

Article

A Novel Attention-Mechanism Based Cox Survival Model by Exploiting Pan-Cancer Empirical Genomic Information

Xiangyu Meng ^{1,†}, Xun Wang ^{1,2,†}, Xudong Zhang ¹, Chaogang Zhang ¹, Zhiyuan Zhang ¹, Kuijie Zhang ¹ and Shudong Wang ^{1,*}

¹ College of Computer Science and Technology, Qingdao Institute of Software, China University of Petroleum, Qingdao 266580, China; xiangyumeng@s.upc.edu.cn (X.M.); wangxun@upc.edu.cn (X.W.); bigdongsir@163.com (X.Z.); s20070030@s.upc.edu.cn (C.Z.); flyeagle237@163.com (Z.Z.); z20070009@gmail.com (K.Z.)

² China High Performance Computer Research Center, Institute of Computer Technology, Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: shudongwang2013@sohu.com

† These authors contributed equally to this work.



Citation: Meng, X.; Wang, X.; Zhang, X.; Zhang, C.; Zhang, Z.; Zhang, K.; Wang, S. A Novel Attention-Mechanism Based Cox Survival Model by Exploiting Pan-Cancer Empirical Genomic Information. *Cells* **2022**, *11*, 1421. <https://doi.org/10.3390/cells11091421>

Academic Editors: An Pan, Baoli Yao, Chao Zuo, Fei Liu, Jiamiao Yang and Liangcai Cao

Received: 22 March 2022

Accepted: 19 April 2022

Published: 22 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Cancer prognosis is an essential goal for early diagnosis, biomarker selection, and medical therapy. In the past decade, deep learning has successfully solved a variety of biomedical problems. However, due to the high dimensional limitation of human cancer transcriptome data and the small number of training samples, there is still no mature deep learning-based survival analysis model that can completely solve problems in the training process like overfitting and accurate prognosis. Given these problems, we introduced a novel framework called SAVAE-Cox for survival analysis of high-dimensional transcriptome data. This model adopts a novel attention mechanism and takes full advantage of the adversarial transfer learning strategy. We trained the model on 16 types of TCGA cancer RNA-seq data sets. Experiments show that our module outperformed state-of-the-art survival analysis models such as the Cox proportional hazard model (Cox-ph), Cox-lasso, Cox-ridge, Cox-nnet, and VAECox on the concordance index. In addition, we carry out some feature analysis experiments. Based on the experimental results, we concluded that our model is helpful for revealing cancer-related genes and biological functions.

Keywords: deep learning; survival analysis; neural networks; Cox regression; cancer prognosis

1. Introduction

Since the 20th century, cancer has become a serious threat to human life and health. Multi-angle and multi-level prognostic studies on cancer emerge one after another. However, prognostic research for cancer is still a challenging task. One of the most important factors is that due to the existence of censored patient samples, the traditional analysis models cannot effectively determine the actual time of death [1]. Compared with traditional prognosis, survival analysis models focus more on the survival time point of the patient rather than the death time point. The use of survival analysis models can be very effective in dealing with censored data. The most widely used is the Cox proportional hazards model (Cox-ph) model [2]. The Cox-ph model is a semi-parametric proportional hazards model. The covariates of the model can explain the relative risk of a patient, called hazard ratio, prognostic index or risk score. According to the relative risk score, the risk of changes in various factors can be effectively analyzed, thereby helping doctors to develop effective targeted therapy strategies.

The arrival of the Human Genome Project has raised the survival sub-model to the research category of high-throughput multi-omics [3]. There are many experiments and studies showing that it is very meaningful to do cancer survival analysis on high-throughput transcriptome data [4–6]. Computationally, however, performing survival analysis on

high-dimensional transcriptome gene expression data is equivalent to solving a regression problem of complex nonlinear equations. Using traditional Cox-ph regression models cannot effectively handle high-dimensional feature representation and accurately predict prognosis. In the past two decades, some researchers have used some machine learning methods to modify the original Cox survival analysis model, and the regression effect of the model has been improved to a certain extent. Some methods use the support vector machine (SVM) algorithm to perform feature extraction and dimensionality reduction for high-dimensional gene expression data [7,8]. The result after mechanized dimensionality reduction of SVM fuses the original high-dimensional gene features, and the use of Cox-ph model to represent low-dimensional gene expression features can effectively implement survival analysis prediction. Some methods will use an ensemble learning method such as Cox-Boost, which divides the parameters into several independent partitions for ensemble training and fitting [9]. Some methods replace the original proportional hazards model by using the ensemble mean cumulative hazard function (CHF) with the help of the nonlinear ensemble method of random forests [10].

With the maturity of deep learning methods in different fields [11,12], the Cox model, based on an artificial neural network, has received extensive attention from researchers. To the best of our knowledge, the earliest application of artificial neural networks for survival analysis is Faraggi et al. [13]. They used four diagnoses as inputs to model the learning of a survival analysis for prostate cancer. Then Ching et al. designed a Cox-nnet composed of two-layer neural networks, and successfully used Cox-nnet to make reasonable survival analysis recommendations for 10 different cancer gene expression data [14]. Katzman et al. proposed a Cox regression model constructed by a multi-layer neural network, while formulating corresponding treatment recommendations based on the trained Cox model [15]. Huang et al. [16] proposed a survival analysis model for multi-omics data of breast cancer. They first constructed a co-expression feature matrix of mRNA data and miRNA data through gene co-expression analysis to alleviate the learning difficulties and overfitting problem caused by high-dimensional data. They finally proposed a multi-group student survival analysis model for breast cancer [16]. Kim et al. used a transfer learning approach, first pre-training a VAE model, and then using the VAE model for fine-tuning training on 20 TCGA datasets [17]. Using the above method effectively alleviates the overfitting problem caused by the high genetic dimension and the small number of training patients. Ramirez et al. [18] focused on the degree of correlation between different genes. They constructed gene association graphs through correlation coefficient scores, PPI networks and other methods, introduced the correlation map as a prior knowledge into the training of the GCN survival analysis model, and explained the important biological significance of the GCN model in survival analysis [18].

The high dimensionality and complex semantics of genetic data bring many challenges to feature extraction. So far, there is a lot of work that is trying to use new ideas to ensure rich feature extraction. The most classic method in deep learning was multilayer perceptron (MLP), which learns linear correlations between different data. It was considered to be the optimal choice for processing sequential data. There are a lot of works to extract high-dimensional genetic data based on MLP [14,16,19]. In the past decades, convolutional neural networks (CNN) have shown excellent results in computer vision. More and more studies have demonstrated the powerful ability of CNNs to deal with spatial structural features. Many recent works have attempted to extract features from high-dimensional genetic data using CNNs and demonstrated the transferability of CNNs in multi-omics data. Rehman et al. proposed densely connected neural network based N4-methylcytosine site prediction (DCNN-4mC), and this framework obtains the greatest performance for 4mC site identification in all species [20]. Chen et al. used Lasso and CNN as a target model and studied the trade-off between the defense power against MIA and the prediction accuracy of the target model under various privacy settings of DP [21]. Torada et al. proposed a CNN-based program, called ImaGene, on genomic data for the detection and quantification of natural selection [22]. Hao et al. proposed a biologically interpretable deep learning model

(PAGE-Net) that integrates histopathological images and genomic data [23]. Jeong et al. proposed a new tool called GMStool-based CNN for selecting optimal marker sets and predicting quantitative phenotypes [24]. Rehman et al. proposed the m6A-NeuralTool to extract the important features from the one-hot encoded input sequence based on CNN [25]. At the same time, some works [18,26] take advantage of novel feature extraction methods for sequence data and exhibit remarkable results.

Since the Generative Adversarial Network (GAN) proposed by Ian et al. [27], a lot of research has used this training strategy for data generation, reconstruction and dimensionality reduction tasks. GAN adopts an adversarial training strategy, which effectively learns the distribution of high-dimensional data and effectively generates results sampled from the overall data distribution. In recent years, GAN has been widely used in protein and gene sequence research. Repecka et al. [28] developed a variant of attention-based GAN and called it ProteinGAN. ProteinGAN learns the evolutionary relationships of protein sequences directly from the complex multidimensional amino acid sequence space and creates highly diverse new sequence variants with natural physical properties, which demonstrates the potential of GANs to rapidly generate highly diverse functional proteins within the biological constraints allowed by the sequence space. LIN et al. [29] proposed the DR-A framework, which implements dimensionality reduction for scRNA-seq data based on an adversarial variational autoencoder approach. Compared with traditional methods, this method can obtain a low-dimensional representation of scRNA-seq more accurately. Jiang et al. [30] introduced a novel GAN framework for predicting disease genes from RNA-seq data. Compared to state-of-the-art methods, the model improves the identification accuracy of disease genes.

