Application of Machine Learning for Tracing the Origin of Metastatic Lung Cancer Tissues

Tingting Wang, Qi Fan, Hongzhi Cai and Beier Zhang

***Abstract*—Accurate identification of the primary site of metastatic lung cancer (LC) is critical to the design of effective treatments that can assist physicians in diagnosis and improve prognosis. We collected genetic information from The Cancer Genome Atlas (TCGA) database of 3808 samples with clear tumor types; differential expression analysis was used for feature gene selection and a machine learning model was built to locate the primary tumor site. Finally, the 59 differentially expressed genes screened were well-characterized for metastatic LC primary tissue localization (lung adenocarcinoma, lung squamous carcinoma, thyroid carcinoma, breast invasive carcinoma, renal clear cell carcinoma, and renal papillary cell carcinoma).Comparing logistic regression analyses, the K-nearest neighbor and support vector machine revealed that the random forest (RF) model increased the average accuracy, precision, sensitivity, specificity, F1 score, macro-F1, micro-F1, and weighted-F1 from 99.06% to 99.78%, 96.42% to 99.37%, 96.77% to 98.90%, 99.44% to 99.86%, 96.63% to 99.12%, 96.63% to 99.12%, 97.18% to 99.32%, and 97.28% to 99.41%, respectively. The highest classification accuracy of 100% was achieved for thyroid cancer origin for all indicators under different algorithms.In conclusion, the RF algorithm was used to construct a model to trace the origin of metastatic LC tissues, which could potentially assist physicians in diagnosis, treatment, follow-up, and effective improvement of prognosis.**

***Index Terms*—Metastatic lung cancer, organization origin tracing, machine learning, differential expression, random forest.**

I. Introduction

METASTATIC lung cancer (LC) with unknown primary site is metastatic disease without any evidence of a primary tumor in the lymph nodes of the lung after detailed examination [1], [2]. About 60% or more of malignant tumors are accompanied by metastases at the time of initial diagnosis, of which 30-50% metastasizes to the patient's lungs [1], [2]. Metastatic LC accounts for 1.5-5% of all tumor cases [1], [3], and ranks 8th [4] among common malignant tumors and 4th in mortality [5]. A meta-study reported that patients with metastatic LC had a median survival time of 4.5 months, with 1- year and 5-year survival rates of 20% and 4.7%, respectively, after receiving chemotherapy; the prognosis was influenced by biological characteristics of the primary tumor [6]. Therefore, accurate identification of the primary site of metastatic LC is essential to design effective treatment and to aid physicians in diagnosis and improve prognosis.

Metastatic tumors are often heterogeneous, which can make clinicopathologic diagnosis and treatment particularly challenging. Immunohistochemistry is currently a key method for determining the site of origin of tumors and can be used to identify the tissue of origin of metastatic LC [7]. However, immunohistochemistry is labor-intensive, suitable for small sample size data, difficult to overcome classification accuracy bottlenecks, and needs urgent improvement [8]. Positron emission tomography (PET) and computed tomography (CT) are effective medical imaging tools for identifying the primary site of a tumor [9]. However, the accuracy of PET and CT in identifying the origin of tissues was 24-40% and 20-27%, respectively, with poor diagnostic performance [10]. Therefore, a new and effective method to identify the primary site of metastatic LC is urgently required.

It is known that metastatic tumors retain the gene expression profile of the tissue at the primary site. Wang et al. constructed a molecular marker containing 96 tissue-specific genes by gene expression profiling to determine 22 common tumor types and tissue origins [11]. Lu et al. developed three machine learning models: random forest (RF), support vector machine (SVM), and neural network; 80 differentially expressed genes (DEGs) were screened to distinguish four types of squamous cell carcinoma to locate the tissue of origin of metastatic cervical cancer [12]. Zhao et al. trained primary tumor expression data using a one-dimensional convolutional neural network model to infer the tissue origin of unknown primary cancer (CUP) [2]. The use of machine learning to construct medical models and analyze gene expression profiles to infer tumor tissue origins has gained popularity, but targeted studies for the tissue origins of metastatic LC are still lacking. The incidence of metastasis to the lung varies among different sites, of which thyroid, breast, and kidney cancers and osteosarcoma exhibit the highest incidence, reaching 60-90%, and are the primary sites of metastatic LC [1]. Therefore, this study will bioinformatically analyze the gene expression data of primary LC and major primary sites of metastatic LC, and construct four machine learning models to explore effective tools for localizing the tissue of LC origin.

II. Materials and Methods

A. Data collection and preparation

Samples with clear tumor type were selected from The Cancer Genome Atlas (TCGA) database. The screening

Manuscript received July 10, 2022; revised February 7, 2023. This study was supported by the National Natural Science Foundation of China (62006091), the Anhui New Era Education Quality Project "Application of Double weighted Random Forest Multi-label Algorithm in Predicting the direction of breast cancer Metastasis" (2022xscx090), and the Natural Science Research Project of Universities in Anhui Province in 2017 (key) "Endocrine personalized therapy for breast cancer patients based on Data Mining technology" (KJ2017A390).

Tingting Wang is a postgraduate student of School of Computer Science and Technology, Huaibei Normal University, China (e-mail: 2128296365@qq.com).

Qi Fan is a professor of School of Computer Science and Technology, Huaibei Normal University, China (corresponding author to provide phone: 13966139306; e-mail: 8073592@qq.com).

Hongzhi Cai is a postgraduate student of School of Life Sciences, Anhui Agricultural University, China (e-mail: 489544690@qq.com).

Beier Zhang is a postgraduate student of School of Life Sciences, Anhui Agricultural University, China (e-mail: 2530066750@qq.com).

TABLE I
GENE EXPRESSION PROFILING DATABASE

Cohort Abbreviation	Tumor type	Total	Case code=01	Control Code=11
LUAD	lung adenocarcinoma	572	513	59
LUSC	lung squamous cell carcinoma	550	501	49
THCA	thyroid carcinoma	560	502	58
BRCA	breast invasive carcinoma	1204	1091	113
KIRC	kidney renal clear cell carcinoma	602	530	72
KIRP	kidney renal papillary cell carcinoma	320	288	32
Total	6	3808	3425	383

criteria: Disease Type=squamous cell neoplasms, Program Name=TCGA or TARGET, Data Category=transcriptome profiling, and Experimental Strategy=RNA-Seq. The screened expression matrix included data from both primary tumor samples (tumor type code = "01") and conventional samples (tumor type code = "11"). Osteosarcoma contained only primary tumor samples and was therefore excluded. Subsequently, the mapping information of Gene Symbol and ENSG-ID was extracted from GENCODE (https://www.gencodegenes.org/human/) and ENSG-ID was uniformly mapped to Gene Symbol [11].

