

An improved random forest algorithm for tracing the origin of metastatic renal cancer tissues

HaiDong Li¹, Tao Xie²

¹Department of Urology, Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College), Chengdu, China

²Department of Imaging, West China Second Hospital, Sichuan University, Sichuan, China

Submitted: 27 February 2023; **Accepted:** 28 June 2023

Online publication: 11 July 2023

Corresponding author:

Tao Xie

Department of Imaging
West China Second Hospital
Sichuan University
Sichuan, China

E-mail: 2126260185@qq.com

Arch Med Sci

DOI: <https://doi.org/10.5114/aoms/168973>

Copyright © 2023 Termedia & Banach

Abstract

Introduction: Tracing the histological origin of metastatic renal cancer (MRC) and locating the pathological root cause lead to precise treatment and improved prognosis.

Material and methods: A total of 3336 patient cases with clear tissue origins from The Cancer Genome Atlas (TCGA) database were screened as experimental data material and feature selection was performed using the differential expression method; the random forest (RF) algorithm was improved to establish a medical retrospective heterogeneous filtered feature selection random forest weighted (ReliefFk_RFw) model to locate tissue origins.

Results: The differential expression analysis method screened 60 signature genes with good differential expression for tracing tissue origins (kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, liver hepatocellular carcinoma, pancreatic adenocarcinoma). Compared with traditional machine learning (support vector machine, decision tree, RF) models, the ReliefFk_RFw algorithm increased the average accuracy from 98.65%, 98.79% and 98.57% to 99.53%, the average precision from 95.58%, 96.40% and 96.54% to 99.36%, and the average sensitivity from 97.03%, 96.61% and 96.76% to 98.89%, mean specificity from 99.50%, 99.39% and 99.35% to 99.90%, and mean F1 score from 96.30%, 96.50% and 96.64% to 99.11%. The highest accuracy in localizing the origin of primary pancreatic cancer was achieved with 100.00% for different models of retrospective metrics.

Conclusions: The improved ReliefFk_RFw model is best for comprehensive assessment and can be used to trace the origin of MRC tissue to assist physicians in diagnosis and treatment.

Key words: metastatic renal cancer, tissue origin tracing, ReliefFk_RFw algorithm, differential expression analysis, TCGA database.

Introduction

Metastatic renal cancer (MRC) remains unlocated after a face-to-face examination and there are no primary tumour features in the kidney [1]. More than 60% of tumours are associated with malignant metastases after the first consultation, of which approximately 10% metastasise to the patient's kidney [2]. MRC accounts for 1% to 1.5% of all malignant clinical cases [3], ranking 13th in malignancy [4] and 7th in mortality and recurrence rates [5]. Afrit *et al.* reported that the average

survival time of patients with MRC after multiple consultations was 3.8 months, with a 1-year survival rate of 18.7%; 5-year survival rate of 4.1% and 10-year survival rate of 2.8%. The authors concluded that prognostic status was governed by the tumour's domination by the biology of the primary tissue [6]. Metastasectomy and/or targeted therapies may play role in prolonging short-term survival rates but unfortunately cannot ensure a complete cure from metastatic cancer [7, 8]. Therefore, tracing the tissue origin of MRC is crucial for physicians to design treatment plans, enabling precise treatment and improving patient prognosis.

Metastatic kidney cancer is histologically heterogeneous and physicians are unable to improve the pathological mechanism of the tumour, which often leads to misdiagnosis and underdiagnosis. Immunohistochemistry is not only used as a concomitant immunosuppressant in steroid-dependent nephrotic syndrome [9] but is also the most commonly used clinical method to locate the primary tumour and can provide a practical reference for physicians in diagnosis and treatment [10]. However, at present, immunohistochemistry methods rely on small sample data, are labour-intensive, have moderate classification accuracy, and have high breakthrough costs, which need to be improved [11]. Computed tomography (CT) and positron emission tomography (PET) can be used to identify the origin of cancer tissue using medical imaging techniques with relative ease and speed, but with an accuracy of only 20–27% and 24–40%, the performance needs to be improved [12]. Therefore, the development of a new, efficient and convenient method to probe the pathological origin of cancer appears urgent and necessary in clinical applications.

A related study reported that metastatic cancer foci could not replicate the structural features of the primary tumour but could still carry the gene expression profile of the original tissue. Lu *et al.* screened 80 differentially expressed genes (DEGs) by raw signal analysis and constructed three medical prediction models – random forest

(RF), support vector machine (SVM) and neural network – then used them to distinguish four types of squamous cell carcinoma to trace the tissue origin of metastatic cervical cancer [13]. By analysing gene expression profiles, Wang *et al.* constructed a database containing 96 specific genes; the database was screened for key genetic information to produce molecular markers, followed by inference of the tissue origin of 22 common metastatic malignancies [14]. Zhao *et al.* trained a one-dimensional convolutional neural network medical prediction model to classify the primary tumour database to locate the tissue origin of unknown primary carcinoma (CUP) by constructing a one-dimensional convolutional neural network [15]. It is evident that the use of machine learning methods to construct medical tracing models and map gene expression profiles to infer the tissue principles of malignant tumours has become a popular aid to diagnosis and treatment. The incidence of metastasis to the kidney varies among different sites of malignancy, with lung adenocarcinoma, lung squamous cell carcinoma, liver hepatocellular carcinoma and pancreatic adenocarcinoma having a total metastasis rate of 45–65% and being the main malignant origin of MRC [16]. Therefore, the study will analyse data on gene expression of primary and MRCs from a bioinformatics perspective and construct five machine learning models to pinpoint the malignant origin of cancer.

Material and methods

Data collection and preparation

The study screened tumour samples of clear malignant origin from The Cancer Genome Atlas (TCGA) with the following data selection terms: Disease Type = Kidney renal clear cell carcinoma, Program Name = TCGA or TARGET, Data Category. The gene expression library (Table I) was constructed from samples identified as primary malignant tumours (sample type number="01") and routine samples (sample type number="11"). Osteosarcoma routine samples are missing, so they are not

Table I. Gene expression library

| Cohort abbreviation | Tumour type | Total | Case | Control |
|---------------------|---------------------------------------|-------|-----------|-----------|
| | | | Code = 01 | Code = 11 |
| KIRC | Kidney renal clear cell carcinoma | 602 | 530 | 72 |
| KIRP | Kidney renal papillary cell carcinoma | 320 | 288 | 32 |
| LUAD | Lung adenocarcinoma | 572 | 513 | 59 |
| LUSC | Lung squamous cell carcinoma | 550 | 501 | 49 |
| LIHC | Liver hepatocellular carcinoma | 1204 | 1091 | 113 |
| PAAD | Pancreatic adenocarcinoma | 88 | 64 | 24 |
| total | 6 | 3336 | 2987 | 349 |

included in the gene expression library. Finally, the mapping data of Gene Symbol and ENSG_ID were downloaded from GENCODE (<https://www.genCODE.org/human/>) and ENSG_ID was mapped to Gene Symbol uniformly [14].

RNA-Seq genetic data from patients should be standardised and normalised to include as many pathological subtypes as possible when tissue information with clear heterogeneity is present. The gene expression library (Table I) includes 6 (kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, liver hepatocellular carcinoma, pancreatic adenocarcinoma) tissue types and 3336 tissue samples of clear origin (including routine samples around the primary cancer), ranging from 88 to 1024 for each tissue sample. Each tissue sample was manually grouped, with the primary tissue screened as the case group (sample type number = "01") and the conventional tissue at the edge of the primary cancer as the control group (sample type number = "11").

