Computational and Structural Biotechnology Journal 18 (2020) 2639–2646

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A deep learning and similarity-based hierarchical clustering approach for pathological stage prediction of papillary renal cell carcinoma

Sugi Lee [a,b,1], Jaeeun Jung [b,1], Ilkyu Park [a,b], Kunhyang Park [c], Dae-Soo Kim [a,b,*]

[a] Department of Bioinformatics, KRIBB School of Bioscience, Korea University of Science and Technology (UST), 217 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
[b] Department of Environmental Disease Research Centers, Korea Research Institute of Bioscience & Biotechnology (KRIBB), 125 Gwahak-ro, Yuseong-gu, Daejeon, Republic of Korea
[c] Department of Core Facility Management Center, Korea Research Institute of Bioscience & Biotechnology (KRIBB), 125 Gwahak-ro, Yuseong-gu, Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 29 June 2020
Received in revised form 16 September 2020
Accepted 16 September 2020
Available online 24 September 2020

Keywords:
Deep learning
Papillary renal cell carcinoma
Pathological tumour stage
Similarity-based hierarchical clustering

ABSTRACT

Papillary renal cell carcinoma (pRCC), which accounts for 10–15% of renal cell carcinomas, is the second most frequent renal cell carcinoma. pRCC patient classification is difficult because of disease heterogeneity, histologic subtypes, and variations in both disease progression and patient outcomes. Nevertheless, symptom-based patient classification is indispensable in deciding treatment options. Here we introduce a prediction method for distinguishing pRCC pathological tumour stages using deep learning and similarity-based hierarchical clustering approaches. Differentially expressed genes (DEGs) were identified from gene expression data of pRCC patients retrieved from TCGA. Thirty-three of these genes were distinguished based on expression in early or late stage pRCC using the Wilcoxon rank sum test, confidence interval, and LASSO regression. Then, a deep learning model was constructed to predict tumour progression with an accuracy of 0.942 and area under curve of 0.933. Furthermore, pathological substage information with an accuracy of 0.857 was obtained via similarity-based hierarchical clustering using 18 DEGs between stages I and II, and 11 DEGs between stages III and IV, identified through Wilcoxon rank sum test and quantile approach. Additionally, we offer this classification process as an R function. This is the first report of a model distinguishing the pathological tumour stages of pRCC using deep learning and similarity-based hierarchical clustering methods. Our findings are potentially applicable for improving early detection and treatment of pRCC and establishing a clearer classification of the pathological stages in other tumours.

1. Introduction

The development of next generation sequencing techniques (NGS) has enabled analysis and modeling using information from cancer patients. The Cancer Genome Atlas (TCGA) projects have RNA-seq data with various tissue types and clinical information for patients [1]. Cancer data largely varies with each patient, making patient classification based on specific properties difficult as this does not often show a clear and definite difference in disease status [2]. Therefore, results vary greatly depending on which variable is selected and which analysis method is used. In other words, when constructing a classification model, the outcome could vary greatly depending on the markers, patient groups, and model algorithms used.

There are various forms of kidney cancer, the most common being renal cell carcinoma (RCC), which also has various types based on histological differences [3]. RCCs include clear cells, papillary, chromophobe, cystic-solid, and collecting ducts renal cell carcinoma. Cancers classified according to their form and characteristics are treated differently according to their types [4,5]. Among these, papillary renal cell carcinoma (pRCC), which accounts for 10–15% of RCCs is the second most prevalent RCC after clear cell renal cell carcinoma (ccRCC) [6]. pRCC has the form of a papilla and is divided into two types, type 1 and type 2, depending on the size, appearance, prognosis, and biological differences [7]. Although most RCCs, including pRCC, exhibit characteristic morphologies that enable easy categorisation, they can show considerable morphological heterogeneity, and it is not uncommon for there to be difficulty in assigning a tumour type [8]. According to recent report, 12% of pRCC patients remain unclassified [9].

* Corresponding author at: Korea Research Institute of Bioscience and Biotechnology, 125 Gwahak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea.
E-mail address: kds2465@kribb.re.kr (D.-S. Kim).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2020.09.029

Kidney cancer is often free of symptoms for a considerable period after tumour development. When the cancer is small, there are very few symptoms until it grows and becomes large enough to push surrounding organs. Consequently, diagnosis is often delayed, and it may be found in an already metastasized state [10]. As the disease progresses, the size of the cancerous tumour grows and the chances of metamorphosis increase; according to the tumour, node and metastasis (TNM) staging system, which was revised in 2009, there are cases of lymph node transfer from stage III [7,11,12]. Cancer stage prediction is a process for estimating the likelihood that the disease has spread before treatment is administered to the patient. Fewer studies exist on genetic markers and therapeutic agents for the relatively recently defined pRCCs than for ccRCC, which has been extensively studied. Although pRCC has a relatively better prognosis than ccRCC, genetic studies for treatment are essential because of the high risk of recurrence [13]. Therefore, diagnosis using molecular markers is imperative at preoperative biopsy to allow clinicians to determine the best approach for treating and managing the disease.

In this study, we established a method for predicting the pathological stage of pRCC for early diagnosis and proper therapy of pRCC. This is based on the concept of a deep learning prediction model and similarity-based hierarchical clustering using TCGA transcriptome dataset. In addition, we designed an automated pathological tumour stage prediction R function consisting of two key modules; a deep learning prediction model module to distinguish early and late stage pRCC, and a similarity calculation module to classifying pathological sub-stage information.