In this paper, we propose a novel deep Self Attention Variational Autoencoder Cox Survival Analysis Model (SAVAE-Cox). This model takes advantages of adversarial transfer learning strategy. In the adversarial pretraining stage, the generator was a variational autoencoder (VAE), which is jointly trained with the discriminator. Meanwhile, we introduce a novel self-attention mechanism [31] to enhance semantically relevant features extraction of the encoder from high-dimensional data. After the pretraining stage, the generator was able to learn the common features of 33 cancer transcriptome data. Next, the encoder of the generator was used to learn survival analysis on 16 cancers. By comparison with state-of-the-art models such as Cox-nnet and VAEcox, our model achieved the highest concordance index on 10 TCGA cancer datasets. Finally, we performed feature analysis of SAVAE-Cox. We select oncogenes and compute correlations with hidden layer nodes in which we find that our hidden layer nodes are highly correlated with oncogenes. We used these nodes to draw Kaplan–Meier plots, and found that these nodes significantly affected the survival of patients. Based on the correlation of hidden layer nodes with genes, we selected leader genes, which we found enriched on cancer-related pathways. According to our experiments, we conclude that our proposed SAVAE-Cox model has significant cancer prognostic ability. Our source code of SAVAE-Cox is available at <https://github.com/menggerSherry/SAVAE-Cox> (Last visited on 21 March 2022).

2. Materials and Methods

2.1. Dataset Preparation

In this work we used 17 datasets from the TCGA database. These 17 datasets are bladder carcinoma (BLCA), breast carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal cell carcinoma (KIRC), brain lower-grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), stomach adenocarcinoma (STAD), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), sarcoma (SARC), uterine corpus endometrial carcinoma (UCEC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM) and pan-cancer (PANCAN). The detailed description of the dataset and download way were in Appendix A.1. Since there are a large number of empty genes and noise genes, it is necessary to perform some

preprocessing options to exclude some redundant noise genes. In Table 1, we present the statistics of the 17 datasets used in this work, including the number of samples in each dataset and the clinical information of 16 cancer types.

Table 1. Statistics of RNA-seq datasets (Pan-Cancer and 16 cancer types) used to train SAVAE-cox.

Cancer Type	Data Attribute		
	Total Samples	Censored Samples	Time Range
PANCAN	9895	#	#
BLCA	397	227	13–5050
BRCA	1031	896	1–8605
HNSC	489	302	1–6417
KIRC	504	347	2–4537
LGG	491	302	1–6423
LIHC	359	183	1–3765
LUAD	491	290	4–7248
LUSC	463	327	1–5287
OV	351	95	8–5481
STAD	345	227	1–3720
CESC	283	215	2–6408
COAD	415	239	1–4270
SARC	253	116	15–5723
UCEC	524	404	1–6859
PRAD	477	289	23–5024
SKCM	312	239	14–1785

#: Not measured in this experiment.

We use the PANCAN dataset to draw scatter plots of 56,716 genes and observe the statistical distribution. Figure 1 shows the distribution of mean and standard deviation of RNA-seq data. From the standard deviation distribution of Figure 1, we observe that there is a valley between 0.278–0.403, and plenty of genes are with zero variance. Therefore, we defined gene expression with a standard deviation in the range (0, 0.4) as noise genes, and removed those noise genes whose standard deviation satisfies this range. At the same time, in order to eliminate the influence of empty data, we removed the RNA-seq genes whose mean value satisfies (0, 0.8). We processed each RNA-seq dataset following these two strategies described above and selected 20,034 intersection genes of 16 cancer types. Before feeding the genes into our module, we performed feature wise min-max normalization of each gene of 16 cancer types.

Distribution of Statistic in Pan-Cancer Dataset

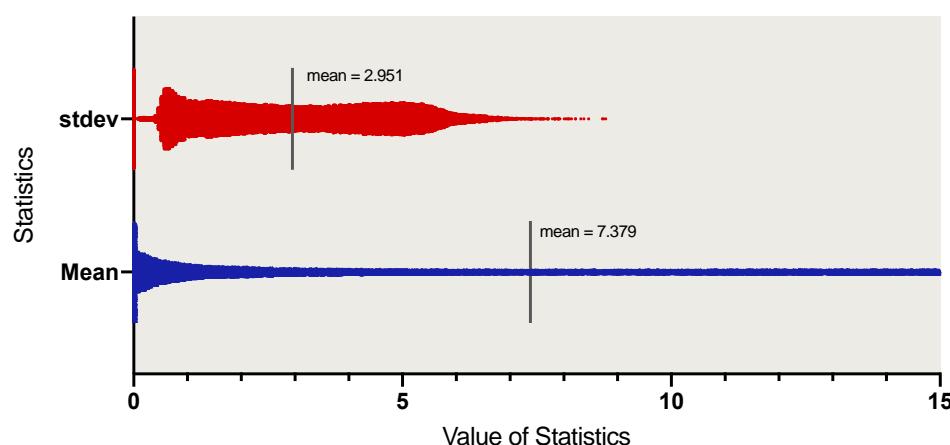


Figure 1. Scatter plot of Pan-Cancer data statistics distribution. The width of the scatter plot represents the number of patient samples. The solid black line represents the mean of the statistic.

2.2. Dimensionality Reduction Pretraining Using GAN

We adopted the strategy of a generative adversarial network (GAN) [27] to design the pre-training stage. The pre-training process is shown in Figure 2a. The generator G takes the genes x_{in} as input and generates the reconstructed genes x_{rec} . The discriminator D takes the x_{in} or x_{rec} as the input and outputs a value which reflects the authenticity of the gene. Through adversarial training with the D , the encoding module E of G gradually improves the feature extraction ability to generate a low-dimensional feature z for x_{rec} generation. After training, E can be used for dimensionality reduction of x_{in} . Compared with computational dimensionality reduction methods, our dimensionality reduction based E was a data-driven method and can be adaptively adjusted according to different data characteristics.

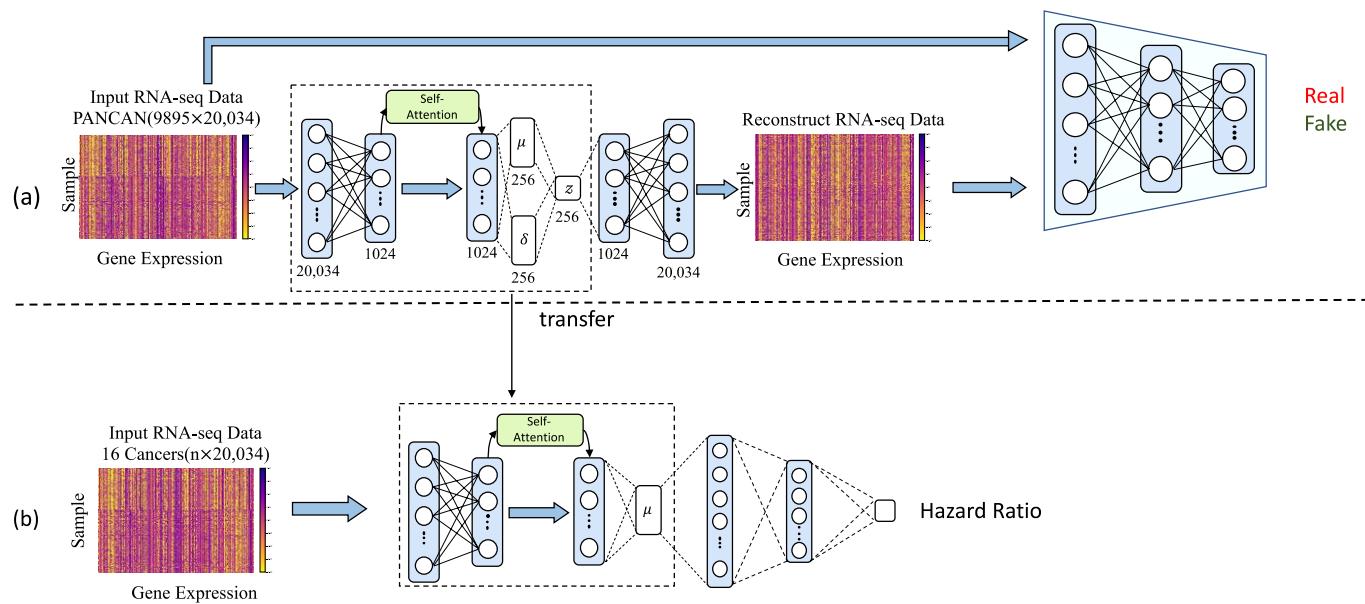


Figure 2. Overview of the SVAE module. (a) Dimensionality reduction pretraining stage using GAN. (b) Survival analysis based on transfer learning.

The network structure of the G is a self-attention VAE(SVAE) framework. For each x_{in} , SVAE is described as:

$$\mu(x_{in}) = w_\mu(\mathcal{L}^\alpha(\delta(w_h x_{in} + b_h))) + b_\mu \quad (1)$$

$$\sigma(x_{in})^2 = \exp(w_\nu(\mathcal{L}^\alpha(\delta(w_h x_{in} + b_h))) + b_\nu) \quad (2)$$

$$z = \mu\zeta + \sigma\zeta \sim N(0, 1) \quad (3)$$

$$x_{rec} = w_h z + b_h, \quad (4)$$

where $\mu(x_{in})$ and $\sigma(x_{in})^2$ are the mean and the variance of the Gaussian distribution. δ is the activate function. ζ is randomly sampled from the standard Gaussian distribution.

