biology

Article

An Ensemble Deep Learning Model with a Gene Attention Mechanism for Estimating the Prognosis of Low-Grade Glioma

Minhyeok Lee

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Korea; mlee@cau.ac.kr

Simple Summary: This paper proposes a deep learning model for prognosis estimation and an attention mechanism for gene expression. In the prognosis estimation of low-grade glioma (LGG), the proposed model, Gene Attention Ensemble NETwork (GAENET), demonstrated superior performance compared to conventional models, where GAENET exhibited an improvement of 7.2% compared to the second-best model. By the proposed gene attention, *HILS1* was discovered as the most significant prognostic gene for LGG. While *HILS1* is classified as a pseudogene, it functions as a biomarker for predicting the prognosis of LGG and has been shown to have the ability to regulate the expression of other prognostic genes.

Abstract: While estimating the prognosis of low-grade glioma (LGG) is a crucial problem, it has not been extensively studied to introduce recent improvements in deep learning to address the problem. The attention mechanism is one of the significant advances; however, it is still unclear how attention mechanisms are used in gene expression data to estimate prognosis because they were designed for convolutional layers and word embeddings. This paper proposes an attention mechanism called gene attention for gene expression data. Additionally, a deep learning model for prognosis estimation of LGG is proposed using gene attention. The proposed Gene Attention Ensemble NETwork (GAENET) outperformed other conventional methods, including survival support vector machine and random survival forest. When evaluated by C-Index, the GAENET exhibited an improvement of 7.2% compared to the second-best model. In addition, taking advantage of the gene attention mechanism, *HILS1* was discovered as the most significant prognostic gene in terms of deep learning training. While *HILS1* is known as a pseudogene, *HILS1* is a biomarker estimating the prognosis of LGG and has demonstrated a possibility of regulating the expression of other prognostic genes.

Keywords: survival estimation; prognosis estimation; deep learning; attention mechanism; gene expression; low-grade glioma; *HILS1*

check for updates

Citation: Lee, M. An Ensemble Deep Learning Model with a Gene Attention Mechanism for Estimating the Prognosis of Low-Grade Glioma. *Biology* **2022**, *11*, 586. https://doi.org/10.3390/biology11040586

Academic Editor: Zhijun Wang

Received: 28 February 2022
Accepted: 11 April 2022
Published: 12 April 2022

1. Introduction

Owing to the rapid development of deep learning algorithms, artificial intelligence models based on deep learning have become mainstream in various domains [1–3]. In biomedicine, such models have been extensively studied for medical imaging [4–6], diagnosis [7–9], and genome sequencing [10–12]. The prognosis of diseases also has been estimated with deep learning models in recent studies [13–15]. Generally, clinical data, such as age and sex, have been used in those models to estimate the prognosis [16,17].

Recently, gene expression data have been regarded as an alternative approach for estimating prognosis since gene expression is highly related to the prognosis of various diseases [18–20]. For these reasons, deep learning models handling gene expression data for estimating prognosis have been proposed in several studies [21–23]. For instance, Wong et al. [21] proposed a deep learning model for discovering prognostic genes related to glioblastoma. They introduced partial likelihood that signifies survival risk as a loss function to train a deep learning model called multilayer perceptron (MLP). Chaudhary et al. [22] employed multi-omics gene expression data with RNA sequencing, miRNA sequencing, and methylation data to train a deep learning model for the survival prediction of hepatocellular carcinoma. An auto-encoder (AE) deep learning model was introduced in this work.

However, these studies have commonly used basic deep learning architectures, such as MLP and AE [21,22]. In recent years, various superb deep learning architectures have been proposed for computer vision and natural language processing, resulting in the rapid advance in such domains [24–26]. While not all these architectures and methods can be applied to the prognosis estimation with gene expression data, it is true that it has not been extensively studied to address such advances thus far.

The attention mechanism is one of the recent advances in deep learning architecture, which has been widely used in computer vision and natural language processing [27–29]. An attention layer regularizes feature maps and word embeddings to make deep learning models more focused on specific regions related to the targets of the models, whereas the other regions are multiplied by approximately zero; thereby, the layer produces valid values only for the specific regions. The attention layer can be successfully adopted for computer vision and natural language processing because, in images and natural language, there are definite regions that determine the estimations. For example, in object detection in an image, an area where a particular object exists becomes the specific region on which the attention layer focuses [30,31].

Similarly, in gene expression data, it is considered that there are specific genes related to prognosis and survival, and the other genes are less associated with prognosis [32–34]. Therefore, it is expected that deep learning models with gene expression can enjoy the advantages of the attention mechanism. Since attention layers can learn which genes are related to the target (i.e., prognosis), it is also possible to determine prognostic genes with a nonlinear perspective of deep learning. However, the conventional attention mechanism cannot be directly applied to a deep learning model with gene expression because existing attention layers commonly use convolutional operations and multiplications between word embeddings [27,28,30], which cannot be done in models for gene expression.

This paper proposes a modified attention mechanism for gene expression, namely gene attention, to address this limitation. A gene attention layer using the attention mechanism is employed in the first layer of the model. In the gene attention layer, prognostic genes are determined by learning from data with a deep learning architecture. Then, high weights are multiplied for estimated prognostic genes, whereas the other genes are multiplied by low weights, approximately zero. In this process, it is possible to enhance the efficiency and accuracy of the proposed estimation model; additionally, the gene attention layer identifies which gene is associated with the prognosis and survival, which corresponds to one of the advantages of the attention mechanism.

Furthermore, recently developed deep learning architectures, which are now conventional in computer vision and natural language processing, are employed in the proposed model. The residual learning scheme [35] is adopted to enable the training of a deep architecture; layer normalization [36] is employed for smooth gradient flows in the back-propagation of the deep architecture; the ensemble learning method [37] is used to prevent overfitting and generalize the estimations of the model.

As the gene expression for the input variables of the proposed model, RNA sequencing (RNA-seq) data are used. Given that the transcriptome generates proteins, and proteins define the phenotypes of patients, it is essential to consider the transcriptome of patients in order to predict prognosis. RNA-seq is a technique for expressing the transcriptome in an organism that has been widely employed in recent years. While a few studies have introduced multi-omics, which utilizes additional types of transcriptome such as miRNA, or multi-modal schemes, which employ medical images [15,22], in practice, it is highly uncertain that such multiple data types can be simultaneously obtained for each patient. These models with multi-omics or multi-modal schemes cannot perform the estimation if any of the data types used in the models is not obtained.