The RNA-Seq information of the samples was standardized and normalized to include as many tissue subtypes as possible for tumor types with significant heterogeneity. After preprocessing, the gene expression profile database was constructed (Table I). The database included six tumor types (lung adenocarcinoma, lung squamous cell carcinoma, thyroid carcinoma, breast invasive carcinoma, kidney renal clear cell carcinoma, and kidney renal papillary cell carcinoma) with 3808 tumor samples (including routine samples of primary tumor margin); the sample size of each tumor ranged from 320 to 1204 cases. The six tumor samples were grouped separately; the screened primary patients were recorded as the case group (tumor type code="01") and the routine samples around the tumor were recorded as the control group (tumor type code="11").

B. Modeling method

1) Logistic regression (LR): LR predicts the probability of future outcomes occurring from the performance of historical data. LR can establish a functional relationship between n ($n \geq 2$) independent variables and one dependent variable, and the output function model is as follows [13]:

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} = \frac{1}{1+e^{-(bT\times X)}} \quad (1)$$

where β_0, β_1, β_2 are the regression coefficients, and x_1, x_2,....., x_n is the independent variable, and is the probability that the primary tissue of metastatic LC is localized to a specific retrospective source [14]. Vector b was determined by logistic regression, and vector b associated each patient with metastatic LC with a specific retrospective probability.

2) K-nearest neighbors (KNN): When the KNN algorithm predicts a new input instance, the majority class of the k nearest points to this instance is used as the class of the new input instance. This idea is also a result of empirical risk minimization [15]. The training samples are (x_i, y_i), when the input instance is x, labeled as c, $N_k(x)$ is the set of KNN training samples of the input instance. Therefore, the training error rate is the proportion of KNN training sample tokens that do not agree with the input tokens. The training error rate is expressed as the following equation [16]:

$$\frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i = c_j) \quad (2)$$

Therefore, to minimize the empirical risk, i.e., the error rate, it is necessary to make the maximum in equation (2) $\frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i \neq c_j)$, i.e., the marker values of KNN match; the input marker values as far as possible; hence, the majority voting rule is equivalent to the empirical risk minimization [17].

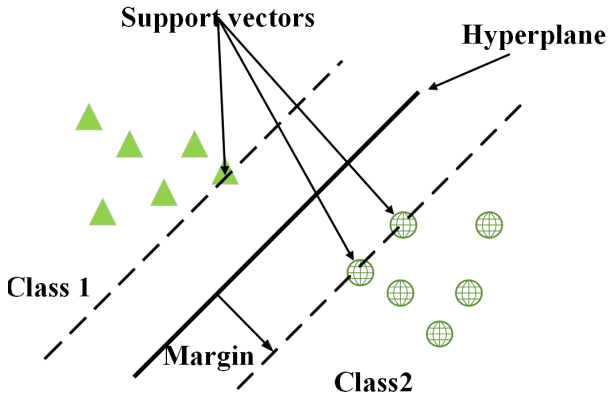

Fig. 1. SVM schematic

3) SVM: The SVM algorithm maximizes the geometric edges while minimizing the empirical error. As illustrated in Fig. 1, the vector (point) that limits the edge width is the support vector, and the sum of the distances from two different classes of support vectors to the hyperplane is called the interval (2*margin). The basic idea of SVM is to treat the input data as two sets of vectors in an n-dimensional space; the boundary between the two sets of data is maximized by creating a separating hyperplane within that space to classify the input data [18].

4) RF: The RF is based on the bagging algorithm, with two modifications: (1) m subsamples are randomly selected from the original dataset, and then when training each base learner, instead of selecting the best features from all features to slice the nodes, k features are randomly selected; the best features are selected from these k features to slice the nodes, thus further reducing the variance of the model; (2) the base learner used in the RF is a decision tree [19]. As in Fig. 2, the RF algorithm contains multiple DTs, and its output category is determined by the multitree of the categories output by the individual DTs [20].

The models were adjusted for super-parameters using 10-fold cross-validation and grid search methods. In 10-fold cross-validation, the initial sample (sample set X, Y) is split

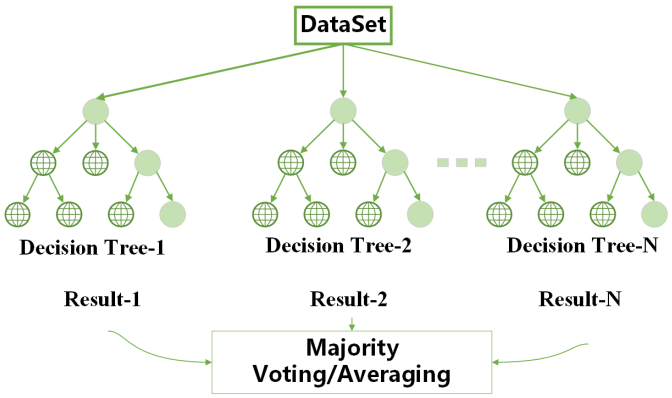

Fig. 2. RF structure

into 10 copies; 1 copy is retained as the data for validating the model (test set) and the other nine copies are used for training (training set). The cross-validation is repeated 10 times, once for each copy, and the results are averaged over 10 times or using other combinations to obtain a single estimate [14]. The grid search method presets multiple parameter permutations and uses the cross-validation method for multiple evaluations to select hyperparameters [21].

C. Model evaluation indicators

Machine learning model prediction performance is commonly measured by accuracy, precision, sensitivity, specificity, F1 score, $macro-F_1$, $micro-F_1$, and $weighted-F_1$. Each metric is greater than 0 and less than 1, with larger values representing higher classification accuracy. The formulae are as follows [21], [22], [23], [24]:

$$accuracy = \frac{TP + TN}{(TP + FN + FP + TN)} \tag{3}$$

$$precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$sensitivity = \frac{TP}{(TP + FN)} \tag{5}$$

$$specificity = \frac{TN}{(TN + FP)} \tag{6}$$

$$F_1 = \frac{2(precision \cdot sensitivity)}{(precision + sensitivity)} \tag{7}$$

$$macro - F_1 = \frac{\sum_{i=1}^{n} F_i}{(n)} \tag{8}$$

$$precision_m = \frac{\sum_{i=1}^{n} TP_i}{(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i)} \tag{9}$$

$$sensitivity_m = \frac{\sum_{i=1}^{n} TP_i}{(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i)} \tag{10}$$

$$micro - F_1 = \frac{2(precision_m \cdot sensitivity_m)}{(precision_m + sensitivity_m)} \tag{11}$$

