch size to 64, and epochs to 16. This autoencoder was optimized using the Adam algorithm to learn the model parameters, and the loss function was mean square error.

2.3. Classifier Construction

The fully connected layer learns the global patterns in feature space, but convolution layer applies filters in the form of convolution operations to learn local patterns from the image (Baek et al., 2018). Inspired by the visual world, CNN has two interesting properties, translation invariant and spatial hierarchies of patterns, which allow a convolution network to efficiently learn increasingly complex and abstract visual concepts (Chollet, 2015, 2017). These properties are specialized for image data and also show outstanding performance in sequence processing (Le et al., 2017, 2019b). The same input transformation was performed on every subsequence; a pattern learned at a certain position in a sequence was later recognized at a different position, making 1D convnets translation invariant. A 1D convolution layer could catch local patterns in a sequence, making it competitive with recurrent neural networks (RNN) on sequence-processing at a considerably cheaper computational cost.

CNA_origin reshaped the 100 features of the sample into a 100 × 1 vector; each input tensor was 100 in width, 1 in height, and 1 in depth. The 1D convolution was used to extract local subsequences with D filters, and each filter was of $k \times 1$ in size, which means the filter was k in width and 1 in height. CNA_origin utilized multi-scale convolution kernels, such as 1 × 1, 3 × 1, 5 × 1, 7 × 1, and 9 × 1, to extract high-order features of different levels and increase the diversity of feature extraction. Among them, the 1 × 1 convolution kernel changed the number of channels, increased the non-linear transformation of features, and improved the generalization ability of the network. The number 48 or 64 in parentheses behind $k \times 1$ meant convolution with 48 or 96 filters. CNA_origin padded the features by adding k/2 columns with elements being zero to the head and tail of the sequence; therefore, the width of the new sequence after convolution with stride 1 was still the same.

The Concat operation in **Figure 1** meant that the layer stacked features from each branch together. Different convolution layers and max-pooling layers concatenated like the Inception module, which increased the depth of the network and improved the robustness of the CNN. At the beginning of the network, a larger convolution kernel was used to reduce the number of parameters and computation, as illustrated in **Figure 1B**. In the last, the network connected two full connection layers, with a dropout layer to avoid overfitting. Usually, the number of hidden units was far larger than the obtained data, resulting in overfitting. The dropout layer helped alleviate this problem by removing some of the connections in the network (Baek et al., 2018). Output such as 50 × 1 × 128 meant that the feature maps were 50 in width, 1 in height, and 128 in depth. The final result was the probability that the sample belonged to each class and was found with the "softmax" activation function, which is often used in solving multi-classification problems. It was defined as Equation (4).

$$P_k = \frac{exp(\alpha_k)}{\sum_{i\,=\,1}^{m} exp(\alpha_i)} \tag{4}$$

P_k was the probability that the sample belonged to class k. exp(x) represented an exponential function, α_k was the input value of class k, and m was the number of tumor classes. The categorical cross-entropy loss corresponding with the "softmax" activation function was used, which was a variant of binary cross-entropy and was defined as Equation (5).

$$loss = -\sum_{i\,=\,1}^{n} y_{i1} logP_{i1} + y_{i2} logP_{i2} + \cdots + y_{im} logP_{im} \tag{5}$$

P_{im} was the predicted probability, n was the number of samples, and y_{im} was the true label.

For the classification learning, the number of multi-scale convolution kernels was set to 64, batch size to 16, and epochs to 12. The learning rate was dynamically adjusted according to the loss value of the test dataset, and the initial value was 0.01. The dropout rate was set to 0.4, and the loss function was sparse categorical crossentropy.