We introduce a residual self-attention module (Figure 3) in the hidden layer of the encoder to enhance the fitting ability of VAE [32]. The self-attention mechanism [31] can effectively learn the semantic correlation of high-dimensional features. It is denoted as:

$$\mathcal{L}^\alpha(x) = sa(x) + \alpha \times x. \quad (5)$$

In Equation (5), α represents a learnable parameter, which can adaptively adjust the weight of the residual connection. sa stands for Self-Attention Module [31]. It is denoted as:

$$sa(x) = \text{softmax}(Q(x)^T \times K(x)) \times V(x), \quad (6)$$

where Q , K , and V represent the query, key and value obtained by performing three Dense layers on the input x .

D is a simple binary classification network whose framework is represented as follows:

$$D(x) = \mathcal{L}_{out}^d \odot \mathcal{L}_n^d \odot \mathcal{L}_{n-1}^d \odot \cdots \odot \mathcal{L}_1^d(x), \quad (7)$$

where the $\mathcal{L}_n^d : \mathbb{R}^N \rightarrow \mathbb{R}^{\frac{N}{2}}$ maps high-dimensional features to low-dimensional features through linear transformation. Finally, the output feature of \mathcal{L}_n^d was fed to a classification layer $\mathcal{L}_{out}^d : \mathbb{R}^{\frac{N}{2}} \rightarrow \mathbb{R}$. The classification layer generates a numerical value, which represents the judgment of the input gene.

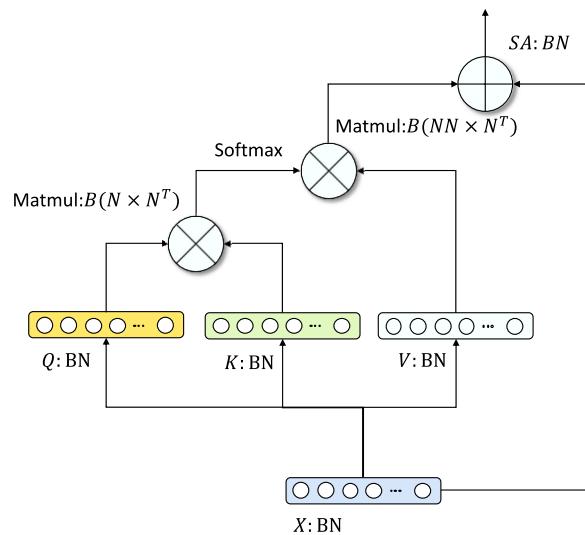


Figure 3. Network framework of residual self-attention module. This network structure can learn the latent semantic correlation of genes.

We introduce the Wasserstein loss [33] to train G and discriminator D jointly. For G , the goal is to synthesize more similar reconstructions x_{rec} so that the discriminator cannot describe whether it is real or fake. For D , the goal is to distinguish between true genes x_{in} and reconstructed genes x_{rec} synthesized by G . The loss of G and D is computed as:

$$\mathcal{L}_{gan} = \mathbb{E}_{x_{in} \sim P_{data}(x_{in})}[D(x_{in})] - \mathbb{E}_{x_{in} \sim P_{data}(x_{in})}[D(G(x_{in}))] - \lambda_p \mathbb{E}_{x \sim \mathcal{X}}[\|\nabla_x D(x)\|_2 - 1]^2, \quad (8)$$

where λ_p represents the hyper parameter for setting the gradient penalty, and \mathcal{X} represents the overall Sample Space of x_{in} and x_{rec} . The Wasserstein loss was calculated to minimize the Wasserstein distance between x_{in} and $G(x_{in})$ by the D , which made the overall sample distribution of x_{in} and $G(x_{in})$ more similar. Furthermore, we introduce the Kullback–Leibler divergence [32] used in training the VAE. Given the input Genes x_{in} , it is denoted as:

$$\mathcal{L}_{KL} = \sum_{i=0}^n (\mu(x_{in})^2 + \sigma(x_{in})^2 - \log(\sigma(x_{in})^2) - 1), \quad (9)$$

where n represents the dim of z . This error measures the distance between the real latent code z under standard Gaussian distribution and the posterior latent variable $P(z|x_{in})$ generated by encoder. The introduction of Kullback–Leibler divergence into the generator can guarantee the similarity between low-dimensional latent variables, which significantly improves the authenticity of the distribution of x_{in} and $G(x_{in})$. At the same time, we introduce $L1$ loss to measure the similarity of each sample. The $L1$ loss is computed as:

$$\mathcal{L}_{L1} = \|x_{rec} - x_{in}\|_1. \quad (10)$$

Unlike the Wasserstein loss and KL divergence, the L1 loss focuses on making G learn to synthesize x_{rec} more similar to x_{in} on genes-wise. Therefore, the overall loss function for G can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{gan} + \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_{L1}, \quad (11)$$

where λ_1 and λ_2 represent the hyper parameters. We therefore aim to solve:

$$\theta_G^* = \operatorname{argmin}_G \operatorname{max}_D \mathcal{L}_{total}. \quad (12)$$

2.3. Survival Analysis Based on Transfer Learning

After the pre-training stage, we transfer the weights learned by the encoder in the pre-training stage to the survival analysis stage as shown in Figure 2b. At the same time, we set an additional classification module to learn the hazard ratio. The hazard ratio measures the likelihood a patient has of dying, and a higher hazard ratio indicates a higher likelihood that a patient will die.

We adopt the training strategy of Cox-ph [2] to train our module, which is denoted as:

$$h(t|x_i) = h_0(t) \exp(wx_i), \quad (13)$$

where $h_0(t)$ was baseline hazard function, w was the trainable parameters of the module, and x_i represents the risk factors of patients. At this stage, x_i is the low-dimensional feature $\mu(x_{in})$ that output by SAVAE encoder. We aim to solve:

$$\theta^* = \operatorname{argmin}_\theta \sum_{C(i)=1} (wx_i - \log \sum_{t_j \geq t_i} wx_j), \quad (14)$$

where t is the the survival time of patient sample, $C(i)$ indicates whether the patient sample i is censored.

2.4. Experiment Settings

We implemented our model using the PyTorch framework. We train and validate our model using an Nvidia Tesla V100 (32GB) GPU. We first pre-train SAVAE using the pan-cancer database. To validate the pre-trained reconstruction results, we randomly divide the training dataset and the test dataset with a ratio of 9:1. For survival analysis on 16 cancer datasets, we trained our model by dividing the dataset into five-fold cross-validation. Both stages are trained using the Adam optimizer. In the pre-training stage, the learning rate of the optimizer is 0.0001, the total epochs are 300, and the batch size is 256. At the same time, since the loss fluctuation problem often occurs in the training process of GAN, it is necessary to set the learning rate decay strategy. Therefore, we stipulate that the learning rate remains constant for the first 150 epochs, and decays linearly to 0 for the last 150 epochs. In the SAVAE-Cox training phase, the learning rate of the optimizer is 0.001, the total training epochs are 20, and the batch size is 512. Note that there are some hyperparameters such as learning rate, λ_p , λ_1 and λ_2 . The specific selection methods and optimal parameter settings are in Appendix A.2. To ensure the fairness of training, we use the same dataset settings to train and evaluate Cox-nnet, Cox-lasso, Cox-ridge and VAE-Cox.

2.5. Evaluation Metric

In this work, the evaluation method we mainly use is the concordance index [34], which is widely used in survival analysis models. It ranges from 0 to 1. When the concordance index ≤ 0.5 means that the model has completed an ineffective survival analysis prediction. When the concordance index > 0.5 and higher this indicates that the prediction effect of the model is better.

3. Results

3.1. Performance of Dimensionality Reduction

In Section 3.1, we evaluated the generator performance in the pre-training stage. All 990 samples in the pan-cancer test dataset were used in this section. We fed the test set of the pan-cancer dataset to the generator, and the generator synthesized the reconstruction results. We then use UMap to perform visualization of the real genes and the reconstructed genes. Figure 4 is the visualization using UMap. We found that the distribution of reconstructed genes closely coincided with the distribution of real genes, which shows that:

1. Generator can reconstruct x_{rec} that are consistent with x_{in} ;
2. The reconstruction of the generator is based on the latent encoding z , indicating that the encoder of the generator can effectively generate z , which retains rich features that can represent x_{in} .

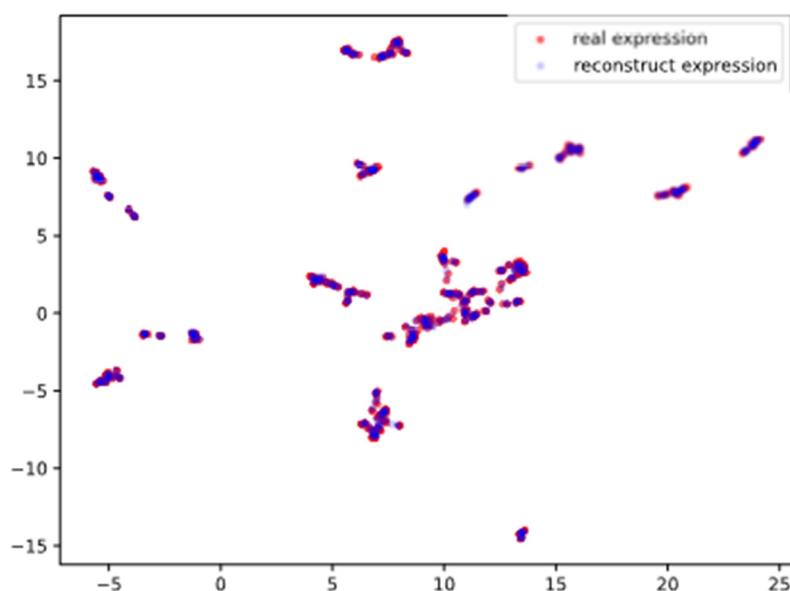


Figure 4. UMap plot of real genes and reconstructed genes. The reconstructed genes and the real genes are highly coincident under low dimension.