In this paper, the prognosis of low-grade glioma (LGG) is estimated by the proposed model. Generally, without appropriate intervention treatments, LGG is eventually aggravated to high-grade glioma (HGG) and leads to death [38]. Therefore, it is of great importance to estimate the prognosis of LGG in order to determine if urgent treatment is required. For instance, it is possible to minimize side effects and treatment costs by prescribing moderate treatments to patients with a favorable estimated prognosis, while extensive treatments may improve survival rates for patients with a poor estimated prognosis. Additionally, it has been reported that the prognosis of LGG is highly correlated with gene expression compared to other cancers [39]; thus, the prognosis of LGG can be a suitable target to verify the performance of the model.

2. Materials and Methods

2.1. Deep Learning with Gene Attention

A conventional deep learning model consists of multiple layers, each of which has a weight matrix and a nonlinear function. In each layer, the input values are multiplied by a weight matrix, and then a nonlinear operation is performed [40]. The result values become the input values of the next layer. This process continues to the final layer. Appendix A describes the fundamentals of deep learning models in further detail. The conventional deep learning model can be summarized by the following equation:

$$\hat{Y} \; := \; \sigma_N(\mathbf{W}_N \cdot \sigma_{N-1}(\mathbf{W}_{N-1} \cdots \sigma_2(\mathbf{W}_2 \cdot \sigma_1(\mathbf{W}_1 \cdot X)))), \tag{1}$$

where $\hat{Y}$ is an estimation vector by the deep learning model; σ_k and $\mathbf{W}_k$ are the nonlinear function and the weight matrix for the kth layer, respectively; X is an input vector of the model; N is the number of layers in the model; all bold terms in the equations represent a matrix.

The attention mechanism corresponds to multiplying specific values to the layer outputs [27]. Generally, an attention value from zero to one is multiplied by each output feature. The attention value indicates how much the corresponding feature is delivered to the next layer. For instance, if an attention value of one is used, the corresponding feature is fully conveyed to the next layer, whereas information in a feature would not be transferred if the value is zero. This process in the kth layer can be represented by the following equation:

$$X_{k+1} := A^T \odot \sigma_k(\mathbf{W}_k \cdot X_k), \tag{2}$$

where X_k is the intermediate output from the $k-1$th layer; A is an attention vector containing the attention values for each feature; $\odot$ denotes component-wise multiplication. The attention values are also determined by learning from data. The attention layer produces the attention vector A and has a weight vector to be trained; it is simultaneously trained with the other deep learning architecture by the backpropagation algorithm to estimate the target.

Various structures exist according to datatypes and deep learning architectures to implement the attention layer. However, most structures are developed and used for a convolutional layer for images as well as embeddings of sequential data, such as natural language, which cannot be applied to gene expression data [28,29]. In this study, a modified squeeze-and-excitation (SE) block [41] for gene expression data, namely the gene attention layer, is proposed. While the SE block is developed for a convolutional layer with three-dimensional image data, it is modified to be able to be used for one-dimensional gene expression data in this study. The proposed structure of the gene attention layer can be represented as follows:

$$A := Sigmoid(\mathbf{W}_{GA-2} \cdot ReLU(\mathbf{W}_{GA-1} \cdot X_0)), \tag{3}$$

$$X_1 := A^T \odot X_0, \tag{4}$$

where $\boldsymbol{W}_{GA-1} \in \mathbb{R}^{q \times p}$, $\boldsymbol{W}_{GA-2} \in \mathbb{R}^{p \times q}$ are weight matrices of the gene attention layer in which p is the number of genes, $\mathbb{R}$ is the real coordinate space, and $q \ll p$; *ReLU* represents the rectified linear unit; X_0 is the gene expression; X_1 is the output of the gene attention layer, which is the input of the following deep learning structures. The dimension of the hidden layer in gene attention, i.e., q, is set to a significantly low value (<100), while the number of genes, i.e., p, is approximately 20,000 in RNA-seq data.

Because the attention vector A can have a value between 0 to 1 per gene, Equation (4) signifies that the gene expression data are not fully conveyed to the next layer. Specifically, a lower attention value for a gene regulates the gene; thereby, information in the gene is not further used in deep learning. The weight matrices in the gene attention layer $\boldsymbol{W}_{GA-1}$ and $\boldsymbol{W}_{GA-2}$ are trained during the training process in deep learning with the backpropagation algorithm. This process implies that the attention values are determined through training with data, not manually. Thus, the attention values are a result of the training and are not controllable by users.

Consequently, the attention value for a gene indicates how much the gene is significant in the deep learning model as well as the target of the model. Therefore, when the target of the model is the prognosis of a disease, the attention values can represent the prognostic genes for the disease. This factor of the gene attention layer is the main advantage of the attention mechanism in which significant features with respect to deep learning can be explained, while deep learning has been generally considered a black box.

2.2. Residual Deep Learning

Residual learning is a method to create a skip connection that links backward and forward of one or more layers [35]. Because gradients from the final layer of a deep learning model can also flow the skip connection, it is more appropriate to train initial layers that are far from the final layer. Thus, in recent years, the residual learning framework has become essential to train deep learning models since the method can alleviate the vanishing gradient problem, which has been a significant hindrance in deep learning. The structure of residual learning can be represented as follows:

$$X_{k+m+1} := X_k + \sigma_{k+m}(\boldsymbol{W}_{k+m} \cdots \sigma_k(\boldsymbol{W}_k \cdot X_k)), \quad (5)$$

where m is the number of layers that the skip connection bypasses.

2.3. Layer Normalization

Layer normalization (*LN*) is a method to maintain the learning space of a layer [36]. During the training of a deep learning model, without a normalization method, the distribution of inputs of a layer fluctuates, where the means and variances of the distribution change in each iteration. Since the output of the layer becomes the input of the next layer, the fluctuation yields another instability in the next layer. Such instability is accumulated throughout a deep learning model and causes a problem to be handled.