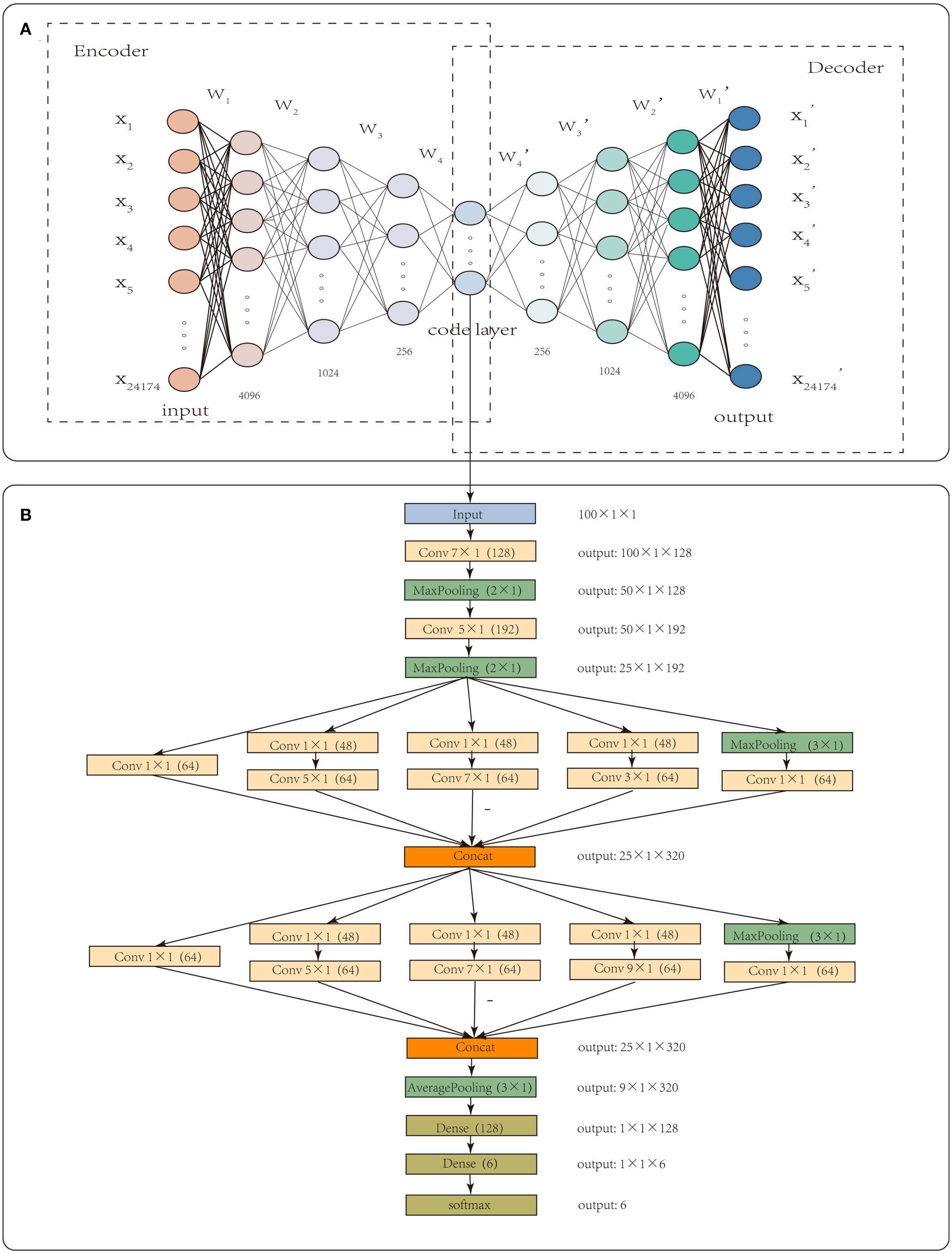

FIGURE 1 | The workflow of CNA_origin. CNA_origin applied a stacked autoencoder to extract the feature of CNA values, which was composed of a symmetrical encoder and decoder network, and 4,096, 1,024, and 256 were the neuron numbers in symmetrical hidden layers **(A)**. A 1D CNN with multi-scale convolution kernels (1 × 1, 3 × 1, 5 × 1, 7 × 1, 9 × 1) was used to construct a classifier model, and the number 48 or 64 in parenthesis behind *k* × 1 meant convolution with 48 or 96 filters. The Concat layer stacked features from each branch together; the output denoted the dimensions of feature maps for each layer **(B)**.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation Metrics

The six tumor datasets were used to train CNA_origin. To understand the generalization performance, CNA_origin was also tested by independent datasets. In this work, the precision (P), recall (R), accuracy (ACC), and F1-score were adopted to assess the performance of the corresponding method; they have been used as measurement metrics in previous works (Le et al., 2018, 2019a). They are defined as Equation (6).