To further verify the superiority of our proposed dimensionality reduction method, we compare the performance with other dimensionality reduction methods. The comparison results were shown in Table 2 and the optimal hyperparameter settings were shown in Table A1. First, we choose Autoencoder (AE) and denoising Autoencoder (denoise-AE) to compare with our model. At the same time, we also compare some classical dimensionality reduction methods based on feature selection such as Chi2, Pearson, mutual information and maximal information coefficient (MIC) and principal component analysis (PCA). Note that unlike data-driven dimensionality reduction methods, the feature selection method does not use pan-cancer data for pre-training but directly applies statistical methods on 16 cancer types to select high-correlation features. We used five-fold cross-validation on 16 types of data to train these models and calculated the mean concordance index.

By comparing these six methods on 16 cancer types, our dimensionality reduction method performed best in nine of them. Interestingly, using the Chi2 feature selection-based dimensionality reduction method outperforms data-driven methods in LUSC, CESC, and SKCM datasets. As shown in Table 2, the samples in the CESC and SKCM datasets are small. Meanwhile, we found significant overfitting problems that emerged while training used the data-driven dimensionality reduction methods on the LUSC dataset. These characteristics reflect the shortcomings of data-driven dimensionality reduction methods that are highly dependent on data.

Table 2. Mean Concordance Index on 16 cancer types using different dimensionality reduction methods.

Cancer Type	Dimensionality Reduction Method						
	AE	Denoise-AE	Chi2	Pearson	MIC	PCA	SAVAE
BLCA	0.642	0.643	0.582	0.552	0.624	0.545	0.654
BRCA	0.704	0.709	0.651	0.500	0.492	0.488	0.724
HNSC	0.649	0.642	0.522	0.590	0.531	0.489	0.651
KIRC	0.725	0.731	0.620	0.698	0.673	0.547	0.723
LGG	0.844	0.843	0.712	0.820	0.786	0.673	0.857
LIHC	0.704	0.696	0.467	0.627	0.401	0.423	0.713
LUAD	0.617	0.635	0.627	0.595	0.571	0.570	0.647
LUSC	0.552	0.559	0.605	0.534	0.529	0.496	0.575
OV	0.608	0.621	0.550	0.512	0.517	0.471	0.620
STAD	0.602	0.616	0.556	0.572	0.531	0.476	0.610
CESC	0.690	0.722	0.724	0.565	0.598	0.398	0.663
COAD	0.631	0.638	0.489	0.533	0.521	0.496	0.728
SARC	0.698	0.700	0.558	0.647	0.637	0.511	0.720
UCEC	0.677	0.701	0.640	0.591	0.611	0.472	0.698
PRAD	0.724	0.649	0.687	0.751	0.586	0.606	0.774
SKCM	0.684	0.655	0.863	0.652	0.531	0.512	0.734

3.2. Performance of Survival Analysis

We use the concordance index to evaluate the performance of SAVAE-cox models. Figure 5 shows the comparisons of performance on 16 cancer datasets. In Figure 5, we selected four models to compare with SAVAE-Cox, which include the classic model like Cox-lasso and Cox-ridge methods, as well as some state-of-the-art methods such as Cox-net and VAE-cox. Each model chooses the optimal parameter settings to ensure the fairness of the experiment. Table A1 shows the optimal parameters of these models. Then we divided the 16 cancer types into train and validation datasets by five-fold cross-validation. For each cancer type, we compute the mean concordance index for the validation set. Finally, we draw boxplots according to the concordance index of the five models. From the experimental results we can see that the predicted concordance index using our model on 12 cancer types is significantly higher than the other four models.

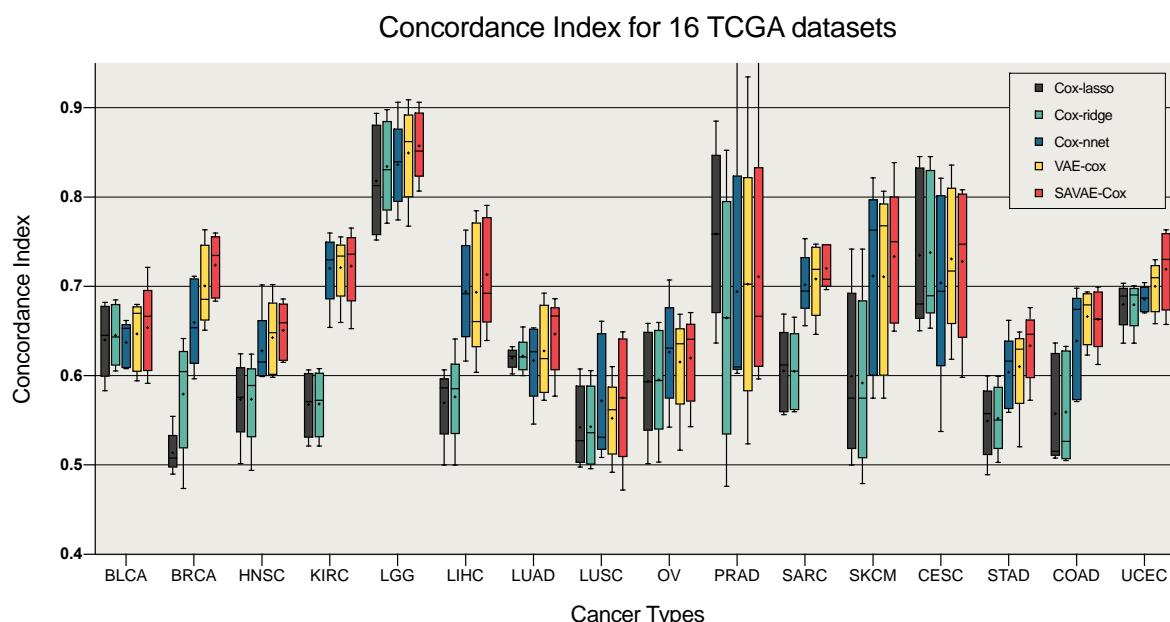


Figure 5. Performance comparison of survival analysis on 16 cancer types. The “+” of each box plot denotes the mean concordance index. The mean concordance index of hazard ratios predicted using our model was best on 12 cancer types.

We performed further survival analysis on 12 cancer types. Based on the hazard ratios of patient samples predicted by SAVAE-Cox, we can calculate the mean reference hazard ratios for 12 cancer types. For each cancer type, we defined patients with predicted hazard ratios above the average to be in the high-risk group, and patients with predicted hazard ratios below the average to be in the low-risk group. In this way, we divided patient samples into high-risk and low-risk groups for each cancer type. Therefore, we draw Kaplan–Meier (KM) survival curves for different cancer types according to these two groups. At the same time, we adopted the same strategy to plot the KM survival curve of the Cox-nnet prediction results. Figure 6 shows the comparison of KM survival curves of SAVAE-Cox and Cox-nnet on 12 cancer types. Based on the results in Figure 6, we found that the hazard ratio predicted by SAVAE-Cox significantly affected patient survival. At the same time, by analyzing the *p*-value, we can find that the hazard ratios predicted by SAVAE-Cox have a more significant impact on patient survival than that predicted by Cox-nnet.