2. Materials and methods

2.1. Data description

We extracted the transcriptome ('HTSeq-FPKM') and associated clinical data of TCGA Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) via the 'GDC-client' from the Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/) and then assembled it using *TCGAbiolinks* R package. "Solid Tissue Normal" and "Primary Solid Tumour" data were selected from the datasets. Clinical tumour stage data were retrieved from clinical data files, and data for 29 tumour patients without tumour stage information were excluded from a total of 288 patients. We selected 259 patients which had both gene expression and tumour stage clinical data to screen for genes involved in tumour progression. Group labels were applied to the initial dataset to distinguish early from late stage samples. Among the 259 transcriptomes with tumour stage information for pRCC, there were 172 stage I, 21 stage II, 51 stage III, and 15 stage IV samples. In stages I and II, the tumour is still confined to the papillary renal cell and has not spread to the central lymph node compartment, increasing the chances of survival. Whereas stages III and IV have more lymphatic metastasis, and decreased survival indices [14]. Therefore, we combined stages I and II as early stage, and stages III and IV as late stage.

Next, the patients were divided into training and validation datasets to build and fit the prediction model. In the past, various studies employed the 80:20 ratio for the partitioning of a dataset into training and validation datasets [1,15,16]. Therefore, we also applied this standard protocol using 80% data as the training dataset for model training and the remaining 20% data as validation dataset for final model validation. The validation dataset was not included in downstream analyses. The distribution of patients across training and validation datasets based on the clinical tumour stage is presented in Table 1. We generated a multidimensional scaling (MDS) plot with normal and tumour samples using whole genes (pre-processed), showing a clear separation between normal and tumour samples (Fig. S1a). However, samples at different stages of tumour development (stages I, II, III, and IV) formed an ensemble without clear separation (Fig. 1a, Fig. S1b).

2.2. Normalization of RNA expression

We used FPKM (fragments per kilobase of transcript per million mapped reads) values of expression quantification for 57,035 RNA transcripts. We normalized the wide range of variation in FPKM values using log2-transformation after adding 1 as a constant number to each FPKM value. Before normalizing the data, we removed the low expression genes to ensure the reliability of the gene sets and reduce the possibility of false positives. Genes with FPKM = 0 for >75% of the patients or with a maximum expression of <1 in all patients were considered low expression.

2.3. Identification of differentially expressed genes

To identify DEGs, we first performed differential expression analysis by comparing tumour and normal samples. Although the RNA expression values were normalized, the large number of patients used in this study may have resulted in one-sided bias due to outliers. Therefore, we determined whether the difference in gene expression values in early and late stage samples was statistically significant or not using the Wilcoxon rank sum test rather than the *t*-test. This is because the *t*-test uses the mean value and is affected by a few large outliers, whereas the Wilcoxon test minimizes the impact of outliers and compares the distributions of both groups to address the threshold of the *t*-test. Only genes with *p*-value of 0.05 or less are selected, and we compared early and late stages using the same method. Secondly, we computed the CI based on a Wilcoxon rank sum test to improve the performance of the gene sets. It is common to use the mean or the median to obtain a fold-change, but to reflect the difference in overall distribution, the CIs were computed and compared. The CI determines a range of expression values from the statistics of the observed data. The range has an associated confidence level that the true parameter is in the proposed interval. After determining gene expression intervals (CIs), genes showing differences in expression between the early and late stage samples were extracted by selecting genes that differed from normal samples and did not overlap with each other. Since the 95% confidence level is most commonly used [17], we selected the genes whose 95% CIs in the early and late stage samples were distinct from each other.

Table 1
Summary of datasets.

Status	Sample Size			Training set	Validation set
Solid Tissue Normal			32		
Primary Solid Tumor	Stage I	Early	172	139	33
	Stage II		21	15	6
	Stage III	Late	51	42	9
	Stage IV		15	11	4
	Unknown		29		
Total			288	207	52

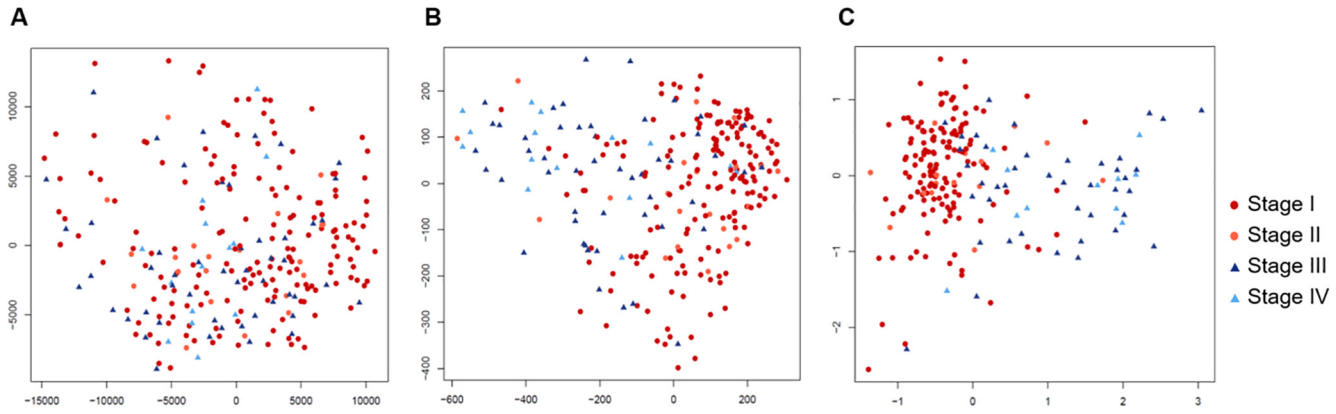

Fig. 1. MDS plot for pRCC patients with (A) whole genes, (B) DEGs filtered by confidence interval, and (C) the 33-feature gene expression. MDS, multidimensional scaling; TCGA-KIRP, The Cancer Genome Atlas – kidney renal papillary cell carcinoma; pRCC, papillary renal cell carcinoma; DEGs, differentially expressed genes.