$$
\begin{aligned}
P &= \frac{T_P}{T_P + F_P} \\
R &= \frac{T_P}{T_P + F_n} \\
ACC &= \frac{T_P + T_n}{T_P + F_p + F_n + T_n} \\
F1 - score &= \frac{2 \times P \times R}{P + R}
\end{aligned}
\tag{6}
$$

where T_P, T_n, F_P, and F_n were the numbers of true positives, true negatives, false positives, and false negatives, respectively. $P \in [0,1]$, $R \in [0,1]$, $ACC \in [0,1]$, and $F1 - score \in [0,1]$. P = 0 indicated that all predicted positive results were actually negative. When all results were incorrect, $T_P = 0$ and $T_n = 0$; therefore, P = 0, R = 0, ACC = 0, and F1-score = 0. When all results were correct, $F_P = 0$ and $F_n = 0$; therefore, P = 1, R = 1, ACC = 1, and F1-score = 1. Precision and recall are two contradictory metrics. Generally speaking, when the precision is high, the recall is often low, while when the recall is high, the precision is often low.

3.2. CNA_Origin Performance

Ten-fold cross-validation was utilized to evaluate our algorithm with the extracted 100-dimensional features. The datasets were randomly divided into ten subsets of approximately equal size. Our network was trained 10 times; nine of the 10 subsets were used as the training datasets, and the remaining one was the test dataset. All of the above evaluation indices of our algorithm, that is, P, R, ACC, and F1-score, were calculated according to the results in our work. The average values of four metrics P, R, ACC, and F1-score defined in Equation (6) over ten test datasets are listed in **Table 2**.

TABLE 2 | CNA_origin performance measured by three metrics via 10-fold cross-validation.

Cancer	Precision	Recall	F1-score
BRCA	0.8750	0.9231	0.8984
COADREAD	0.8158	0.7381	0.7750
GBM	0.9310	0.8438	0.8852
KIRC	0.8889	0.9600	0.9231
OV	0.8980	0.8672	0.8800
UCEC	0.6792	0.7200	0.6990

3.3. Performance Comparison With Other Algorithms

The performance of our algorithm was compared with four other classical classification algorithms with the same benchmark datasets. Random forest (RF) is an ensemble classifier that produces multiple decision trees using a randomly selected subset of training samples and variables (Liu et al., 2019). XGBoost is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning and has been used in many bioinformatics fields (Chen and Guestrin, 2016; Deng et al., 2020; Hu et al., 2020). Long Short-Term Memory (LSTM) is an artificial RNN architecture that is well-suited to classifying, processing, and making predictions based on time series data (Hochreiter and Schmidhuber, 1997). Zhang proposed a method to computationally classify cancer types by using CNA

TABLE 3 | Comparison of CNA_origin predictions with those of other algorithms.

Cancer	Predictor	Precision	Recall	F1-score
BRCA	CNA_origin	**0.8750**	**0.9231**	**0.8984**
	LSTM	0.8713	0.8462	0.8585
	RF	0.8556	0.8645	0.8601
	XGboost	0.8214	0.8846	0.8519
	CNA_zhang	0.7916	0.8735	0.8306
COADREAD	CNA_origin	0.8158	0.7381	0.7750
	LSTM	**0.8571**	**0.8077**	**0.8317**
	RF	0.7659	0.6923	0.7272
	XGboost	0.7959	0.7500	0.7723
	CNA_zhang	0.6000	0.7346	0.6605
GBM	CNA_origin	0.9310	0.8438	0.8852
	LSTM	0.8913	0.8913	0.8913
	RF	0.8627	0.8627	0.8627
	XGboost	**0.9535**	**0.8913**	**0.9213**
	CNA_zhang	0.8870	0.8593	0.8730
KIRC	CNA_origin	0.8889	**0.9600**	**0.9231**
	LSTM	0.8837	0.9268	0.9048
	RF	**0.9056**	0.8571	0.8807
	XGboost	0.8780	0.8780	0.8780
	CNA_zhang	0.8085	0.9268	0.8636
OV	CNA_origin	**0.8980**	0.8627	**0.8800**
	LSTM	0.7843	**0.9091**	0.8421
	RF	0.7826	0.9000	0.8372
	XGboost	0.7551	0.8409	0.7957
	CNA_zhang	0.8461	0.7586	0.8000
UCEC	CNA_origin	0.6792	**0.7200**	**0.6990**
	LSTM	0.6897	0.6557	0.6723
	RF	0.6451	0.6060	0.6250
	XGboost	0.7407	0.6557	0.6957
	CNA_zhang	**0.7419**	0.4693	0.5750

