frontiers
in Plant Science

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fpls.2022.839044

CRIA: An Interactive Gene Selection Algorithm for Cancers Prediction Based on Copy Number Variations

*Qiang Wu and Dongxi Li**

College of Data Science, Taiyuan University of Technology, Taiyuan, China

Genomic copy number variations (CNVs) are among the most important structural variations of genes found to be related to the risk of individual cancer and therefore they can be utilized to provide a clue to the research on the formation and progression of cancer. In this paper, an improved computational gene selection algorithm called CRIA (correlation-redundancy and interaction analysis based on gene selection algorithm) is introduced to screen genes that are closely related to cancer from the whole genome based on the value of gene CNVs. The CRIA algorithm mainly consists of two parts. Firstly, the main effect feature is selected out from the original feature set that has the largest correlation with the class label. Secondly, after the analysis involving correlation, redundancy and interaction for each feature in the candidate feature set, we choose the feature that maximizes the value of the custom selection criterion and add it into the selected feature set and then remove it from the candidate feature set in each selection round. Based on the real datasets, CRIA selects the top 200 genes to predict the type of cancer. The experiments' results of our research show that, compared with the state-of-the-art related methods, the CRIA algorithm can extract the key features of CNVs and a better classification performance can be achieved based on them. In addition, the interpretable genes highly related to cancer can be known, which may provide new clues at the genetic level for the treatment of the cancer.

Keywords: gene selection, correlation-redundancy analysis, interaction analysis, copula entropy, copy number variations (CNVs), cancers prediction

Edited by:
*Weihua Pan,
Agricultural Genomics Institute at Shenzhen (CAAS), China*

Reviewed by:
*Lin Wang,
Tianjin University of Science and Technology, China
Jiazhou Chen,
South China University of Technology, China*

****Correspondence:***
*Dongxi Li
dxli0426@126.com*

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *19 December 2021*
Accepted: *19 January 2022*
Published: *21 March 2022*

Citation:
Wu Q and Li D (2022) CRIA: An Interactive Gene Selection Algorithm for Cancers Prediction Based on Copy Number Variations. Front. Plant Sci. 13:839044. doi: 10.3389/fpls.2022.839044

INTRODUCTION

The occurrence of many diseases is associated with genome structural variations. Human genome variations include single nucleotide polymorphisms (SNPs), copy number variations (CNVs), etc. The copy number variations refer to the amplification, deletion, and more complex mutations in the genome of DNA fragments longer than 1 kb in length (Redon et al., 2006). SNPs account for 0.5% of the human genome, and nearly 12% of the human genome often undergoes copy number variations (Redon et al., 2006). Copy number variations have become an important genomic variation, and their role in the pathogenesis of complex human diseases is still being revealed.

The close relationship between CNVs and diseases has been widely recognized. Numerous studies have demonstrated that not a few human diseases involved copy number variations that could change the diploid status of particular locus of the genome (Zhang et al., 2016). The Flierl research team found that the higher vulnerability of Parkinson's disease and stress sensitivity of neuronal precursor cells carry an α-synuclein gene triplication (Flierl et al., 2014).

Grangeon et al. (2021) discovered that early-onset cerebral amyloid angiopathy and Alzheimer Disease (AD) were related to an amyloid precursor protein (App) gene triple amplification. Breunis et al. (2008) reported that the copy number variations of FCGR2C gene promoted idiopathic thrombocytopenic purpura. Zheng et al. (2017) found that the low copy number of FCGR3B was associated with lupus nephritis in a Chinese population. And Pandey et al. (2015) revealed that there was both direct and indirect evidence suggesting abnormalities of glycogen synthase kinase (GSK)-3β and β-catenin in the pathophysiology of bipolar illness and possibly schizophrenia (SZ). Moreover, several neuro-developmental relevant genes, such as A2BP1, IMMP2, and AUTS2, were reported with mutational CNVs (Elia et al., 2010). In 2006, a research team composed of researchers from the United Kingdom, Japan, the United States, Canada and other countries studied 270 individuals in 4 groups of the HapMap project, and constructed the first-generation copy number variations map of the human genome, and obtained 144 CNVs region (about 12% of the size of the human genome). Among them, 285 CNVs regions were related to the occurrence of known diseases (Redon et al., 2006). Compared with SNPs, CNVs regions contained more DNA sequences, disease sites and functional elements, which could provide more clues for disease research. The publication of this map has become an important tool for studying the complex structural variations of the human genome and human diseases.

Cancer is a kind of diseases which involves uncontrolled abnormal cell growth and can spread to other tissues (Du and Elemento, 2015). The formation and development of cancer are also associated with copy number variations (Frank et al., 2007). Van Bockstal et al. (2020) discovered that HER2 gene amplification had a relationship with a bad result in invasive breast cancer and the amplification of heterogeneous HER2 had been described in 5–41% of breast cancer. The experimental results of Buchynska et al. (2019) shown that assessment of copy number variations of HER-2/neu, c-MYC and CCNE1 genes revealed their amplification in the tumors of 18.8, 25.0 and 14.3% of endometrial cancer patients, respectively. Heo et al. (2020) pointed out that CNVs were related to the mechanism of lung cancer development through a comparative experiment. Moreover, Tian et al. (2020) found that CNVs of CYLD, USP9X and USP11 were significantly associated with the risk of colorectal cancer. A latest global cancer burden data released by the International Agency for Research on Cancer(IARC) of the WHO showed that the number of patients with new cancer and cancer deaths in China ranked first around the world with 4.57 million patients with new cancer and 3 million cancer deaths, accounting for 23.7 and 30%, respectively. It is of great significance to investigate cancer causes and its treatment. Because the gene expression patterns in cancer tumor have high specificity (Liang et al., 2020), studying the relationship between these genetic information and cancer can provide a new idea for investigating the causes of cancer and help in early cancer diagnosis.

However, few studies have utilized machine learning (ML) or deep learning (DL) methods to use copy number variations data for the prediction of various cancer types. Zhang et al. (2016) used the mRMR and IFS methods to select 19 features from the 24,174 gene features of the copy number variations data set, which contained a total of 3,480 samples of 6 cancer types. They applied the Dagging algorithm with ten-fold cross-validation to classify cancer. But the accuracy of final result only reached 75%. Liang et al. (2020) used CNA_origin for cancer classification on the same data set. CNA_origin was an intelligent combined deep learning network, which was composed of two parts—a stacked autoencoder and a one-dimensional convolutional neural network with multiscale convolutional kernels. CNA_origin eventually had an overall accuracy of 83.81% on ten-fold cross-validation. But it could not identify which gene features were more important and more closely associated with cancer classification.

Here, we present an improved novel computational algorithm named CRIA, which can successfully classify cancer based on the information of gene CNVs levels from the same dataset. CRIA can not only effectively perform dimensionality reduction operation on high-dimensional gene CNVs data, which can improve the efficiency of the experiment, but also selects specific gene features closely related to cancer, making it clear which genes are more important in cancer classification. And the final results had higher classification accuracy than the state-of-the-art methods.

The rest sections of this paper are structured as follows: Section Background describes the theoretical background and related work. Section The Proposed Method-CRIA introduces the collection of CNVs dataset, the implementation details and performance of the proposed algorithm. Section Results and Discussions demonstrates the experimental results on CNVs dataset and the performance comparison with the recent methods. In section Conclusions, we summarize the conclusions and point out our future work.

BACKGROUND

In section Information Theory, we introduce some basic information theory knowledge, which is the core of our proposed algorithm. Before proposing our algorithm, we summarize some related work on gene selection methods and point out their drawbacks in section Related Work.

Information Theory

As early as 1948, Shannon's information theory had been proposed (Shannon, 2001), providing an effective method for measuring random variables' information. The entropy can be understood as a measure of the uncertainty of a random variable (Cover and Thomas, 1991). The greater the entropy of a random variable, the greater its uncertainty. If $X = \{x_1, x_2, ..., x_l\}$ is a discrete random variable, its probability distribution is $p(x) = P(X = x), x \in X$. The entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{l} p(x_i)\log p(x_i) \tag{1}$$

where $p(x_i)$ is the probability of x_i. Here the base of log is 2 and specified that $0 \log 0 = 0$.

If $Y = \{y_1, y_2, ..., y_m\}$ is a discrete random variable, $p(x_i, y_j)$ is the joint probability of X and Y. Then, their joint entropy is defined as:

$$H(X, Y) = -\sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j) \tag{2}$$

If the random variable X is in a given situation, the uncertainty measure of the variable Y can be defined by conditional entropy as follows:

$$H(Y|X) = H(X, Y) - H(X) = -\sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i) \tag{3}$$

where $p(y_j|x_i)$ is the conditional probability of Y under the condition of X.