Traditional modelling methods

SVM model

An SVM is a binary classification model whose basic model is a linear classifier defined by maximising the interval on the feature space, which distinguishes it from a perceptron; SVMs also include kernel tricks, which make them essentially non-linear classifiers. The learning algorithm for SVM is the optimisation algorithm for solving con-

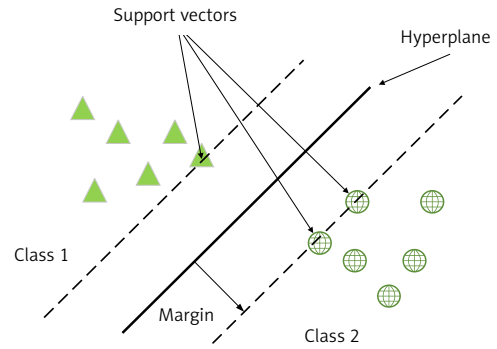


Figure 1. SVM schematic diagram

vex quadratic programming. As in Figure 1, support vectors (SV) are used to limit the width of the model edges. The core idea of the SVM is to divide the input fields into 2 sets of vectors in a multidimensional space and construct a hyperplane to separate the input vectors, which maximizes the boundary between the 2 sets of input vectors [17].

Decision tree (DT) model

A DT is a model that presents decision rules and classification results in a tree-like data structure. As an inductive learning algorithm, the focus is on taking seemingly disordered and disorganised known data and transforming them by some technical means into a tree model that can predict unknown data. Each path in the tree from the root node (the attribute that contributes most to the final classification result) to a leaf node (the final classification result) represents a rule for making a decision. The DT diagram (Figure 2) is as follows [18, 19].

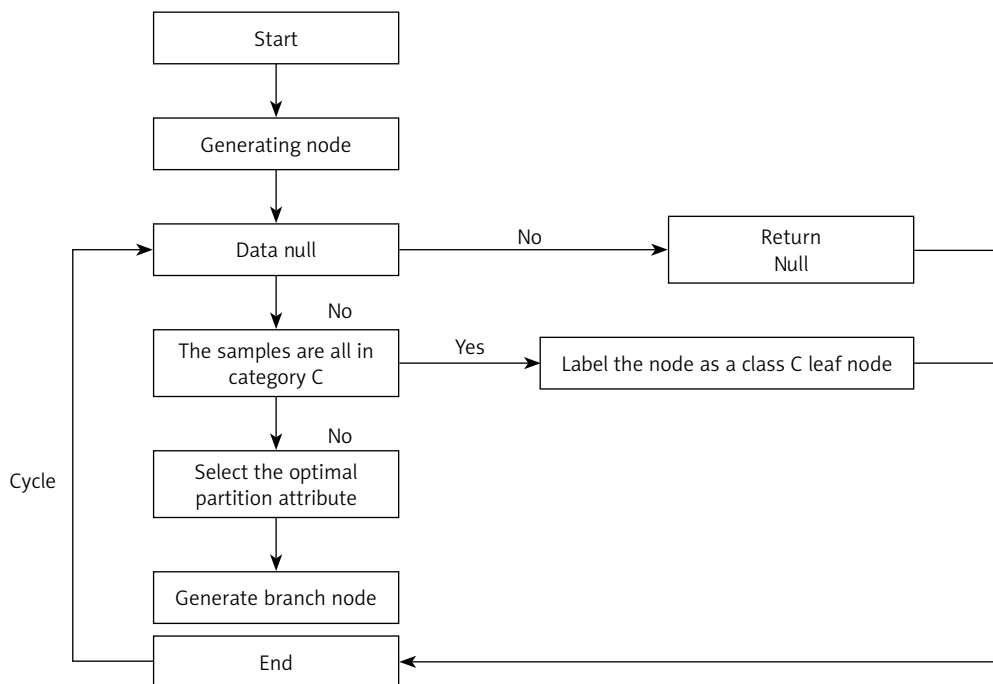


Figure 2. DT schematic diagram

1. Firstly, from the start position, divide all the data into one node, the root node.
2. Then go through the two steps in orange, with the orange indicating the judgement condition.
3. If the data are the empty set, jump out of the loop. If the node is the root node, return null; if the node is an intermediate node, mark the node as the class with the most classes in the training data.
4. If the samples all belong to the same class, skip the loop and the node is marked as that class.
5. If none of the judgment conditions marked in orange jump out of the loop, the node is considered for division. Since this is an algorithm, the division should not be done arbitrarily, but with efficiency and accuracy, choosing the best attribute division under the current conditions.
6. After going through the division in the previous step, a new node is generated, and then the judgment condition is cycled, and new branching nodes are continuously generated until all nodes have jumped out of the loop.
7. End. This results in a DT.

RF model

RF is an ensemble learning algorithm of the bagging type, which combines multiple weak classifiers (DTs) and the final result is obtained by voting or taking the mean, making the overall model result highly accurate and generalisable, as shown in the structure below (Figure 3) [20]. Its good results can be attributed to ‘random’, which allows it to resist over-fitting, and ‘forest’, which makes it more accurate [21].

Improved RF modelling approach

Improved Relief (ReliefFk) algorithm

The binary filtered feature selection (Relief) algorithm can be used to solve the feature weight calculation in binary classification problems. ReliefF is an optimisation of Relief that can be used in multi-classification scenarios [22]. In this section, the ReliefF algorithm is incorporated into the feature selection step of the RF construction DT, which can initially remove some of the features that negatively affect the model, and can also be used to alleviate the problem of false high or low classification accuracy caused by imbalance in medical data.

The core idea of the Relief algorithm is to use the correlation between positive (1) and negative classes (0) and features as a reference basis for assigning specific weights to each feature attribute. The basic idea is that the feature is preferred, a random sample is selected from the test set, the most similar sample of the same type (same spe) and the most similar sample of the different class (diff spe) are selected, and the mean distance between the feature and the same spe and diff spe samples is calculated successively. If there is a difference in the mean distance between the two classes, it means that the feature has a strong ability to differentiate between the samples of the current class. Conversely, if the mean distance values are the same or similar, it means that the distinguishing ability needs to be improved, and the weight of the feature should be reduced. The formula for calculating the weights can be expressed as [22, 23]:

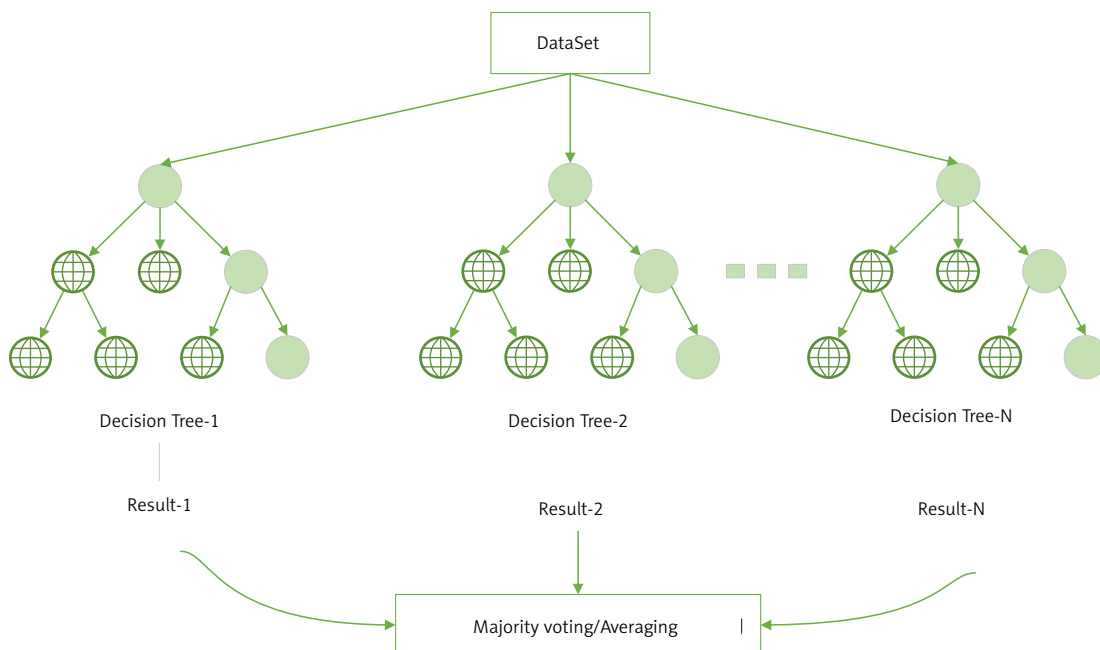


Figure 3. RF structure

$$W(a) = W(a) - \frac{\text{diff}(a, r, s)}{m} + \frac{\text{diff}(a, r, d)}{m}, \quad (1)$$

where $W(a)$ is the feature a weight.

The ReliefF algorithm is a modification of the Relief algorithm, based on which a multi-categorisation strategy is added. The basic idea is that instead of taking positive and negative samples, a sample is taken from each category and the weights of the features are calculated and updated. This not only solves the multi-category problem, but also indirectly reduces external noise interference and improves the stability of the algorithm. The weighting formula is updated to [22]:

$$W(a) = W(a) - \frac{\sum_{j=1}^k \text{diff}(a, r, H_j)}{mk} + \sum_{C \notin \text{class}(r)} \left[\frac{p(C)}{1 - p(\text{class}(r))} \sum_{j=1}^k \text{diff}(a, r, M_j(C)) \right], \quad (2)$$

$p(C)$ is the proportion of class C samples in the original data, $M_j(C)$ is the j nearest neighbour sample in class $C \notin \text{class}(r)$, and $\text{diff}(a, r_1, r_2)$ represents the difference in distance between samples r_1 and r_2 on feature a , which can be expressed as [23]:

$$\text{diff}(a, r_1, r_2) = \begin{cases} \frac{|r_1[a] - r_2[a]|}{\max(a) - \min(a)}, & \text{if } a \text{ is continuous} \\ 0, & \text{if } a \text{ is discrete and } r_1[a] = r_2[a] \\ 1, & \text{if } a \text{ is discrete and } r_1[a] \neq r_2[a] \end{cases}, \quad (3)$$

From Equation (3), the ReliefF algorithm averages the distance differences between the k samples closest to x_i in heterogeneous class C on feature a and multiplies this by the proportion of heterogeneous class C samples to all heterogeneous samples from x_i . This operation is repeated for all heterogeneous samples from x_i , resulting in the mean value of the distance differences between samples in heterogeneous class C on feature a . $W = \{w_1, w_2, w_3, \dots, w_n\}$ is the weight vector obtained by the ReliefF algorithm, and the features are sorted in reverse order by weight value.

The improved idea of the ReliefF algorithm is for the multi-category problem. The algorithm uses random function for random sampling among different categories of samples, which can achieve a better anti-noise interference effect and maintain certain stability under normal circumstances. In practice, however, the mean distance between heterogeneous samples and features a is large, and the mean distance between similar samples and features a is small; therefore, it is mainly the heterogeneous sample values that open up the weight gap (mean distance difference) and guar-

antee the stability of the weight gap. Therefore it is also necessary to slightly improve the sampling steps of the algorithm to obtain the ReliefFk algorithm: reduce the sampling weight of similar samples, increase the sampling weight of each class of heterogeneous samples, stabilize and refine the distance mean between heterogeneous samples and feature a , in order to allow a variety of heterogeneous samples with high decision power. The specific implementation steps are: Assume that n is the number of samples in the medical training set, where the number of positive class samples is n_+ and the number of negative class samples is n_- . Let the initial number of nearest neighbour samples in the training set be k , the number of nearest neighbour like samples be k_1 , the number of nearest neighbour dissimilar samples be k_2 and the initial $k_1 = k, k_2 = k$. If sample r is a positive class, then k_1 and k_2 can be expressed as:

$$k_1 = k * \frac{n_+}{n_+ + n_-}, \quad (4)$$

$$k_2 = k \frac{n_+ + n_-}{n_-}, \quad (5)$$

When the sample taken r is a negative class, k_1 and k_2 can be expressed as

$$k_1 = k * \frac{n_-}{n_+ + n_-}, \quad (6)$$

$$k_2 = k \frac{n_+ + n_-}{n_+}, \quad (7)$$

When calculating the weights of the characteristics a , the number of samples of the two categories of similarity and dissimilarity is changed from k to k_1 and k_2 respectively, and $k_1 < k, k_2 > k$; this theoretically reduces the proportion of samples of similarity and increases the proportion of samples of dissimilarity in each category, giving more decision power to the dissimilar samples and guaranteeing the stability of the distance difference (characteristic weights). The improved weighting formula is as follows.

$$W(a) = W(a) - \frac{\sum_{j=1}^{k_1} \text{diff}(a, r, H_j)}{mk_1} + \sum_{C \notin \text{class}(r)} \left[\frac{p(C)}{1 - p(\text{class}(r))} \sum_{j=1}^{k_2} \text{diff}(a, r, M_j(C)) \right], \quad (8)$$

Heterogeneous filtered feature selection random forest (ReliefFk_RF) algorithm

As shown in Figure 4, the ReliefFk algorithm is incorporated into the RF algorithm to form the new algorithm ReliefFk_RF algorithm: first-

ly, the formula for calculating the weights of the ReliefF algorithm is modified (reducing the sampling weight of similar samples k_1 and increasing the sampling weight of each class of dissimilar samples k_2), then some of the attribute features that have no effect on the classification results are initially screened out, and the remaining attribute features are ranked by their weights, and the training subset of the attributes of the training subset are divided equally into three intervals of high, medium and low weights. When RF enters the DT construction step, a subset of features with uniform classification effect is used to train the DT, and the set of features in the three intervals is drawn evenly, which can effectively avoid too many redundant features being repeatedly drawn, reduce the trouble of poor classification performance and enhance DT stability.

As shown in Figure 4 above, the model performs bootstrap sampling of the dataset to obtain various training subsets and OOB test sets (the out-of-bag dataset can be used for final model evaluation); the extracted dataset is then sorted

by the weights of the ReliefFk features to obtain the set of features with the attribute weights. The larger the weights of the features, the higher is the classification performance. The sorted attribute features are divided into three classes (high, medium and low) according to their classification performance. When the DT is constructed, the traditional form of random sampling is discarded and the features are drawn equally in the set of three classes to evenly consider the classification performance of each class of attributes and indirectly guarantee DT stability. The DT is constructed by continuously extracting features in sequence for multiple training subsets to form the final RF, and the algorithm1 pseudo-code is as follows (Algorithm 1).