The bold values are the best performance among counterparts.

level values; this was denoted as CNA_zhang here because the authors did not give the method a name (Zhang et al., 2016). CNA_zhang used minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) to select features and the Dagging algorithm to give the final classification. The input of LSTM, RF, and XGboost was the extracted features from the autoencoder, and the GridSearchCV function in the sklearn package was used to select the optimal super-parameters that, were promised in the best condition.

Table 3 shows that the performance of CNA_origin was superior to LSTM, RF, XGboost, and CNA_zhang for BRCA, KIRC, OV, and UCEC. For BRCA, compared with LSTM and CNA_zhang, the F1-score was increased by 4.6 and 8.1%, respectively, and the recall (R) was increased by 9.08 and 5.67%, respectively. For GBM, CNA_origin performed slightly worse than the best, XGboost, with reductions of 2.35% in precision, 5.32% in recall, and 3.91% in F1-score. For KIRC, compared with LSTM and CNA_zhang, the F1-score was increased by 2.02 and 6.88%, respectively, and the recall was increased by 3.58%. For UCEC, compared with LSTM and CNA_zhang, the F1-score was increased by 3.97 and 21.56%, respectively, and the recall was increased by 9.80 and 53.41%, respectively. For COADREAD, CNA_origin performed slightly worse than the best LSTM algorithm, with reductions of 4.81% in precision, 8.61% in recall, and 6.81% in F1-score, respectively. For OV, the F1-score of CNA_origin was increased by 4.50% and 10.00% compared with LSTM and CNA_zhang; the recall was worse than the best, LSTM, by 5.10%, and precision was better than LSTM and CNA_zhang by 14.49 and 6.13%, respectively. CNA_origin exhibited perfect performance for the tumor classification.

The macro-averages of precision, F1-score, recall, and accuracy of six types of tumors were utilized to evaluate our predictor. Ten-fold cross-validation was run 100 times to test CNA_origin, LSTM, RF, XGboost, and CNA_zhang. For precision, CNA_origin had a mean value of 0.8369, which was increased by 0.70 and 6.87% compared with LSTM and CNA_zhang. For recall, the mean value of CNA_origin was 0.8345, which was increased by 0.91 and 8.68% compared with LSTM and CNA_zhang, respectively. For the F1-score, the mean value of CNA_origin was 0.8339, which was increased by 0.77 and 8.22% compared with LSTM and CNA_zhang, respectively. For accuracy, the CNA_origin had a mean value of 0.8381, which was increased by 0.92 and 7.75% compared with LSTM and CNA_zhang, respectively. The results are shown in **Figure 2**.

The results showed that the sensitivity, accuracy, and specificity of UCEC were significantly lower than those of other tumors. The results of UCEC were further analyzed, and it was found that about 48–76% of UCEC samples were predicted