Definition 1: Mutual information (MI) (Cover and Thomas, 1991) is a measure of useful information in information theory. It can be regarded as the amount of information shared by two random variables. MI can be defined as:

$$\begin{aligned} I(X;\,Y) &= \sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \\ &= H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) \end{aligned} \tag{4}$$

Definition 2: Conditional mutual information (CMI) (Cover and Thomas, 1991) can be defined as the amount of information that shared by variables X and Y, if a discrete random variable $Z = \{z_1, z_2, ..., z_n\}$ is known.

$$\begin{aligned} I(X;\,Y\,|Z) &= \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n} p(z_k)p(x_i, y_j|z_k) \log \frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)} \\ &= H(Y|Z) - H(Y\,|X, Z) \end{aligned} \tag{5}$$

Definition 3: Joint mutual information (JMI) (Cover and Thomas, 1991) measures the amount of information shared by a joint random variable $(X_1, X_2, \cdots X_q)$ and Y and it can be defined as:

$$\begin{aligned} &I(X_1, X_2, ..., X_q;\,Y) = \sum_{x_1 \in X_1}\sum_{x_2 \in X_2} \cdots \sum_{x_q \in X_q}\sum_{y \in Y} p(x_1, x_2, ..., x_q, y) \\ &\log \frac{p(x_1, x_2, ..., x_q, y)}{p(x_1, x_2, ..., x_q)p(y)} \\ &= H(X_1, X_2, ..., X_q) - H(X_1, X_2, ..., X_q|Y) \end{aligned} \tag{6}$$

Definition 4: Interaction gain (IG) had been introduced by Jakulin (2003), Jakulin and Bratko (2004) to measure the amount of information shared by three random variables at the same time. Mutual information can be regarded as a two-way interaction gain. IG is defined as follows:

$$IG(X;\,Y;\,Z) = I(X;\,Y;\,Z) = I(X, Y;\,Z) - I(X;\,Z) - I(Y;\,Z) \tag{7}$$

Related Work

The irrelevant features and redundant features existed in high-dimensional data will damage the performance of the learning algorithm and reduce the efficiency of the learning algorithm. Therefore, the dimensionality reduction of features is one of the most common methods of data preprocessing (Orsenigo and Vercellis, 2013) and its purpose is to reduce the training time of the algorithm and improve the accuracy of final results (Bennasar et al., 2015). In recent years, the research of gene selection methods based on mutual information has received wide attention from scholars. Best individual gene selection (BIF) (Chandrashekar and Sahin, 2014) is the simplest and fastest filtering gene selection algorithm, especially suitable for high-dimensional data.

Battiti utilized the mutual information (MI) between features and class labels $[I(f_i;\,c)]$ to measure the relevance and the mutual information between features $[I(f_i;\,f_s)]$ to measure the redundancy (Battiti, 1994). He proposed the Mutual Information Gene selection (MIFS) criterion and it is defined as:

$$J_{MIFS}(f_i) = I(f_i;\,c) - \beta \sum_{f_s \in \Omega_S} I(f_i;\,f_s),\ f_i \in F - \Omega_S \tag{8}$$

where F is the original feature set, Ω_S is the selected feature subset, F $- \Omega_S$ is the candidate feature subset and c is the class label. β is a configurable parameter to determine the trade–off between relevance and redundancy. However, β is set experimentally, which results in an unstable performance.

Peng et al. (2005) proposed the Minimum-Redundancy Maximum-Relevance (MRMR) criterion and its evaluation function is defined as:

$$J_{mRMR}(f_i) = I(f_i;\,c) - \frac{1}{|n_s|} \sum_{f_s \in \Omega_S} I(f_i;\,f_s),\ f_i \in F - \Omega_S \tag{9}$$

where $|n_s|$ is the number of selected features.

Similarly, other gene selection methods that consider relevance between features and the class label and redundancy between features are concluded, such as Normalized Mutual Information Gene selection (NMIFS) and Conditional Mutual Information (CMI), and they were proposed by Estévez et al. (2009) and Liang et al. (2019) respectively. Their evaluation function are defined as follows:

$$J_{NMIFS}(f_i) = I(f_i;\,c) - \frac{1}{|n_s|} \sum_{f_s \in \Omega_S} \frac{I(f_i;\,f_s)}{\min(H(f_i), H(f_s))}, f_i \in F - \Omega_S \tag{10}$$

$$J_{CMI}(f_i) = I(f_i;\,c) - \frac{H(f_i\,|c)}{H(f_i)} \sum_{f_s \in \Omega_S} \frac{I(f_s;\,c)I(f_i;\,f_s)}{H(f_s)H(c)},\ f_i \in F - \Omega_S \tag{11}$$

where $H(f_i)$ is the information entropy and $H(f_i\,|c)$ is the conditional entropy.

Many gene selection algorithms based on information theory tend to use mutual information as a measure of relevance, which will bring a disadvantage that mutual information tends to select features with more discrete values (Foithong et al., 2012). Thus, the symmetrical uncertainty (Witten and Frank, 2002) (a normalized form of mutual information, *SU*) is adopted to solve this problem. The symmetrical uncertainty can be described as:

$$SU(f_i; c) = \frac{2I(f_i; c)}{H(f_i) + H(c)} \tag{12}$$

The *SU* can redress the bias of mutual information as much as possible and scale its values to [0,1] by penalizing inputs with large entropies. It will make the performance of gene selection better. Same as MI, for any two features f_{i1} and f_{i2}, if $SU(f_{i1}; c) > SU(f_{i2}; c)$, due to more information can be provided by the former, f_{i1} and c are more relevant. If $SU(f_{i1}; f_s) > SU(f_{i2}; f_s)$, owing to the information shared by f_{i1} and f_s being more and providing less information, f_{i1} and f_s have greater redundancy.

Additionally, these gene selection algorithms mentioned above fail to take the feature interaction into consideration. After relevance and redundancy analysis, one feature deemed useless may interact with other features to provide more useful information. Especially in complicated biology systems, molecules interacting with each other, they work together to express physiological and pathological changes. If we only consider relevance and redundancy but ignore the feature interaction in data analysis, we may miss some useful features and affect the analysis results (Chen et al., 2015).

Sun et al. (2013), Zeng et al. (2015), and Gu et al. (2020), respectively proposed a gene selection method using dynamic feature weights: Dynamic Weighting-based Gene selection algorithm (DWFS), Interaction Weight based Gene selection algorithm (IWFS) and Redundancy Analysis and Interaction Weight-based gene selection algorithm (RAIW). All of them employ the symmetric uncertainty to measure the relevance between features and the class label, and exploit the three-dimensional interaction information (mentioned at **Information Theory Definition 4**) to measure the interaction between two features and the class label. The evaluation functions are defined as follow:

$$J_{DWFS}(f_i) = SU(f_i; c) \times w_{DWFS}(f_i),\ f_i \in -\Omega_S \tag{13}$$

$$J_{IWFS}(f_i) = w_{IWFS}(f_i) \times [1 + SU(f_i; c)],\ f_i \in F - \Omega_S \tag{14}$$

$$\begin{aligned} J_{RAIW}(f_i) &= SU(f_i; c) \times [1 - \alpha SU(f_i; f_s)] \\ &\quad \times w_{RAIW}(f_i), f_i \epsilon F - \Omega_s \end{aligned} \tag{15}$$

where $w(f_i)$ is the weight of each feature and its initial value is set to 1, α is a redundancy coefficient and the value is relevant to the number of dataset's features, f_s is one of features in the selected feature subset. In each round, the feature weight $w(f_i)$ is updated by their interaction weight factors.

$$\begin{aligned} w_{DWFS}(f_i) &= w_{DWFS}(f_i{}') \times [1 + CR(f_i, f_s)] = w_{DWFS}(f_i{}') \\ &\quad \times [1 + 2\tfrac{I(f_i; c|f_s) - I(f_i; c)}{H(f_i) + H(c)}] \\ &= w_{DWFS}(f_i^{'}) \times [1 + 2\tfrac{I(f_i; f_s; c)}{H(f_i) + H(c)}] \end{aligned} \tag{16}$$

$$\begin{aligned} w_{IWFS}(f_i) &= w_{IWFS}(f_i^{'}) \times IW(f_i, f_s) \\ &= w_{IWFS}(f_i{}') \times [1 + \frac{I(f_i; f_s; c)}{H(f_i) + H(f_s)}] \end{aligned} \tag{17}$$

$$\begin{aligned} w_{RAIW}(f_i) &= w_{RAIW}(f_i^{'}) \times [1 + If(f_i, f_s, c)] \\ &= w_{RAIW}(f_i{}') \times [1 + \frac{2I(f_i; f_s; c)}{H(f_i) + H(f_s) + H(c)}] \end{aligned} \tag{18}$$

where $w(f_i{}')$ denotes the feature weight of the previous round, $I(f_i; c|f_s)$ is the conditional mutual information of f_i and c when f_s is given. $I(f_i; f_s; c)$ is three-dimensional interaction information. However, we can find that although DWFS and IWFS take into account relevance and interaction, they ignore the redundancy between features. Correlation, redundancy and interaction are all taken into account by RAIW, but there is a no reasonable value for α in a specific dataset.

Furthermore, some other gene selection methods about three-way mutual information are listed and their evaluation function are defined as follows, such as Composition of Feature Relevance (CFR) (Gao et al., 2018a), Joint Mutual Information Maximization (JMIM) (Bennasar et al., 2015), Dynamic Change of Selected Feature with the class (DCSF) (Gao et al., 2018b) and Max-Relevance and Max-Independence (MRI) (Wang et al., 2017).

$$J_{CFR}(f_i) = \sum_{f_s \in \Omega_S} I(f_i; c|f_s) + \sum_{f_s \in \Omega_S} I(f_i; f_s; c),\ f_i \in F - \Omega_S \tag{19}$$

$$J_{JMIM}(f_i) = \max[\min_{f_s \in \Omega_S}(I(f_i, f_s; c))],\ f_i \in F - \Omega_S \tag{20}$$

$$\begin{aligned} J_{DCSF}(f_i) &= \sum_{f_s \in \Omega_S} I(f_i; c|f_s) + \sum_{f_s \in \Omega_S} I(f_s; c|f_i) \\ &\quad - \sum_{f_s \in \Omega_S} I(f_i; f_s),\ f_i \in F - \Omega_S \end{aligned} \tag{21}$$

$$\begin{aligned} J_{MRI}(f_i) &= I(f_i; c) + \sum_{f_s \in \Omega_S} I(f_i; c|f_s) \\ &\quad + \sum_{f_s \in \Omega_S} I(f_s; c|f_i),\ f_i \in F - \Omega_S \end{aligned} \tag{22}$$

where $I(f_i, f_s; c)$ is the joint mutual information of f_i, f_s and c. $I(f_s; c|f_i)$ is the conditional mutual information of f_s and c when f_i is given. However, these algorithms only take into account three-way mutual information among features and the class label, and none of them considers relevance, redundancy and three-dimensional mutual information between features at the same time, which will affect the performance of these algorithms.