Normalization methods are required between layers in a deep learning model to address this problem. These normalization methods maintain the means and variances of inputs during the training process and stabilize the deep learning training. Among the normalization methods, *LN* is employed in the proposed model in this study because gene expression data generally lack the number of samples (i.e., the number of patients) compared to other datatypes for deep learning training. Thus, the number of samples in a mini-batch should be less than that of conventional deep learning. When the number of samples in a mini-batch is small, it has been demonstrated that *LN* outperforms the batch normalization (BN) method [42]. The *LN* method used in this study can be computed as follows:

$$LN(X_k) := \frac{X_k - \mu}{\sigma}, \quad (6)$$

$$\mu = \frac{1}{q}\sum\nolimits_{x \in X_k} x, \ \ \sigma = \sqrt{\frac{1}{q}\sum\nolimits_{x \in X_k} (x - \mu)^2}, \tag{7}$$

where q is the number of nodes in the layer. This method is a modified *LN* where adaptive parameters, i.e., gains and biases, are used in the conventional *LN*.

2.4. Ensemble Learning

Overfitting is one of the significant issues in the training of deep learning. Complex machine learning models, including deep learning, have more training parameters than the number of samples, which causes the overfitting problem. Machine learning models and deep learning models with overfitting produce accurate answers to training sets, while the performance in test sets is relatively inferior. The ensemble learning method has been used to alleviate this overfitting problem [43,44]. In ensemble learning, many models are trained and produce estimations. Then, the final estimation is made by a weighted sum or the average of the estimations by the models.

In this study, an ensemble of multiple deep learning models is employed. In deep learning, models can yield different results even if the same structure is used because the random initialization of weight parameters and mini-batch sampling cause such a difference in training results. Integrated with the overfitting problem, this inconsistency of training results is also one of the significant issues in deep learning. The ensemble learning can handle these problems since the noises from the overfitting and randomness are offset, and estimations are overlapped and added. The ensemble deep learning models used in this study yield the final result as follows:

$$\hat{Y} := \frac{1}{n}\sum\nolimits_{i \in \{1,\ldots,n\}} DL_i(X), \tag{8}$$

where n is the number of sub-models; DL_i is ith sub-model. Each DL_i has the identical structure, but the random initialization of weight parameters causes different training results and estimations.

2.5. The Gene Attention Ensemble Network

The proposed deep learning model, namely the Gene Attention Ensemble NETwork (GAENET), uses the proposed gene attention layer, residual learning, *LN*, and the ensemble structure. The proposed structures are illustrated in Figure 1. The residual learning is employed in the hidden layer of the GAENET, in which three fully connected layers are used in each residual block. *LN* is used between an activation function and a fully connected layer, which is a conventional method to use *LN*. As mentioned, the proposed gene attention layer becomes the first layer of the GAENET, thereby discovering prognostic genes. Dropout layers are introduced in the middle of the residual block as well as prior to the last fully connected layer to alleviate overfitting. Since the GAENET consists of an ensemble of sub-models, each of which is a deep learning model (Figure 1A), the estimation of the GAENET is the average of n number of sub-models, while they share the weights parameters in the gene attention layer.

During the training of the GAENET, an appropriate loss for the backpropagation algorithm to optimize the weights of the model is required. To estimate the prognosis, log values of survival days of samples become the target of the model. In order to train relative survival according to gene expression, the target value of a sample is determined considering relative survival days in a mini-batch of a single backpropagation iteration instead of targeting the absolute value of survival days. Then, the GAENET aims at minimizing the difference between estimation and the relative survival days. The loss to minimize the difference is computed as follows:

$$\mathcal{L} := -\frac{1}{s_m}\left(\log Y - \frac{1}{s_m}\sum\nolimits_{i\in\{1,\ldots,s_m\}}\log Y_i\right)^T\left(\hat{Y} - \frac{1}{s_m}\sum\nolimits_{i\in\{1,\ldots,s_m\}}\hat{Y}_i\right), \tag{9}$$

where $\mathcal{L}$ is the loss to minimize; s_m is the number of samples in a mini-batch; Y_i and Y are the survival days of ith sample and the vector of survival days, respectively; $\hat{Y}$ represents the vector of prognosis estimations for a mini-batch. Since the output activation of the GAENET is the hyperbolic tangent, the model tries to estimate 1 if a sample has a higher survival day than the average of the mini-batch; otherwise, −1 is targeted. Therefore, the output value of a trained GAENET becomes low if the prognosis of the input sample is likely to be poor.