2.4. Feature selection

Feature selection is an important task that determines the success of deep learning classification. Feature selection reduces the dimensionality of feature space by removing non-useful features and helps in improving the accuracy of classifiers for learning and prediction [15,18].

To reduce the dimensionality of the datasets and identifying relevant features for building an efficient deep learning model, we implemented the feature selection algorithm, LASSO, a regression model for preventing the built model from being over-fitting by minimizing the sum of the absolute values of the weights. When the model is over-fitting, the size of the model coefficient tends to increase excessively. Therefore, constraints are generally a way to limit the size of the coefficients. LASSO selects only a few important variables and features selection by reducing the other coefficients to zero. Finally, the gene is selected using a LASSO regression model that minimizes error during cross-validation in the R package, 'glmnet' (version 4.0). The LASSO equation is:

$$\sum_{i=1}^{N}\left(y_i - \frac{1}{1+\exp(-x_i^T\beta)}\right) \text{ subject to } \sum_{j=1}^{P}|\beta_j| \le t$$

where $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ are covariates of the ith observation and t > 0 is a tuning parameter.

When using LASSO, slight variations in the statistical data, such as adding or removing a few observations, will lead to significant changes in the coefficient estimates (β) [19]. To obtain reliable features from different distributions of samples due to the heterogeneity of pRCC, ten sub-sample groups comprising 80% of the training dataset were extracted randomly and tested 100 times by LASSO under various conditions. We tried to select more stable features not affected by variations in the samples through iteration-like bootstrap. Accordingly, more than 20% of repetitive genes in more than eight groups were finally selected.

2.5. Construction of the deep learning model

We applied the deep learning technique to build a model that distinguishes the pathological stages (early and late stage) of primary tumours. Compared to other machine learning methods, deep learning yields better outcomes with larger patient groups and the data to be used for stage classification can be learned separately. Since each cancer has different characteristics, deep learning can be modelled to match the characteristics of various cancer types, and adjusted by setting various hyperparameters, including epochs, activation functions, and hidden layers. The optimal hyperparameters for each TCGA project's prediction model are not the same. Therefore, we interrogated the optimal hyperparameters by applying these hyperparameters randomly with a grid search. For epochs, 10, 50, and 100 were used, respectively. We used 'Rectified Linear', 'Maxout' and 'Tanh' as the activation function, and the equations are stated in Table 2 [20]. The hidden layer is composed of one or two layers. Using these hyperparameters randomly or in combination, we built 400 Feedforward Neural Networks (FNNs) using the h2o package (version 3.26.0.2). FNNs are the most fundamental part of artificial neural networks. In FNNs, the neurons are arranged in the form of layers, primarily input, hidden, and output layers. Connections also exist between the neurons of one layer and those of the next layer [21]. Fig. 2a represents the structure of single neuron and the equation of the neural network is:

$$\hat{y} = f\left(\sum_i w_i x_i + b\right)$$

where $f(\cdot)$ is activation function and input x_i, weight w_i, and bias b [20]. Fig. 2b shows a feedforward neural network. The performance of the hyperparameters was evaluated via a grid search, the AUC and accuracy values were assessed, and the best optimal model was selected based on the AUC and accuracy values.

2.6. Classification of pathological stages using similarity

We attempted to develop an automated method for predicting the pathological sub-stages of early and late stage pRCC patients by evaluating the similarities among tumour samples using the -DEGs in the different stages (stage I vs. II and stage III vs. IV).

We identified DEGs for each tumour stage using the Wilcoxon rank sum test and quantile approach. In accordance with the pre-method for identification of DEGs, we performed the Wilcoxon rank sum test to obtain statistically significant differences between stages I and II and between stages III and IV. To obtain the genes which were clearly distinguished between stages, we used the quantiles of gene expression. Although, CI approach is suitable for groups with obvious discrepancies, such as early and late

Table 2
Activation function of deep learning method.

Function	Formula	Range
Tanh	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f(\cdot) \in [-1, 1]$
Rectified Linear	$f(x) = \max(0, x)$	$f(\cdot) \in \mathbb{R}_+$
Maxout	$f(x_1, x_2) = \max(x_1, x_2)$	$f(\cdot) \in \mathbb{R}$

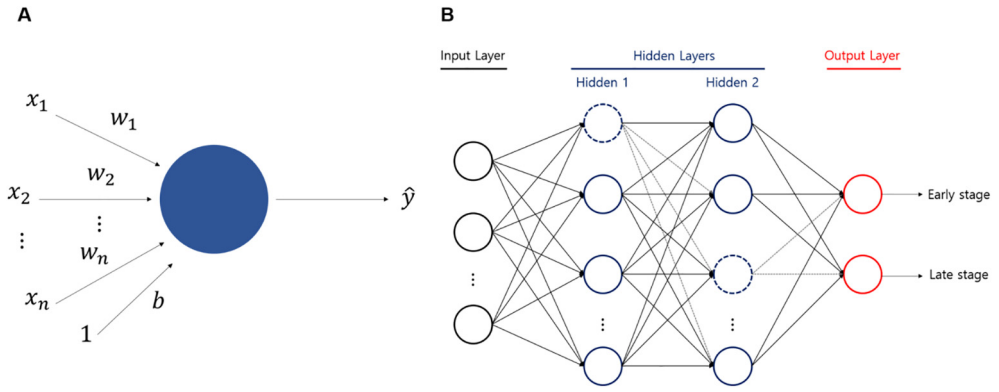

Fig. 2. Feedforward artificial neural network. (A) Structure of a single neuron that obtains output $\hat{y}$ by input x_i, weight w_i, and bias b. (B) Feedforward artificial neural network. Each circular node represents an artificial neuron and an arrow represents data flow from one neuron to another.