The aim of the ReliefFk_RF algorithm is to modify the formula for calculating the weights of the ReliefF algorithm to remove the features that are not useful for the overall classification effect; then the remaining features are evenly differentiated according to their weights, so that the features in each weight range can be evenly sampled during

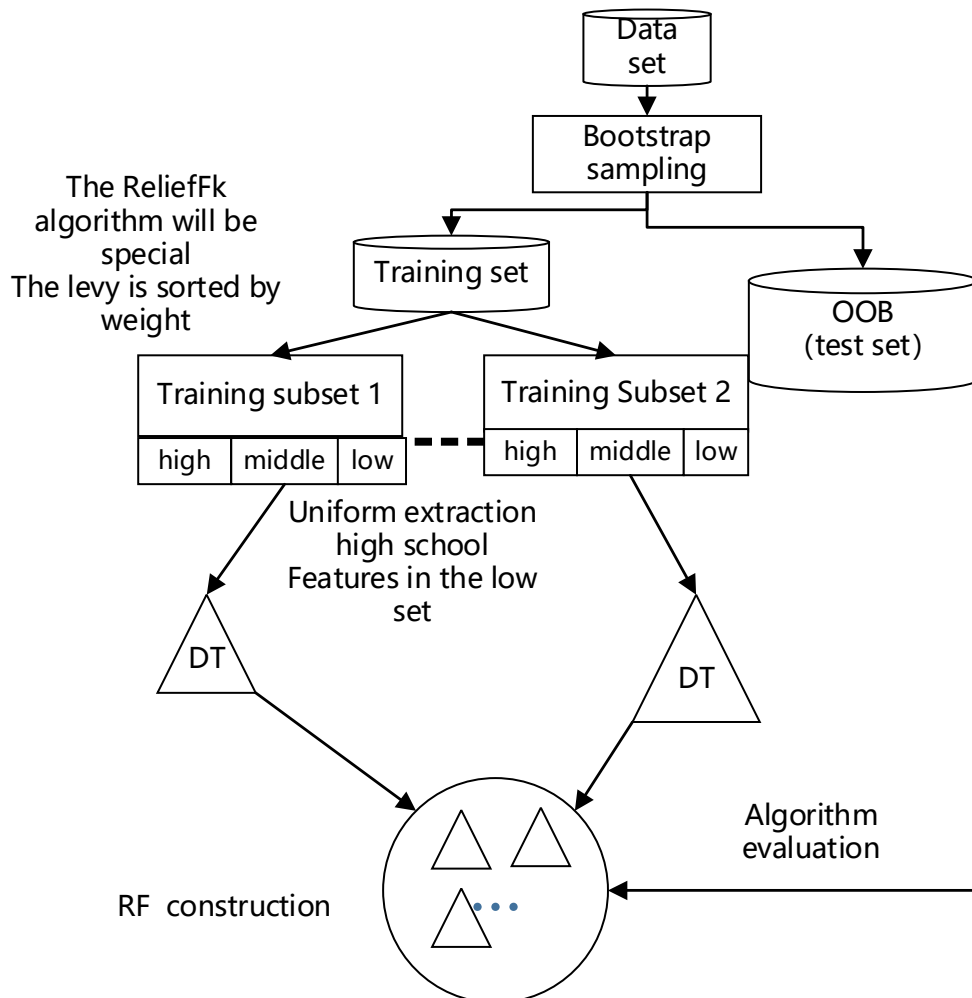


Figure 4. Flowchart of ReliefFk algorithm integrated into RF algorithm

Algorithm1: ReliefFk_RF Core steps pseudo-code

Input: Training set D ;
 Sampling frequency m , initial $m = 10$;
 Characteristic dimension n , feature weight set W_i
 Initially null;
 k_1, k_2 are the number of nearest neighbour
 samples drawn from each of the D categories.
 Output: Set of highly differentiated features W_h ;
 Feature set of medium distinction W_m ;
 Low distinction feature set W_l .

1. Set all feature weights $W_i = 0, i = 1, 2, \dots, n$ to 0;
2. for $i = 1$ to m do
 - 1) A single sample is randomly selected from D set r ;
 - 2) Find k_1 nearest neighbour $H_j(j = 1, 2, \dots, k)$ samples from the similar sample set of r , k_2 nearest neighbour $M_j(C)$ samples are found from the heterogeneous sample set of r ;
 - 3) for $a = 1$ to n (all features) do
 Updating feature weights with ReliefF algorithm:

$$W(a) = W(a) - \frac{\sum_{j=1}^{k_1} \text{diff}(a, r, H_j)}{mk_1} + \frac{\sum_{C \in \text{class}(r)} \left[\frac{p(C)}{1 - p(\text{class}(r))} \sum_{j=1}^{k_2} \text{diff}(a, r, M_j(C)) \right]}{mk_2}$$
- 4) end;
3. end;
4. Delete the feature with a weight of 0;
5. $W_h = W_p, i = 1, 2, \dots, n/3$;
 $W_m = W_p, i = n/3 + 1, \dots, 2n/3$;
 $W_l = W_p, i = 2n/3 + 1, \dots, n$;

DT construction, effectively avoiding too many redundant features being repeatedly selected, and also indirectly alleviating the problem of false high or false low classification accuracy, and stabilizing the DT accuracy.

Heterogeneous filtered feature selection random forest weighted (ReliefFk_RFw) algorithm

The traditional RF assumes equal weights for each DT, ignoring the variability among DTs. Based on this, this paper proposes a DT error weighting of the previous ReliefFk_RF algorithm to obtain the ReliefFk_RFw algorithm, with the weighting formula for a single DT as [24]:

$$w(i) = \frac{1/l^2(i)}{\sum_{j=1}^T 1/l^2(j)}, \quad (9)$$

where $w(i)$ is the weight of the i -th DT and $l^2(i)$ is the variance of the difference between the predicted and actual points in the training set of the i -th DT. A larger value of $l^2(i)$ represents a lower stability of DT prediction, so the weight is small-

er. The DT weighting parameters proposed in the study satisfy the normalisation feature [24].

$$\sum_{i=1}^T w(i) = \sum_{i=1}^T \left(\frac{1/l^2(i)}{\sum_{j=1}^T 1/l^2(j)} \right) = \frac{\sum_{i=1}^T 1/l^2(i)}{\sum_{j=1}^T 1/l^2(j)} = 1, \quad (10)$$

The ReliefFk_RFw algorithm regression algorithm process is:

- (1) Generate multiple stable DTs using the ReliefFk_RF algorithm.
- (2) When training the i -th DT, the distance variance $l^2(i)$ between all predicted and actual points in the DT is calculated and the weight values are obtained as:

$$w(i) = \frac{1/l^2(i)}{\sum_{j=1}^T 1/l^2(j)}, \quad (11).$$

- (3) The weighted predicted value $\hat{u}(x)$ of the ReliefFk_RFw algorithm is expressed as:

$$\hat{u}(x) = \sum_{i=1}^T w(i)Y_i, \quad (12).$$

Where Y_i is the initial predicted value of the i -th tree DT.