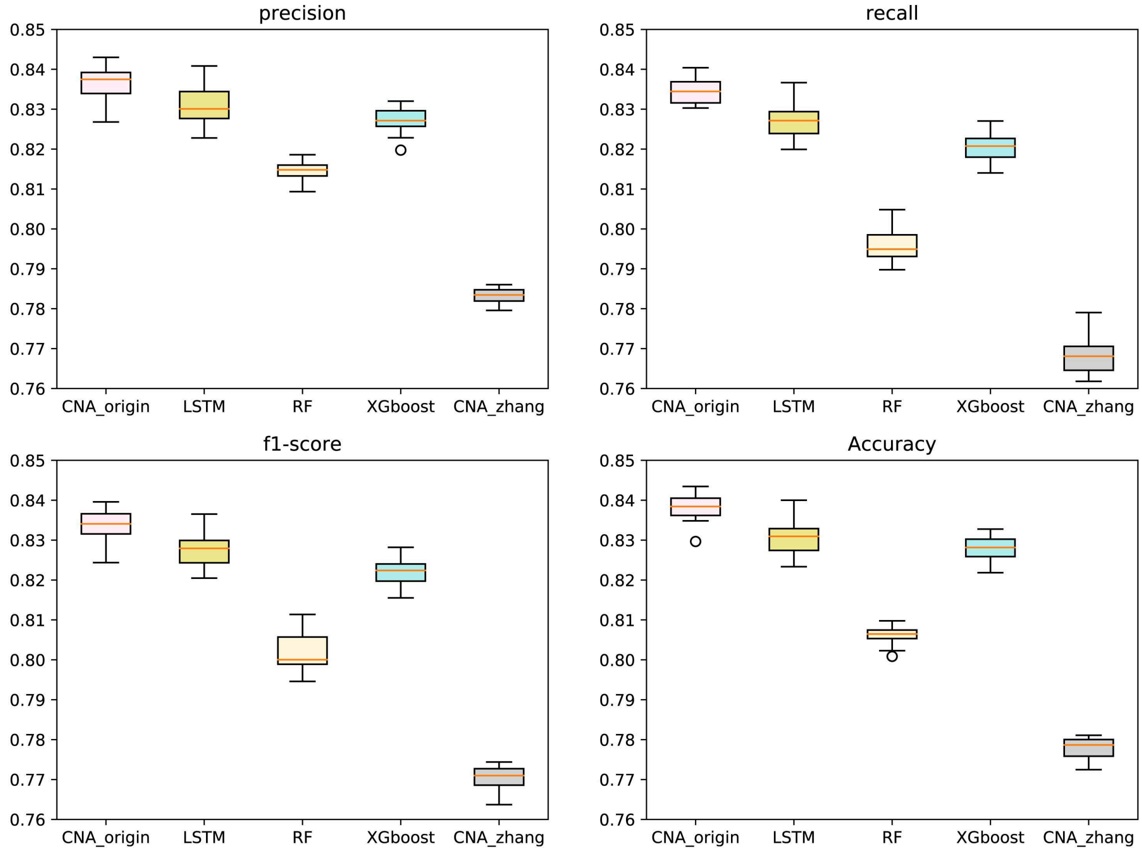

FIGURE 2 | Performance comparison between CNA_origin and other algorithms (basic LSTM, RF, XGboost, and CNA_zhang) for the macro-averages of precision, F1-score, recall, and accuracy from 10-fold cross-validation 100 times.

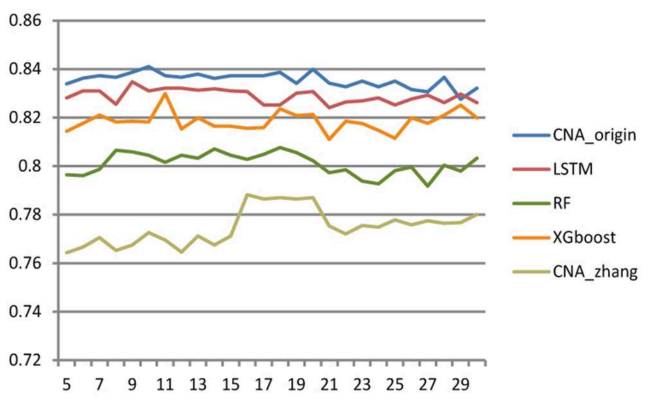

FIGURE 3 | Effect of cross-validation fold k value on classifier performance. When the value of k became larger, the performance of classifiers was improved, but a small sample size of the test set had a negative impact on model evaluation.

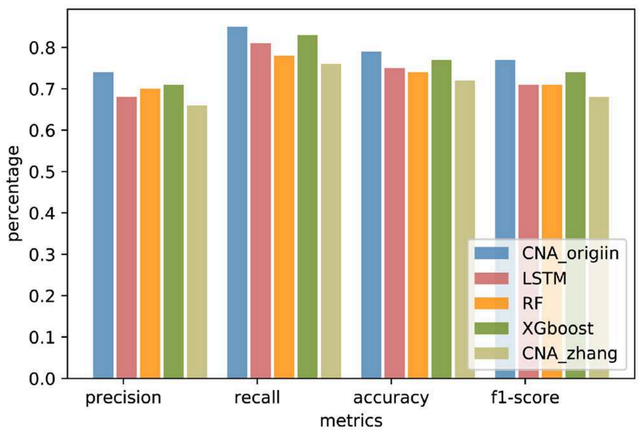

FIGURE 4 | Performance comparison of CNA_origin and other algorithms (basic LSTM, RF, XGboost, and CNA_zhang) for independent datasets from the TCGA.