stages, it is difficult to find enough differences between stages I and II and between stages III and IV. The quantile approach is a highly valuable complement to the mean approach for detecting differential gene expressions [22]. Therefore, in the early stage, we selected DEGs within the 30th percentile of stage I which was larger than the 70th percentile of stage II or, the 30th percentile of stage II was larger than the 70th percentile of stage I. Even though the first and third quartiles are used generally, since stages I and II were very similar, the 30th and 70th percentiles were used. For the late stage, DEGs were selected in the same manner as above using the first and third quartiles.

With DEGs of the pathological tumour stages, the correlation between samples (early or late stage) was calculated, and the top 20% of samples were extracted to compute the Euclidian distance among samples. To perform similarity-based hierarchical clustering, only the Euclidian distance values were necessary; however, to overcome pRCC heterogeneity and to reduce noise during clustering, we acquired a homogeneity set [23,24]. In other words, for early stage cases, the correlation was calculated between all 154 samples (stage I and stage II patients from the training dataset), and then the 30 most closely correlated samples were selected to measure distance.

Then, these samples were hierarchically clustered using P-Values via Multiscale Bootstrap ('pvclust' R package, version 2.2.0, https://github.com/shimo-lab/pvclust) with the Ward.D2 method. The numeric value (vector) of the number of bootstrap replications was set as 1000 and the initialized seed was set as 20. Hierarchical clustering shows the samples with close pathological stages. Based on the hierarchical clustering plot, the pathological tumour stage was annotated as the stage information of the nearest sample. All the analyses were performed in R (version 3.6.0).

In order to use this classification model, we have implemented an R function which shows the final result in a hierarchical clustering plot with the correlation value, AU (approximately unbiased) p-value, and BP (bootstrap probability) value for each cluster in a dendrogram. A detailed code is provided in Supplementary File 2.

2.7. Evaluation of the deep learning model performance compared with machine learning methods

In addition, early and late stage classifications were performed using well-known machine learning techniques and the R caret package. Multiple techniques, such as RF, SVM, cforest and GLM were applied to generate appropriated classification models. The hyperparameters of machine learning models were tuned through the Randomised Grid Search Cross-Validation approach (Table S1). We computed AUC and accuracy through 10 cross validation processes, and then compared the performances between four machine and deep learning methods.

3. Results

3.1. Identification of differentially expressed genes

To identify the genes that distinguish the tumour stages, we first extracted genes that differed from normal, then conducted a Wilcoxon rank sum test on all genes to find differentially expressed genes (DEGs) between normal and tumour samples, and 18,985 genes with p-values < 0.05 were selected. Among the normal and tumour DEGs, 6,687 genes capable of distinguishing early and late stage tumours were further selected (p-value $\leq$ 0.05).

To improve effectiveness, genes with different expressions between early and late stages were filtered through the 95% confidence interval (CI). Use the 95% CI to distinguish the pattern as shown in the Fig. S2, and the expression is as follows.

$$Early_{upper} \leq Late_{lower} \text{ and } Late_{upper} \leq Normal_{lower} \quad (1)$$

or

$$Late_{upper} \leq Early_{lower} \text{ and } Early_{upper} \leq Normal_{lower} \quad (2)$$

or

$$Normal_{upper} \leq Late_{lower} \text{ and } Late_{upper} \leq Early_{lower} \quad (3)$$

or

$$Normal_{upper} \leq Early_{lower} \text{ and } Early_{upper} \leq Late_{lower}. \quad (4)$$

The number of cases using CI is as shown above. Where, *upper* is the upper boundary of the 95% CI and *lower* is lower boundary of the 95% CI. To compare the datasets, the lower boundary of one group should be larger than the upper boundary of the other group. Above all, since entire tumour samples must be distinct with normal samples, equations (1) and (2) show patterns similar to tumour-suppressed genes with the lower boundary of the normal 95% CI being higher than the upper boundary of the tumours. Equations (3) and (4) show patterns similar to tumour-derived genes with the lower boundary of the tumour 95% CI being higher than the upper boundary of the normal. Using equations (1) and (4), we extracted genes that were more highly expressed in the late stage than early stage samples. The opposite was the case for equations (2) and (3). Except for the above expression patterns, the

Table 3
Number of differentially expressed genes.

	Wilcoxon rank sum test (p-value < 0.05)		95% CI	LASSO regression
	Normal vs. Tumour	Early vs. Late Stage		
Number of DEGs	18,985	6,687	1,624	33

Abbreviations: CI, confidence interval.

Table 4
Hyperparameters of deep learning model.

Hyperparameters		Values
Epochs		10
Layers	Input	33
	Hidden 1	5
	Hidden 2	5
	Output	2
Activation function	a1	Rectifier
	a2	Rectifier
	aO	Softmax
Drop out	d0	0%
	d1	0%
	d2	0%
L1-regularization	L11	0.0001
	L12	0.0001
	L1O	0.0001
L2-regularization	L21	0.001
	L22	0.001
	L2O	0.001

remaining were excluded because normal samples acting as noise. To identify non-overlapping genes, we examined 95% CI for every gene, resulting in the selection of 1624 genes (Fig. 1b).