THE PROPOSED METHOD-CRIA

In section CNVs Dataset, we firstly introduce the collection of datasets and the process of data processing specifically. Subsequently, we redress other methods' shortcomings and propose an improved gene selection algorithm called CRIA

TABLE 1 | The number of samples for each cancer type in this dataset.

Class label	Histology	Samples	Percentage
1	UCEC (Uterine corpus endometrial carcinoma)	443	12.73%
2	KIRC (Kidney renal clear cell carcinoma)	490	14.08%
3	OV (Ovarian serous cystadenocarcinoma)	562	16.15%
4	GBM (Glioblastoma multiforme)	563	16.18%
5	COAD/READ (Colon adenocarcinoma/Rect-um adenocarcinoma)	575	16.52%
6	BRCA (Breast invasive carcinoma)	847	24.34%
Total		3,480	100%

in section The Proposed Algorithm and give it a specific implementation in section Algorithm Implementation. Finally, in section Verify the Performance of CRIA, we verify the performance of CRIA by comparing the experimental results of CRIA and other 8 algorithms on 5 datasets.

CNVs Dataset

The datasets of copy number variations in different cancer types used in this paper comes from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/, Release 2/4/2013) (Cerami et al., 2012; Ciriello et al., 2013; Gao et al., 2013). The copy number values in the dataset are generated by Affymetrix SNP 6.0 arrays for the set of samples in the cancer genome atlas (TCGA) study (Liang et al., 2020). The preprocessing analysis of the dataset is performed with GISTIC (Beroukhim et al., 2007). There are 11 cancer types in the cBioPortal database with the largest sample number was 847 and the smallest sample was 135. In order to avoid affecting the experimental results due to the large difference in the number of samples of cancer types, we only select six cancer types with more than 400 samples as our experimental data. The details of six cancer types are listed in **Table 1**, and totally there are 3480 samples in our experimental dataset.

In this dataset, each sample consists of labels for 24174 genetic cytobands. The CNV spectrum is divided into five regions/labels by setting four thresholds in cancer algorithm (Mermel et al., 2011). Then, the CNV values are discretized into 5 different values—"-2," "-1," "0," "1," "2," where "-2" denotes the deletion of both copies (possibly homozygous deletion), "−1" means the deletion of one copy (possibly heterozygous deletion), "0" corresponds to exactly two copies, i.e., no gain/loss (diploid), "1" denotes a low-level copy number gain and "2" means a high-level copy number amplification (Ciriello et al., 2013).

The CNVs values are preprocessed to the range of [−1,1] with Equation (23).

$$val' = \frac{val}{|val|_{\max}} \tag{23}$$

where *val* is the value of gene copy number variations of each sample, $|val|_{\max}$ is the maximum absolute value of gene CNVs among samples and val' is the recalculated value.

The Proposed Algorithm

In section Related Work, we analyze the 11 gene selection methods and point out their shortcomings. In view of the defects of these algorithms, we propose an improved gene selection algorithm to redress their shortcomings: Correlation-Redundancy and Interaction Analysis based gene selection algorithm (CRIA). This method uses the symmetric uncertainty (SU) to measure the correlation between features and the class label and the redundancy among features. In addition, copula entropy is introduced to measure the feature interaction information. Different from the three-way interaction of DWFS, IWFS and RAIW, the proposed algorithm considers the interaction between the candidate feature and the entire set of selected features, instead of being limited to the three-dimensional interaction.

As we know, Shannon's definition of mutual information aims at a pair of random variables, and it measures the correlation between two random variables. Therefore, naturally, many researchers have tried to study how to extend the definition of mutual information from two variables to multivariate situations. In 2011, Ma and Sun published a paper (Ma and Sun, 2011), which contributed to the entropy of information theory. They defined a new concept of entropy in that paper, called Copula Entropy. Copula Entropy is defined on a set of random variables and conformed to symmetry. Therefore, it is a multivariate extension of mutual information, which can be utilized to measure the full-order, non-linear correlation among random variables. They proved the equivalence between copula entropy and the concept of mutual information, which was, mutual information was equal to negative copula entropy (Ma and Sun, 2011).

The copula entropy of $\overrightarrow{x} = (x_1, x_2, ..., x_N) \in R^N$ is defined as:

$$H_c(\overrightarrow{x}) = -\int c(\overrightarrow{u}) \log c(\overrightarrow{u}) d\overrightarrow{u} \tag{24}$$

where $\overrightarrow{x}$ are random variables with marginal functions $\overrightarrow{u} = [F_1, F_2, ..., F_N]$ and copula density $c(\overrightarrow{u}) = \frac{d^N C(\overrightarrow{u})}{du_1 du_2 ... du_N}$.

Thus, we can use interaction factor IF_{CRIA}, which is defined in Equation (25) to measure the interaction between the candidate feature and the selected feature subset. The meaning of IF_{CRIA} is that, after adding a random candidate feature f_i into the selected feature subset Ω_S, the amount of interaction information increased relative to the original selected feature subset. So, the bigger value of IF_{CRIA}, the bigger value of interaction between f_i and Ω_S. In each round of calculation, we are supposed to choose the variable that maximizes the IF_{CRIA} value.

$$IF_{CRIA} = \frac{H_c(\Omega_S, f_i, c)}{H_c(\Omega_S, c)} \tag{25}$$

Where c is the target class label.

Integrating the correlation between the features and the class label and the redundancy between features that we improved, and the interaction factor we proposed, we can define the evaluation

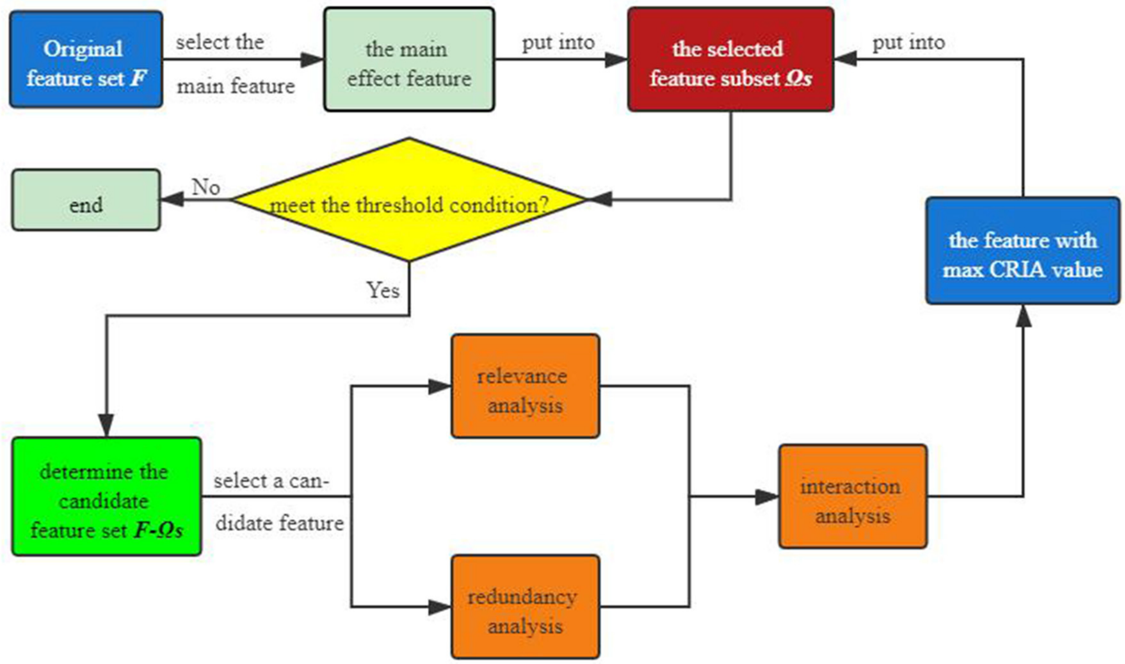
Original feature set F
select the main feature
the main effect feature
put into
the selected feature subset Ωs
put into
end
No
meet the threshold condition?
Yes
the feature with max CRIA value
relevance analysis
determine the candidate feature set F-Ωs
select a candidate feature
interaction analysis
redundancy analysis

FIGURE 1 | General flow chart of the proposed algorithm.

criterion of a candidate feature as follows:

$$
\begin{aligned}
J_{CRIA}(f_i) &= \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \tfrac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times IF_{CRIA}\} \\
&= \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \tfrac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times \tfrac{H_c(\Omega_S,f_i,c)}{H_c(\Omega_S,c)}\}
\end{aligned}
\tag{26}
$$

For the equation (26), we can see that the proposed algorithm can take into account the relevance between the candidate feature and the class label, redundancy and multi-dimensional interaction among the candidate feature and the selected features at the same time. The formula $SU(f_i, c)$ can denote the relevance and $\frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i, f_s)$ calculates the redundancy. Also, the formula $\frac{H_c(\Omega_S,f_i,c)}{H_c(\Omega_S,c)}$ denotes the interaction among the features.

According to the definition of copula entropy and Equation (24), there is a theorem.

Theorem 1: The mutual information of random variables is equivalent to their negative copula entropy (Ma and Sun, 2011):

$$
I(\vec{x}) = -H_c(\vec{x}) \tag{27}
$$

According to Theorem 1, the value of copula entropy can be calculated by the MI of multivariates. The definition of mutual information extended from two variables to multivariate is described as follows:

$$
\begin{aligned}
I(X_m,c) &= \iint p(X_m,c)\log\tfrac{p(X_m,c)}{p(X_m)p(c)}dX_m dc \\
&= \iint p(X_{m-1},x_m,c)\log\tfrac{p(X_{m-1},x_m,c)}{p(X_{m-1},x_m)p(c)}dX_{m-1}dx_m dc \\
&= \int\ldots\int p(x_1,\ldots,,x_m,c)\log\tfrac{p(x_1,\ldots,x_m,c)}{p(x_1,\ldots,x_m)p(c)}dx_1,\ldots,dx_m dc
\end{aligned}
\tag{28}
$$

TABLE 2 | Datasets for comparison between CRIA algorithm and other algorithms.