$$weighted-F_1 = \sum_{i=1}^{n} F_i w_i \tag{12}$$

TP represents the number of correct scores into positives. FP represents the number of incorrectly classified positives. TN represents the number of correct classifications as negative. FN stands for the number of incorrectly classified as negative [25]. Precision and sensitivity are a pair of contradictory metrics. Usually when precision is high, the sensitivity value tends to be low; whereas when precision is low, the sensitivity value tends to be high. In order to consider these two metrics together, the F1 score (a weighted summed average of precision and sensitivity) is proposed [24]. $macro - F_1$ averages the precision and sensitivity of each category; thus, it does not take into account the sample size of each category; thus, it is more affected by categories with higher precisions and sensitivities. $micro-F_1$ takes into account the number of categories and is not affected by high precisions and sensitivities; however, it is more affected by categories with higher numbers. $macro-F_1$ and $micro-F_1$ complement each other to evaluate the model performance. where w_i represents the weight of cancer category in the total sample, and F_i represents the F1 score value of category. The $macro - F_1$ is the arithmetic mean of F1 scores of each category, which treats all categories equally, regardless of the importance of different categories. $micro - F_1$ is also calculated on the basis that each sample has the same weight. The $weighted-F_1$ takes into account the number of samples per category in the total sample, which can effectively mitigate the effect of data imbalances [23], [24].

III. RESULTS

A. Characteristic gene selection

The initial tissue genes of 58,938 were not well represented. The R package t.test function was used to assess the significance of the difference between each gene in the primary tumor (group1=case) and normal control (group2=control); then, the p.adjust function was called to calculate the significant false discovery rate of each gene, and finally the difference information of each gene was obtained. A volcano map (Fig. 3 and Fig. 4) was drawn to visualize the distribution of the expressed genes, with the horizontal axis representing the fold change, and the farther away from the center 0, the greater the difference. The vertical axis indicates p-value, and the point closer to the top indicates that the expression difference between the two samples is more significant. The black dots in the middle and lower part of the figure indicate that genes are not differentially expressed, and the red and green dots indicate upregulated and downregulated genes (upregulated + downregulated = total DEGs) respectively. The Venn diagram (Fig. 5) shows the information of DEGs among the six cancers with 7447, 10530, 3430, 6253, 11493 and 4433 DEGs for LUAD, LUSC, THCA, BRCA, KIRC, and KIRP, respectively. In the differential set, LUAD (1024), LUSC (2780), THCA (740), BRCA (1405), KIRC (4905) and KIRP (537) "characteristic" DEGs (significantly expressed in only one type of cancer; Fig. 5 non-overlapping part) were identified.

A total of 572 "characteristic" DEGs (52 LUAD, 139 LUSC, 37 THCA, 71 BRCA, 246 KIRC, and 27 KIRP) were screened for differential expression in the top 5% of