3.2. Feature selection

In order to verify the correlation between the gene expression level and the tumour projection, the effect of overfitting was excluded from the interaction between the genes. Therefore, we used the variable method of the LASSO (Least Absolute Shrinkage and Selection Operator) regression algorithm to solve the multicollinearity problem, which has strong correlation between independent variables. To reduce the influence of pRCC heterogeneity on feature selection, we applied the feature selection algorithm to each of the ten sub-sample groups, and then the selected genes were combined to ensure consistently stable features (see methods). Consequently, we identified 33 efficient features which could distinguish early and late stage pRCC (Table 3, Table S2), and generated an MDS plot with each pRCC stage sample using these features. The plot shows a definite separation between early and late stage samples, although some samples were mixed with other stage samples (Fig. 1c). Among the 33 features, 16 genes had been already investigated previously in tumour studies from kidney or various carcinoma types (Table S2).

Furthermore, to confirm the selected features were representative of pRCC and were not biased towards the training dataset, we extracted five independent datasets from whole samples; all samples belong to at least one dataset. All the feature selection algorithms (Wilcoxon rank sum test, confidence interval, LASSO) were applied to the five datasets under the same conditions. Then, each selected feature of the five datasets was shared over 90% with the 33 features (31, 30, 33, 33, and 31 features, respectively) (Fig. S3).

3.3. Construction of deep learning model to distinguish early and late stage pRCC

To predict early and late stage pRCC, we built a deep learning model, using the selected features (33 genes). However, due to the difficulty of finding a suitable hyperparameter for the deep learning model, we designed 400 random deep learning models using random hyperparameters with a grid search, and the optimal model was selected based on the area under curve (AUC) and accuracy. The best hyperparameters of the optimal deep learning stage prediction model for pRCC are shown in Table 4.

To evaluate the performance of deep learning, we compared the results against the random forest (RF), support vector machine (SVM), cforest, generalized linear model (GLM) which have been used in previous machine learning studies. The performance of the models was compared using different measures, including not only accuracy and AUROC, but also balanced accuracy and F_1 scores, which are more informative in evaluating estimates on imbalanced datasets. These results show that the accuracy and AUC of deep learning models yielded the best early and late stage pRCC classification outcomes. The deep learning prediction accuracy and AUROC of pRCC were 0.942 and 0.933, respectively, which are much higher than the results of five other machine learning methods (Table 5). Asides the deep learning outcome, the random forest (RF) method performed better with an accuracy of 0.846 and AUROC of 0.914 (PR-AUC of 0.891).

3.4. Classification of pathological sub-stages of pRCC using similarity

Between early stages I and II and between late stages III and IV, patients at different stages of pRCC were not clearly classified, possibly due to tumour heterogeneity or differences in tumour type. Therefore, we identified specific DEGs that distinguished pathological sub-stages, 18 genes between stages I and II and 11 genes

Table 5
Overall performance table for stage prediction of pRCC using deep learning and four machine learning methods.

Model	Training dataset		Validation dataset			
	Accuracy	SD	Accuracy	AUROC	Balanced Accuracy	F_1 Score
DL	0.922	0.082	0.942	0.933	0.885	0.963
RF	0.864	0.062	0.846	0.914	0.692	0.907
SVM	0.883	0.058	0.788	0.720	0.731	0.857
cforest	0.848	0.063	0.808	0.804	0.641	0.884
GLM	0.839	0.073	0.769	0.692	0.692	0.846

Abbreviations: DL, deep learning; RF, random forest; SVM, support vector machine; GLM, generalized linear model; SD, standard deviation.

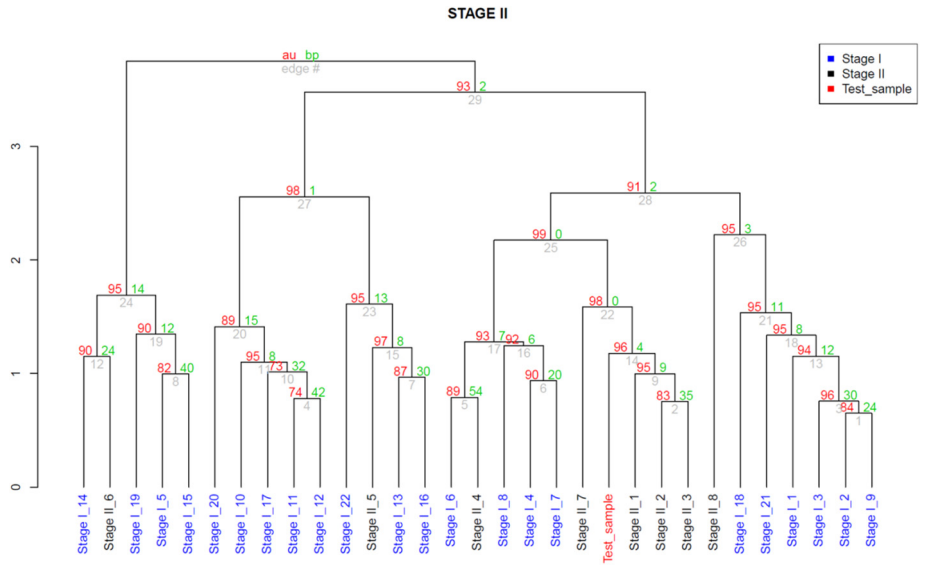

Fig. 3. The similarity-based hierarchical clustering plot shows the relationships found within stage data. At each edge, red and blue letters represent approximately unbiased *p*-values and bootstrap probability, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between stages III and IV, using the Wilcoxon rank sum test and quantile approach.