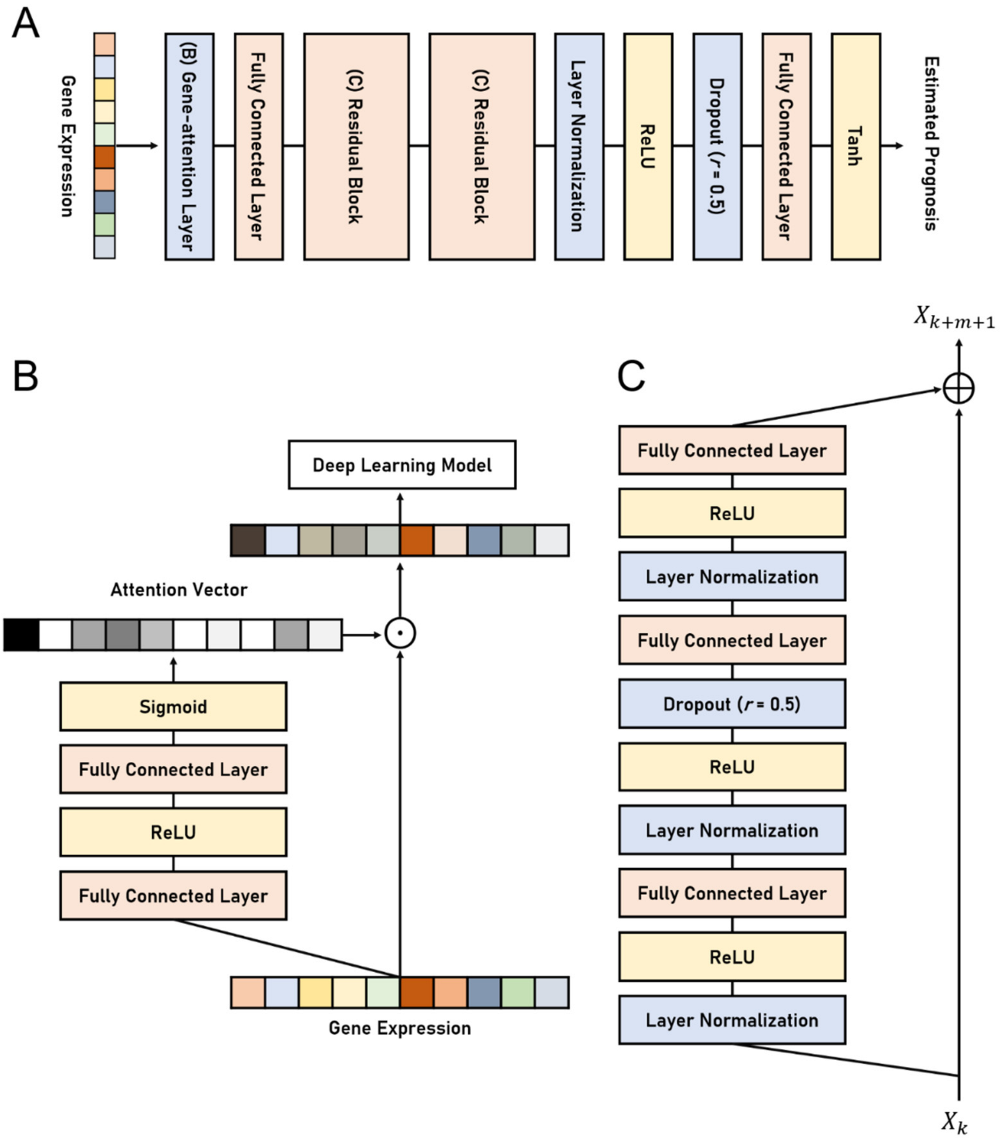

Figure 1. The architectures of the GAENET. (**A**) The architecture of a sub-model in the GAENET. (**B**) The proposed gene attention layer. (**C**) Residual block in the sub-model. The gene attention layer and residual blocks in (**A**) are illustrated in (**B**) and (**C**), respectively. Each purpose of the process is indicated by its color: Orange denotes a deep learning layer; yellow denotes an activation function; blue denotes regularization methods for layer outputs. The GAENET is composed of an ensemble of n number of the sub-models in (**A**), whereas the models share the gene attention layer. The r in the dropout layers indicates the dropout rate.

Thus, the GAENET is trained with survival days. By using the survival days as the target, these estimation values from -1 to 1 produced by the GAENET become correlated with survival days, since Equation (9) aims to maximize the correlation between survival days and the values. In this manner, it is possible to infer that a patient seems to have a poor prognosis if his or her gene expression has a lower estimation value approaching -1.

After the training of the GAENET, the attention vector in the gene attention layer signifies how much the genes are used to perform the estimation. Thus, the values of the attention vector indicate prognostic genes, where the prognostic genes have high attention values while the others have low values. In order to determine prognostic genes, attention vectors used for each sample are averaged as follows:

$$P_G := \frac{1}{s} \sum\nolimits_{i \in \{1,\ldots,s\}} GA\left(X^{(i)}\right), \tag{10}$$

where P_G is the prognostic gene score; $X^{(i)}$ indicates gene expression of a sample; s is the number of samples; $GA(\cdot)$ is the sigmoid outputs in the gene attention layer.

2.6. Hyperparameters for Training and Experimental Settings

The number of sub-models in the GAENET, i.e., n in Equation (8), was set at 10. The number of nodes of the layers in the GAENET was 16; however, the number of nodes was reduced to 8 at the first and second layers in each residual block to alleviate overfitting. The number of nodes in the hidden layer of gene attention was set at 10. The Adam optimization algorithm [45] with a learning rate of 0.0001 was used to train the model. In the experiments, the model was trained with 100 epochs; 5 mini-batches were randomly composed for each epoch.

The model was evaluated using cross-validation with 20% of the samples as a test set. The test set was randomly selected from the samples in the dataset, and this cross-validation was performed 30 times with random test sets. The performance of the model was compared to conventional methods, ridge regression, survival support vector machine (SurvivalSVM) [46], random survival forest (RSF) [47], and Coxnet [48]. The performance was measured by C-Index and the log-rank test [49].

The Cancer Genome Atlas for Low-Grade Glioma dataset (TCGA-LGG) was used in the experiments. The uncensored 125 samples were employed for cross-validation; thus, 25 samples became a test set in each cross-validation. The RNA-seq data were normalized by RNA-Seq by Expectation-Maximization (RSEM) [50]; then, the gene expression data were transformed by log normalization. If the zero value of a gene exceeds 50%, the gene was excluded from the dataset.

3. Results

3.1. The Prognosis Estimation of the GAENET

The performance comparison and the top-50 prognostic genes of LGG estimated by the GAENET are shown in Figure 2. As a result, *HILS1* was estimated to be the most significant prognostic gene with an attention value of 0.8033. While the linear relationship between genes and prognosis has conventionally been studied, deep learning models can estimate the prognosis in a nonlinear manner. This indicates that the prognostic genes estimated by the GAENET are significant considering combinations with the expression of other genes. Since genes regulate and promote with each other by composing gene networks, such combinational effects can be noticeable in estimating prognosis.

HILS1 has been reported as a pseudogene [51], which is considered that it is not likely encoded to a protein. However, recent studies have demonstrated that some pseudogenes regulate the expression of other genes, verifying the relationships with the prognosis of cancers [52]. In addition, previous studies have reported that *HILS1* is a prognostic gene of LGG, considering the linear relationships with the prognosis [53]. In addition, the second and third significant prognostic genes, *DACH1* and *GLIS1*, have been widely reported that they are associated with the regulation of tumors and cancers [54,55]. Conversely, it has not yet been reported whether *FARP2*, the fourth prognostic gene, is associated with LGG, which should be further studied in terms of nonlinear and combinational perspectives with other genes since the prognostic genes estimated by the GAENET are selected in such a view.