to be OV, while 24–52% of UCEC samples were predicted to be BRCA. This may be because BRCA, OV, and UCEC are hormone-dependent tumors, which have a close relationship in tumorigenesis. Many reports have pointed out that BRCA, OV, and UCEC are related to changes in estrogen and estrogen receptors (Rodriguez et al., 2019; Scherbakov et al., 2019; Sehouli et al., 2019). Moreover, the physical location of ovary and uterus is very close, which may lead to contamination of tissue samples and difficulty in distinguishing UCEC from OV samples.

3.4. Impact of Sample Size

Different cross-validation fold k values were used to study the effect of sample number on the performance of the classifier. The larger k was, the more samples there were in the training set, and then the fewer samples there were in the test set, and vice versa. The range of k ranged from 5 to 30 with step size = 1, and **Figure 3** shows the accuracy of CNA_origin, LSTM, RF, XGboost, and CNA_origin with the different fold k values. With increasing k value, the performance of CNA_origin was gradually improved at first, which could be due to a bigger k including more training samples. But, as k became larger, the number of samples in the test set became smaller, and the performance of the classifiers was weakened. The results indicated that the performance of CNA_origin would be further improved if the training samples were expanded and that sufficient test samples were also very important for model evaluation.

3.5. Performance Comparison of Independent Datasets

In order to compare generalization performance on the independent data, experiments were performed with CNA datasets released by TCGA in 2016 downloaded from http://gdac.broadinstitute.org/. The TCGA datasets had 1080 BRCA samples, 611 COADRAD samples, 577 GBM samples, 528 KIRC samples, 552 OV samples, and 533 UCEC samples, respectively. The preprocessing analysis of 24776 gene CNA values was performed with GISTIC2 (Mermel et al., 2011). The TCGA datasets were reasonably independent of the training data because of preprocessing analyses such as quality control, alignment, and variation detection, which had a different systematic bias. The genes involved in both MSKCC datasets and TCGA datasets were selected, and the TCGA samples existing in MSKCC datasets were removed. There were 19895 common genes present in the MSKCC and TCGA datasets, and the independent datasets contained 234 BRCA samples, 50 COADRAD samples, 25 GBM samples, 41 KIRC samples, 21 OV samples, and 99 UCEC samples (see **Supplementary Material** for details). The independent datasets were used to evaluate the performance of CNA_origin. As shown in **Figure 4**, the overall performance of CNA_origin in terms of precision, recall, accuracy, and F1-score was the highest among the tools, at 0.74, 0.85, 0.79, and 0.77, respectively (see **Supplementary Material** for details). According to the results shown in **Figure 4**, it was concluded that CNA_origin performed successfully in the independent datasets.

4. CONCLUSIONS

Patients with CUPS often have a low median survival time of 9–12 months. Precise diagnosis of the tissue origin for metastatic CUPS is essential for determining the treatment scheme to improve patient prognosis. A lot of studies have tried to use cancer biomarkers to predict the primary tumor site for CUPS so as to provide important guidelines for timely patient care and cancer therapy. CNA provides a new way to identify and classify tumor types. In this study, a computational method, CNA_origin, was proposed to predict the tissue of origin from information on gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a CNN. This predictor successfully learned the inherent information of gene copy number and exhibited superior performance to the classical algorithms on k-fold cross-validations and independent datasets.

At present, the accuracy of using only CNA as the biomarker for tumor traceability is not very high. Integrating multiple biomarkers, such as CNA and DNA methylation or gene expression data, to trace tumor is our future goal.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YL conceived of the algorithm, develop the program, and wrote the manuscript. JY, BW, and GT helped with manuscript editing, designed, and performed experiments. PS and YW prepared the datasets. XL and CD carried out analyses and helped with the program design. HW designed of the work and participated in revising articles. All authors read and approved the final manuscript.