With these DEGs, we calculated the correlation among early or late stage pRCC 0samples determined via the deep learning prediction model, computed Euclidian distances between samples to indicate similarities, and performed the unsupervised hierarchical clustering with bootstrap support to show the relationship among patients based on pathological tumour stage similarities. Therefore, when a pRCC patient was predicted as "early stage" via the deep learning algorithm, the relationship between stage I and stage II was also displayed through the hierarchical clustering plot to visually represent which stage was more likely (Fig. 3).

We evaluated the performance of our methods via a validation dataset for 52 patients which was entirely isolated from our initial analysis. The dataset contained 39 early (33 stage I, 6 stage II) and 13 late (9 stage III, 4 stage IV) stage patients. Of these, a total of 49 patients (excluding only three patients) were predicted as reliable (correctly annotated) early or late stage with an accuracy of 0.94 via the deep learning prediction model (Table 5). The tumour stages for 41 of these 49 patients were accurately classified: 33 of 39 early-stage patients were closer to stage I and 6 were closer to stage II, with true positive rates (TPR) of 85% (Table S3). Of 10 late-stage patients, 9 were closer to stage III and 1 were closer to stage IV, with TPRs of 80% (Table S4). Thus, the similarity calculation approach for classifying the pathological sub-stages of pRCC successfully discriminated between stages with TPR of 83.7% (Table S3).

We further tried to develop an automated tumour stage prediction method to annotate the pathological stages of pRCC patients by constructing an R function. This allowed the classification of early and late stage pRCC and their pathological sub-stage by calculating the distance between samples and performing hierarchical clustering using gene expression data derived from RNA-seq. The R function developed in this study consists of two key modules: a deep learning prediction model to distinguish early and late stage pRCC and a similarity calculation for detailed pathological sub-stage classification. Further, it shows the relationship between patients and tumour stage as displayed in the hierarchical clustering plot (Fig. 3). Typically, the user needs to provide gene expression (FPKM) values of biomarker genes for every patient. The output includes a patient list and the corresponding predicted pRCC stage (early or late), and also provides pathological sub-stage information about which stage is closer to the clustering group. Each patient's result will be provided as a dendrogram (Supplementary File 1). The R function source code is reported in Supplementary File 2.

3.5. Performance assessment of classification methods

Furthermore, we evaluated our pathological tumour stage prediction methods using additional test dataset. Since, there are no publicly available pRCC RNA-seq datasets, to the best of our knowledge, we used the 29 unknown stage patients excluded from the initial dataset. The American Joint Committee on Cancer (AJCC) clinical stage information was available for 11 of these patients (seven stage I and four stage II patients). We examined these patients using our automatic analysis process and found that one stage II patient had been wrongly classified as late stage in the first module, and two stage II patients predicted as stage I in the second module, indicating the high performance and accuracy (0.80) of our process (Table S4). The remaining 18 samples without stage information were classified by our methods (Table S6).

4. Discussion

Reliable tumour stage prediction is important as it is used as a criterion for determining physical or chemical treatment methods and disease prognosis. Most gene expressions associated with tumour characteristics are distinguished by differences in normal tissue [25,26]. Among the genes that reveal tumour characteristics, genes that distinguish stage characteristics are rare. Unlike the differences in gene expression levels in normal and tumour cells, gene expression based on stages in tumour tissues are not significantly different [1,27]. Nevertheless, identifying differences in gene expression for various cancer stages is a valid verification process, in terms of diagnosis and treatment, especially since early findings greatly affect the patient's welfare. Thus, it is still valuable to identify differences in gene expression by pathological tumour stage.

In this study, we focus on analysing the dataset of poorly studied pRCCs, which are well-known to have a high risk of recurrence. Therefore, characterization of pRCC stages is necessary for early detection and effective treatment. In a previous study, 17 hub genes were identified through network analysis to distinguish pRCC pathological stages with AUCs > 0.7 using TCGA data for 106 patients [28]. Another recent study reported 104 genes identified through machine learning methods with an PR-AUC of 0.804 and accuracy of 88% using TCGA data for 161 patients [29]. These studies only used a subset of the available TCGA dataset (from a total of 288 primary solid tumour patients, 259 with stage information) and obtained low performances. Although several machine learning approaches exist for distinguishing tumour stages based on gene expression data, there remains a critical need to improve accuracy. Recent advances in the machine learning community have shown great promise for the application of deep learning to cancer classification [30]. Furthermore, several supervised and unsupervised deep learning-based classification methods have been proposed for cancer detection and diagnosis, and these have demonstrated superior performance over classical methods, such as SVM and RF [31,32].

In the present study, we constructed an optimum deep learning model which distinguished early and late stage pRCC patients using gene expression data. We used 259 (207 for biomarker identification and model training, and 52 for model validation) of the 288 pRCC patient RNA-seq data available in TCGA, only excluding 29 patients with unknown stages. In selecting features for distinction between early and late stage, differentially expressed genes between early and late stage patients were selected using the Wilcoxon rank sum test and CIs, multicollinearity was eliminated using LASSO regression, and 33 genes were finally selected for use in the prediction model. To predict tumour patient' pathological stage, the deep learning method was used; we set up 400 models to optimize hyperparameters, and selected models based on their predictability and suitability. The predictive performance of the optimum deep learning model was tested by comparing the AUC and accuracy values with the results of four other machine learning methods using validation datasets. As expected, deep learning showed the best predictive power with an accuracy of 0.942 and AUROC of 0.933 (PR-AUC of 0.891), mainly because the optimum parameter was identified by applying multiple random parameters. Considering the relatively poor prediction of late stage classification in the validation dataset, it is likely that overfitting occurred to some extent for early stages due to the unequal number of patients. The number of pRCC patients with late stage disease (53) used for stage classification modelling was relatively small compared to those with early stage disease (154). As the number of late stage samples that can be analysed increases, the performance of the stage classification model will improve.