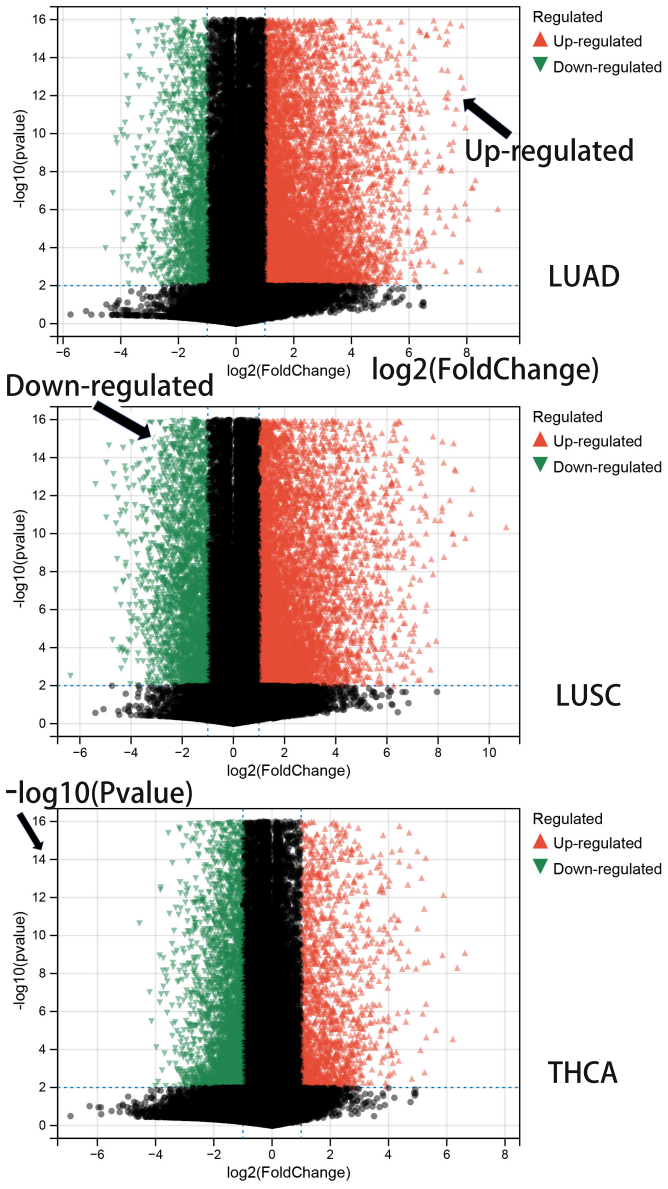

Fig. 3. DEGs volcano map1

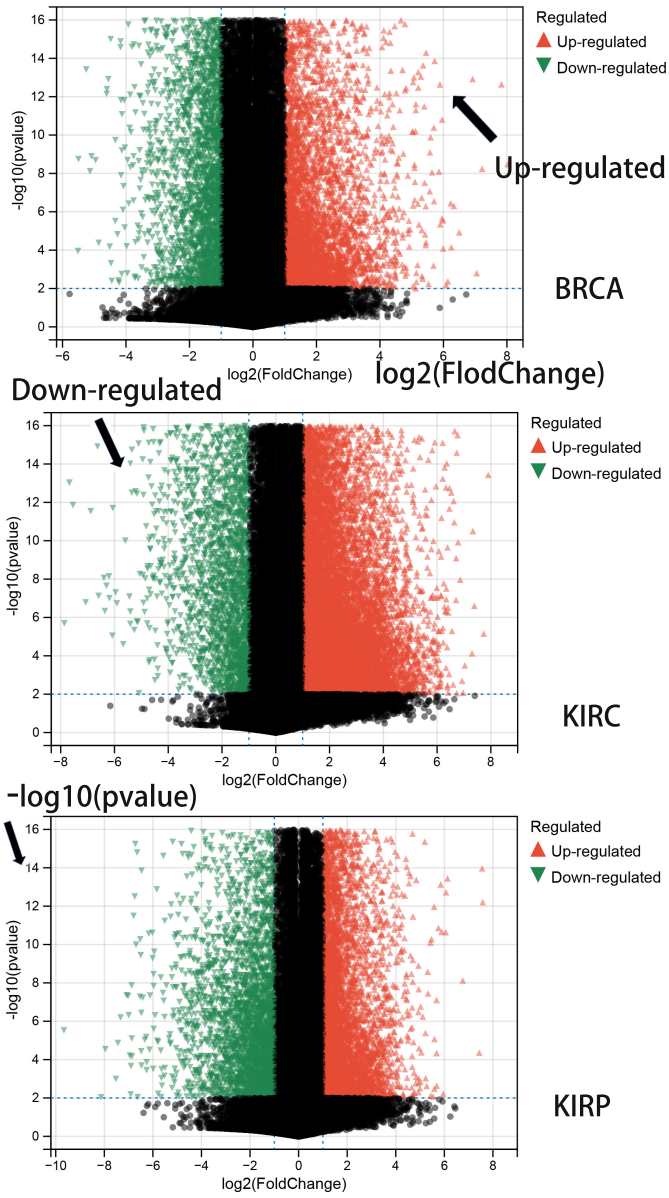

Fig. 4. DEGs volcano map2

each cancer. The post-screening information is for genes that are differentially expressed between each primary tumor and normal control, and are mutually exclusive between cancer types.

B. Model construction

Model variable selection: The independent variables included 572 "characteristic" DEGs, all of which were numerical variables. The dependent variable was the tumor sample type and was named a multicategorical variable, coded as LUAD=1, LUSC=2, THCA=3, BRCA=4, KIRC=5, and KIRP=6 in that order.

The matrix data was transposed using the T function and called the train_test_split function to create a 7/3 balanced split of the data (70% as the training set and 30% as the validation set) [26]. Various machine learning methods [LR, KNN, SVM and RF] were used to construct the models, and performance of the classifiers in the validation set was evaluated regarding accuracy, precision, sensitivity, specificity, F1 score, $macro-F_1$, $micro-F_1$ and $weighted-F_1$ metrics. Finally, the optimal retrospective model and the "characteristic" DEG attributes that affected the identification of metastatic LC tissue were filtered and derived.

C. Parameter tuning results

After multiple selections, the range of the RF split dataset metrics [entropy, gini], the range of maximum depth of each DT was [7,8,9,10,11], the range of minimum split sample size of each DT leaf was [4,8,12,16,20,24], and the range of number of DTs was [13,15,17,19]. An exhaustive search was performed to select the optimal parameters criterion: gini, max_depth: 8, min_samples_split:16, n_estimators:15.

D. Analysis of model results

The study used an RF algorithm to train a real-world dataset to obtain an importance score table for the "characteristic" DEGs attribute for locating the tissue of origin of

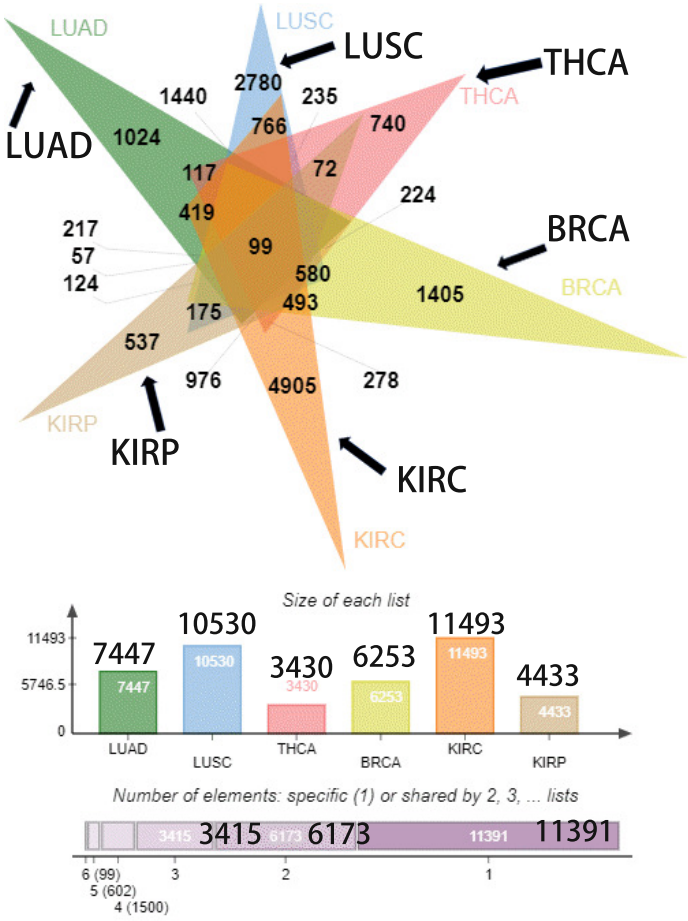

Fig. 5. DEGs Venn diagram

metastatic LC (Table II). The top 10 ”characteristic” DEGs of LUAD were ACY3, AL121949.1, DDIT4L, ZNF239, FGB, FAM53A, HMGB3, KRTAP5-1, DOK5, and RASGEF1A in that order. The top 10 ”characteristic” DEGs of LUSC were TRIM16L, AP3B2, TMEM117, SERPINB13, RN7SL399P, SFTA3, AC009118.2, ETNK2, CERS3, and NAP1L4P2 in that order. The top 9 ”characteristic” DEGs of THCA were AC046143.1, CELF4, ADAMTS9, MAMLD1, OR4D10, LRG1, AC023490.6, AC016405.2, and KLK11 in that order. The top 10 ”characteristic” DEGs of BRCA were SMYD3, KCNK15, CIT, TNFSF4, CLIP4, CRYAB, GNAL, CEMIP, BOC, and DMD in that order. The top 10 ”characteristic” DEGs of KIRC are ANXA2R, ATG16L2, RP11-69E11.4, AL358340.1, TNFRSF4, SAMD3, DPRXP4, COL4A5, AC008735.1, and PRRT2 in that order. The top 10 ”characteristic” DEGs of KIRP were RASGRF2, ASAP2, DNM1, NR2F1, RASD1, PCDH1, TMEM204, TBX2, THY1, and MEIS3, in that order. The above 59 ”characteristic” DEGs exhibit good representation of the origin of metastatic LC organs and are useful for the diagnosis and treatment of the origin of metastatic cancer.