Datasets type	No.	Datasets	Samples	Features	Classes	Types
Biological data	1	leukemia	72	7,070	2	Discrete
	2	Carcinoma	174	9,182	11	Continuous
	3	colon	62	2,000	2	Discrete
	4	TOX_171	171	5,748	4	Continuous
Digit recognition	5	Gisette	7,000	5,000	2	Continuous

where $X_m = \{x_1, x_2, ..., x_{m-1}, x_m\} = \{X_{m-1}, x_m\}$.

According to the Equation (28), we have:

$$
\begin{aligned}
H(X_{m-1},x_m) &= H(X_m) = \sum_{i=1}^{m} H(x_i) - I(X_m) \\
H(X_{m-1},x_m,c) &= H(X_m,c) = H(c) + \sum_{i=1}^{m} H(x_i) - I(X_m,c)
\end{aligned}
\tag{29}
$$

Therefore,

$$
\begin{aligned}
& I(X_m,c) = H(x_1) + \ldots + H(x_m) + H(c) - H(x_1,\ldots,x_m,c) \\
& I(X_m,\hat{X}_s,c) = H(x_1) + \ldots + H(x_m) + H(\hat{X}_s) + H(c) \\
& -H(x_1,\ldots,x_m,\hat{X}_s,c)
\end{aligned}
\tag{30}
$$

According to Equation (26), (27) and (30), we have:

$$
\begin{aligned}
J_{CRIA}(f_i) &= \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \tfrac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times IF_{CRIA}\} \\
&= \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \tfrac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times \tfrac{I(\Omega_S,f_i,c)}{I(\Omega_S,c)}\}
\end{aligned}
\tag{31}
$$

Let $\Omega_S = \{f_1, f_2, ..., f_m\}$, Since,

$$\frac{I(\Omega_S, f_i, c)}{I(\Omega_S, c)} = \frac{\sum_{k=1}^{m} H(f_k) + H(f_i) + H(c) - H(\Omega_S, f_i, c)}{\sum_{k=1}^{m} H(f_k) + H(c) - H(\Omega_S, c)} \quad (32)$$

Therefore,

$$J_{CRIA}(f_i) = \max_{f_i \in F - \Omega_S} \{[SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)] \times \frac{\sum_{k=1}^{m} H(f_k) + H(f_i) + H(c) - H(\Omega_S, f_i, c)}{\sum_{k=1}^{m} H(f_k) + H(c) - H(\Omega_S, c)}\} \quad (33)$$

The general flow chart of the proposed algorithm is presented in **Figure 1** we can see that an original feature set F is first given, from which we select the main effect feature that maximizes the value of (12). Then the main feature is put into the selected feature subset Ω_S. For each feature in the candidate feature set, after conducting correlation and redundancy analysis, we are next supposed to use (25) to perform interaction analysis on it. Choose the feature that maximizes the value of (33), which then is put into the selected feature set. If the number of the selected features meets the threshold condition, the above steps will be executed again, otherwise the program ends directly.

Algorithm Implementation

We propose a gene selection method based on correlation-redundancy and interaction analysis. The pseudo code of CRIA algorithm is described as follows.

Here, for the CNVs dataset, we set the value of the threshold M to be 200 to reduce the calculation time and avoid curse of dimensionality. In addition, we need to control the number of selected features to be same as the method proposed by Zhang et al. (2016).

The CRIA algorithm consists of two stages:

Stage 1 (lines 1–7): In this part, the selected feature subset Ω_S and the original feature set F are first initialized. For each feature in the original feature set f_i, the symmetrical uncertainty $SU(f_i; c)$ between f_i and class label c is calculated. The feature whose value of symmetrical uncertainty with class label is the maximum is selected out and added into the selected features subset Ω_S, which we name "the main effect feature."

Stage 2 (lines 8-18): The second stage mainly calculates the correlation measure $SU(f_i; c)$ and the redundancy measure $\frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)$. Then the interaction value IF_{CRIA} between Ω_S, f_i and c is updated. $J_{CRIA}(f_i)$ is calculated and the feature with the maximum value is added into the selected feature subset Ω_S. This procedure terminates until the number of selected features is no less than predefined threshold M.

According to **Algorithm 1**, when the size of the feature subset reaches the set threshold M, the procedure will be terminated. The value of the threshold setting should be determined by different datasets. A small M can reduce the amount of calculation but may also lose many effective features that are

TABLE 3 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with J48 classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**95.00 ± 1.17** **(1)**	93.08 ± 0.67 (4)	93.09 ± 0.64 (3)	92.51 ± 1.14 (7)	92.05 ± 0.73 (8)	93.36 ± 0.59 (2)	92.70 ± 0.98 (6)	93.07 ± 1.10 (5)	91.50 ± 1.44 (9)
Carcinoma	**75.17 ± 1.67** **(1)**	71.73 ± 1.80 (3)	71.78 ± 2.26 (2)	69.79 ± 1.58 (4)	64.47 ± 1.82 (9)	64,84 ± 1.68 (5)	68.65 ± 2.20 (6)	68.01 ± 1.97 (7)	65.53 ± 1.89 (8)
colon	**79.68 ± 2.54** **(1)**	76.55 ± 3.86 (5)	77.14 ± 3.31 (4)	74.32 ± 2.48 (8)	76.32 ± 2.07 (6)	77.28 ± 3.77 (3)	77.30 ± 4.73 (2)	74.71 ± 3.96 (7)	73.31 ± 3.77 (9)
TOX_171	62.01 ± 1.61 (4)	62.21 ± 2.04 (2)	56.61 ± 2.21 (8)	60.94 ± 3.08 (5)	62.09 ± 2.14 (3)	57.49 ± 3.01 (9)	59.78 ± 2.02 (7)	60.18 ± 1.91 (6)	**62.26 ± 2.71** **(1)**
gisette	**93.71 ± 0.17** **(1)**	92.40 ± 0.08 (6)	92.02 ± 0.08 (8)	92.66 ± 0.10 (5)	92.05 ± 0.12 (7)	91.18 ± 0.12 (9)	92.84 ± 0.08 (3)	92.82 ± 0.07 (4)	93.35 ± 0.08 (2)
Avg.acc	81.11	79.19	78.53	78.04	77.40	77.63	78.25	77.76	77.19
Avg.rank	1.60	4.00	5.00	5.80	6.60	5.60	4.80	5.80	5.80
Improved rate	–	2.42%	3.29%	3.93%	4.79%	4.48%	3.65%	4.31%	5.08%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

TABLE 4 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with IB1 classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**99.44 ± 0.97** **(1)**	97.03 ± 0.87 (2)	96.16 ± 0.43 (4)	95.56 ± 0.90 (7)	88.75 ± 1.88 (9)	96.61 ± 0.78 (3)	96.13 ± 0.90 (5)	95.73 ± 0.95 (6)	94.22 ± 0.80 (8)
Carcinoma	**86.84 ± 0.50** **(1)**	82.45 ± 1.33 (2)	81.39 ± 1.02 (4.5)	82.32 ± 1.07 (3)	76.88 ± 1.97 (9)	81.10 ± 1.06 (7)	81.39 ± 1.16 (4.5)	81.35 ± 1.08 (6)	80.96 ± 1.18 (8)
colon	**86.77 ± 1.27** **(1)**	78.60 ± 2.10 (2)	78.22 ± 1.54 (3)	75.94 ± 2.32 (5)	70.87 ± 1.97 (9)	76.69 ± 2.35 (4)	71.77 ± 3.41 (8)	72.24 ± 2.17 (7)	74.87 ± 1.80 (6)
TOX_171	84.56 ± 0.52 (3)	85.13 ± 1.26 (2)	78.14 ± 1.36 (8)	**85.19 ± 1.30** **(1)**	82.59 ± 1.78 (5)	76.68 ± 1.68 (9)	81.69 ± 1.64 (7)	82.05 ± 1.28 (6)	84.04 ± 1.48 (4)
gisette	**93.75 ± 0.14** **(1)**	91.88 ± 0.08 (6)	91.26 ± 0.09 (7)	92.26 ± 0.07 (5)	91.05 ± 0.15 (8)	90.20 ± 0.10 (9)	92.70 ± 0.06 (3)	92.58 ± 0.05 (4)	93.13 ± 0.10 (2)
Avg.acc	90.27	87.02	85.03	86.25	82.03	84.26	84.74	84.79	85.44
Avg.rank	1.40	2.80	5.30	4.20	8.00	6.40	5.50	5.80	5.60
Improved rate	–	3.73%	6.16%	4.66%	10.05%	7.13%	6.53%	6.46%	5.65%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

TABLE 5 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with Naïve Bayes classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**99.58 ± 0.67** **(1)**	97.44 ± 0.66 (2)	96.27 ± 0.30 (7)	97.03 ± 0.70 (4)	95.48 ± 1.96 (9)	96.18 ± 0.39 (8)	97.15 ± 0.80 (3)	96.79 ± 0.71 (5)	96.70 ± 0.58 (6)
Carcinoma	81.61 ± 1.11 (2)	**82.02 ± 0.83** **(1)**	80.23 ± 1.33 (7)	81.58 ± 0.95 (3)	76.38 ± 1.85 (9)	80.19 ± 1.28 (8)	80.41 ± 0.86 (5)	80.34 ± 0.87 (6)	80.46 ± 1.09 (4)
colon	**88.71 ± 0.00** **(1)**	82.97 ± 1.34 (2)	82.70 ± 1.20 (3)	80.72 ± 1.47 (6)	74.39 ± 4.37 (9)	81.66 ± 1.45 (5)	79.84 ± 2.40 (7)	78.95 ± 1.72 (8)	82.32 ± 2.08 (4)
TOX_171	69.53 ± 0.70 (3)	**70.74 ± 1.07** **(1)**	63.73 ± 1.61 (8)	68.68 ± 1.26 (4)	65.64 ± 1.63 (7)	60.41 ± 2.35 (9)	66.96 ± 1.56 (6)	67.04 ± 1.74 (5)	70.28 ± 1.60 (2)
gisette	**93.16 ± 0.05** **(1)**	90.46 ± 0.13 (2)	88.26 ± 0.02 (4)	87.69 ± 0.08 (5)	86.23 ± 0.24 (8)	86.01 ± 0.05 (9)	87.60 ± 0.04 (6)	87.48 ± 0.03 (7)	89.46 ± 0.05 (3)
Avg.acc	86.52	84.73	82.24	83.14	79.62	80.89	82.39	82.12	83.84
Avg.rank	1.60	1.80	5.80	4.40	8.40	7.80	5.40	6.20	3.80
Improved rate	–	2.11%	5.20%	4.07%	8.67%	6.96%	5.01%	5.36%	3.20%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