After dividing the early and late stages, we classified the pathological tumour stage by calculating the similarity and cluster analogous approach using the 18 DEGs for stages I and II, 11 DEGs for stages III and IV and quantile filtering. The quantity and quality of TCGA molecular data have been lauded by a large number of scientists, and these data have resulted in studies that have significantly advanced our understanding of cancer biology [33]. In addition, numerous independent investigators have used TCGA as a resource to support their own studies and to help interpret molecular testing of individual patients in clinical settings [33–35]. Therefore, when a patient's tumour was predicted as early stage using the deep learning model, the patient was compared with stages I and II patients to identify the pathological sub-stage, using TCGA as a reference database. Then, we applied an unsupervised hierarchical clustering method, using similarity and distance measures to cluster most similar data points into the same cluster [36], yielding an accuracy of 0.84 for the validation dataset and accuracy of 0.80 for the additional test dataset.

Furthermore, we implemented an automated pRCC pathological stage prediction R function, which can analyse the gene expression data from a sample and predict whether it is an early or late stage patient and the pathological sub-stage.

pRCCs are frequently asymptomatic, and are often only incidentally detected on imaging related with other clinical causes. Consequently, diagnosis is often delayed, and it may be found in an already metastasized state [10]. In addition, 12% of pRCC patients remain unclassified [9]. Surgery is effective for localized pRCC (early stage); however, once pRCC becomes metastatic (late stage) the survival rate of patients drops sharply. Patients with higher pathological stages tend to have worse prognoses. Therefore, the pathological stage of pRCC is the most effective prognosis factor. The standard treatment for localized pRCC is surgery, including radical or partial nephrectomy, due to its insensitivity to radiotherapy and chemotherapy [37]. Targeted therapies have better results and fewer side effects compared with immunotherapy. However, targeted therapies are still limited and liable to drug resistance [38,39]. In order to overcome the current limitations in diagnosis, therapy and follow-up of pRCC, biomarkers for use during biopsy have been proposed as essential components for precision medicine. Moreover, 29 samples (10%) of pRCC provided by TCGA did not have pathological stage information. These could be exactly identified by pathological stage with molecular investigations. Therefore, molecular investigations with the biomarkers characterized in this study may help to identify the precise stage of disease, which could improve the prediction of oncological outcomes and lead to optimized target therapies.

Since this study was designed and performed using only TCGA dataset due to the non-existence of publicly available pRCC RNA-seq data, we recommend future confirmatory studies using larger datasets. Nevertheless, to the best of our knowledge, this is the first report of a model distinguishing the pathological tumour stages of pRCC using deep learning and similarity-based hierarchical clustering methods. Our findings are potentially applicable for improving early detection and treatment of pRCC and establishing a clearer classification of the pathological stages in other tumours.

CRediT authorship contribution statement

Sugi Lee: Conceptualization, Methodology, Software, Writing - original draft. **Jaeeun Jung:** Validation, Investigation, Formal analysis, Writing - review & editing. **Ilkyu Park:** Formal analysis. **Kunhyang Park:** Data curation, Investigation. **Dae-Soo Kim:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2019R1I1A2A01060140, 2014M3A9A5034349, 2018M3A9H3023077) and by the KRIBB Research Initiative Program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.09.029.

References

[1] Rahimi A, Gönen M. Discriminating early- and late-stage cancers using multiple kernel learning on gene sets. Bioinformatics 2018;34:i412–21. doi: 10.1093/bioinformatics/bty239.

[2] Fedele C, Tothill RW, McArthur GA. Navigating the challenge of tumor heterogeneity in cancer therapy. Cancer Discovery 2014;4(2):146–8.

[3] Muglia VF, Prando A. Renal cell carcinoma: histological classification and correlation with imaging findings. Radiol Bras 2015;48:166–74. doi: 10.1590/0100-3984.2013.1927.

[4] Arik D, Açikalin MF, Can C. Papillary renal cell carcinoma and collecting duct carcinoma combination. A case report and review of synchronous renal cell carcinoma subtypes in the same kidney. Arch Med Sci 2015;11:686–90. doi: 10.5114/aoms.2015.52378.

[5] Idikio HA. Human cancer classification: a systems biology- based model integrating morphology, cancer stem cells, proteomics, and genomics. J Cancer 2011;2:107–15. https://doi.org/10.7150/jca.2.107.

[6] Ravaud A, Oudard S, De Fromont M, Chevreau C, Gravis G, Zanetta S, Theodore C, Jimenez M, Sevin E, Laguerre B, Rolland F, Ouali M, Culine S, Escudier B. First-line treatment with sunitinib for type 1 and type 2 locally advanced or metastatic papillary renal cell carcinoma: a phase II study (SUPAP) by the French Genitourinary Group (GETUG). Ann Oncol 2015;26(6):1123–8. https://doi.org/10.1093/annonc/mdv149.

[7] Delahunt B, Eble JN. Papillary renal cell carcinoma: a clinicopathologic and immunohistochemical study of 105 tumors. Mod Pathol 1997;10:537–44.

[8] Warren AY, Harrison D. WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. World J Urol 2018;36(12):1913–26. https://doi.org/10.1007/s00345-018-2447-8.

[9] Ricketts CJ, De Cubas AA, Spellman PT, Kimryn Rathmell W, Linehan WM. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell Rep 2018;23:313–26. doi: 10.1016/j.celrep.2018.03.075.