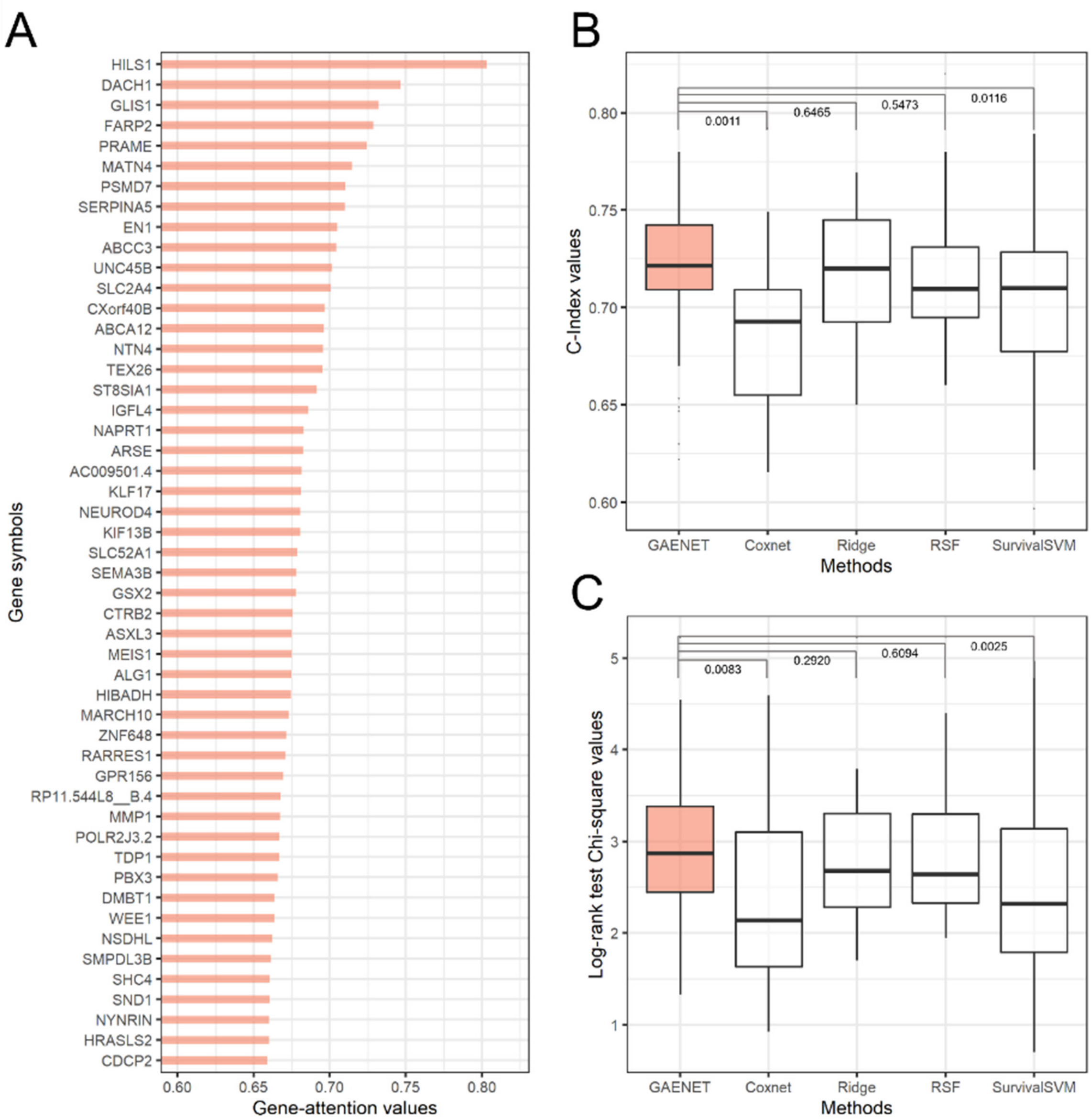

Figure 2. Prediction results of the GAENET. (**A**) Estimated prognostic genes for low-grade glioma. (**B**) Comparison of estimation performance with C-Index. (**C**) Comparison of estimation performance with the log-rank test. The estimated prognostic genes were discovered with test sets. The values above the boxplots are the Wilcoxon signed rank test *p*-values. The results of log-rank test are exhibited with square root values of Chi-square values. The estimation performances were measured with 30 times of cross-validations. The proposed gene attention algorithm ranked *HILS1*, a pseudogene, as the most significant prognostic gene. The log-rank tests indicated that the proposed GAENET had a median χ value of 2.870, a difference of 7.2 percent from the second-best model, ridge regression.

In the performance evaluation, the GAENET outperformed the other methods, considering C-Index and Chi-square values. For example, the GAENET exhibited a median χ value of 2.870, demonstrating a 7.2% difference compared to the second-best model, ridge regression. As shown in the Figure 2, it should be noted that the variance of performance of the GAENET is relatively low, which is one of the advantages of ensemble structures.

3.2. Combinational Analysis of Estimated Prognostic Genes

As mentioned in the previous section, the estimated prognostic genes were discovered from a perspective of deep learning, which was nonlinear and combinational. Such combinational effects are significant, but prognosis has conventionally been analyzed in a single gene with a linear relationship considering the nature of gene expression. In this study, a combinational analysis of gene expression with the estimated prognostic genes was performed. The gene expression of two groups, poor prognosis and good prognosis, were analyzed. The good prognosis group and the poor prognosis group are in the first quartile and the fourth quartile of survival days, respectively. The maximum survival of the poor prognosis group was 438 days, while the minimum survival of the good prognosis group was 1578 days.

The gene expression of the prognostic genes according to the group was distinct from each other. As shown in Figure 3, the gene expression of two groups with respect to *HILS1* was noticeable; in the poor prognosis group, *HILS1* was up-regulated while it was down-regulated in the other group. In addition, it is demonstrated that the expression of *HILS1* is correlated with that of the other prognostic genes; among the 19 genes, it is confirmed that 17 genes are correlated (p-value < 0.05) with *HILS1* while eight genes are highly correlated (p-value < 0.001), which implies that *HILS1* is a biomarker likely to regulate the expression of other prognostic genes. This factor accords with the characteristic of *HILS1*, in which a pseudogene is not encoded into a protein while they may regulate other genes [56].