Model evaluation indicators

Accuracy, precision, sensitivity, specificity and F1 Score were used to measure the classification ability of a machine learning model. The larger the value of each index is, the better the model evaluation effect will be. The index value interval is [0,1]. The formula is as follows [25, 26]:

$$\text{accuracy} = (TP + TN)/(TP + FN + FP + TN), \quad (13)$$

$$\text{precision} = TP/(TP + FP), \quad (14)$$

$$\text{sensitivity} = TP/(TP + FN), \quad (15)$$

$$\text{specificity} = \frac{TN}{(TN + FP)}, \quad (16)$$

$$F_1 = 2(\text{precision} \times \text{sensitivity})/(\text{precision} + \text{sensitivity}), \quad (17).$$

Where TP is the number of positive cases predicted correctly, TN is the number of negative cases predicted correctly, FP is the number of positive cases predicted incorrectly, and FN is the number of negative cases predicted incorrectly [27]. F1 Score is the weighted average of the precision rate and sensitivity of the machine learning model, taking into account the precision rate and sensitivity of the model classification [28].

Results

Characteristic gene selection

At the initial stage of the study, there were 69,672 genes, which was of low representativeness. The *T. est* function was used to calculate the difference between each gene in primary tissue (group1=case) and normal control tissue (group2=control not determined). After that, the *p.adjust* function was called to calculate the fixed significance FDR of each gene and evaluate the differences of each gene. The Venn diagram (Fig-

ure 5) shows the DEG data relationship among six types of cancer. KIRC, KIRP, LUAD, LUSC, LIHC and PAAD contain 11493, 4433, 7447, 10530, 6253 and 3439 DEGs respectively. A total of 4930 (KIRC), 495 (KIRP), 1060 (LUAD), 2752 (LUSC), 1466 (LIHC), and 1234 (PAAD) “specific” DEGs were included in the differential expression information (only significantly expressed in one malignancy, Figure 5).

The “specific” DEGs in the top 5% of cancer differential expression values were extracted (247 KIRC, 25 KIRP, 53 LUAD, 138 LUSC, 74 LIHC

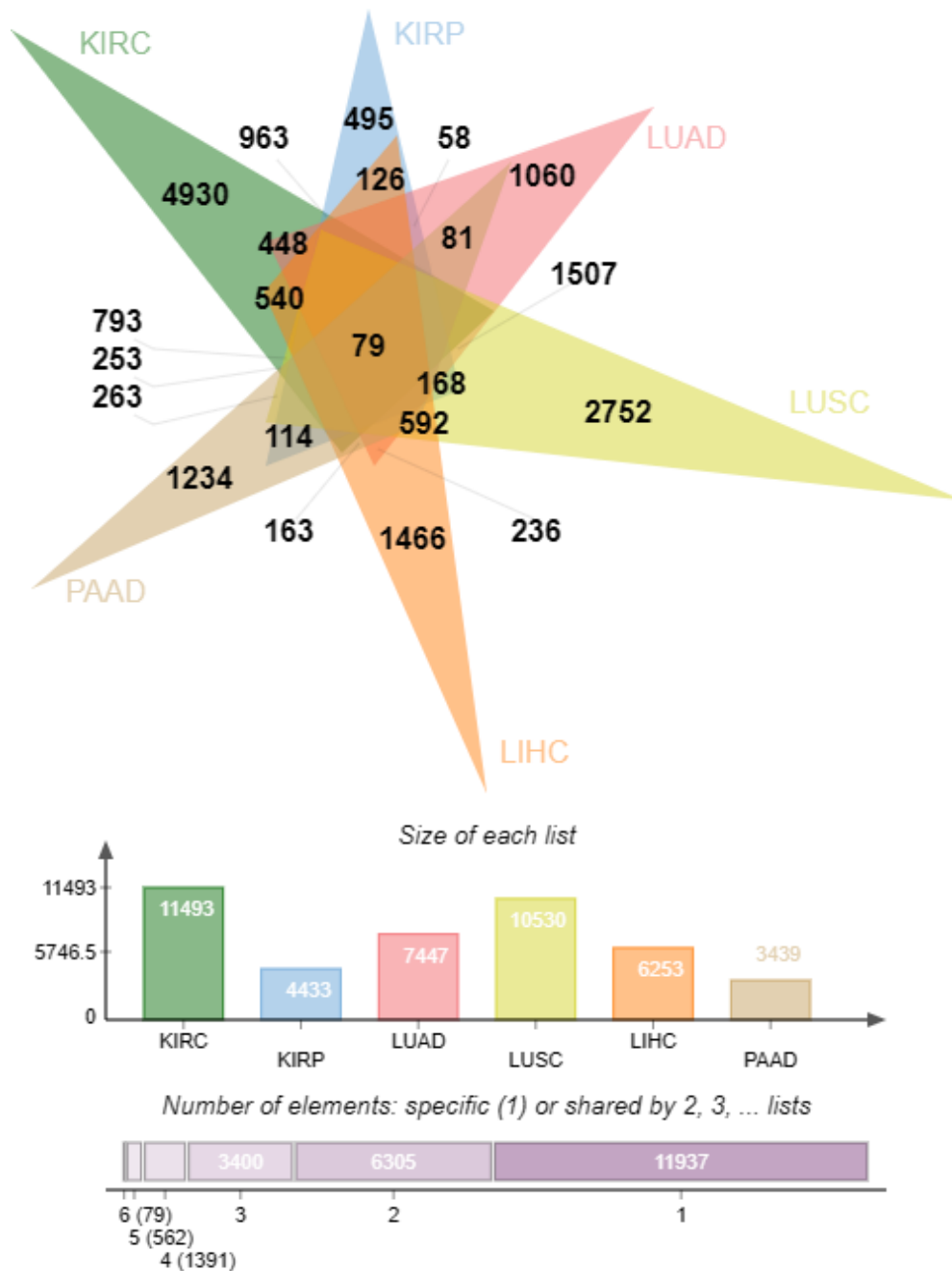


Figure 5. Venn diagram of DEGs

and 62 PAAD, for a total of 599 “specific (the “DEGs”). The genetic information following signature gene selection was genetic data for significant differences between each primary cancer and conventional controls, and the ‘signature’ DEGs were mutually exclusive between cancer types.

Model construction

Variable definitions: The independent variables consist of 599 ‘specific’ DEGs, all numeric variables. The dependent variable is cancer type and is named cancerType, which is a multicategorical variable, coded as KIRC = 1, KIRP = 2, LUAD = 3, LUSC = 4, LIHC = 5, PAAD = 6 in that order.

The initial genetic attribute information is in vertical coordinates and the data relationships are inverted. The information matrix was transposed using the T function and the train_test_split function was called 7/3 (70% as the training set and 30% as the validation set [29]) to split the dataset and serve subsequent experiments. Medical tracing models were constructed using SVM, DT, RF, and ReliefFk_RFW models, and the classification performance of the models was evaluated by accuracy, precision, sensitivity, specificity and F1 Score metrics. Finally, the best medical tracing models were selected, the reasons for their superiority explained, and “specific” DEGs influencing the tissue origin of MRC derived.