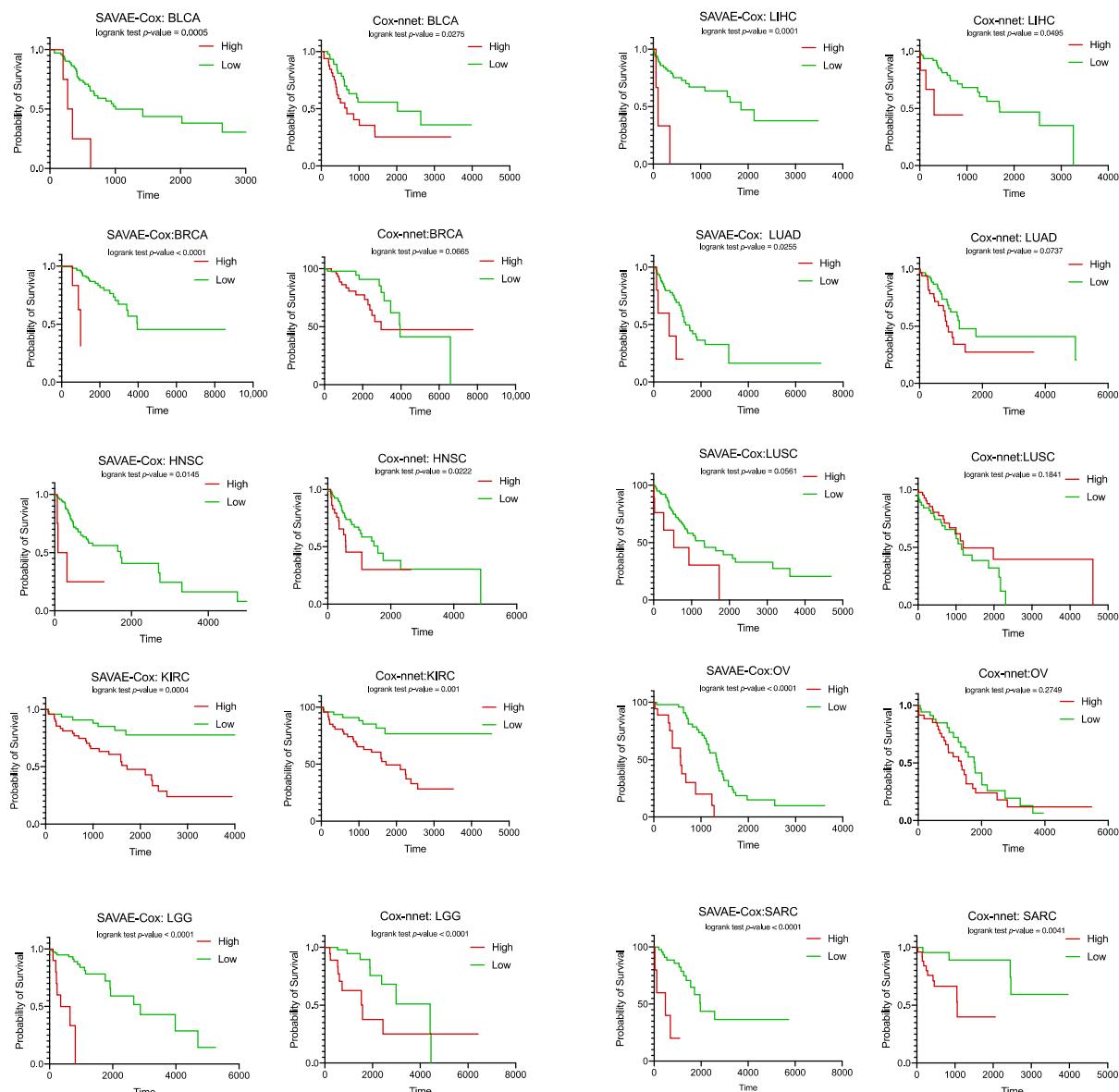


Figure 6. Kaplan–Meier survival curves using SAVAE-Cox and Cox-nnet on 12 cancer types. The smaller the *p*-value, the more significant the risk difference between the two groups predicted by the model.

3.3. Feature Analysis of SAVAE-Cox

The incidence of BRCA is very high, so there were a lot of related studies on this malignancy. At the same time, benefiting from the professionalism of the TCGA project, the researchers collected abundant samples. Choosing BRCA for survival analysis can obtain the most stable survival analysis results and conclusions. Therefore, we take BRCA as an example to conduct experiments for further correlation analysis between model features and genes. All of the 1031 patient samples in the BRCA dataset were tested in this study. First, we analyzed the hidden layer nodes that contribute the most to the prognosis according to the mean and variance. After analysis, we selected the top 20 key prognostic hidden layer nodes. Finally, we calculated a Pearson correlations matrix for each node and gene expression across the patient samples. According to the calculated correlation matrix, we can analyze the correlation between hidden layer nodes and different genes.

We selected 34 cancer-related genes from DISEASE (<https://diseases.jensenlab.org/>, Last visited on 21 March 2022) and plotted them in Figure 7. These genes have high correlation scores in DISEASE, and some genes are oncogenes in ovarian cancer and significantly affect patient survival. The meta-analysis strongly supports the prognostic role of BCL2 as assessed by immunohistochemistry in breast cancer [35]. Germline variation NEK10 is associated with breast cancer incidence [36]. The progesterone receptor (PgR) is one of the most important prognostic and predictive immunohistochemical markers in breast cancer [37]. The CCDC170 gene affects both breast cancer risk and progression [38]. ESR1 amplification may be a common mechanism in proliferative breast disease and a very early genetic alteration in a large subset of breast cancers [39]. The SLC4A7 variant rs4973768 is associated with breast cancer risk [40]. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles [41]. RAD51 is a potential biomarker and attractive drug target for metastatic triple negative breast cancer [42]. CTLA-4 was expressed and functional on human breast cancer cells through influencing maturation and function of DCs in vitro [43]. MYC deregulation contributes to breast cancer development and progression and is associated with poor outcomes [44]. The CDH1 mutation frequency affecting exclusively lobular breast cancer [45]. Inherited mutations of BRCA1 are responsible for about 40–45% of hereditary breast cancers [46]. EGFR is an oncogene in breast cancer [47]. ERBB2 is an oncogene in breast cancer [48]. Six different germline mutations in breast cancer families are likely to be due to BRCA2 [49]. Rare mutations in XRCC2 increase the risk of breast cancer [50]. Amino acid substitution variants of XRCC1 and XRCC3 genes may contribute to breast cancer susceptibility [51]. Overexpression of an ectopic H19 gene enhances the tumorigenic properties of breast cancer cells [52]. CYP19A1 genetic variants in relation to breast cancer survival in a large cohort of patients [53]. BARD1 mutations may be regarded as cancer risk alleles [54]. Master regulators of FGFR2 signalling and breast cancer risk [55]. Interestingly, from Figure 7 we can see that 20 key nodes are significantly associated with these oncogenes, which suggests that exploring patient survival using our proposed model could serve as a new avenue for the discovery of oncogenes.

To further explore the contribution of hidden layer nodes to patient survival, we plot Kaplan–Meier survival curves with 20 key nodes. In this experiment, we also selected all patient samples in BRCA for evaluation. By analyzing the variation of each node in the overall patient sample, we divided them into two groups according to the mean value of nodes. Finally, we calculated the log-rank *p*-value for each node and drew the Kaplan–Meier survival curves. Figure 8 shows the survival curves of the first four key nodes. From the survival curve, we found that patient samples can be significantly divided into risk groups and safety groups according to key nodes, and the larger the value of the node, the lower the survival probability of the patient, which proves that our hidden layer nodes can be used as a key prognostic factor.

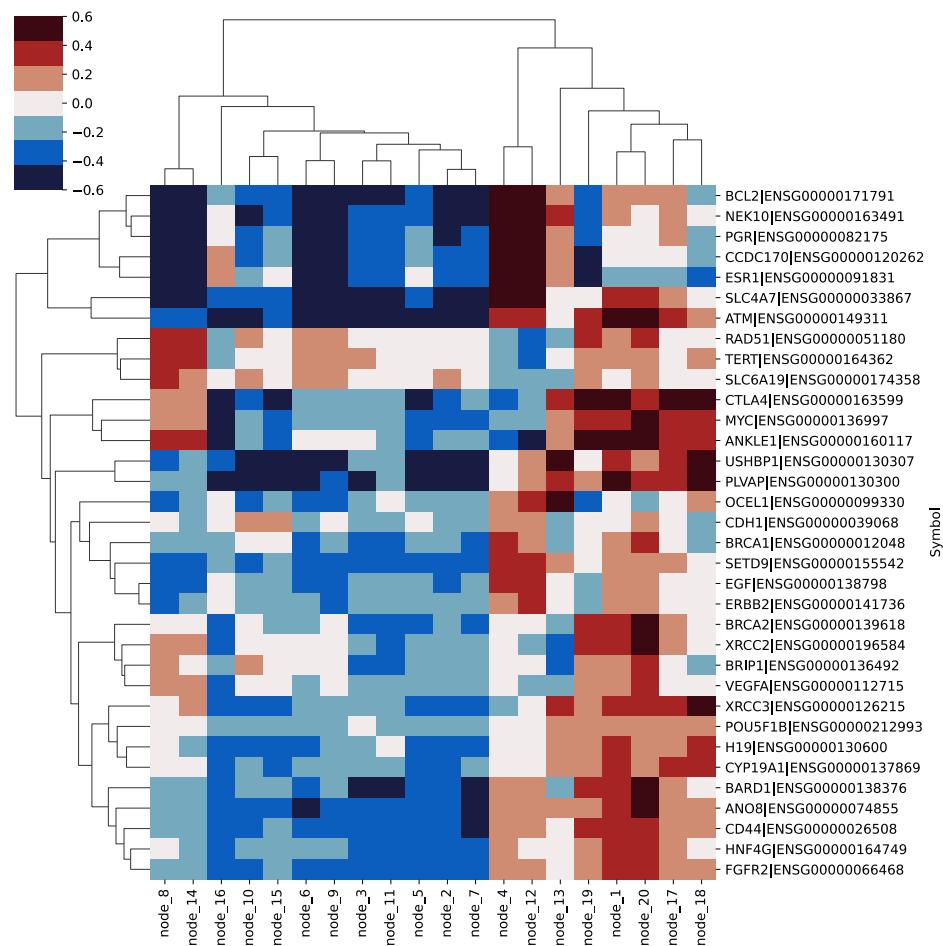


Figure 7. Pearson correlation heatmap of 34 cancer-related genes and 20 key nodes in the BRCA study. All of the 34 genes are highly associated with breast cancer.