Algorithm 1 | CRIA: correlation-redundancy and interaction analysis based gene selection algorithm.

```
Input N: the number of original features, M: the number of features to be selected, $n_s$: the number of selected features.
Output: the selected feature subset ($\Omega_S \subseteq \mathbb{F}$).
First initializes $\Omega_S = \varnothing, \mathbb{F} = \{f_1, f_2, ..., f_N\}$;
for each $f_i \in \mathbb{F}$ do:
    calculate $SU(f_i, c)$;
end for
select the feature $f_{i\max} \in \mathbb{F}$ with the largest value of $SU(f_i, c)$;
$\Omega_S = \Omega_S \cup \{f_{i\max}\}$;
$F = F - \{f_{i\max}\}$;
while $n_s \le M$ do:
    for $f_i \in \mathbb{F}$ do:
        calculate $SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)$;
        calculate $IF_{CRIA} = \frac{H_c(\Omega_S, f_i, c)}{H_c(\Omega_S, c)}$;
        calculate $J_{CRIA}(f_i) = [SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)] \times IF_{CRIA}$ and append it into a list;
    end for
    select the feature $f_{k\max} \in \mathbb{F}$ with the largest value of $J_{CRIA}(f_i)$ from list;
    $\Omega_S = \Omega_S \cup \{f_{k\max}\}$;
    $F = F - \{f_{k\max}\}$;
end while
output $\Omega_S$.
```

useful; a large M will increase the amount of calculation but may improve the accuracy of final result (Foithong et al., 2012). Actually, when the threshold exceeds a certain value, the accuracy of the final result will not only not increase, but may decrease, and it will bring computational complexity. The selected features are ranked according to the value of the evaluation function $J_{CRIA}(f_i)$ from largest to smallest.

Verify the Performance of CRIA

Eight gene selection algorithms—JMIM (Bennasar et al., 2015), mRMR (Peng et al., 2005), DWFS (Sun et al., 2013), IWFS (Zeng et al., 2015), RAIW (Gu et al., 2020), CFR (Gao et al., 2018a), DCSF (Gao et al., 2018b), and MRI (Wang et al., 2017) are used to compare with CRIA to examine the performance of our proposed method.

The datasets used in validation experiment come from Arizona State University (ASU) datasets (Li et al., 2017), which include four biological data and one other type of data (digit recognition). They are all high-dimensional data. The smallest feature number is 2000 and the largest feature number is 9182 among them. The specific details of these datasets are shown in **Table 2**. We only use minimum description length method (Fayyad and Irani, 1993) for gene selection and utilize it to convert these numerical features.

The number of features N used in the experiment is reduced to 50 and three classifiers—IB1, J48 and Naïve Bayes are exploited. The parameters of the classifiers are set to the default parameters of Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009). We use 10 times of ten-fold cross-validation to avoid the influence of randomness on experimental results. Then mean value and Standard Deviation (STD) are taken as the comparison indices of performance of each algorithm and STD is defined as follows:

$$STD = \sqrt{\frac{1}{n_{run}} \sum_{i=1}^{N} (ACC_i - u)} \tag{34}$$

where n_{run} is the number of times of our experiments, here we set $n_{run} = 10$, ACC is the classification accuracy, u represents the average value of ACC, and N denotes the number of samples. The bigger ACC, the better performance, and the smaller STD, the higher stability.

The comparison results between the proposed algorithm and other gene selection algorithms are shown in **Tables 3–5**. As shown in **Table 3**, for the five data sets in the experiment, we can see that the classification results of CRIA in four data sets are better than other eight algorithms, which ranking first, except TOX_171, ranking fourth. Compared with other algorithms, the average accuracy of CRIA is increased by 2.42–5.08%. In **Table 4**, CRIA also outperforms the other 8 algorithms on four data sets except TOX_171, on which the experimental results of CRIA ranking third. The biggest improved rate of the proposed algorithm is 10.05% and the smallest one is 3.73%. From **Table 5**, we can find that the results of CRIA on the three data sets are superior to other algorithms, ranking first. However, on the datasets of Carcinoma and TOX_171, compared with the maximum values, the experimental accuracies of CRIA are slightly decreased by 0.50 and 1.71%, ranking second and third respectively. From the perspective of average accuracy, CRIA's result is better than other algorithms, and it is improved by 2.11–8.67%.

RESULTS AND DISCUSSIONS

Evaluation Metrics of Experimental Results

Four evaluation metrics—precision, recall, accuracy and F1-score are utilized to evaluate the performance of the corresponding method and values of these criteria are defined as equation (35).

$$\begin{aligned} precision &= \frac{T_P}{T_P + F_P} \\ recall &= \frac{T_P}{T_P + F_N} \\ accuracy &= \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \\ F1 - score &= \frac{2 \times precision \times recall}{precision + recall} \end{aligned} \tag{35}$$

TABLE 6 | The top 15 feature genes chosen by CRIA defined as equation (26).

Ranked order	Official name	Official full gene name	Category	CRIA value
1	RPS15	ribosomal protein S15	Protein Coding	0.168
2	TBC1D5	TBC1 Domain Family Member 5	Protein Coding	0.089
3	CUL2	Cullin 2	Protein Coding	0.093
4	SMPD3	Sphingomyelin Phosphodiesterase 3	Protein Coding	0.089
5	CTAGE10P	CTAGE Family Member 10, Pseudogene	Pseudogene	0.071
6	C1orf98	Chromosome 1 Open Reading Frame 98	Protein Coding	0.043
7	ZNF281	Zinc Finger Protein 281	Protein Coding	0.061
8	CDKN2A	Cyclin Dependent Kinase Inhibitor 2A	Protein Coding	0.161
9	EGFR	Epidermal Growth Factor Receptor	Protein Coding	0.121
10	TMEM98	Transmembrane Protein 98	Protein Coding	0.103
11	CTBP2	C-Terminal Binding Protein 2	Protein Coding	0.083
12	SEMA6A	Semaphorin 6A	Protein Coding	0.081
13	MIR1208	MicroRNA 1208	RNA Gene	0.077
14	RBFOX1	RNA Binding Fox-1 Homolog 1	Protein Coding	0.069
15	CDC25A	Cell Division Cycle 25A	Protein Coding	0.066

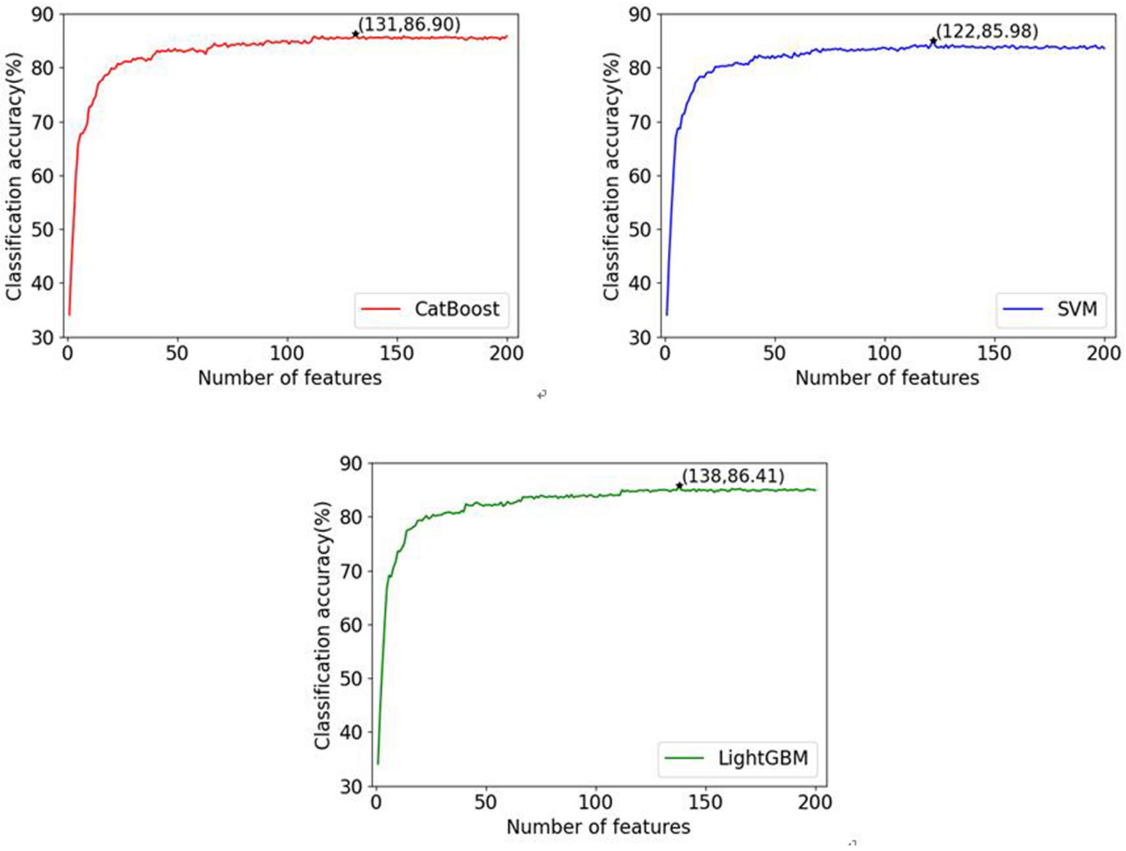

FIGURE 2 | Classification accuracies of three classifiers (CatBoost, LightGBM and SVM) with different numbers of features during the IFS procedure. The top 200 feature genes are selected by CRIA method.

where T_P, T_N, F_P, and F_N denotes the numbers of true positives, true negatives, false positives, and false negatives respectively.