FUNDING

This study was supported by Jiangxi Provincial Department of Education Science and Technology plan Project (Grant No. GJJ180197) and Jiangxi Provincial Natural Science Fund (No. 20192ACB21004).

ACKNOWLEDGMENTS

The authors thank the reviewers for their useful feedback and constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00701/full#supplementary-material

REFERENCES

Bae, J. M., Kim, K., Chae, H. J., Wen, X., Kim, K. Y., Gwon, H. K., et al. (2018). Abstract 3312: Identification of tissue-of-origin in cancer of unknown primary site (cups) using methylation-specific targeted resequencing: a pilot study. *Cancer Res.* 78(13 Suppl.), 3312–3312. doi: 10.1158/1538-7445.AM2018-3312

Baek, J., Lee, B., Kwon, S., and Yoon, S. (2018). LncRNAnet: long non-coding RNA identification using deep learning . *Bioinformatics* 34, 3889–3897. doi: 10.1093/bioinformatics/bty418

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20007–20012. doi: 10.1073/pnas.0710052104

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., et al. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am. J. Pathol.* 164, 9–16. doi: 10.1016/S0002-9440(10)63090-8

Chen, F., Zhang, Y., Bossé, D., Lalani, A.-K. A., Hakimi, A. A., Hsieh, J. J., et al. (2017). Pan-urologic cancer genomic subtypes that transcend tissue of origin. *Nat. Commun.* 8, 1–15. doi: 10.1038/s41467-017-00289-x

Chen, T., and Guestrin, C. (2016). "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA).

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258.

Chollet, F. (2015). *keras*. GitHub repository. Available online at: https://github.com/keras-team/keras

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567. doi: 10.1038/s41591-018-0177-5

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., et al. (2020). Developing computational model to predict protein-protein interaction sites based on the xgboost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms21072274

Ferracin, M., Pedriali, M., Veronese, A., Zagatti, B., Gafà, R., Magri, E., et al. (2011). Microrna profiling for the identification of cancers with unknown primary tissue-of-origin. *J. Pathol.* 225, 43–53. doi: 10.1002/path.2915

Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., et al. (2019). Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* 2:e192597. doi: 10.1001/jamanetworkopen.2019.2597

Handorf, C. R., Kulkarni, A., Grenert, J. P., Weiss, L. M., Rogers, W. M., Kim, O. S., et al. (2013). A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. *Am. J. Surg. Pathol.* 37:1067. doi: 10.1097/PAS.0b013e31828309c4

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for qsar prediction. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009. [Epub ahead of print].

Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2018). Classifying the molecular functions of rab gtpases in membrane trafficking using deep convolutional neural networks. *Anal. Biochem.* 555, 33–41. doi: 10.1016/j.ab.2018.06.011

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., Chua, M. C. H., and Yeh, H.-Y. (2019a). Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture. *Comput. Struct. Biotechnol. J.* 17, 1245–1254. doi: 10.1016/j.csbj.2019.09.005

Le, N. Q. K., Yapp, E. K. Y., and Yeh, H.-Y. (2019b). Et-gru: using multi-layer gated recurrent units to identify electron transport proteins. *BMC Bioinform.* 20:377. doi: 10.1186/s12859-019-2972-5

Li, M., Li, H., Hong, G., Tang, Z., and Guo, S. (2019). Identifying primary site of lung-limited cancer of unknown primary based on relative gene expression orderings. *BMC Cancer* 19:67. doi: 10.1186/s12885-019-5274-4

Liang, Y., Qiu, K., Liao, B., Zhu, W., Huang, X., Li, L., et al. (2016). Seeksv: an accurate tool for somatic structural variation and virus integration detection. *Bioinformatics* 33, 184–191. doi: 10.1093/bioinformatics/btw591

Liu, X., Liu, X., Lai, Y., Yang, F., and Zeng, Y. (2019). "Random decision dag: An entropy based compression approach for random forest," in *International Conference on Database Systems for Advanced Applications* (Springer), 319–323.