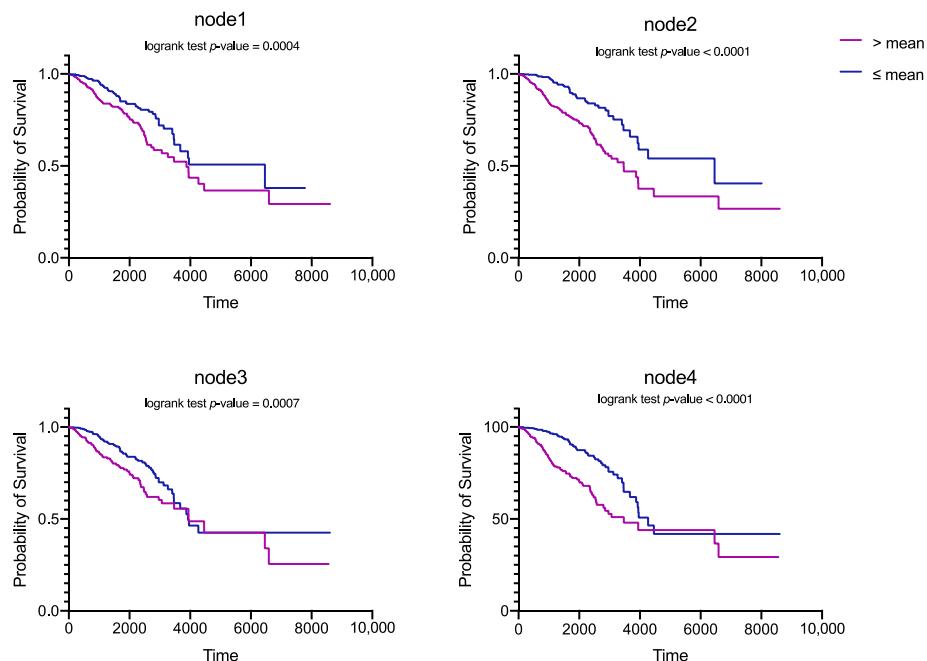


Figure 8. Kaplan–Meier survival curves for four key nodes in a hidden layer. The smaller the *p*-value, the more significant the effect of the node on the survival of the patient.

3.4. Biological Function Analysis of Hidden Nodes

To further explore the biological relevance of hidden layer nodes, we performed a gene set enrichment analysis (GSEA) using the KEGG pathway. We ranked genes according to the Pearson correlation and selected the leader gene for each key node. Based on the leader genes, we created the pathway association network (Figure 9). In Figure 9, each point represents a pathway, and the size of the point represents the number of genes enriched in this pathway, which indicates that the hidden layers of our module can effectively learn the biological functions associated with diseases.

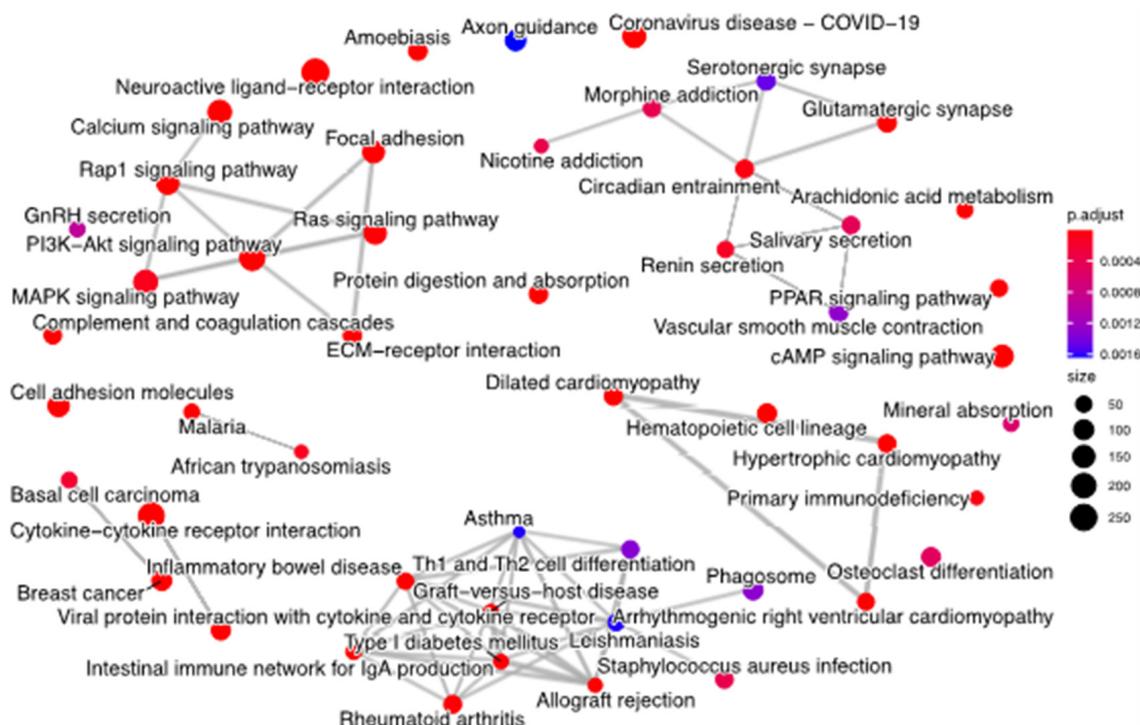


Figure 9. Pathway association network of leader genes. Each point represents a pathway signal, and the gray solid represents the association between pathways. The size of points represents the number of genes enriched in this pathway.

3.5. Ablation Study for SAVAE-Cox

To verify the contribution of our proposed model to the effect of survival prognosis, we performed ablation experiments on SAVAE-Cox. Briefly, we divided the model into four groups: Cox-nnet, SAVAE-Cox without pretrain, SAVAE-Cox without attention, and SAVAE-Cox. We trained four models on 16 cancer types and divided ranks 1, 2, 3, and 4 in descending order according to the performance of the four models in each cancer and plotted Figure 10. Note that optimal parameter settings were selected for all four models in the ablation study. In Table A1 we listed the parameter settings of these four models.

By comparing SAVAE-Cox and SAVAE-Cox without attention, we find that using the self-attention module can significantly improve the model prognosis accuracy. This result shows that there is a potential feature semantic correlation in high-dimensional gene expression, and this correlation cannot be effectively learned using traditional fully connected layers. This latent relationship can be found to a large extent using attention-based methods. By comparing SAVAE-Cox and SAVAE-Cox without pretrain, we find it interesting that the module that used transfer learning on the five datasets HNSC, KIRC, LUAD, LUSC, and OV fails to improve the model's prognostic results. In general, using the transfer learning strategy improves the C-index of the model.

	SAVAE-Cox	1	1	2	2	1	1	2	2	3	1	1	1	1	1	2	2
without pretrain		2	2	1	1	2	3	1	1	2	2	2	2	4	3	3	3
without attention		3	3	3	3	3	2	3	4	4	3	3	3	3	4	1	1
Cox-nnet		4	4	4	4	4	2	4	3	1	4	4	4	2	2	4	4
	BLCA	BRCA	HNSC	KIRC	LGG	LIHC	LUAD	LUSC	OV	STAD	SARC	UCEC	PRAD	SKCM	CESC	COAD	

Figure 10. Ablation study on 16 cancer types. The performance results of four models were divided to ranks 1, 2, 3, 4 in descending order.

4. Discussion

In this work, we introduced a novel survival analysis model for different cancer types, which is the first attempt to improve the overall survival analysis accuracy with the help of the self-attention mechanism. At the same time, we designed a data-driven dimensionality reduction method regarding the idea of transfer learning and GAN [27] to further improve the prediction effect. Our results in Figure 5 suggest that the best performance can be achieved using SAVAE-Cox in the prediction of survival analysis for 12 cancer types including BLCA, BRCA, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, SARC, SKCM, STAD, and UCEC. There are multiple factors or complex genetic associations in most of these 12 cancer types that potentially influence patient survival, which shows that the SAVAE-Cox can effectively discover such latent semantic correlations. However, for some small sample datasets like CESC, our model still cannot achieve optimal results. Using transfer learning can indeed alleviate the overfitting problem to a certain extent. However, due to the deepening of the network, it is not comparable to some classical methods in the cancer dataset with sparse samples. Besides, for some cancer datasets like PRAD with an extremely uneven number of positive and negative samples, using SAVAE-Cox for survival analysis is also not the best choice. We analyze that the cause for this problem is also that the complexity of the network increases, which leads to a larger differentiation of the fitting to the two samples, resulting in a loss of accuracy.