The confusion matrices of LR, KNN, SVM and RF on the TCGA dataset are shown in Fig. 6. The models all display excellent prediction results, and the quantitative evaluation metrics are shown in Tables III, IV and V. Longitudinally observing the model evaluation effect, RF all scored the highest in the six cancer accuracy measurement results, with an average score of 99.78%, higher than LR (average score of 99.19%), KNN (average score of 99.06%) and SVM (average score of 99.09%). In the precision measurement results, RF scored the highest except BRCA, with an average score of 99.37%, and increased by 2.61%, 2.8%, and 2.95% compared with LR, KNN and SVM, respectively. In the specificity measurement results, all RF scored the highest except BRCA, with an average score of 99.86%, and increased by 0.33%, 0.42%, and 0.38% compared to LR, KNN, and SVM, respectively. RF scored the highest in all six cancers sensitivity and F1 score metrics, with 98.9% and 99.12%, respectively. RF had the highest $macro - F_1$ (99.12%). In addition, RF had the highest $micro - F_1$ (99.32%) and $weighted{-}F_1$ (99.41%), with an increase of 1.75% and 1.74%, 2.14%, and 2.13%, 2.04%, and 2.03% compared to LR, KNN and SVM, respectively. Overall, RF performed well in all eight metrics evaluated, with LR, KNN, and SVM being second. The reason is that LR, KNN, and SVM are based on regression and cannot handle highly correlated and nonlinear data; the ”characteristic” DEGs have biological correlation among data variables, so the model prediction effect is somewhat disturbed. The RF algorithm can handle highly correlated data, and also effectively overcome the structural limitation of single tree easy to fit; therefore, the strong classifier RF algorithm consisting of multiple trees worked best.

Cross-sectional observation of the model assessment effect showed the highest classification accuracy for THCA tissue origin, with 100.00% for all metrics under different algorithms. BRCA followed, especially under LR and SVM algorithms, with precision and specificity up to 100%, which are higher than 99.39% and 99.71% of the prediction effect of the RF algorithm, respectively. However, for LUSC and KIRP, about 95% precision, sensitivity, and F1 score was observed under all types of algorithms, even as low as 90.60%. From the confusion matrix (Fig. 6), it can be found that there is difficulty in distinguishing between some LUAD and LUSC due to tissue similarity (prediction error data available off the diagonal); this is similar for KIRC and KIRP. Since this embedding is heavily dependent on the type and number of input samples, some misclassification is inevitable in the absence of a larger sample cohort.

IV. DISCUSSION

In this study, four machine learning methods: LR, KNN, SVM and RF algorithms were used to construct models aimed at multiclassification tracing of primary tissue origins of metastatic LC of patients in the TCGA database. A high value of the model evaluation index strongly validates the hypothesis. When a patient with metastatic LC is being treated, the physician can select the retrospective model that best assesses the effect. Patient-specific DEG field values (60 influencing factors for metastatic LC tissue identification) are then entered sequentially in order of importance to obtain output results to trace the patient’s tissue of cancer cell origin.For example, when the model retrospective result is type=3 (tissue origin is THCA), the physician can feasibly combine thyroid cancer and LC treatment plans and be alert to the disease status of the cancer origin site during the follow-up period. The clinical treatment plan is based on a combination of patient condition, follow-up, and physician experience, and the results of this study only serve as an auxiliary reference.

TABLE II
"CHARACTERISTIC" DEGS ATTRIBUTE IMPORTANCE SCORE

LUAD(DEGs)	score	LUSC(DEGs)	score	THCA(DEGs)	score
ACY3	1.25E-02	TRIM16L	9.05E-02	AC046143.1	9.52E-02
AL121949.1	6.18E-04	AP3B2	6.21E-02	CELF4	1.31E-02
DDIT4L	5.61E-04	TMEM117	3.02E-02	ADAMTS9	1.86E-03
ZNF239	4.66E-04	SERPINB13	2.34E-02	MAMLD1	1.44E-03
FGB	3.50E-04	RN7SL399P	2.31E-02	OR4D10	1.24E-03
FAM53A	3.33E-04	SFTA3	1.79E-02	LRG1	9.30E-04
HMGB3	2.75E-04	AC009118.2	1.60E-02	AC023490.6	3.05E-04
KRTAP5-1	2.66E-04	ETNK2	1.06E-02	AC016405.2	2.07E-04
DOK5	1.71E-04	CERS3	8.98E-03	KLK11	7.30E-05
RASGEF1A	9.94E-05	NAP1L4P2	8.45E-03	**GALE**	**0.00E+00**
BRCA(DEGs)	score	KIRC(DEGs)	score	KIRP(DEGs)	score
SMYD3	2.02E-02	ANXA2R	7.20E-02	RASGRF2	2.72E-01
KCNK15	1.90E-02	ATG16L2	1.92E-02	ASAP2	8.69E-03
CIT	1.55E-02	RP11-69E11.4	1.22E-02	DNM1	4.74E-03
TNFSF4	9.39E-03	AL358340.1	4.69E-03	NR2F1	1.34E-03
CLIP4	8.89E-03	TNFRSF4	4.21E-03	RASD1	8.70E-04
CRYAB	5.57E-03	SAMD3	4.00E-03	PCDH1	4.60E-04
GNAL	1.76E-03	DPRXP4	2.87E-03	TMEM204	3.21E-04
CEMIP	7.75E-04	COL4A5	2.11E-03	TBX2	2.83E-04
BOC	7.72E-04	AC008735.1	1.11E-03	THY1	2.81E-04
DMD	7.50E-04	PRRT2	1.10E-03	MEIS3	2.24E-04

TABLE III
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Type	**Accuracy (%)**				**Precision (%)**			
	LR	**KNN**	**SVM**	**RF**	**LR**	**KNN**	**SVM**	**RF**
LUAD	99.03	98.54	98.74	**99.81**	95.00	93.71	94.90	**99.35**
LUSC	98.74	98.35	98.35	**99.81**	90.60	94.04	94.04	**99.33**
THCA	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
BRCA	99.61	99.22	99.61	**99.81**	**100**	99.38	**100**	99.39
KIRC	98.83	99.22	98.93	**99.61**	96.82	98.09	97.44	**98.14**
KIRP	98.93	99.03	98.93	**99.61**	92.13	94.19	92.13	**100**
Ave	99.19	99.06	99.09	**99.78**	96.76	96.57	96.42	**99.37**