The most significant correlation was observed between *ABCC3* and *SERPINA5*; they were commonly up-regulated in the poor prognosis group. The joint distribution of gene expression exhibited that *TEX26* and *ABCC3* are correlated (p-value < 0.01) only in the poor prognosis group, whereas no correlation is observed in the other group. This implies that these two genes may jointly affect the poor prognosis of LGG. Such correlations only in the poor prognosis group (p-value < 0.01) were also observed between *HILS1* and *EN1*, *GLIS1* and *NAPRT1*, *PSMD7* and *CXorf40B*, and *SLC2A4* and *IGFL4*.

The most important finding of this analysis is that *HILS1* expression is substantially linked with the expression of other prognostic genes. This suggests that *HILS1* may regulate other prognostic genes, resulting in prognostic disparities. As a result, *HILS1* is a likely candidate for down-regulation, given its high expression in patients with a poor prognosis. Additionally, some genes were shown to be correlated exclusively in the group with a poor prognosis but not in the other group. This implies that a few genes identified in this investigation may contribute to the poor prognosis of LGG in combination.

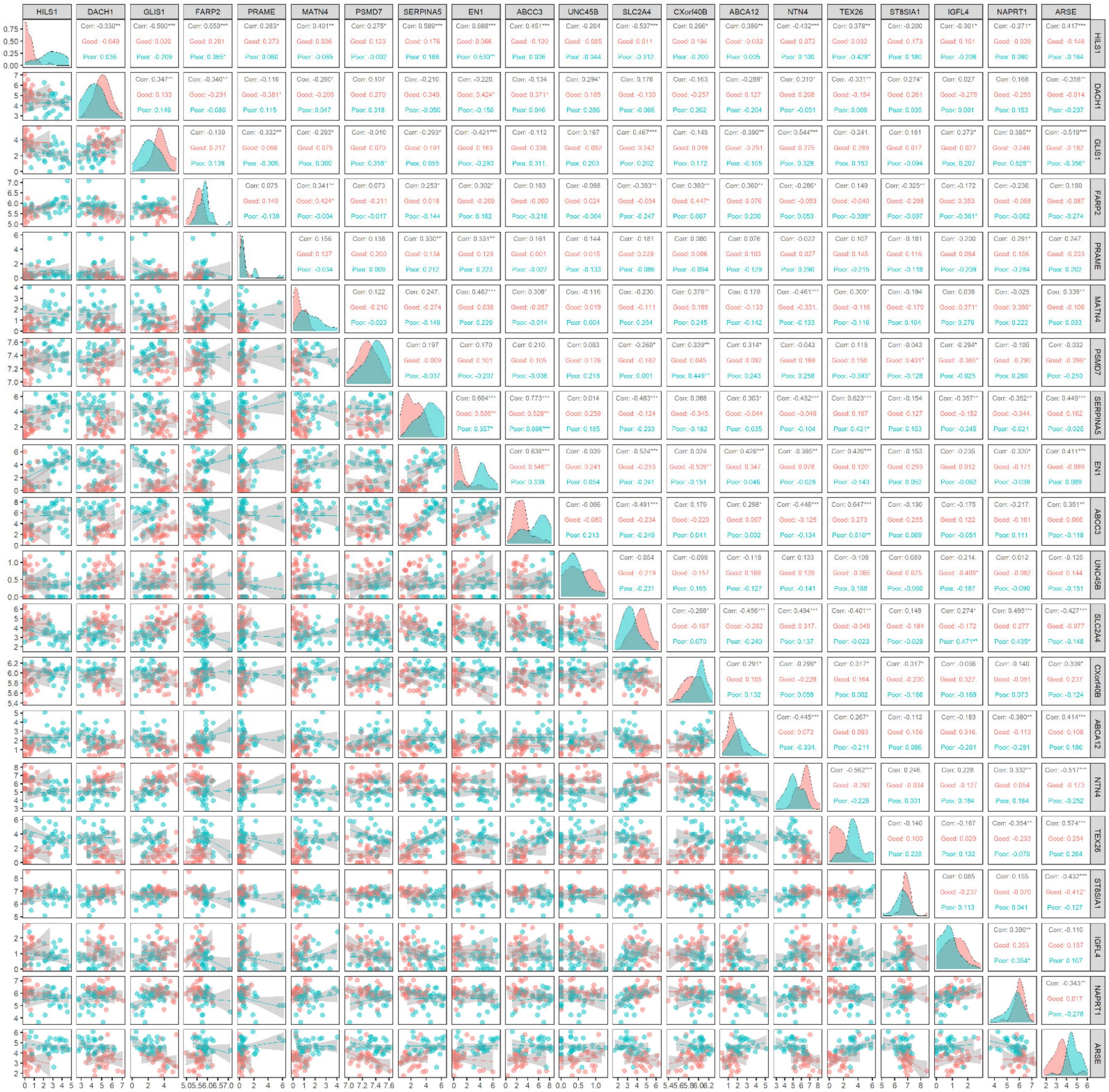

Figure 3. Joint distributions of gene expression of the estimated prognostic genes. The good prognosis group and the poor prognosis group are in the first quartile and the fourth quartile of survival days, respectively. Diagonal elements represent the marginal distribution of each gene. Lower-triangle elements exhibit the sample distribution for the groups. Upper-triangle elements display the correlation in each group. In the correlation, the symbols *, **, and *** indicate $p < 0.05$, < 0.01, and < 0.001, respectively. It is established that the expression of *HILS1* is correlated with the expression of other prognostic genes; 17 genes are correlated with *HILS1* ($p < 0.05$), while 8 genes are highly correlated ($p < 0.001$), implying that *HILS1* is a biomarker likely to regulate the expression of other prognostic genes.

3.3. Ablation Study of the GAENET

An ablation study was performed to investigate the effects of the methods used in the GAENET. In the ablation study, the performance without each method consisting of the GAENET was measured with C-Index. The models, each of which does not use one of gene attention, *LN*, residual learning, and ensemble learning, were compared. Additionally, the performance considering sub-models in the ensemble in the GAENET was analyzed. The performance is summarized in Figure 4.