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2015). Tumortracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med. Genomics* 8:58. doi: 10.1186/s12920-015-0130-0

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Mohsen, H., El-Dahshan, E.-S. A., El-Horbaty, E.-S. M., and Salem, A.-B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Comput. Inform. J.* 3, 68–71. doi: 10.1016/j.fcij.2017.12.001

Mueller, W. C., Spector, Y., Edmonston, T. B., Cyr, B. S., Jaeger, D., Lass, U., et al. (2011). Accurate classification of metastatic brain tumors using a novel microrna-based test. *Oncologist* 16, 165–174. doi: 10.1634/theoncologist.2010-0305

Poduri, A., Evrony, G. D., Cai, X., and Walsh, C. A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science* 341:1237758. doi: 10.1126/science.1237758

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

Rodriguez, A. C., Blanchard, Z., Maurer, K. A., and Gertz, J. (2019). Estrogen signaling in endometrial cancer: a key oncogenic pathway with several open questions. *Horm. Cancer* 10, 51–63. doi: 10.1007/s12672-019-0358-9

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.* 26, 462–469. doi: 10.1038/nbt1392

Rosenwald, S., Gilad, S., Benjamin, S., Lebanony, D., Dromi, N., Faerman, A., et al. (2010). Validation of a microrna-based qrt-pcr test for accurate identification of tumor tissue origin. *Mod. Pathol.* 23, 814–823. doi: 10.1038/modpathol.2010.57

Scherbakov, A., Shestakova, E., Galeeva, K., and Bogush, T. (2019). Brca1 and estrogen receptor α expression regulation in breast cancer cells. *Mol. Biol.* 53, 442–451. doi: 10.1134/S0026893319030166

Sehouli, J., Braicu, E. I., Richter, R., Denkert, C., Jank, P., Jurmeister, P. S., et al. (2019). Prognostic significance of ki-67 levels and hormone receptor expression in low-grade serous ovarian carcinoma: an investigation of the tumor bank ovarian cancer network. *Hum. Pathol.* 85, 299–308. doi: 10.1016/j.humpath.2018.10.020

Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., and Wang, T. (2016). Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing* 194, 87–94. doi: 10.1016/j.neucom.2016.01.074

Søkilde, R., Vincent, M., Møller, A. K., Hansen, A., Høiby, P. E., Blondal, T., et al. (2014). Efficient identification of mirnas for classification of tumor origin. *J. Mol. Diagn.* 16, 106–115. doi: 10.1016/j.jmoldx.2013.10.001

Staub, E., Buhr, H., and Gröne, J. (2010). Predicting the site of origin of tumors by a gene expression signature derived from normal tissues. *Oncogene* 29, 4485–4492. doi: 10.1038/onc.2010.196

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2017). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., Van Laar, R. K., et al. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res.* 65, 4031–4040. doi: 10.1158/0008-5472.CAN-04-3617

Wang, A., An, N., Chen, G., Liu, L., and Alterovitz, G. (2018). Subtype dependent biomarker identification and tumor classification from gene expression profiles. *Knowl.Based Syst.* 146, 104–117. doi: 10.1016/j.knosys.2018.01.025

Wang, Q., Xu, M., Sun, Y., Chen, J., Chen, C., Qian, C., et al. (2019). Gene expression profiling for diagnosis of triple-negative breast cancer: a multicenter, retrospective cohort study. *Front. Oncol.* 9:354. doi: 10.3389/fonc.2019.00354

Wu, A. H., Drees, J. C., Wang, H., VandenBerg, S. R., Lal, A., Henner, W. D., et al. (2010). Gene expression profiles help identify the tissue of origin for metastatic brain cancers. *Diagn. Pathol.* 5:26. doi: 10.1186/1746-1596-5-26

Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., et al. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Mod. Pathol.* 29, 546–556. doi: 10.1038/modpathol.2016.60

Yang, X., Wu, W., Chen, Y., Li, X., Zhang, J., Long, D., et al. (2019). An integrated inverse space sparse representation framework for tumor classification. *Pattern Recogn.* 93, 293–311. doi: 10.1016/j.patcog.2019.04.013

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860(11 Pt B), 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zheng, Y., Ding, Y., Wang, Q., Sun, Y., Teng, X., Gao, Q., et al. (2019). 90-gene signature assay for tissue origin diagnosis of brain metastases. *J. Transl. Med.* 17:331. doi: 10.1186/s12967-019-2082-1

Conflict of Interest: JY, CD, GT, and BW were employed by the company Geneis Beijing Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.