The CRIA and IFS Results

As mentioned in section Evaluation Metrics of Experimental Results, each sample is represented by 24,174 features, each of

TABLE 7 | Average performance of precision, recall and F1-score on 10 test datasets with three classifiers *via* ten-fold cross-validation (%).

Metrics		UCEC	KIRC	OV	GBM	COAD/ READ	BRCA
Precision	CRIA_CatBoost	74.31	93.74	84.59	94.63	89.67	84.48
	CRIA_SVM	70.47	90.33	85.40	95.64	88.54	84.76
	CRIA_LightGBM	71.46	93.37	82.84	95.72	90.23	84.71
Recall	CRIA_CatBoost	73.14	91.63	87.90	90.76	86.09	88.67
	CRIA_SVM	73.81	89.59	88.43	89.70	83.30	87.96
	CRIA_LightGBM	72.91	92.04	89.32	91.30	83.48	87.01
F1-score	CRIA_CatBoost	73.72	92.67	86.21	92.65	87.84	86.52
	CRIA_SVM	72.13	89.96	86.89	92.57	85.84	86.33
	CRIA_LightGBM	72.18	92.70	85.96	93.46	86.72	85.84

which indicates the expression level of genes. The 24174 feature genes are sorted by CRIA value in descending order. However, we only select the top 200 features in this work for the consideration of computational time and curse of dimensionality. The top 15 key feature genes chosen by CRIA defined by equation (26) are listed in **Table 6**.

We use the Incremental Gene selection (IFS) (Yang et al., 2019) to determine the optimal feature set. The first 200 features are added one by one to a feature subset in order. Each time a feature is added, a classifier is trained and examined. So, 200 classifiers are constructed. We use the criteria of accuracy to evaluate the performance of all the 200 classifiers and then we choose the classifier with the highest accuracy as the final one. The corresponding feature subset that the final classifier used is deemed to be the optimal feature set.

In this paper, three commonly used classifiers are adopted to verify the generalization performance of the proposed gene selection method on different classifiers. ten-fold cross-validation is used to evaluate our algorithm with the selected features. The complete data set is randomly split into 10 parts of approximately equal size. The three classifiers are trained 10 times; nine of the 10 subsets are used as the training datasets, and the remaining one is the test dataset. The average values of accuracy for each classifier are calculated and the IFS results are shown in **Figure 2**. Here, we name our methods as CRIA_CatBoost, CRIA_SVM and CRIA_LightGBM. From **Figure 2**, it can be seen that the highest accuracy of 86.90% for CRIA_CatBoost method followed by 86.41% for CRIA_LightGBM and 85.98% for CRIA_SVM method, with only using the CNVs of 131 genes, 138 genes and 122 genes respectively.

The Proposed Algorithm Performance

For the different classifiers used in this work, after determining the optimal numbers of features according to the CRIA and IFS results, the classification performance can be further analyzed. The average values of three metrics-precision, recall and F1-score defined in Equation (35) on 10 test datasets are listed in **Table 7**.

TABLE 8 | Performance comparison of the proposed algorithm predictions with those of other methods (%).

Cancer	Predictor	Precision	Recall	F1-score
UCEC	CRIA_CatBoost	**74.31**	**73.14**	**73.72**
	CRIA_SVM	70.47	**73.81**	**72.13**
	CRIA_LightGBM	71.46	**72.91**	**72.18**
	CNA_origin	67.92	72.00	69.90
	mRMR_Dagging	74.19	46.93	57.50
KIRC	CRIA_CatBoost	**93.74**	91.63	**92.67**
	CRIA_SVM	**90.33**	89.59	89.96
	CRIA_LightGBM	**93.37**	92.04	**92.70**
	CNA_origin	88.89	**96.00**	92.31
	mRMR_Dagging	80.85	92.68	86.36
OV	CRIA_CatBoost	84.59	**87.90**	86.21
	CRIA_SVM	85.40	**88.43**	86.89
	CRIA_LightGBM	82.84	**89.32**	85.96
	CNA_origin	**89.80**	86.72	**88.00**
	mRMR_Dagging	84.61	75.86	80.00
GBM	CRIA_CatBoost	**94.63**	**90.76**	**92.65**
	CRIA_SVM	**95.64**	**89.70**	**92.57**
	CRIA_LightGBM	**95.72**	**91.30**	**93.46**
	CNA_origin	93.10	84.38	88.52
	mRMR_Dagging	88.70	85.93	87.30
COADREAD	CRIA_CatBoost	**89.67**	**86.09**	**87.84**
	CRIA_SVM	**88.54**	**83.30**	**85.84**
	CRIA_LightGBM	**90.23**	**83.48**	**86.72**
	CNA_origin	81.58	73.81	77.50
	mRMR_Dagging	60.00	73.46	66.05
BRCA	CRIA_CatBoost	84.48	88.67	86.52
	CRIA_SVM	84.76	87.96	86.33
	CRIA_LightGBM	84.71	87.01	85.84
	CNA_origin	**87.50**	**92.31**	**89.84**
	mRMR_Dagging	79.16	87.35	83.06

Performance Comparison With Other Methods

After selecting important features, we use three common classifiers—CatBoost, SVM and LightGBM to predict cancer samples. The performance of our methods are compared with other two classification methods published before whose experimental dataset is the same as ours. Liang et al. (2020) used a method called CNA_origin which was composed of

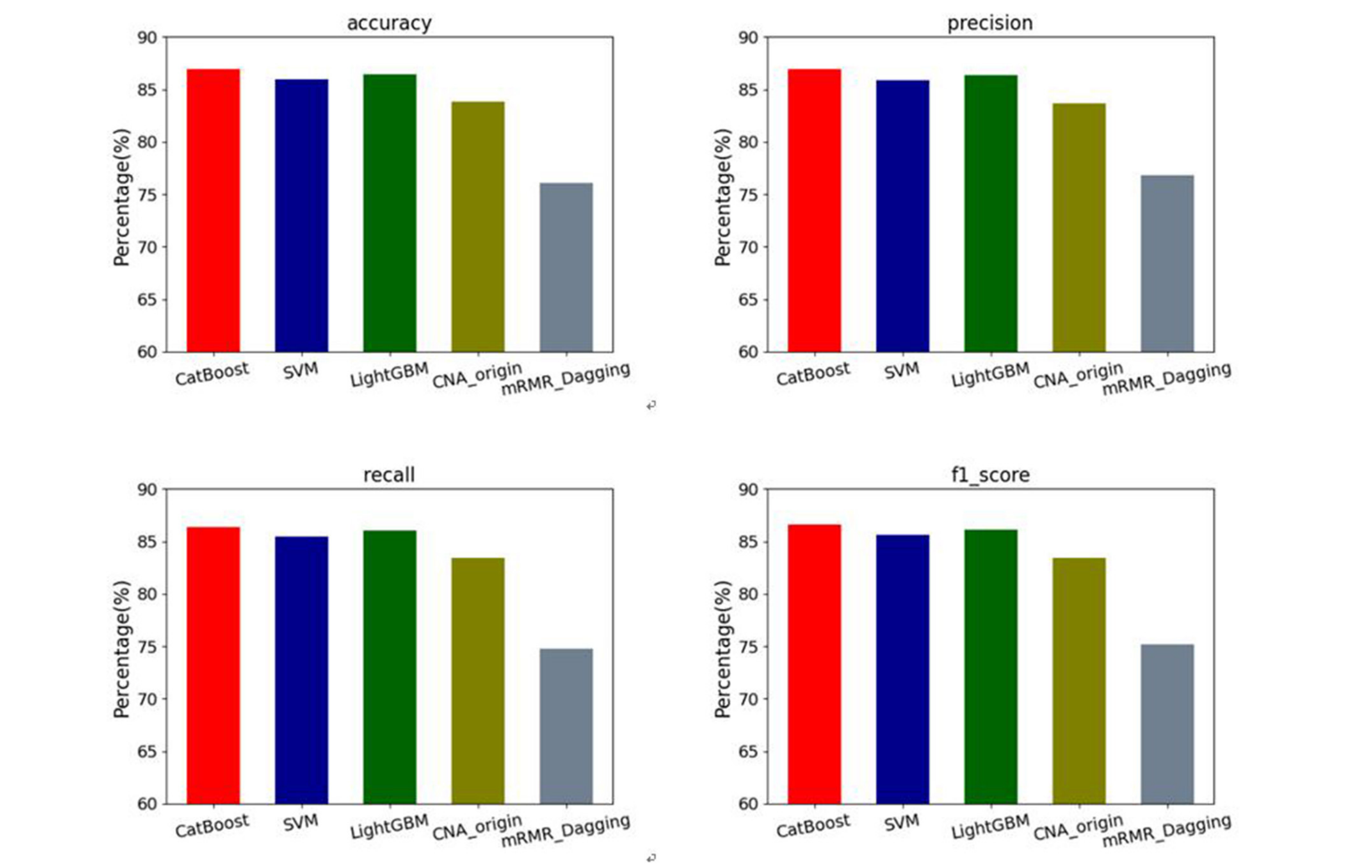

FIGURE 3 | Performance comparison of 4 evaluation metrics: accuracy, precision, recall and F1-score among our methods and the other two algorithms (CNA_origin and mRMR_Dagging).

a stacked autoencoder and an one-dimensional convolutional neural network. The 24,174 gene features were extracted to 100 genes by the autoencoder, and then these 100 gene features were put into the 1D CNN for classification (Liang et al., 2020). A computationally method for cancer types classification proposed by Zhang et al. (2016) was named as mRMR_Dagging here because there was no specific method name given by authors. It first used mRMR and IFS to select 19 of the 24,174 genes as classification features, and then used the Dagging algorithm to give the final results.

