Predicting the outcome after a cancer diagnosis is critical. Advances in high-throughput sequencing technologies provide physicians with vast amounts of data, yet prognostication remains challenging because the data are greatly dimensional and complex. We evaluated Wnt/β-catenin, carbohydrate metabolism, and PI3K-Akt signaling pathway-related genes as predictive features for classifying tumors and normal samples. Using differentially expressed genes as controls, these pathway-related genes were assessed for accuracy using support-vector machines and three other recommended machine learning models, namely, the random forest, decision tree, and k-nearest neighbor algorithms. The first two outperformed the others. All candidate pathway-related genes yielded areas under the curve exceeding 95.00% for cancer outcomes, and they were most accurate in predicting colorectal cancer. These results suggest that these pathway-related genes are useful and accurate biomarkers for understanding the mechanisms behind cancer development.

1. Introduction

Cancer, associated with high mortality, is indeed a serious threat to public health. One cause for the high mortality rate is nonspecific symptoms in the early stages, resulting in a poor prognosis and a high fatality rate. Thus, accurately predicting cancer is a most critical and urgent task for physicians. Because cancer is fundamentally caused by gene malfunction, utilizing their expression levels as relatively direct methods of diagnoses has attracted a great deal of research attention. To date, analyses of gene expression level data have greatly benefited cancer diagnoses and treatments [1–3]. However, the high dimensionality and noise associated with the data can make these analyses and applications challenging. To reduce these challenges, data are initially processed to identify a small subset of genes primarily responsible for the disease [4, 5]. Feature selection is reportedly a very effective method for reducing the high dimensionality of gene expression datasets [6].

Cancer biology research is rapidly finding the recurring roles of a small set of signaling cascades: the Wnt cascade, metabolism, PI3K/AKT signaling pathway, and so on. The Wnt signaling pathway is prevalent in carcinogenesis, playing an essential role in the development of various tumors [7, 8]. Indeed, current evidence suggests that up to 80% of colorectal cancers are driven by an activating mutation in the Wnt cascade [9]. Altered energy metabolism is believed to be a hallmark characteristic of cancer [10, 11]. Even in the presence of oxygen, cancer cells can reprogram their glucose metabolisms to produce energy, thus largely limiting energy metabolism to glycolysis [12]. In addition, glycolysis provides cancer cells with various metabolic
precursors that promote the synthesis of amino acids, nucleotides, and lipids, leading to cancer development. The PI3K-Akt signaling pathway is most frequently activated in a variety of cancer lineages [13–15]. A range of malignancies, including ovarian, breast, colorectal, and endometrial cancers, frequently exhibit activation of the PI3K pathway through various mechanisms, including genomic mutations or alterations involving PIK3CA, PIK3R1, PTEN, AKT, TSC1, TSC2, LKB1 (also known as STK11), MTOR, and other oncogenes or tumor suppressor genes [16, 17]. This regulates key biological processes, including proliferation, the cell cycle, motility, metabolism, and genomic instability, all of which support the survival, expansion, and dissemination of cancer [18].

In conjunction with the rapidly increasing amount of gene expression data, state-of-the-art data analysis tools are being developed. Of them, machine learning (ML) methods such as random forest (RF), support-vector machine (SVM), decision tree (DT), and k-nearest neighbor (KNN) have been successfully applied to various areas of genomics research [19, 20]. Included are the expression profiles of genes [21], predicting the functional activity of genomic sequences [22], and predicting the intrinsic molecular subtypes of breast cancer [23]. Notably, RF uses highly dimensional data and data that are unbalanced and missing values [24]. An SVM is an ML algorithm that separates entities into appropriate classes using a hyperplane [25]. In cancer research, it has been used successfully to classify people as those with and without cancer based on microarray expression data [26].

These methods were used in this study to predict the cancer state from gene expression data from various types of cancer. Given the significant roles of these cancers, pathway-related genes were used as alternative features.

2. Materials and Methods

2.1. Data Acquisition. Genetic data were downloaded from The Cancer Genome Atlas, a publicly accessible dataset (https://cancergenome.nih.gov/). The microarray expression data included colorectal cancer (1222 samples, 1109 tumorous), gastric cancer (407 samples, 375 tumorous), and breast cancer (440 samples, 410 tumorous). Detailed information about the data is shown in Table 1, and the number of pathway-related genes in the candidate cancers is shown in Table 2.