TABLE IV
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Type	**Sensitivity (%)**				**Specificity (%)**				**F1 Score (%)**			
	LR	**KNN**	**SVM**	**RF**	**LR**	**KNN**	**SVM**	**RF**	**LR**	**KNN**	**SVM**	**RF**
LUAD	98.70	96.75	96.75	**99.35**	99.08	98.86	99.08	**99.89**	96.81	95.21	95.82	**99.35**
LUSC	94.67	94.67	94.67	**99.33**	99.43	98.97	98.97	**99.89**	95.63	94.35	94.35	**99.33**
THCA	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
BRCA	98.78	98.17	98.78	**100**	**100**	99.71	**100**	99.71	99.39	98.77	99.39	**99.69**
KIRC	95.60	96.86	95.60	**99.37**	99.42	**99.65**	99.54	**99.65**	96.21	97.47	96.51	**98.75**
KIRP	**95.35**	94.19	**95.35**	**95.35**	99.26	99.47	99.26	**100**	93.71	94.19	93.71	**97.62**
Ave	97.18	96.77	96.86	**98.90**	99.53	99.44	99.48	**99.86**	96.96	96.67	96.63	**99.12**

TABLE V
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Evaluation	LR	KNN	SVM	RF
$macro-F_1$	96.96%	96.67%	96.63%	**99.12%**
$micro-F_1$	97.57%	97.18%	97.28%	**99.32%**
$weighted-F_1$	97.67%	97.28%	97.38%	**99.41%**

A combination of differential expression analysis and machine learning was used to extract "characteristic" DEGs and present model data relationships. The original machine learning methods can perform feature extraction but have limited understanding of raw information. Differential expression analysis allowed the selection of a small number of 572 "signature" DEGs from the initial tissue of 58,938 genes based on upregulation and downregulation information of the genes; these DEGs had significant differential expression. The selected "characteristic" DEGs control biomorphological and physiological processes and are central to the regulation of cellular life processes [27]; this is an important observation regarding the origin of metastatic LC tissue in this study. The pre-experimental differential expression analysis can distinguish the genetic differences between different primary cancer tissues and surrounding conventional tissues; the post-machine learning model construction can distinguish the genetic differences between different cancers in a clear and articulated manner. Therefore, in feature engineering, differential expression analysis is used to extract "characteristic" DEGs for multiclassification problems, which can provide a reference for machine learning engineers.

This study evaluates the multicategory classification capability of machine learning methods based on eight metrics. The RF average accuracy, precision, sensitivity, specificity,

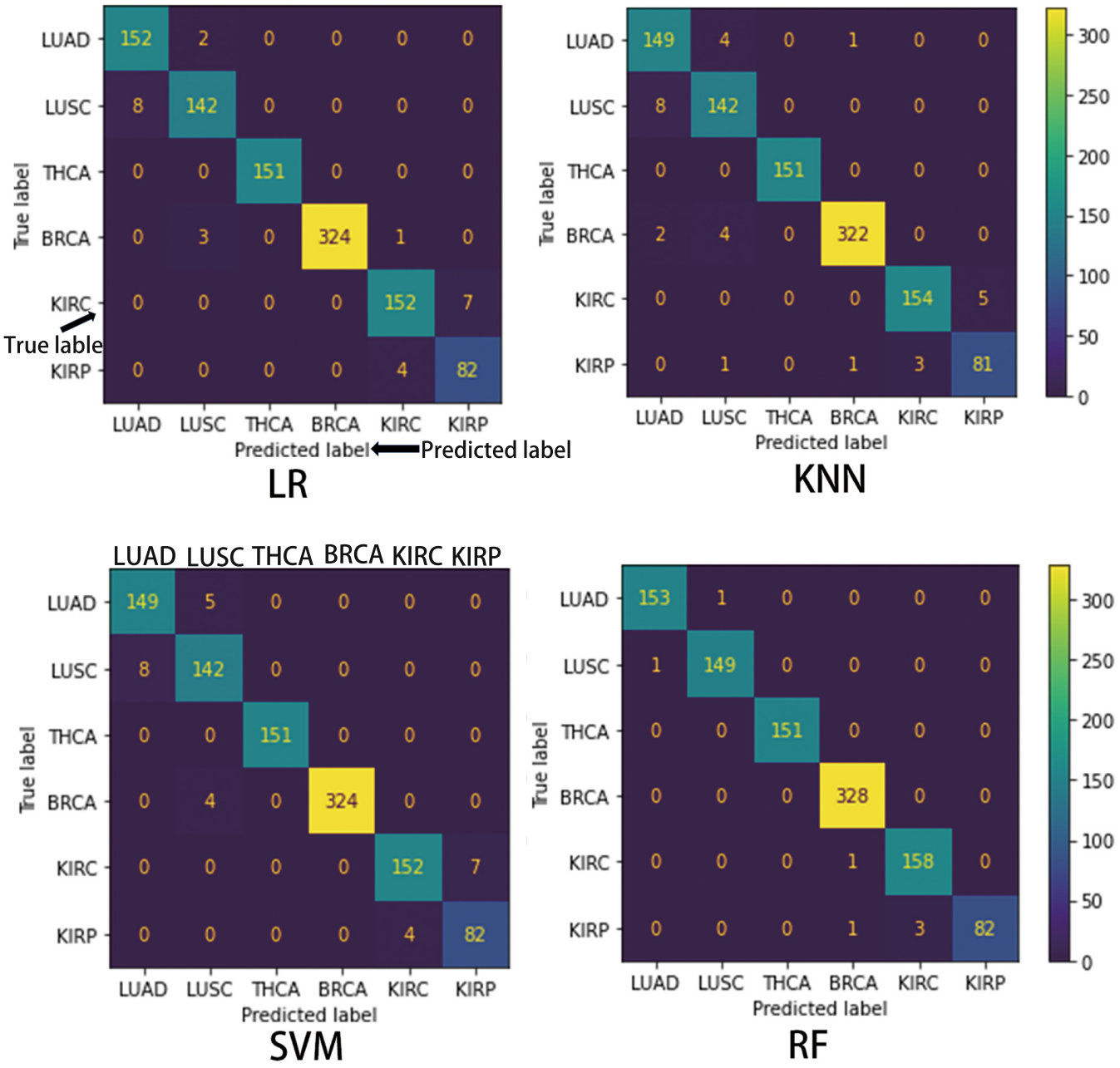

Fig. 6. Model Confusion Matrix

F1 score, $macro - F_1$, $micro - F_1$, and $weighted{-}F_1$ are as high as 99.78%, 99.37%, 98.90%, 99.86%, 99.12%, 99.12%, 99.32%, and 99.41%; the prediction results are even better than the popular deep learning or image processing disease research methods in recent years. For example, Shin et al. trained a deep learning model with foreign body surface-enhanced Raman spectroscopy signals from normal and LC cell lines; their classification accuracy was 95% lower than the RF model (99%) in this study [28]. Nasrullah et al. proposed a new deep learning multi-strategy model for the accurate diagnosis of malignant nodules in the lung; its sensitivity (94%) and specificity (91%) were assessed lower than 98% and 99% in the present study [29]. Shaffie et al. extracted the appearance and shape features of LC from CT images and later used a classifier to classify the feature data, which could accurately assist in diagnosis; however, its accuracy (92.55%) sensitivity (91.70%) and specificity (93.40%) were lower than 99.78%, 98.90% and 99.86% in this study [30]. Muhammad et al. developed a multilayer feed-forward neural network to identify 10 amino acid biomarkers in saliva that can affect gastric cancer to aid in diagnosis and treatment [31]. The accuracy, sensitivity, and specificity of this method were 92.27%, 94.8% and 90.2%, which were lower than 99.78%, 98.90% and 99.86% in this study [31].