As a result, while the GAENET showed the best performance, the models without gene attention and *LN* were marginally inferior to the GAENET. Conversely, the residual learning exhibited a significant difference in performance where the difference of C-Index was 3.14×10^{-2}, which indicated the effectiveness of the residual learning in terms of the deep architecture of neural networks. Similarly, the performance was inferior to the GAENET when a single model was used instead of the ensemble in the GAENET.

As shown in Figure 4B, all ensemble models demonstrated superior performances compared to the single model. Additionally, the best performance (C-Index: 0.7176) was observed when the number of models was five. Although it was expected that the performance increased as the number of models increased, the GAENET with ten sub-models was marginally inferior to five sub-models due to various random effects in deep learning, such as the random initialization of weight parameters and random mini-batch sampling.

This result indicates that each of the deep learning techniques employed in the GAENET contributes to the performance of the model. Among the techniques, residual learning and ensemble learning are the most effective methods for dramatically improving performance. As a result, in this study, it is demonstrated that modern deep learning techniques can be adapted for gene expression and are more successful than ordinary deep learning models.

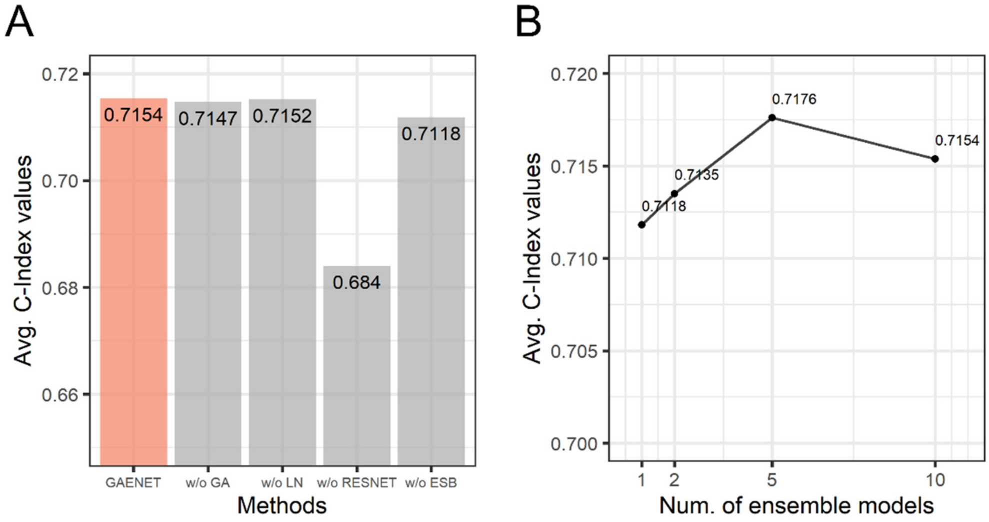

Figure 4. The results of ablation study of the GAENET. (**A**) C-Index values without each method. (**B**) C-Index values with respect to number of sub-models in the ensemble of the GAENET. GA: gene attention; LN: layer normalization; RESNET: residual learning; ESB: ensemble learning.

4. Discussion

A deep learning model for prognosis estimation with gene expression, the GAENET, and gene attention was proposed in this study. While other studies have employed ordinary deep learning models, i.e., MLP and AE, without recent developments [21,22], the GAENET demonstrated that the recent developments, such as attention mechanism and residual learning, can be modified for gene expression and exhibited superior performances compared to the models without the developments.

A few biomedical findings were exhibited by the proposed gene attention since gene attention can find prognostic genes through deep learning training. The proposed gene attention is informative in a nonlinear view because prognostic genes are considered with combinational effects and polynomial relationships. While conventional studies commonly focused on linear relationships with linear models, such as the Cox proportional-hazards model and Kaplan-Meier estimation, this study proposed a new method to investigate prognostic genes in a nonlinear view. As a result, *HILS1* was found as the most significant prognostic gene. This result accorded with other studies that reported *HILS1* was likely correlated with the prognosis of LGG [53].

Other studies have reported several prognostic genes of LGG. For instance, a study investigated the relationships between pseudogenes and the prognosis of LGG and reported *PKMP3*, *AC027612.4*, *HILS1*, *RP5-1132H15.3*, and *HSPB1P1* as candidates of prognostic pseudogenes of LGG [53]. A meta-analysis has found that high expression of *HIP1R* and low expression of *TBXAS1* were correlated with increased overall survival of diffuse LGG after investigating eight independent microarray datasets [57]. A recent study using the Cox model reported that *CTSZ*, *EFEMP2*, *ITGA5*, *KDELR2*, *MDK*, *MICALL2*, *MAP2K3*, *PLAUR*, *SERPINE1*, and *SOCS3* were related to the prognosis of glioblastoma multiforme and LGG, while the same dataset (TCGA) was used in this study [58].

Since these studies were based on the linear relationships between gene expression and prognosis as well as univariate analyses, the prognostic genes in these studies can be different from the result of gene attention. However, the most significant prognostic gene estimated by gene attention, i.e., *HILS1*, was also reported in the other study [53], which indicated that *HILS1* was a distinct prognostic gene not only in a nonlinear view but also in a linear view.

Generally, *IDH1* mutation and 1p/19q codeletion has been known as prognostic biomarkers in glioma [59]. Since RNA-seq gene expression data are used in this study, the GAENET has a limitation in that the model is not able to detect such biomarkers. However, it has been reported that nearly all 1p/19q codeleted glioma has *IDH1* mutation [60]; therefore, such biomarkers can be represented by the gene expression of *IDH1* in a mediate manner, although the mutation cannot be directly detected. As a result, the gene attention found that *IDH1* was the top 2.2% prognostic gene, which was 443rd out of 19,990 genes. Given that the expression of *IDH1* was investigated instead of its mutation, it can be said that gene attention partially considered *IDH1* as a potential prognostic gene.

5. Conclusions

In this study, an ensemble deep learning model, namely the GAENET, was proposed to estimate the prognosis of LGG with gene expression. The GAENET employed the proposed attention mechanism for gene expression data, called gene attention. Owing to the advantage of the gene attention mechanism, it is possible to discover prognostic genes in a nonlinear and combinational perspective, which is the nature of deep learning training. While deep learning training has been considered a black box, it was demonstrated that prognostic genes in such a view can be determined by the gene attention layer. Additionally, a few conventional training methods for deep learning, including residual learning, were introduced.