2.2. Data Preprocessing. Data preprocessing is a crucial step in ML, and errors at this stage can lead to misleading prediction results. This study included the following preprocessing steps: Data were normalized for each sample by first transforming the data using the log ratio base 2 and then, for each probe, calculating the median of the log-summarized values from all samples and subtracting it from each sample. Missing values were replaced with the attribute mean.

2.3. Feature Selection. For clinical use, the number of cancer samples was unbalanced in comparison with the number of features, possibly leading to a high risk of overfitting and degrading the classification performance, thus significantly affecting predication accuracy. However, effective feature selection is a method used to address this challenge [27]. Considering the importance of pathways in tumorigenesis, three pathway-related genes were selected as candidate features. They were the Wnt/β-catenin, carbohydrate metabolism, and PI3K-Akt signaling pathways. Simultaneously, significantly differentially expressed genes (DEGs) were used as controls for comparing the features used for cancer classification. These DEGs have been previously employed in cancer prediction studies, and the findings support their use as valid features. The DESeq R package [28] was used to identify DEGs. Our criteria were a P value of less than 0.001 and a log 2 fold change of 4 or more. Notably, the pathway-related genes were derived from the Kyoto Encyclopedia of Genes and Genomes (http://www.kegg.jp/) analysis.

2.4. Conventional Machine Learning Algorithms. All four widely used classification methods (SVM, RF, DT, and KNN) were adopted. In the SVM method, the parameter C was assigned a value of either 0.1, 1, 10, or 100, and the kernel function was either “linear,” “rbf,” “poly,” or “sigmoid.”

In the KNN method, the number of neighbors was assigned as 3, 5, or 7, and the Euclidean distance, Manhattan distance, and Minkowski distance were combined to train the model.

In the DT algorithm, CART was used, and the maximum tree depth was 5 or 10. In the RF model, the numbers of DTs were 5, 10, or 50 and the numbers of features were 2, 4, 10, or 20.

3. Results

3.1. General Classification Workflow. Data were extracted from the Kyoto Encyclopedia of Genes and Genomes database. Specifically, 142, 356, and 350 elements (pathway-related genes) were found for the Wnt, carbohydrate metabolism, and PI3K-Akt signaling pathways, respectively. In addition, 314, 241, and 133 DEG parameters were included for colorectal, breast, and gastric cancer, respectively. To evaluate the cancer predictive ability of these pathway-related genes, the workflow shown in Figure 1 was implemented. Before training the model, all data were subjected to pretraining the model using an autoencoder without labels. This step was designed to improve model performance, avoid random initialization of the weights, and select the candidate model architecture associated with the minimum mean square error.

3.2. Wnt Pathway-Related Genes Score as High as DEGs in Predicting Colorectal Cancer. Detailed information about the relative sample and pathway-related genes is shown in Tables 1 and 2. The prediction performances of the entire set of Wnt pathway-related genes and of the DEGs were evaluated using three common metrics: precision, recall, and accuracy. Results are shown in Tables 3 and 4. Scores using
Wntpathway-relatedgenesarecomparabletothosefound
using DEGs, achieving approximately 95% accuracy for
classifying colorectal cancer regardless of the ML method
used (Figure 2).

3.3. Wnt Pathway-Related Genes Are Efficient Predictors of
Cancer. Based on these results, we hypothesized that the
Wnt pathway is potentially a feature that can be adopted for
cancer detection. To test this, it was evaluated with common
cancers such as breast and gastric cancers. Similar
procedures and algorithms were selected, and DEGs were
used as controls. Not surprisingly, results using the Wnt
pathway-related genes were similar to those using the
control group: the area under the curve (AUC) exceeded
94.00%. It is worth noting that Wnt pathway-related genes
in breast cancer outperformed those in gastric cancer
(AUC values of approximately 98% and 95%, respectively
Figure 3).