Clinical practice mainly uses established medical instruments to determine the origin of cancer tissue. Even though machine learning is beginning to be applied to primary tissue detection, origins are rarely traced for specific metastatic cancers. This study is the first to propose the use of machine learning algorithms to trace the origin of metastatic LC, and then to detect the relationship between ”characteristic” DEGs and the tissue of origin to assist physicians in diagnosis and treatment, which has theoretical and medical implications.

Currently, clinical medicine often targets various antigenic substances (proteins, peptides, enzymes, etc.) by immunohistochemistry to visualize in situ cellular tissue and assist in localizing the origin of cancer [1], [32]. This method is highly sensitive to establish and even develop more practical immunoenzymatic techniques, but it is still labor-intensive and only applicable to small-scale data, making it difficult to overcome the accuracy bottleneck [8]. Although PET and CT have been applied in medical imaging diagnosis to accurately distinguish benign and malignant cancer lesions; their low accuracy of 24-40% and 20-27% of origin retrospectively has been a pressing problem [10]. Therefore, new inspection technology is urgently required to replace such labor-intensive or low-accuracy inspections. Avanzo et al. applied machine learning models for automatic segmentation of organs at risk for LC radiotherapy; stratifying patients according to local and distant recurrence risk to identify candidates for molecularly targeted therapy and immunotherapy, demonstrating high accuracy, convenience, generalizability, noninvasiveness, and reproducibility [33]. Machine learning algorithms assist in the diagnosis and treatment of LC with precision and convenience, which can effectively assess the urgency of patients' conditions and provide a reference for clinical practice.

Although the present study performed well in tracing the tissue of origin of metastatic LC, it still has some limitations. (1) Osteosarcoma is an important tissue of origin of metastatic LC, but the TCGA database does not include routine tissue samples for osteosarcoma, resulting in models predicting y values in only six categories (LUAD, LUSC, THCA, BRCA, KIRC, and KIRP). (2) This study was limited to tracing the organ of origin of the cancer and did not consider primary aggregates outside the organ. For example, a study based on unsupervised analysis of TCGA data from multiple genomics platforms showed that tumors can aggregate outside the tissue of origin, between unrelated cancer types, or as a heterogeneous group independent of known tumor types [34]. (3) This study only used the cross-validation internal validation method. Although the reproducibility of the model development process was effectively verified, its generalizability and portability are yet to be verified. Dual internal and external validation needs to be taken at the same time; external validation such as time-period validation, spatial validation, and domain validation is performed based on the good performance of the internal validation of the developed model.

The study only traced the tissue of origin of metastatic LC and did not map the overall metastatic trajectory of the cancer. Although the primary cancer tissue is the key diagnostic target in clinical practice, the historical and future metastatic trajectories are also important references for treatment and prognostic follow-up. Medeiros et al. found that understanding the mechanisms driving cancer metastasis is essential for identifying new biomarkers and therapeutic targets, and that lung metastasis from breast cancer is associated with 60-70% mortality, with metastatic trajectories and organophilia as important bases for diagnosis and treatment [35]. Physicians need to develop individualized treatment plans for different cancers and patient conditions. Therefore, our next research work will use machine learning methods to create models for mapping historical metastasis trajectories and future metastasis directions in patients with metastatic LC, and to fully consider signs of metastasis beyond organs.

ACKNOWLEDGMENT

We would like to thank Editage (www.editage.cn) for English language editing.

REFERENCES

[1] J. M. Carney, A. M. Kraynie, and V. L. Roggli, "Immunostaining in lung cancer for the clinician. commonly used markers for differentiating primary and metastatic pulmonary tumors." *Annals of the American Thoracic Society*, vol. 123, pp. 429–35, 2015.

[2] Y. Zhao, Z. Pan, S. Namburi, A. Pattison, A. Posner, S. Balachander, C. A. Paisie, H. V. Reddi, J. Rueter, A. J. Gill, S. Fox, K. P. Raghav, W. F. Flynn, R. W. Tothill, S. Li, R. K. M. Karuturi, and J. George, "Cup-ai-dx: A tool for inferring cancer tissue of origin and molecular subtype using rna gene-expression data and artificial intelligence," *EBioMedicine*, vol. 61, p. 103030, 2020.

[3] Y. Zhang, "Management of patients with tumors with unknown primary foci (in chinese)," *Foreign Medical Sciences (section of Surgery)*, vol. 29, no. 5, pp. 282–285, 2002.

[4] N. Pavlidis and K. Fizazi, "Cancer of unknown primary (cup)," *Critical Reviews in Oncology/Hematology*, vol. 54, no. 3, pp. 243–250, 2005.

[5] K. Kamposioras, G. Pentheroudakis, and N. Pavlidis, "Exploring the biology of cancer of unknown primary: breakthroughs and drawbacks. eur j clin invest," *Eur J Clin Invest*, vol. 43, no. 5, pp. 491–500, 2013.

[6] A. Richardson, R. Wagland, R. Foster, J. Symons, C. Davis, L. Boyland, C. Foster, and J. Addington-Hall, "Uncertainty and anxiety in the cancer of unknown primary patient journey: A multiperspective qualitative study," *BMJ Supportive and Palliative Care*, vol. 5, pp. 366–72, 11 2013.

[7] J. Selves, E. Long-Mira, M.-C. Mathieu, P. Rochaix, and M. Ilié, "Immunohistochemistry for diagnosis of metastatic carcinomas of unknown primary site," *Cancers*, vol. 10, no. 4, p. 108, 2018.

[8] X. Liu, L. Li, L. Peng, B. Wang, J. Lang, Q. Lu, X. Zhang, Y. Sun, G. Tian, H. Zhang, and L. Zhou, "Predicting cancer tissue-of-origin by a machine learning method using dna somatic mutation data," *Frontiers in Genetics*, vol. 11, p. 674, 2020.