We proposed an adversarial VAE-based [32] pre-training method for dimensionality reduction of high-dimensional genes. Unlike classical feature selection methods, our proposed dimensionality reduction method is data-driven. SAVAE-Cox can adaptively extract useful features from high-dimensional genes based on this novel method. This dimensionality reduction method is applicable to any type of data distribution. From Table 2, we found that using the data-driven method showed the best mean concordance index on 13 cancer types, and the effect of using the data-driven dimensionality reduction method was more significant. However, in some aspects, traditional feature selection dimensionality reduction methods may be more effective. For example, in dimensionality reduction tasks for small sample datasets, data-driven methods cannot predict the whole low-dimensional latent space based on a small number of sample distributions. According to the results in Table 2, we can find that using the Chi2 method for dimensionality reduction achieved the best results on two datasets with small sample sized datasets such as LUSC, CESC and SKCM. Meanwhile, we combine GAN [27] and VAE [32] to design a more powerful data-driven dimensionality reduction strategy. This method introduces distribution constraints in GAN, which improves the stability of dimensionality reduction under the condition of ensuring strong generation and fitting capabilities. By comparing the state-of-the-art data-driven dimensionality reduction methods in Table 2, the performance using our method is better.

Through the analysis of BRCA, our model can discover cancer-related genes and reveal biological functions. From Figures 7 and 8 we find that each key node of the hidden layer

of the SAVAE-Cox model is a prognostic features affecting patient survival. Furthermore, our model helps to explain and discover new cancer-related genes. Meanwhile, according to Figure 9, numerous node-related genes are enriched in cancer pathways such as the breast cancer pathway, the PI3K-Atk signaling pathway, the Rap1 signaling pathway, and the MPKA signaling pathway, in which we confirmed that the hidden layer of the model is highly related to biological functions and reveal rich biological function signals.

However, we found that the overfitting of the SAVAE-Cox is still a very serious problem. At the same time, the performance of the model is not good enough on datasets with an imbalanced number of positive and negative samples. In future work, we will study some data augmentation methods to solve these problems and explore some novel multi-head attention-based survival analysis frameworks.

5. Conclusions

We introduced a brand new survival analysis model and performed survival prognosis on 16 cancer types. The prognosis was significantly improved when self-attention and transfer learning was integrated to the SAVAE-Cox. With the further analysis of the hidden layer features of SAVAE-Cox, we confirmed that the hidden layer features of this model play a significant role in cancer prognosis and the revealing of biological function. In conclusion, the SAVAE-Cox combines a self-attention mechanism with transfer learning, and feature selection provides a new prospect for future deep cancer prognosis.

Author Contributions: Conceptualization, X.M., X.W. and S.W.; methodology, X.M.; software, X.M.; validation, X.W., X.Z. and C.Z.; formal analysis, Z.Z.; investigation, K.Z.; resources, X.W.; data curation, X.M.; writing—original draft preparation, X.M.; writing—review and editing, X.W.; visualization, X.Z.; supervision, X.M.; project administration, X.W.; funding acquisition, X.W. and S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [Grant Nos. 61873280, 61873281, 61972416] and Natural Science Foundation of Shandong Province [No. ZR2019MF012].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study are accessible from UCSC Xena (<https://xenabrowser.net/datapages/>, Last visited on 21 March 2022). At the same time, we disclosed access links of the 16 cancer types and pan-cancer datasets processed using the method selected in this paper: <https://drive.google.com/drive/folders/1KuDVRkPJZWYfQ2Z4YRxDo6e8lo5s68za> (Last visited on 21 March 2022).

Acknowledgments: The authors are grateful to Shanchen Pang and Tao Song for advice and excellent technical assistance.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Descriptions and Download Details of the Datasets

In this work, we used TCGA mRNA expression and clinical datasets of 16 cancer types. The 16 cancer types of data are: bladder carcinoma (BLCA), breast carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal cell carcinoma (KIRC), brain lower-grade glioma (LGG), liver hepatocellular carcinoma carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), stomach adenocarcinoma (STAD), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), sarcoma (SARC), uterine corpus endometrial carcinoma (UCEC), prostate adenocarcinoma (PRAD), and skin cutaneous melanoma (SKCM). At the same time, the TCGA Pan-Cancer (PANCAN) dataset covering 33 cancer types was used in this study. The Pan-Cancer dataset is from the Pan-

Cancer Atlas project. Thirty-three cancer types from the TCGA database were analyzed in this program. The GDC mRNA quantification analysis pipeline measures gene-level expression with STAR as raw read counts. Subsequently, the counts are augmented with several transformations including fragments per kilobase of transcript per million mapped reads (FPKM), upper quartile normalized FPKM (FPKM-UQ), and transcripts per million (TPM). The main purpose of using this normalization is to remove technical bias from different sequencing data. More information on the GDC pipeline used to generate this data is at: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/ (Last visited on 21 March 2022). In this work, we mainly use the data normalized by FPKM-uq. UCSC Xena did some preliminary preprocessing on the TCGA dataset, all of which was open access, which can save a lot of preprocessing work. All the datasets used by our work can be download at <https://xenabrowser.net/datapages/> (Last visited on 21 March 2022). The mRNA dataset obtained by Xena includes 56,716 genes was normalized by $\log(FPKM + 1)$, many of which are empty. So further preprocessing work is required.

Appendix A.2. Hyperparameter Selection

In this work, we mainly chose the optimal hyperparameters based on empirical knowledge and Bayesian optimization methods. In the training process of SAVAE, the optimal hyperparameters were: $lr = 0.0001$, $\lambda_p = 10$, $\lambda_1 = 10$, $\lambda_2 = 10$, epoch = 300, batch_size = 256. In training SAVAE-Cox, the optimal hyperparameter in survival analysis was shown in Table A1.

Table A1. Optimal Hyperparameter Settings for Survival Analysis.

	Survival Analysis Models	Learning Rate	Epoch	Batch Size
dimensionality reduction	Cox-Chi2	0.0005	15	1024
	Cox-Pearson	0.001	15	1024
	Cox-MIC	0.0005	15	1024
	Cox-PCA	0.0005	15	1024
	Cox-AE	0.001	15	1024
	Cox-dnoiseAE	0.0005	15	1024
Comparative Experiment	Cox-lasso	0.0005	15	1024
	Cox-ridge	0.001	15	1024
	Cox-nnet	0.001	15	1024
	VAECox	0.001	15	1024
Ablation Study	Without pretrain	0.001	20	512
	Without attention	0.001	20	512
Ours	SAVAE-Cox	0.001	20	512

References

- Nicholson, R.I.; Gee, J.M.W.; Harper, M.E. EGFR and cancer prognosis. *Eur. J. Cancer* **2001**, *37*, 9–15.
- Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–202.
- Broder, S.; Subramanian, G.; Venter, J.C. The human genome. *Pharm. Search Individ. Ther.* **2002**, *9*, 34.
- Lussier, Y.A.; Li, H. Breakthroughs in genomics data integration for predicting clinical outcome. *J. Biomed. Inform.* **2012**, *45*, 1199.
- Valdes-Mora, F.; Handler, K.; Law, A.M.; Salomon, R.; Oakes, S.R.; Ormandy, C.J.; Gallego-Ortega, D. Single-cell transcriptomics in cancer immunobiology: The future of precision oncology. *Front. Immunol.* **2018**, *9*, 2582.
- Nagy, Á.; Munkácsy, G.; Győrffy, B. Pancancer survival analysis of cancer hallmark genes. *Sci. Rep.* **2021**, *11*, 6047.
- Ding, Z. The application of support vector machine in survival analysis. In Proceedings of the 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Zhengzhou, China, 8–10 August 2011; pp. 6816–6819.
- Evers, L.; Messow, C.-M. Sparse kernel methods for high-dimensional survival data. *Bioinformatics* **2008**, *24*, 1632–1638.
- Bin, R.D. Boosting in Cox regression: A comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Comput. Stat.* **2016**, *31*, 513–531. [CrossRef]
- Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860.