Analysis of model results

We used the modified best trace ReliefFk_RFW model to train the real tissue origin dataset to obtain a score table of the importance of “specific” DEGs’ expression in determining the primary location of MRC (Table 2). The top 10 “specific” DEGs of KIRC are ANXA2R, RP11-69E11.4, ATG16L2, TNFRSF4, DPRXP4, SAMD3, AL358340.1, COL4A5, PRRT2, and AC008735.1. The top 10 “feature” DEGs of KIRP are RASGRF2, DNM1, ASAP2, NR2F1, RASD1, TMEM204, PCDH1, THY1, MEIS3, TBX2 in order. The top 10 “specific” DEGs of LUAD are ACY3, ZNF239, DDIT4L, AL121949.1, FAM53A, FGB, KRTAP5-1, HMGB3, DOK5, and RASGEF1A. The top 10 “specific” DEGs of LUSC are TRIM16L, AP3B2, RN7SL399P, SERPINB13, TMEM117, AC009118.2, SFTA3, ETNK2, NAP1L4P2 and CERS3. The top 10 “characteristic” DEGs of LIHC are KCNK15, SMYD3, CIT, CLIP4, TNFSF4, GNAL, CRYAB, BOC, CEMIP, DMD. The top 10 “feature” DEGs of PAAD are DZIP1, FLRT3, NEK11, ZNF185, KCNJ16, SFRP1, NPR1, HAGHL, GALNT14, NR1H4. The above 60 “specific” DEGs have a strong influence on the localization of primary tumour in MRC, and can guide the diagnosis and treatment of metastatic disease (Table II).

In the genetic data set, SVM, DT, RF and ReliefFk_RFW models all had good predictive effects, and the quantitative evaluation indexes are shown in

Table II. Importance score of “feature” DEGs’ attribute

| KIRC(DEGs) | Score | KIRP(DEGs) | Score | LUAD(DEGs) | Score |
|--------------|----------|------------|-----------|------------|----------|
| ANXA2R | 7.20E-02 | RASGRF2 | 2.78E-01 | ACY3 | 1.77E-02 |
| RP11-69E11.4 | 1.82E-02 | DNM1 | 8.29E-03 | ZNF239 | 6.21E-04 |
| ATG16L2 | 1.32E-02 | ASAP2 | 4.54E-03 | DDIT4L | 5.99E-04 |
| TNFRSF4 | 4.39E-03 | NR2F1 | 1.38E-03 | AL121949.1 | 4.57E-04 |
| DPRXP4 | 4.28E-03 | RASD1 | 8.28E-04 | FAM53A | 3.59E-04 |
| SAMD3 | 4.01E-03 | TMEM204 | 4.55E-04 | FGB | 3.24E-04 |
| AL358340.1 | 2.79E-03 | PCDH1 | 3.81 E-04 | KRTAP5-1 | 2.88E-04 |
| COL4A5 | 2.04E-03 | THY1 | 2.52E-04 | HMGB3 | 2.41E-04 |
| PRRT2 | 1.29E-03 | MEIS3 | 2.43E-04 | DOK5 | 1.79E-04 |
| AC008735.1 | 1.01E-03 | TBX2 | 2.03E-04 | RASGEF1A | 1.29E-04 |
| LUSC(DEGs) | Score | LIHC(DEGs) | Score | PAAD(DEGs) | Score |
| TRIM16L | 9.09E-02 | KCNK15 | 2.09E-02 | DZIP1 | 2.12E-02 |
| AP3B2 | 6.76E-02 | SMYD3 | 1.87E-02 | FLRT3 | 1.22E-02 |
| RN7SL399P | 3.44E-02 | CIT | 1.58E-02 | NEK11 | 9.36E-03 |
| SERPINB13 | 2.78E-02 | CLIP4 | 9.34E-03 | ZNF185 | 8.21E-03 |
| TMEM117 | 2.26E-02 | TNFSF4 | 8.91E-03 | KCNJ16 | 7.32E-03 |
| AC009118.2 | 1.89E-02 | GNAL | 5.48E-03 | SFRP1 | 5.09E-04 |
| SFTA3 | 1.77E-02 | CRYAB | 1.59E-03 | NPR1 | 1.76E-04 |
| ETNK2 | 1.09E-02 | BOC | 7.85E-04 | HAGHL | 7.57E-04 |
| NAP1L4P2 | 8.78E-03 | CEMIP | 7.81E-04 | GALNT14 | 7.21E-04 |
| CERS3 | 8.21E-03 | DMD | 7.42E-04 | NR1H4 | 5.42E-04 |

Tables III and IV. The model performance evaluation results showed that the ReliefFk_RFw model had the highest accuracy score (99.53% on average), which increased by 0.88%, 0.74% and 0.96% compared with SVM, DT and RF, respectively. The ReliefFk_RFw model has the highest precision evaluation results except LIHC, and the average score of the ReliefFk_RFw model is up to 3.78%, 2.96% and 2.82% compared with SVM, DT and RF respectively. In the specificity evaluation results, the ReliefFk_RFw model had the highest score (99.90%) except LIHC, which was higher than RF (99.35%), DT (99.39%) and SVM (99.50%). The ReliefFk_RFw model had the highest score in the sensitivity and F1 Score evaluation results of 6 kinds of malignant tumours. In summary, the ReliefFk_RFw model has the best comprehensive performance among the 5 evaluation results, followed by RF, DT and SVM. The reason is that SVM, based on regression, is unable to process nonlinear and highly correlated data information. In this study, there was a biological correlation between genetic attribute variables, so the model was not effective. The RF algorithm can process nonlinear highly correlated data, and effectively overcome the defect of the easy fitting DT single tree, so the overall effect of the model has been improved to a certain extent.

The RF algorithm has several drawbacks when processing data with a large and unbalanced number of features: first, the excessive number of fea-

tures has a certain degree of redundancy, which may have a negative impact on the model prediction results, and thus disturb the training ability of the medical prediction model [22]. Second, it is easy to unbalance the selection of high-weight features or low-weight features during the construction of the model DT, resulting in false high classification and low classification accuracy, and unstable and unrepresentative results [22]. Third, after DT construction by the traditional RF model, the default DT weight is equal and the difference between DTs is ignored, resulting in low overall accuracy [24]. Based on this, the improved ReliefF algorithm was used in this study to sort features according to their weights, extract feature subsets stratified according to the weights of features, construct the CART tree for feature selection, then reweight, and construct the optimized ReliefFk_RFw medical prediction model. Data presented in Tables III and IV show the best evaluation effect.

The results of cancer prediction (horizontal view of Tables III and IV) showed that the tracing effect of PAAD tissue origin was the best, with each evaluation index reaching 100% under different models. LIHC came in second, but LUSC and KIRP needed to improve their precision, sensitivity and F1 Score (as low as 90.20%). According to the comprehensive data results, there were misjudgements between LUSC and LUAD due to the pathological similarities, and KIRP and KIRC were

Table III. Comparison of machine learning classification performance

| Type | Accuracy (%) | | | | Precision (%) | | | |
|------|--------------|--------|--------|--------------|---------------|--------|--------|--------------|
| | SVM | DT | RF | ReliefFk_RFw | SVM | DT | RF | ReliefFk_RFw |
| KIRC | 97.82 | 98.24 | 97.98 | 99.21 | 96.11 | 98.04 | 97.04 | 98.29 |
| KIRP | 97.83 | 98.44 | 97.21 | 99.05 | 92.28 | 94.23 | 92.55 | 100.00 |
| LUAD | 99.02 | 98.24 | 98.38 | 99.52 | 95.04 | 93.51 | 94.80 | 99.25 |
| LUSC | 98.09 | 98.59 | 98.29 | 99.55 | 90.20 | 94.24 | 94.84 | 99.33 |
| LIHC | 99.11 | 99.23 | 99.58 | 99.82 | 100.00 | 98.38 | 100.00 | 99.29 |
| PAAD | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Ave | 98.65 | 98.79 | 98.57 | 99.53 | 95.58 | 96.40 | 96.54 | 99.36 |