In **Table 8**, it can be seen that if the results of our methods are superior to CNA_origin and mRMR_Dagging, they are marked in bold. Similarly, if the largest of CNA_origin and mRMR_Dagging results is better than our method, it is also marked in bold. **Table 8** demonstrated that the performance of our methods is superior to CNA_origin and mRMR_Dagging for UCEC, KIRC, GBM, and COADREAD. For UCEC, the recall and F1-score of our methods (CRIA_Cat- Boost, CRIA_SVM and CRIA_LightGBM) are all superior to CNA_origin and mRMR_Dagging. The best precision of our methods is 0.12 percentage points higher than mRMR_Dagging. SVM and LightGBM are slightly worse than mRMR_Dagging with reductions of 5.28 and 3.82% in precision respectively. For KIRC, the precision and F1-score are all superior to CNA_origin and mRMR_Dagging except the F1-score of SVM, which performs slightly worse than the CNA_origin with reductions of 2.61%. Compared with the best, CNA_origin, the recall of our methods are decreased by 4.77% for CatBoost, 7.15% for SVM and 4.30% for LightGBM. For OV, compared with CNA_origin, the recall of our methods is at least increased by 1.36%. The precision and F1-score are slightly worse than CNA_origin, with reductions at most of 8.40, and 2.37%, respectively. For GBM and COADREAD, our methods are better than CNA_origin and mRMR_Dagging on all evaluation indicators. Compared with the best of the other two algorithms, the worst precision of our methods is increased by 1.64 and 8.53%, respectively, the worst recall is increased by 4.39 and 12.86%, respectively, and the worst F1-score is increased by 4.58 and 10.76%, respectively. For BRCA, the worst among our methods performs slightly worse than the best CNA_origin algorithm, with reductions of 3.57% in precision, 6.09% in recall and 4.66% in F1-score respectively.

In addition, the macro-average results of four evaluation metrics: accuracy, precision, recall and F1-score are used to assess our methods and the other two algorithms on the datasets of six types of cancers. The results can be seen in **Figure 3**. For accuracy, our methods have mean values of

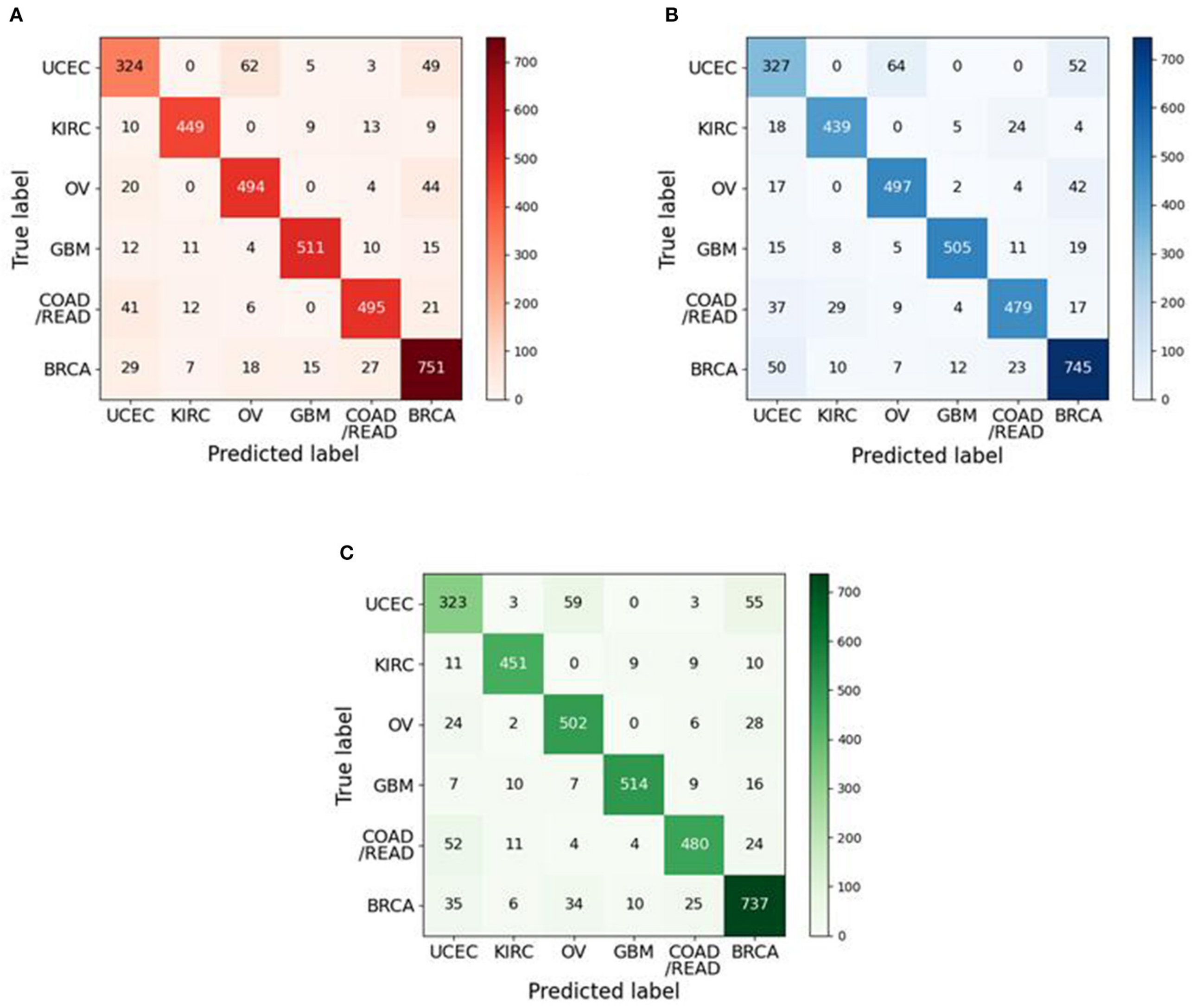

FIGURE 4 | Confusion matrices on the test groups: **(A)** CRIA_CatBoost; **(B)** CRIA_SVM; **(C)** CRIA_LightGBM.

86.90% for CatBoost, 86.41% for LightGBM and 85.98% for SVM respectively, which are increased by 3.69, 3.10, and 2.59% compared with CNA_origin. For precision, the average values of our methods are 86.61, 86.39, and 85.86%, which are increased by 3.49, 3.23, and 2.59%, respectively compared with the best among CNA_origin and mRMR_Dagging. For recall, our methods' mean values are 86.37, 86.01, and 85.47%, which are 2.92, 2.56 and 2.02 percentage points higher than CNA_origin, respectively. For F1-score, compared with our methods, whose average values are 86.60, 86.14, and 85.62%, CNA_origin is decreased by 3.71, 3.19, and 2.60%, respectively.

Further Discussion

In order to study the relationship between the classes, we also summarize the confusion matrices in **Figure 4** for class predictions using our methods. From **Figure 4**, we can find that there existed a high error rate when predicting the samples of UCEC. Regardless of whether it is CRIA_CatBoost, CRIA_SVM or CRIA_LightGBM, more than 10% of the UCEC samples are incorrectly predicted as OV and BRCA. In **Figure 4A**, 14.00% of UCEC samples are predicted as OV, while 11.06% of UCECsamples are predicted to be BRCA. In **Figure 4B**, 14.45 and 11.74% of UCEC samples are predicted as OV and BRCA respectively. In **Figure 4C**, 13.32 and 12.42% of UCEC samples are predicted as OV and BRCA respectively. The reasons may be that UCEC, OV and BRCA are hormone-dependent tumors and they relate closely in tumorigenesis. The 16 and 27 risk regions were identified by an independent genome-wide association study (GWAS) on endometrial cancer and ovarian cancer, respectively (Glubb et al., 2020). Studies have shown that mutations in breast cancer susceptibility genes (BRCA1, BRCA2) have a relationship in hereditary ovarian cancer. Mutations at either end of the BRCA1 gene increase a person's risk of breast cancer, and its probability is higher than ovarian cancer. However, mutations in the middle of the BRCA1 gene put a person at a higher risk of ovarian cancer than breast cancer (Shi et al., 2017). In addition, there is also a study indicated that UCEC, OV and BRCA all have a relationship with the changes in estrogen and estrogen receptors (Rodriguez et al., 2019).

CONCLUSIONS

In this paper, we introduce a gene selection algorithm—CRIA. We firstly apply this algorithm to 5 datasets and verify the effective performance of CRIA through comparison with other eight gene selection algorithms. The proposed algorithm can select features which are closely related to the class label. Then, we use this algorithm to select 200 genes that have a close relationship with cancer types from 24,174 genes features based on the value of copy number variations in the samples, and then combine three common classifiers—CatBoost, SVM and LightGBM to predict the type of cancer. Our experimental results show that our methods have higher accuracies than the state-of-the-art methods for solving this problem. Our research has a certain degree of interpretability for cancer-related researches at the genetic level. As we all know, cancer is closely related to gene structural variations and the appearance of cancer is often accompanied by abnormalities in the deoxyribonucleic acid (DNA) sequence. Because CNVs is one of the most crucial structural variations of genes, studying the relationship between cancers and CNVs is of great significance. Many studies have tried to utilize the genetic information of cancers to predict cancer type, which can provide significant guidance for patient care and cancer therapy in promptly.