3.4. Carbohydrate Metabolism and PI3K-Akt Signaling
Pathways Can Predict Cancer Status. It is unknown whether
other cancer-related pathways can predict cancer status.
Thus, a set of carbohydrate metabolism and PI3K-Akt
signaling pathway-related genes were chosen to test their
abilities to predict our candidate cancers. The carbohydrate
metabolism pathway-related genes scored highest for colo-
orectal cancer followed by breast cancer and gastric cancer.
Similar results were found using ML methods: AUC values
were 98.28%, 97.30%, 96.07%, and 96.31% when using SVM,
RF, DT, and KNN, respectively. Interestingly, the PI3K-Akt
signaling pathway-related genes performed similarly. Both
carbohydrate metabolism and PI3K-Akt signaling pathways
yielded AUCs above 96.00%, implying that both pathways
can detect cancer with great accuracy (Table 5). Of note, the
SVM and RF methods outperformed DT and KNN in cancer
detection (Figure 4). Taken together, these results indicate
that these three pathway-related genes can be vital features
for cancer prediction and that these pathways vary in
predictive power. We believe that most pathway-related
genes are promising features that could be used for early
cancer diagnoses.

Table 1: Clinical features of patients in The Cancer Genome Atlas (TCGA) dataset.

<table>
<thead>
<tr>
<th>Clinical factor</th>
<th>COAD n = 440</th>
<th>BRCA n = 1222</th>
<th>STAD n = 407</th>
<th>PRAD n = 550</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient count (selected/original)</td>
<td>387/410</td>
<td>1089/1109</td>
<td>375/375</td>
<td>493/498</td>
</tr>
<tr>
<td>Age (years, mean ± SD)</td>
<td>65.73 ± 13.06</td>
<td>58.46 ± 13.20</td>
<td>65.83 ± 10.65</td>
<td>65.83 ± 10.65</td>
</tr>
<tr>
<td>Sex (male/female/−)</td>
<td>201/186</td>
<td>12/1077</td>
<td>241/134</td>
<td>493/0</td>
</tr>
<tr>
<td>Death (dead/alive/−)</td>
<td>82/304/1</td>
<td>152/937</td>
<td>150/225</td>
<td>10/483</td>
</tr>
<tr>
<td>Overall survival (months, mean ± SD)</td>
<td>28.46 ± 26.27</td>
<td>40.96 ± 30.17</td>
<td>19.32 ± 18.08</td>
<td>35.76 ± 25.89</td>
</tr>
</tbody>
</table>

Note. Selected patients included those with clinical characteristics after removing normal, replicate, and missing features from the total sample used in the model.

Table 2: Elements of pathway-related genes in candidate cancers.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Typical</th>
<th>Nontypical</th>
<th>DEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wnt</td>
<td>Ca-Me</td>
<td>PI3K</td>
</tr>
<tr>
<td>COAD (total/selected)</td>
<td>143/142</td>
<td>356/356</td>
<td>351/350</td>
</tr>
<tr>
<td>BRCA (total/selected)</td>
<td>143/142</td>
<td>356/356</td>
<td>351/350</td>
</tr>
<tr>
<td>STAD (total/selected)</td>
<td>143/142</td>
<td>356/356</td>
<td>351/350</td>
</tr>
<tr>
<td>PRAD (total/selected)</td>
<td>143/142</td>
<td>356/356</td>
<td>351/350</td>
</tr>
</tbody>
</table>

Note. COAD, colorectal cancer; BRCA, breast cancer; STAD, gastric cancer; PRAD, prostate cancer; Wnt, Wnt signaling pathway; Ca-Me, carbohydrate metabolism signaling pathway; PI3K, PI3K-Akt signaling pathway; TLR, toll-like receptor signaling pathway; TH, thyroid hormone signaling pathway; RPSC, signaling pathways regulating pluripotency of stem cells; DEGs, differentially expressed genes.
4. Discussion

Increasing evidence indicates that colorectal cancer is often initiated by an activating mutation in the Wnt cascade. The correlation between the Wnt pathway and colorectal cancer prompted our investigation into whether Wnt pathway-related genes serve as features for detecting colorectal cancer.

Thus, we designed this study to take advantage of various conventional ML models and cancer-related pathways for predicting cancer. Results show that these three pathway-related genes could be used as features for cancer prediction; they yielded results equal to those of DEGs.