Table IV. Comparison of machine learning classification performance

| Sensitivity (%) | | | | Specificity (%) | | | | F1 score (%) | | | |
|-----------------|--------|--------|---------------|-----------------|--------|--------|---------------|--------------|--------|--------|---------------|
| SVM | DT | RF | Relief-Fk_RFw | SVM | DT | RF | Relief-Fk_RFw | SVM | DT | RF | Relief-Fk_RFw |
| 95.44 | 96.26 | 95.62 | 99.39 | 99.21 | 99.28 | 99.09 | 99.81 | 95.77 | 97.14 | 96.32 | 98.84 |
| 95.09 | 94.23 | 95.12 | 95.22 | 99.24 | 99.44 | 99.21 | 100.00 | 93.66 | 94.23 | 93.82 | 97.55 |
| 98.51 | 96.09 | 96.20 | 99.52 | 99.09 | 98.82 | 99.09 | 99.90 | 96.74 | 94.78 | 95.49 | 99.38 |
| 94.33 | 94.89 | 94.62 | 99.20 | 99.44 | 98.89 | 98.68 | 99.78 | 92.22 | 94.56 | 94.73 | 99.26 |
| 98.78 | 98.21 | 98.99 | 100.00 | 100.00 | 99.88 | 100.00 | 99.92 | 99.39 | 98.29 | 99.49 | 99.64 |
| 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 97.03 | 96.61 | 96.76 | 98.89 | 99.50 | 99.39 | 99.35 | 99.90 | 96.30 | 96.50 | 96.64 | 99.11 |

also difficult to distinguish due to similar tissues. Because the differentiation of such interference depends heavily on the type and amount of input genetic data, misclassification is also common in the absence of multi-class and multi-number experimental queues.

Discussion

In this study, the ReliefFk_RFw model was created by rewriting the pseudo-code at the bottom of the algorithm, and the effects of the primary tissue of MRC were compared with the SVM, DT and RF models. The ReliefFk_RFw model has the best overall evaluation results, which effectively verifies the existing hypothesis that “metastatic tissue retains genetic characteristics of origin”, and can assist in guiding clinical diagnosis and treatment. Faced with diagnosis of MRC, physicians can create an improved ReliefFk_RFw model, input patient-specific “DEGs” in order of importance, and evaluate the output by classification to locate the primary origin of metastatic tissue. Assuming that the output was cancerType=4 (primary origin =LUSC), the physician recommended a combination treatment regimen dominated by lung squamous cell carcinoma and assisted by kidney cancer, and was on high alert for re-metastasis of tissue origin during follow-up. The clinical treatment plan and follow-up plan are developed according to the specific disease occurrence of patients and doctors’ experience. The output results of this study model are only for auxiliary reference.

This study combined machine learning and differential expression analysis techniques to initially screen out “specific” DEGs to reduce the model prediction dimension, increase the data relevance, and effectively improve the accuracy of tracing results. Although the traditional machine learning model can carry out feature screening, its understanding of bioinformatics information is only at the data level, unable to analyse the correlation of genetic data, so the traceability performance is limited. Differential expression analysis could select 599 “specific” DEGs with significant expression from 69,672 genetic genes of the initial malignant tumour according to the up-down-regulated gene scores of fixed cancer. Selected “specific” DEGs can control the physiological mechanism and biological development form of cancer, and are the main driving force for regulating the life activity process of cancer cells [30]. At the initial stage of the study, differential expression analysis can effectively distinguish the biological differences between different primary cancers and surrounding conventional tissues; at the middle stage, the machine learning algorithm can distinguish the genetic differences between

different primary cancers; at the later stage, the improved ReliefFk_RFw model can achieve a leap-frog improvement in the precision of the primary tissue trace procedure. Therefore, in feature engineering, differential expression analysis is adopted to screen “specific” DEGs, which can fill the gap in the biology of machine learning algorithms and provide an experimental reference for machine learning engineers.

In this study, five evaluation indexes (accuracy, precision, sensitivity, specificity and F1 Score) were used to comprehensively measure the multi-category classification ability of the machine learning model. The average accuracy, precision, sensitivity, specificity and F1 Score of the ReliefFk_RFw model are as high as 98.89%, 99.90%, 99.11%, 99.53% and 99.36%, and the retrospective effect is even better than that of the popular medical aid methods of image parsing and deep learning in recent years. For example, Jin *et al.* developed a deep learning system based on preoperative CT images of gastric cancer patients, which was used to predict lymph node metastases of multiple lymph nodes. The results of image analysis showed excellent prediction accuracy in the external validation queue, but the sensitivity (0.743) and specificity (0.936) were still lower than 0.989 and 0.999 of the ReliefFk_RFw model in this study [31]. Doppalapudi *et al.* performed a feature-importance analysis to understand how factors related to lung cancer affected survival in patients. The classification accuracy of the recurrent neural network (RNN) model in the deep learning field is 71.18%, lower than the 99.53% of the ReliefFk_RFw model in this study [32]. Liu *et al.* explored the value of deep learning (DL) based on global digital mammography in the prediction of microcalcification malignancy in the Breast Imaging Reporting and Data System (BI-RADS) 4. The combined DL model showed good sensitivity and specificity of 85.3% and 91.9%, respectively, when predicting the four types of malignant microcalcifications of BI-RADS in the test data set, but these values were still lower than the 98.9% and 99.9% of the ReliefFk_RFw model in this study [33].

In diagnosis and treatment, mature medical instruments are often used to determine the origin of tissues. Although machine learning methods have begun to be used for primary tracing, precise localization of specific cancers has not yet been studied. This study proposed for the first time that the improved ReliefFk_RFw model could be used to predict the origin of MRC tissue, accurately match the specific “DEGs” and fix the genetic relationship between the primary cancer, and assist physicians in diagnosis and treatment, which has considerable medical significance. At present, clinical practices mostly use immunohistochemical methods

to lock specific antigens (such as proteins, peptides, enzymes, etc.) and colour primary sites to assist in tracing the origin [34]. Although suitable immunoenzymatic techniques have been developed to establish a high degree of sensitivity with the primary tissue, they are still only suitable for small-scale data analysis and are labour-intensive, making it difficult to overcome the accuracy bottleneck [11]. CT and PET are the most convenient and rapid primary tissue tracing techniques at the present stage; they are powerful tools in medical imaging and can effectively distinguish benign and malignant tissues, but their low accuracy rates of 20–27% and 24–40% need to be improved [12]. Therefore, a new machine learning model tracing technology is needed to replace such low-accuracy, labour-intensive and data-dependent traditional detection technology. Tian et al. used RF and SVM models to screen biomarkers, and used minimum absolute contraction and selection operator (LASSO) regression analysis to construct multi-gene signatures. Univariate and multivariate Cox regression analysis was used to explore the relationship between clinical features and prognosis, which promoted the personalized management of patients with kidney cancer, showing advantages such as high accuracy, convenience, universality, non-invasiveness and repeatability [35]. It can be seen that the participation of machine learning models in the diagnosis and treatment of kidney cancer is convenient, accurate and efficient, and can effectively guide medical decision-making and assist diagnosis and treatment.