[10] Vogelzang NJ, Stadler WM. Kidney cancer. The Lancet 1998;352 (9141):1691–6. https://doi.org/10.1016/S0140-6736(98)01041-1.

[11] Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med 2016;374:135–45. https://doi.org/10.1056/NEJMoa1505917.

[12] Waldert M, Haitel A, Marberger M, Katzenbeisser D, Ozsoy M, Stadler E, et al. Comparison of type I and II papillary renal cell carcinoma (RCC) and clear cell RCC. BJU Int 2008;102:1381–4. https://doi.org/10.1111/j.1464-410X.2008.07999.x.

[13] Lee J, Chae HK, Lee W, Nam W, Lim B, Choi SY, Kyung YS, You D, Jeong IG, Song C, Hong B, Hong JH, Ahn H, Kim C-S. Comparison of prognosis in types 1 and 2 papillary renal cell carcinoma and clear cell renal cell carcinoma in T1 stage. Korean J Urol Oncol 2018;16(3):119–25.

[14] Liu K, Ren Y, Pang L, Qi Y, Jia W, Tao L, et al. Papillary renal cell carcinoma: A clinicopathological and whole-genome exon sequencing study. Int J Clin Exp Pathol 2015;8:8311–35.

[15] Jagga Z, Gupta D. Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. BMC Proc 2014;8(S6). https://doi.org/10.1186/1753-6561-8-S6-S2.

[16] Kaur H, Bhalla S, Raghava GPS. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. PLoS One 2019;14. doi: 10.1371/journal.pone.0221476.

[17] Zar JH. Biostatistical Analysis – ERRATA. Prentice Hall New Jersey USA 2010:663. doi: 10.1037/0012764.

[18] Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinf 2017;18(1). https://doi.org/10.1186/s12859-016-1423-9.

[19] Melkumova LE, Shatskikh SY. Comparing Ridge and LASSO estimators for data analysis. Procedia Eng 2017;201:746–55. https://doi.org/10.1016/j.proeng.2017.09.615.

[20] Candel A, Parmar V, Lelell E, Arora A, Lanford J. Deep learning with H_2O. 5th ed. CA, USA: Mountain View; 2016.

[21] Alemu H, Wu W, Zhao J. Feedforward Neural Networks with a Hidden Layer Regularization Method. Symmetry (Basel) 2018;10:525. doi: 10.3390/sym10100525.

[22] Wang H, He X. Detecting differential expressions in GeneChip microarray studies: a quantile approach. J Am Stat Assoc 2007;102(477):104–12.

[23] Smid M, Rodríguez-González FG, Sieuwerts AM, Salgado R, Prager-Van Der Smissen WJC, Vlugt-Daane M Van Der, et al. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. Nat Commun 2016;7:1–9. doi: 10.1038/ncomms12910.

[24] Liao W, Ying Yang M, Zhan N, Rosenhahn B. Triplet-Based Deep Similarity Learning for Person Re-Identification 2017:385–93.

[25] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci 2001;98(19):10869–74.

[26] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale A-L, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000;406(6797):747–52. https://doi.org/10.1038/35021093.

[27] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science 2013;339(6127):1546–58.

[28] He Z, Sun M, Ke Y, Lin R, Xiao Y, Zhou S, Zhao H, Wang Y, Zhou F, Zhou Y. Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis. Oncotarget 2017;8(17):27904–14.

[29] Singh NP, Bapi RS, Vinod PK. Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med 2018;100:92–9. https://doi.org/10.1016/j.compbiomed.2018.06.030.

[30] Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, Vermeulen L, Wang X. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis 2019;8(9). https://doi.org/10.1038/s41389-019-0157-8.

[31] Karabulut EM, Ibrikci T. Discriminative deep belief networks for microarray based cancer classification. Biomed Res 2017;28:1016–24.

[32] Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification 2013.

[33] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173:400-416.e11. doi: 10.1016/j.cell.2018.02.052.

[34] Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, Yoshimatsu TF, Pitt JJ, Hoadley KA, Troester M, Ru Y, Lichtenberg T, Sturtz LA, Shelley CS, Benz CC, Mills GB, Laird PW, Shriver CD, Perou CM, Olopade OI. Comparison of breast cancer molecular features and survival by African and European ancestry in the cancer genome atlas. JAMA Oncol 2017;3(12):1654. https://doi.org/10.1001/jamaoncol.2017.0595.

[35] Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110. doi: 0.1016/j.ccr.2009.12.020.

[36] Shirkhorshidi AS, Aghabozorgi S, Wah TY. A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS One 2015;10:e0144059. doi: 10.1371/journal.pone.0144059.

[37] Chen L, Yuan L, Qian K, Qian G, Zhu Y, Wu CL, et al. Identification of biomarkers associated with pathological stage and prognosis of clear cell renal cell carcinoma by co-expression network analysis. Front Physiol 2018;9. doi: 10.3389/fphys.2018.00399.

[38] Ljungberg B, Bensalah K, Canfield S, Dabestani S, Hofmann F, Hora M, Kuczyk MA, Lam T, Marconi L, Merseburger AS, Mulders P, Powles T, Staehler M, Volpe A, Bex A. EAU guidelines on renal cell carcinoma: 2014 update. Eur Urol 2015;67(5):913–24. https://doi.org/10.1016/j.eururo.2015.01.005.

[39] Coppin C, Kollmannsberger C, Le L, Porzsolt F, Wilt TJ. Targeted therapy for advanced renal cell cancer (RCC): A Cochrane systematic review of published randomised trials. BJU Int 2011;108:1556–63. doi: 10.1111/j.1464-410X.2011.10629.x.