11. Meng, X.; Zhang, X.; Wang, G.; Zhang, Y.; Shi, X.; Dai, H.; Wang, Z.; Wang, X. Exploiting full Resolution Feature Context for Liver Tumor and Vessel Segmentation via Fusion Encoder: Application to Liver Tumor and Vessel 3D reconstruction. *arXiv* **2021**, arXiv:2111.13299.
12. Song, T.; Zhang, X.; Ding, M.; Rodriguez-Paton, A.; Wang, S.; Wang, G. DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions. *Methods* **2022**, in press. [CrossRef]
13. Faraggi, D.; Simon, R. A neural network model for survival data. *Stat. Med.* **1995**, *14*, 73–82.
14. Ching, T.; Zhu, X.; Garmire, L.X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **2018**, *14*, e1006076.
15. Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; Kluger, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **2018**, *18*, 24.
16. Huang, Z.; Zhan, X.; Xiang, S.; Johnson, T.S.; Helm, B.; Yu, C.Y.; Zhang, J.; Salama, P.; Rizkalla, M.; Han, Z. SALMON: Survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* **2019**, *10*, 166.
17. Kim, S.; Kim, K.; Choe, J.; Lee, I.; Kang, J. Improved survival analysis by learning shared genomic information from pan-cancer data. *Bioinformatics* **2020**, *36*, i389–i398.
18. Ramirez, R.; Chiu, Y.-C.; Zhang, S.; Ramirez, J.; Chen, Y.; Huang, Y.; Jin, Y.-F. Prediction and interpretation of cancer survival using graph convolution neural networks. *Methods* **2021**, *192*, 120–130.
19. Huang, Z.; Johnson, T.S.; Han, Z.; Helm, B.; Cao, S.; Zhang, C.; Salama, P.; Rizkalla, M.; Yu, C.Y.; Cheng, J. Deep learning-based cancer survival prognosis from RNA-seq data: Approaches and evaluations. *BMC Med. Genom.* **2020**, *13*, 41.
20. Rehman, M.U.; Tayara, H.; Chong, K.T. DCNN-4mC: Densely connected neural network based N4-methylcytosine site prediction in multiple species. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6009–6019.
21. Chen, J.; Wang, W.H.; Shi, X. Differential privacy protection against membership inference attack on machine learning for genomic data. In Proceedings of the BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, Kohala Coast, HI, USA, 3–7 January 2021; pp. 26–37.
22. Torada, L.; Lorenzon, L.; Beddis, A.; Isildak, U.; Pattini, L.; Mathieson, S.; Fumagalli, M. ImaGene: A convolutional neural network to quantify natural selection from genomic data. *BMC Bioinform.* **2019**, *20*, 337.
23. Hao, J.; Kosaraju, S.C.; Tsaku, N.Z.; Song, D.H.; Kang, M. PAGE-Net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, HI, USA, 3–7 January 2020; pp. 355–366.
24. Jeong, S.; Kim, J.-Y.; Kim, N. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Sci. Rep.* **2020**, *10*, 19653. [PubMed]
25. Rehman, M.U.; Hong, K.J.; Tayara, H.; Chong, K. m6A-NeuralTool: Convolution neural tool for RNA N6-Methyladenosine site identification in different species. *IEEE Access* **2021**, *9*, 17779–17786.
26. Ramirez, R.; Chiu, Y.-C.; Herrera, A.; Mostavi, M.; Ramirez, J.; Chen, Y.; Huang, Y.; Jin, Y.-F. Classification of cancer types using graph convolutional neural networks. *Front. Phys.* **2020**, *8*, 203. [PubMed]
27. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Processing Syst.* **2014**, *27*. Available online: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf> (accessed on 15 March 2022).
28. Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **2021**, *3*, 324–333.
29. Lin, E.; Mukherjee, S.; Kannan, S. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinform.* **2020**, *21*, 64.
30. Jiang, X.; Zhao, J.; Qian, W.; Song, W.; Lin, G.N. A generative adversarial network model for disease gene prediction with RNA-seq data. *IEEE Access* **2020**, *8*, 37352–37360.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 5998–6008.
32. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114, preprint.
33. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. *Adv. Neural Inf. Processing Syst.* **2017**, *30*. Available online: <https://www.semanticscholar.org/paper/Improved-Training-of-Wasserstein-GANs-Gulrajani-Ahmed/edf73ab12595c6709f646f542a0d2b33eb20a3f4> (accessed on 15 March 2022).
34. Raykar, V.C.; Steck, H.; Krishnapuram, B.; Dehing-Oberije, C.; Lambin, P. On ranking in survival analysis: Bounds on the concordance index. In Proceedings of the Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 1209–1216.
35. Callagy, G.M.; Webber, M.J.; Pharoah, P.D.; Caldas, C. Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. *BMC Cancer* **2008**, *8*, 153.
36. Bryan, M.S.; Argos, M.; Andrulis, I.L.; Hopper, J.L.; Chang-Claude, J.; Malone, K.E.; John, E.M.; Gammon, M.D.; Daly, M.B.; Terry, M.B. Germline variation and breast cancer incidence: A gene-based association study and whole-genome prediction of early-onset breast cancer. *Cancer Epidemiol. Prev. Biomark.* **2018**, *27*, 1057–1064.
37. Kunc, M.; Biernat, W.; Senkus-Konefka, E. Estrogen receptor-negative progesterone receptor-positive breast cancer—“Nobody’s land” or just an artifact? *Cancer Treat. Rev.* **2018**, *67*, 78–87.

38. Jiang, P.; Li, Y.; Poleshko, A.; Medvedeva, V.; Baulina, N.; Zhang, Y.; Zhou, Y.; Slater, C.M.; Pellegrin, T.; Wasserman, J. The protein encoded by the CCDC170 breast cancer gene functions to organize the golgi-microtubule network. *EBioMedicine* **2017**, *22*, 28–43.
39. Holst, F.; Stahl, P.R.; Ruiz, C.; Hellwinkel, O.; Jehan, Z.; Wendland, M.; Lebeau, A.; Terracciano, L.; Al-Kuraya, K.; Jänicke, F. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nat. Genet.* **2007**, *39*, 655–660.
40. Chen, W.; Zhong, R.; Ming, J.; Zou, L.; Zhu, B.; Lu, X.; Ke, J.; Zhang, Y.; Liu, L.; Miao, X. The SLC4A7 variant rs4973768 is associated with breast cancer risk: Evidence from a case–control study and a meta-analysis. *Breast Cancer Res. Treat.* **2012**, *136*, 847–857.
41. Ahmed, M.; Rahman, N. ATM and breast cancer susceptibility. *Oncogene* **2006**, *25*, 5906–5911.
42. Wiegmans, A.P.; Al-Ejeh, F.; Chee, N.; Yap, P.-Y.; Gorski, J.J.; Da Silva, L.; Bolderson, E.; Chenevix-Trench, G.; Anderson, R.; Simpson, P.T. Rad51 supports triple negative breast cancer metastasis. *Oncotarget* **2014**, *5*, 3261.
43. Chen, X.; Shao, Q.; Hao, S.; Zhao, Z.; Wang, Y.; Guo, X.; He, Y.; Gao, W.; Mao, H. CTLA-4 positive breast cancer cells suppress dendritic cells maturation and function. *Oncotarget* **2017**, *8*, 13703.
44. Xu, J.; Chen, Y.; Olopade, O.I. MYC and breast cancer. *Genes Cancer* **2010**, *1*, 629–640.
45. Corso, G.; Intra, M.; Trentin, C.; Veronesi, P.; Galimberti, V. CDH1 germline mutations and hereditary lobular breast cancer. *Fam. Cancer* **2016**, *15*, 215–219.
46. Rosen, E.M.; Fan, S.; Pestell, R.G.; Goldberg, I.D. BRCA1 gene in breast cancer. *J. Cell. Physiol.* **2003**, *196*, 19–41.
47. Chrysogelos, S.A.; Dickson, R.B. EGF receptor expression, regulation, and function in breast cancer. *Breast Cancer Res. Treat.* **1994**, *29*, 29–40.
48. Revillion, F.; Bonneterre, J.; Peyrat, J. ERBB2 oncogene in human breast cancer and its clinical significance. *Eur. J. Cancer* **1998**, *34*, 791–808.
49. Wooster, R.; Bignell, G.; Lancaster, J.; Swift, S.; Seal, S.; Mangion, J.; Collins, N.; Gregory, S.; Gumbs, C.; Micklem, G. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **1995**, *378*, 789–792.
50. Park, D.; Lesueur, F.; Nguyen-Dumont, T.; Pertesi, M.; Odefrey, F.; Hammet, F.; Neuhausen, S.L.; John, E.M.; Andrulis, I.L.; Terry, M.B. Rare mutations in XRCC2 increase the risk of breast cancer. *Am. J. Hum. Genet.* **2012**, *90*, 734–739.
51. Smith, T.R.; Miller, M.S.; Lohman, K.; Lange, E.M.; Case, L.D.; Mohrenweiser, H.W.; Hu, J.J. Polymorphisms of XRCC1 and XRCC3 genes and susceptibility to breast cancer. *Cancer Lett.* **2003**, *190*, 183–190.
52. Lottin, S.; Adriaenssens, E.; Dupressoir, T.; Berteaux, N.; Montpellier, C.; Coll, J.; Dugimont, T.; Curgy, J.J. Overexpression of an ectopic H19 gene enhances the tumorigenic properties of breast cancer cells. *Carcinogenesis* **2002**, *23*, 1885–1895.
53. Long, J.-R.; Kataoka, N.; Shu, X.-O.; Wen, W.; Gao, Y.-T.; Cai, Q.; Zheng, W. Genetic polymorphisms of the CYP19A1 gene and breast cancer survival. *Cancer Epidemiol. Prev. Biomark.* **2006**, *15*, 2115–2122.
54. Ratajska, M.; Antoszewska, E.; Piskorz, A.; Brozek, I.; Borg, Å.; Kusmierenk, H.; Biernat, W.; Limon, J. Cancer predisposing BARD1 mutations in breast–ovarian cancer families. *Breast Cancer Res. Treat.* **2012**, *131*, 89–97.
55. Fletcher, M.N.; Castro, M.A.; Wang, X.; De Santiago, I.; O'Reilly, M.; Chin, S.-F.; Rueda, O.M.; Caldas, C.; Ponder, B.A.; Markowitz, F. Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **2013**, *4*, 2464. [PubMed]