Given the complexity and high mortality of cancer, the accurate early diagnosis of a cancer type can facilitate clinical management. Only relatively recently has cancer researchers attempted to apply ML for cancer prediction and prognosis [29–31]. Most previous work employed ML methods for modeling cancer progression and then identified informative factors used in a classification scheme and attempted to develop a set of classifiers for feature selection. Conventional ML algorithms require engineering domain knowledge to identify features from raw data, whereas ML automatically extracts simple features from the input data using an all-purpose learning procedure. These simple features are mapped into outputs using a complex architecture composed of a series of nonlinear functions (i.e., “hierarchical representations”) to maximize the predictive accuracy of the model. This measure can be improved using rich information contained in the biological research. We aimed to fill this void by assessing pathway-related genes for their performances in cancer prediction and identification.

We demonstrated that three cancer-related pathways (the Wnt signaling pathway, carbohydrate metabolism signaling pathway, and PI3K-Akt signaling pathway) have high predictive accuracy compared with DEGs for cancer.
prediction and identification. Furthermore, their performances were similar regardless the ML algorithm used. The use of DEGs as features has been previously documented. However, the outcomes suggest that all three pathway-related genes can be used as features for cancer detection. By assessing various cancer types, we observed that the features perform best for colorectal cancer followed closely by breast cancer and then gastric cancer. We speculated that the

Figure 3: Receiver operating characteristic curves for the Wnt signaling pathway-, PI3K-Akt signaling pathway-, and carbohydrate metabolism signal pathway-related genes for the three datasets.

Table 5: Performance of candidate pathway-related genes in cancer prediction.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Pathway-related genes</th>
<th>Wnt</th>
<th>Carbohydrate metabolism</th>
<th>PI3K-Akt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COAD (%)</td>
<td>BRCA (%)</td>
<td>STAD (%)</td>
<td>COAD (%)</td>
</tr>
<tr>
<td>RF</td>
<td>99.49</td>
<td>99.26</td>
<td>98.04</td>
<td>99.26</td>
</tr>
<tr>
<td>DT</td>
<td>89.45</td>
<td>92.56</td>
<td>78.51</td>
<td>99.39</td>
</tr>
<tr>
<td>KNN</td>
<td>99.42</td>
<td>98.85</td>
<td>93.54</td>
<td>97.83</td>
</tr>
</tbody>
</table>

Note. Tumor samples in the positive group versus the normal samples. COAD, colorectal cancer; BRCA, breast cancer; STAD, gastric cancer.
function of pathway-related genes in various cancer types can vary and are more serious in colorectal cancer. Results also show that these three pathway-related genes achieved different performances for one cancer type, and this can result in contributions of their compositions that vary based on the type of tumorigenesis.

Finally, these results demonstrate that the SVM and RF algorithms are superior to those of DT and KNN in genomics research. This variation might be because the classifier differs from one problem to another (e.g., the SVM model tends to meet rule-matching well when hundreds of thousands of dimensions exist, as in this study, whereas DT and KNN depend largely on feature selection in nonlinearly related variables). Unlike studies using other ML methodologies, this study offers additional insights on feature extraction for cancer classification. Each of the novel observations we found are worthy of further investigation.

5. Conclusions

We propose that pathway-related genes have the potential to be used as biomarkers for cancer prediction. We demonstrated that the Wnt signaling pathway, carbohydrate metabolism signaling pathway, and PI3K-Akt signaling pathway can be incorporated into ML models to achieve better prediction performance. The proposed features have the potential to facilitate preoperative care of patients with cancer.

Data Availability

Genetic data were downloaded from The Cancer Genome Atlas, a publicly accessible dataset (https://cancergenome.nih.gov/), and the pathway-related genes were derived from the Kyoto Encyclopedia of Genes and Genomes (http://www.kegg.jp/) analysis.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors’ Contributions

Pengliang Chen and Pengwei Shi contributed equally.

Acknowledgments

This work was supported by Guangdong Provincial Science and Technology Projects (nos. 2015A030313254 and 2016A020215114).

References


Figure 4: 3D bar plots of the three candidate features in various types of cancers. The z-axis indicates percent area under the curve. (a) COAD, (b) BRCA, and (c) STAD.