[9] Z. Fu, X. Chen, X. Yang, and Q. Li, "Diagnosis of primary clear cell carcinoma of the vagina by 18f-fdg pet/ct," *Clin Nucl Med*, vol. 44, no. 4, pp. 332–333, 2019.

[10] V. Ambrosini, C. Nanni, D. Rubello, A. Moretti, G. B. and P Castellucci and M Farsad, L. Rampin, G. Fiorentini, R. F. and R Canini, and S. Fanti, "18f-fdg pet/ct in the assessment of carcinoma of unknown primary origin," *Clin Nucl Med*, vol. 111, no. 8, pp. 1146–55, 2006.

[11] Q. Wang, Q. Xu, J. Chen, C. Qian, X. Liu, and X. Du, "Establishment and evaluation of a novel molecular marker of tumor tissue origin (in chinese)," *China Oncology*, vol. 26, no. 10, pp. 801–812, 2016.

[12] D. Lu, J. Jiang, X. Liu, H. Wang, S. Feng, X. Shi, Z. Wang, Z. Chen, X. sheng Yan, H. Wu, and K. Cai, "Machine learning models to predict primary sites of metastatic cervical carcinoma from unknown primary," *Frontiers in Genetics*, vol. 11, p. 614823, 2020.

[13] Y. Zhong, L. Luo, X. Wang, and J. Yang, "Multi-factor stock selection model based on machine learning," *Engineering Letters*, vol. 29, no. 1, pp. 177–182, 2020.

[14] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, and Y.-L. Kuo, "Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree," *Journal of Medical Systems*, vol. 38, p. 106, 10 2014.

[15] R. Arian, A. Hariri, A. Dehnavi, A. Fassihi, and F. Ghasemi, "Protein kinase inhibitors' classification using k-nearest neighbor algorithm," *Computational Biology and Chemistry*, vol. 86, p. 107269, 04 2020.

[16] M. Tahir, M. Hayat, and M. Kabir, "Sequence based predictor for discrimination of enhancer and their types by applying general form of chou's trinucleotide composition," *Computer Methods and Programs in Biomedicine*, vol. 146, p. 69–75, 07 2017.

[17] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technology and Health Care*, vol. 24, no. 1, pp. 31–42, 2016.

[18] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgemann, V. H. van Berkel, and H. B. Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *International Journal of Medical Informatics*, vol. 108, pp. 1–8, 2017.

[19] J. Gutiérrez-Cárdenas and Z. Wang, "Classification of breast cancer and breast neoplasm scenarios based on machine learning and sequence features from lncrnas–mirnas-diseases associations," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 4, pp. 572–581, 2021.

[20] B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in african women using random forest classifier," *Cancer Treatment and Research Communications*, vol. 28, p. 100396, 2021.

[21] H. Zhou, Z. Zhong, M. Hu, and J. Huang, "Determining the steering direction in critical situations: A decision tree–based method," *Traffic Injury Prevention*, vol. 21, no. 6, pp. 395–400, 2020.

[22] J. Y. Lee, K.-s. Lee, B. K. Seo, K. R. Cho, O. H. Woo, S. E. Song, E.-K. Kim, H. Y. Lee, J. S. Kim, and J. Cha, "Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on mri," *European Radiology*, vol. 32, no. 1, pp. 650–660, 2022.

[23] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.

[24] M. Asif, M. M. Nishat, F. Faisal, R. R. Dip, M. H. Udoy, M. Shikder, R. Ahsan *et al.*, "Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease." *Engineering Letters*, vol. 29, no. 2, pp. 731–741, 2021.

[25] Z. DeVries, E. Locke, M. Hoda, D. Moravek, K. Phan, A. Stratton, S. Kingwell, E. K. Wai, and P. Phan, "Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and f1-score for the assessment of prognostic capability," *The Spine Journal*, vol. 21, no. 7, pp. 1135–1142, 2021.

[26] Y. Xu, L. Jiao, S. Wang, J. Wei, Y. Fan, M. Lai, and E. I.-c. Chang, "Multi-label classification for colon cancer using histopathological images," *Microscopy Research and Technique*, vol. 76, no. 12, pp. 1266–1277, 2013.

[27] R. Ullah, A. Naz, H. S. Akram, Z. Ullah, M. Tariq, A. Mithani, and A. Faisal, "Transcriptomic analysis reveals differential gene expression, alternative splicing, and novel exons during mouse trophoblast stem cell differentiation," *Stem Cell Research and Therapy*, vol. 11, no. 1, pp. 1–17, 2020.

[28] H. Shin, S. Oh, S. Hong, M. Kang, D. Kang, Y.-g. Ji, B. H. Choi, K.-W. Kang, H. Jeong, Y. Park *et al.*, "Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes," *ACS Nano*, vol. 14, no. 5, pp. 5435–5444, 2020.

[29] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated lung nodule detection and classification using deep learning combined with multiple strategies," *Sensors*, vol. 19, no. 17, p. 3722, 2019.

[30] A. Shaffie, A. Soliman, A. Eledkawy, V. van Berkel, and A. El-Baz, "Computer-assisted image processing system for early assessment of lung nodule malignancy," *Cancers*, vol. 14, no. 5, p. 1117, 2022.

[31] M. A. Aslam, C. Xue, M. Liu, K. Wang, and D. Cui, "Classification and prediction of gastric cancer from saliva diagnosis using artificial neural network." *Engineering Letters*, vol. 29, no. 1, pp. 10–24, 2020.

[32] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai, H. H. M. Ng, D. Wu, Y. T. F. Wee, J. C. T. Lim, J. Yeong, and T. K. H. Lim, "Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy," *Cancer Communications*, vol. 40, no. 4, pp. 135–153, 2020.

[33] M. Avanzo, J. Stancanello, G. Pirrone, and G. Sartor, "Radiomics and deep learning in lung cancer," *Strahlentherapie und Onkologie*, vol. 196, no. 10, pp. 879–887, 2020.

[34] E. Taskesen, S. M. Huisman, A. Mahfouz, J. H. Krijthe, J. De Ridder, A. Van De Stolpe, E. Van Den Akker, W. Verheagh, and M. J. Reinders, "Pan-cancer subtyping in a 2d-map shows substructures that are driven by specific combinations of molecular characteristics," *Scientific Reports*, vol. 6, no. 1, pp. 1–14, 2016.

[35] B. Medeiros and A. L. Allan, "Molecular mechanisms of breast cancer metastasis to the lung: clinical and experimental perspectives," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2272, 2019.