Although the improved ReliefFk_RFW model can effectively trace the origin of MRC, there are still some limitations in experimental studies. (1) Due to the lack of conventional tissue sample data of osteosarcoma in the TCGA database, there are only 6 cancerType values predicted by the model (KIRC, KIRP, LUAD, LUSC, LIHC, PAAD). (2) The study was limited to the organ-specific origin of MRC and did not consider the aggregation of origin outside the organ. For example, a study of unsupervised analysis of TCGA multi-genomic data suggested that primary malignant tissue may aggregate outside the organ of origin, or between unrelated organs, or as a heterogeneous tissue independent of existing medical tumour types [36]. (3) In addition, only internal cross-validation was used in this study. Although the repeatability of the ReliefFk_RFW model was verified, the portability and generalization of ReliefFk_RFW need to be verified. Future research will adopt internal and external verification methods. On the basis of ensuring model stability (internal verification), external verification such as spatial verification, time period verification and domain verification will be carried out.

Acknowledgments

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

Conflict of interest

The authors declare no conflict of interest.

References

- Bhanvadia RR, Baky FJ, Ashbrook CQ, et al. Pathologic fracture in metastatic kidney cancer: identifying widening disparities and opportunity for quality improvement. *Urol Oncol* 2022; 40: 384.e1-8.
- Rocca CJ, Cherqui S. Gene transfer to mouse kidney in vivo. *Methods Mol Biol* 2019; 1937: 227-34.
- Labriola MK, George DJ. Setting a new standard for long-term survival in metastatic kidney cancer. *Cancer* 2022; 128: 2058-60.
- Pal SK, Agarwal N. Kidney cancer: finding a niche for girentuximab in metastatic renal cell carcinoma. *Nat Rev Urol* 2016; 13: 442-3.
- Singh P, Agarwal N, Pal SK. Sequencing systemic therapies for metastatic kidney cancer. *Curr Treat Options Oncol* 2015; 16: 316.
- Afrit M, Yahyaoui Y, Bouzouita A, et al. Traitements médicaux des cancers du rein localement avancés et/ou métastatiques [Medical therapies for locally advanced/metastatic kidney cancer]. *Presse Med* 2015; 44: 135-43.
- Banyra O, Tarchynets M, Shulyak A. Renal cell carcinoma: how to hit the targets? *Cent European J Urol* 2014; 66: 394-404.
- Ouzaid I, Capitanio U, Staehler M, et al. Young academic urologists kidney cancer working group of the european association of urology. surgical metastasectomy in renal cell carcinoma: a systematic review. *Eur Urol Oncol* 2019; 2: 141-9.
- Zhong H, Li HY, Zhou T, et al. Rituximab therapy in adults with steroid-dependent nephrotic syndrome. *Arch Med Sci* 2023; 19: 577-85.
- Yuki M, Machida N, Sawano T, et al. Investigation of serum concentrations and immunohistochemical localization of α 1-acid glycoprotein in tumor dogs. *Vet Res Commun* 2011; 35: 1-11.
- Solomon JP, Benayed R, Hechtman JF, et al. Identifying patients with NTRK fusion cancer. *Ann Oncol* 2019; 30 (Suppl_8): viii16-22.
- Kang X, Xia H, Skudder-Hill L, et al. Magnetic resonance imaging (MRI) and positron emission tomography (PET)/computed tomography features of atypical teratoid/rhabdoid tumors: case series and review. *J Child Neurol* 2022; 37: 1003-9.
- Lu D, Jiang J, Liu X, et al. Machine learning models to predict primary sites of metastatic cervical carcinoma from unknown primary. *Front Genet* 2020; 11: 614823.
- Wang Q, Xu Q, Chen J, et al. Establishment and evaluation of a novel molecular marker of tumor tissue origin. *Chin J Cancer* 2016; 26: 801-12.
- Zhao Y, Pan Z, Namburi S, et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *Ebiomedicine* 2020; 61: 103030.
- Carney JM, Kraynie AM, Roggli VL. Immunostaining in lung cancer for the clinician. Commonly used markers for differentiating primary and metastatic pulmonary tumors. *Ann Am Thorac Soc* 2015; 12: 429-35.

17. Kalafi EY, Nor NAM, Taib NA, et al. Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia Biol (Praha)* 2019; 65: 212-20.
18. Park EY, Yi M, Kim HS, et al. A decision tree model for breast reconstruction of women with breast cancer: a mixed method approach. *Int J Environ Res Public Health* 2021; 18: 3579.
19. Gadot R, Anand A, Lovin BD, et al. Predicting surgical decision-making in vestibular schwannoma using tree-based machine learning. *Neurosurg Focus* 2022; 52: E8.
20. Zhou Y, Yu Z, Liu L, et al. Construction and evaluation of an integrated predictive model for chronic kidney disease based on the random forest and artificial neural network approaches. *Biochem Biophys Res Commun* 2022; 603: 21-8.
21. Cheng X, Wang J, Zheng T, et al. Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Arch Med Sci* 2022; 18: 1208-20.
22. Gao R. Improvement of Random forest algorithm and its application in medical diagnosis system. Nanjing University of Posts and Telecommunications, 2020.
23. Wang L, Liu S. Research on improved Random Forest Algorithm based on mixed sampling. *J Nanjing Univ Posts Telecommun (Natural Science Edition)* 2022; 01: 1673-5439.
24. Yanping Xu. Urban air quality prediction model based on improved Random Forest algorithm: a case study of Chongqing Municipality. Chongqing Technology and Business University 2021.
25. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019; 394: 861-7.
26. Dong L, He W, Zhang R, et al. Artificial intelligence for screening of multiple retinal and optic nerve diseases. *JAMA Netw Open* 2022; 5: e229960.
27. Choi E, Kim D, Lee JY, et al. Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram. *Sci Rep* 2021; 11: 10246.
28. Zhu S, Lu B, Wang C, et al. Screening of common retinal diseases using six-category models based on efficient-net. *Front Med (Lausanne)* 2022; 9: 808402.
29. Zheng X, Huang R, Liu G, et al. Development and verification of a predictive nomogram to evaluate the risk of complicating ventricular tachyarrhythmia after acute myocardial infarction during hospitalization: a retrospective analysis. *Am J Emerg Med* 2021; 46: 462-8.
30. Cheng Q, Chen X, Wu H, et al. Three hematologic/immune system-specific expressed genes are considered as the potential biomarkers for the diagnosis of early rheumatoid arthritis through bioinformatics analysis. *J Transl Med* 2021; 19: 18.
31. Jin C, Jiang Y, Yu H, et al. Deep learning analysis of the primary tumour and the prediction of lymph node metastases in gastric cancer. *Br J Surg* 2021; 108: 542-9.
32. Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: deep learning approaches. *Int J Med Inform* 2021; 148: 104371.
33. Liu H, Chen Y, Zhang Y, et al. A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. *Eur Radiol* 2021; 31: 5902-12.
34. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun (Lond)* 2020; 40: 135-53.
35. Tian M, Wang T, Wang P. Development and clinical validation of a seven-gene prognostic signature based on multiple machine learning algorithms in kidney cancer. *Cell Transplant* 2021; 30: 963689720969176.
36. Taskesen E, Huisman SM, Mahfouz A, et al. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci Rep* 2016; 6: 24949.