The future direction of this work can continue to develop from two aspects. First of all, because we only use the datasets of six cancer types and the total number of samples is only 3,480 in this paper, by collecting data sets of other cancer types and optimizing the proposed algorithm, we can continue to conduct further research in the field of cancer classification based on copy number variations. Moreover, integrating non-CNVs features for the samples can be taken into consideration. In addition to using CNVs for cancer prediction, we can also apply other genetic information for cancer prediction, or combine several biomarkers to reduce the error rate of classification as much as possible.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

QW conducted the experiments and wrote the manuscript. DL conceived and provided the main direction of the manuscript and guided the writing and modification of this manuscript. Both authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 11571009) and Applied Basic Research Programs of Shanxi Province (Grant No. 201901D111086).

REFERENCES

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* 5, 537–550. doi: 10.1109/72.298224

Bennasar, M., Hicks, Y., and Setchi, R. (2015). Gene selection using Joint Mutual Information Maximisation. *Expert Syst. Appl.* 42, 8520–8532. doi: 10.1016/j.eswa.2015.07.007

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

Breunis, W. B., van Mirre, E., Bruin, M., Geissler, J., de Boer, M., Peters, M., et al. (2008). Copy number variation of the activating FCGR2C gene predisposes to idiopathic thrombocytopenic purpura. *Blood* 111, 1029–1038. doi: 10.1182/blood-2007-03-079913

Buchynska, L. G., Brieieva, O. V., and Iurchenko, N. P. (2019). Assessment of HER-2/neu, α-MYC and CCN- E1 gene copy number variations and protein expression in endometrial carcinomas. *Exp. Oncol.* 41. doi: 10.32471/exp-oncology.2312-8852.vol-41-no-2.12973

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer geno- mics data: figure 1. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chandrashekar, G., and Sahin, F. (2014). A survey on gene selection methods. *Comput. Electr. Eng.* 40, 16–28. doi: 10.1016/j.compeleceng.2013.11.024

Chen, Z., Wu, C., Zhang, Y., Huang, Z., Ran, B., Zhong, M., et al. (2015). Gene selection with redundancy-complementariness dispersion. *Knowl. Based Syst.* 89, 203–217. doi: 10.1016/j.knosys.2015.07.004

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

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: John Wiley and Sons.

Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* 34, 3215–3225. doi: 10.1038/onc.2014.291

Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Glessner, J. T., et al. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmen- tal genes. *Mol. Psychiatry* 15, 637–646. doi: 10.1038/mp.2009.57

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. A. (2009). Normalized Mutual Information Gene selection. *IEEE Trans. Neural Netw.*20, 189–201. doi: 10.1109/TNN.2008.2005601

Fayyad, U. M., and Irani, K. B. (1993). "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," in *Pro-ceedings of International Joint Conference on Artificial Intel- ligence, pp* 1022–1027

Flierl, A., Oliveira Luís, M. A., Falomir-Lockhart Lisandro, J., Mak Sally, K., Hesley, J., Soldner, F., et al. (2014). Higher vulnerability and stress sensitivity of neuronal precursor cells carrying an alpha-synuclein gene triplication. *PLoS ONE* 9, e112413. doi: 10.1371/journal.pone.0112413

Foithong, S., Pinngern, O., and Attachoo, B. (2012). Feature subset selection wrapper based on mutual information and rough sets. *Expert Syst. Appl.* 39, 574–584. doi: 10.1016/j.eswa.2011.07.048

Frank, B., Bermejo, J. L., Hemminki, K., Sutter, C., Wappenschmidt, B., Meindl, A., et al. (2007). Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis* 28, 1442–1445. doi: 10.1093/carcin/bgm033

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the Cbioportal. *Sci. Signaling* 6, pl1–pl1. doi: 10.1126/scisignal.2004088

Gao, W., Hu, L., and Zhang, P. (2018b). Class-specific mutual information variation for gene selection. *Pattern Recogn*. 79, 328–339. doi: 10.1016/j.patcog.2018.02.020

Gao, W., Hu, L., Zhang, P., and He, J. (2018a). Gene selection considering the composition of feature relevancy. *Pattern Recogn. Lett.* 112, 70–74. doi: 10.1016/j.patrec.2018.06.005

Glubb, D. M., Thompson, D. J., Aben, K. K., Alsulimani, A., Amant, F., Annibali, D., et al. (2020). Cross-cancer genome-wide association study of endometrial cancer and epithelial ovarian cancer identifies genetic risk regions associated with risk of both cancers. *Cancer Epidemiol. Biomarkers Prev*. 30, 217–28. doi: 10.1158/1055-9965.EPI-20-0739

Grangeon, L., Cassinari, K., Rousseau, S., Croisile, B., Formaglio, M., Moreaud, O. et al. (2021). Early-onset cerebral amyloid angiopathy and alzheimer disease related to an app locus triplication. *Neurol. Genet*. 7, e609–e609. doi: 10.1212/NXG.0000000000000609

Gu, X., Guo, J., Li, C., and Xiao, L. (2020). A gene selection algorithm based on redundancy analysis and interaction weight. *Appl. Intell.* 51, 2672–2686. doi: 10.1007/s10489-020-01936-5

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor Newsl*. 11, 10–18. doi: 10.1145/1656274.1656278

Heo, Y., Heo, J., Han, S., Kim, W. J., Cheong, H. S., and Hong, Y. (2020). Difference of copy number variation in blood of patients with lung cancer. *Int. J. Biol. Markers* 36, 3–9. doi: 10.1177/1724600820980739

Jakulin, A. (2003). *Attribute Interactions in Machine Learning (Master thesis). Computer and Information Science, University of Ljubljana.*

Jakulin, A., and Bratko, I. (2004). "Testing the significance of attribute interactions," in *Proceedings of the Twenty-first international conference on Machine learning - ICML'04*. Banff, AL: ACM Press. pp. 409–416.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2017). Gene selection: a data perspective. *ACM Comput. Surv*. 50, 1–45. doi: 10.1145/3136625

Liang, J., Hou, L., Luan, Z., and Huang, W. (2019). Gene selection with conditional mutual information considering feature interaction. *Symmetry* 11, 858. doi: 10.3390/sym11070858

Liang, Y., Wang, H., Yang, J., Li, X., Dai, C., Shao, P., et al. (2020). A deep learning framework to predict tumor tissue-of-origin based on copy number alteration. *Front. Bioeng. Biotech*. 8, 701. doi: 10.3389/fbioe.2020.00701

Ma, J., and Sun, Z. (2011). Mutual information is copula entropy. *Tsinghua Sci. Technol*. 16, 51–54. doi: 10.1016/S1007-0214(11)70008-6

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number al teration in human cancers. *Genome Biol.* 12, 4. doi: 10.1186/gb-2011-12-4-r41

Orsenigo, C., and Vercellis, C. (2013). A comparative study of non-linear manifold learning methods for cancer microarray data classification. *Expert Syst. Appl.* 40, 2189–2197. doi: 10.1016/j.eswa.2012.10.044

Pandey, G. N., Rizavi, H. S., Tripathi, M., and Ren, X. (2015). Region-specific dysregulation of glycogen synthase kinase-3β and β-catenin in the postmortem brains of subjects with bipolar disorder and schizophrenia. *Bipolar Disord*. 17, 160–171. doi: 10.1111/bdi.12228

Peng, H., Long, F., and Ding, C. (2005). Gene selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE T. Pattern Anal*. 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

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

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

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev*. 5, 3–55. doi: 10.1145/584091.584093

Shi, T., Wang, P., Xie, C., Yin, S., Shi, D., Wei, C., et al. (2017). *BRCA1* and *BRCA2* mutations in ovarian cancer patients from China: ethnic-related mutations in *BRCA1* associated with an increased risk of ovarian cancer: *BRCA1/2* mutation in Chinese ovarian cancer. *Int. J. Cancer* 140, 2051–2059. doi: 10.1002/ijc.30633

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., and Wang, K. (2013). Gene selection using dynamic weights for classification. *Knowl. Based Syst.* 37, 541–549. doi: 10.1016/j.knosys.2012.10.001

Tian, T., Bi, H., Liu, Y., Li, G., Zhang, Y., Cao, L., et al. (2020). Copy number variation of ubiquitin- specific proteases genes in blood leukocytes and colorectal cancer. *Cancer Biol. Ther*. 21, 637–646. doi: 10.1080/15384047.2020.1750860

Van Bockstal, M. R., Agahozo, M. C., van Marion, R., Atmodimedjo, P. N., Sleddens, H. F. B. M., Dinjens, W. N. M., et al. (2020). Somatic mutations and copy number variations in breast cancers with heterogeneous *HER2* amplification. *Mol. Oncol*. 14, 671–685. doi: 10.1002/1878-0261.12650

Wang, J., Wei, J. M., Yang, Z., and Wang, S. Q. (2017). Gene selection by Maximizing Independent Classification Information. *IEEE Trans. Knowl. Data Eng.* 29, 828–841. doi: 10.1109/TKDE.2017.2650906

Witten, I. H., and Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *SIGMOD Rec*. 31, 76–77. doi: 10.1145/507338.507355

Yang, Y., Song, S., Chen, D., and Zhang, X. (2019). Discernible neighborhood counting based incremental gene selection for heterogeneous data. *Int. J. Mach. Learn. Cybern*. 11, 1115–1127. doi: 10.1007/s13042-019-00997-4

Zeng, Z., Zhang, H., Zhang, R., and Yin, C. (2015). A novel gene selection method considering feature interaction. *Pattern Recogn*. 48, 2656–2666. doi: 10.1016/j.patcog.2015.02.025

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

Zheng, Z., Yu, R., Gao, C., Jian, X., Quan, S., Xing, G., et al. (2017). Low copy number of FCGR3B is associated with lupus nephritis in a Chinese population. *Exp. Ther. Med*. 14, 4497–4502. doi: 10.3892/etm.2017.5069

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.