As a result, it was demonstrated that the GAENET outperformed conventional methods for estimating prognosis, exhibiting a 7.2% performance difference compared to the second-best model. Additionally, the prognostic genes of LGG, including *HILS1*, *DACH1*, and *GLIS1*, were discovered by the proposed gene attention. In the ablation study of the GAENET, it was confirmed that the methods used in the GAENET contributed to the performance of the GAENET, in which the residual learning showed the most significant difference.

Funding: This research was supported by the Chung-Ang University Research Grants in 2021 as well as the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1050977).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the retrospective nature of the study and the use of a public dataset.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of a public dataset.

Data Availability Statement: The dataset used in this paper is a public dataset, obtained at Genomic Data Commons Data Portal: https://portal.gdc.cancer.gov/ (accessed on 27 February 2022).

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

Appendix A

A deep learning model consists of weight matrices and nonlinear activation functions. Given an input, a deep learning model produces an output with a mathematical computation with the weight matrices and activation functions. The simplest deep learning model has a weight matrix and a nonlinear function, constituting a layer. In this model, the mathematical computation is simply multiplying the input and the weight matrix and applying the nonlinear function, which can be represented as follows:

$$\hat{Y} := \sigma\left(\begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_j \end{bmatrix}\right), \tag{A1}$$

where the matrix with elements of w is the weight matrix; the vector with elements with x is the input; σ is the activation function.

This computation in Equation (A1) produces an output with i number of elements. Such a computation can be repeated by treating the output as a new input for the next layer with a different weight matrix and activation function. The number of repetitions indicates the depth of the model. In this manner, Equations (1) and (3) are deep learning models with multiple layers.

Modifications of Equation (A1) can also be a layer of deep learning models. Numerous modifications for each purpose have been proposed. Equations (2) and (4) regularize the output of a layer by multiplying specific values. When smaller values are multiplied, the resulting outputs become less significant, whereas outputs multiplied by larger values become critical. These multiplied values are also defined by a deep learning model of Equation (3), which is the proposed gene attention layer.

Additionally, Equations (5)–(7) correspond to the modifications necessary for optimal deep learning model training. The residual learning in Equation (5) is for backpropagation with smooth gradient flows. The layer normalization in Equations (6) and (7) also correspond to modifications of layer outputs, as in Equations (2) and (4). It maintains output distributions for more effective training.

References

1. Lee, M.; Tae, D.; Choi, J.H.; Jung, H.-Y.; Seok, J. Improved recurrent generative adversarial networks with regularization techniques and a controllable framework. *Inf. Sci.* **2020**, *538*, 428–443. [CrossRef]
2. Lee, M.; Seok, J. Estimation with uncertainty via conditional generative adversarial networks. *Sensors* **2021**, *21*, 6194. [CrossRef] [PubMed]
3. Park, M.; Lee, M.; Yu, S. HRGAN: A Generative Adversarial Network Producing Higher-Resolution Images than Training Sets. *Sensors* **2022**, *22*, 1435. [CrossRef] [PubMed]
4. Liu, X.; Song, L.; Liu, S.; Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **2021**, *13*, 1224. [CrossRef]
5. Li, Y.; Zhao, J.; Lv, Z.; Li, J. Medical image fusion method by deep learning. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 21–29. [CrossRef]
6. Kim, M.; Lee, B.-D. Automatic lung segmentation on chest X-rays using self-attention deep neural network. *Sensors* **2021**, *21*, 369. [CrossRef]
7. Ahsan, M.; Based, M.; Haider, J.; Kowalski, M. COVID-19 detection from chest X-ray images using feature fusion and deep learning. *Sensors* **2021**, *21*, 1480.
8. Yu, K.; Tan, L.; Lin, L.; Cheng, X.; Yi, Z.; Sato, T. Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health. *IEEE Wirel. Commun.* **2021**, *28*, 54–61. [CrossRef]
9. Liu, X.; Wang, H.; Li, Z.; Qin, L. Deep learning in ECG diagnosis: A review. *Knowl. Based Syst.* **2021**, *227*, 107187. [CrossRef]
10. Vitsios, D.; Dhindsa, R.S.; Middleton, L.; Gussow, A.B.; Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.* **2021**, *12*, 1504. [CrossRef]

11. Liu, Y.; Huang, Y.; Wang, G.; Wang, Y. A deep learning approach for filtering structural variants in short read sequencing data. *Brief. Bioinform.* **2021**, *22*, bbaa370. [CrossRef]
12. Zhang, J.X.; Yordanov, B.; Gaunt, A.; Wang, M.X.; Dai, P.; Chen, Y.-J.; Zhang, K.; Fang, J.Z.; Dalchau, N.; Li, J. A deep learning model for predicting next-generation sequencing depth from DNA sequence. *Nat. Commun.* **2021**, *12*, 4387. [CrossRef]
13. Vale-Silva, L.A.; Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **2021**, *11*, 13505. [CrossRef]
14. Allegra, A.; Tonacci, A.; Sciaccotta, R.; Genovese, S.; Musolino, C.; Pioggia, G.; Gangemi, S. Machine Learning and Deep Learning Applications in Multiple Myeloma Diagnosis, Prognosis, and Treatment Selection. *Cancers* **2022**, *14*, 606. [CrossRef]
15. Arya, N.; Saha, S. Multi-modal advanced deep learning architectures for breast cancer survival prediction. *Knowl. Based Syst.* **2021**, *221*, 106965. [CrossRef]
16. Atlam, M.; Torkey, H.; El-Fishawy, N.; Salem, H. Coronavirus disease 2019 (COVID-19): Survival analysis using deep learning and Cox regression model. *Pattern Anal. Appl.* **2021**, *24*, 993–1005. [CrossRef]
17. Vo, T.-H.; Lee, G.-S.; Yang, H.-J.; Oh, I.-J.; Kim, S.-H.; Kang, S.-R. Survival prediction of lung cancer using small-size clinical data with a multiple task variational autoencoder. *Electronics* **2021**, *10*, 1396. [CrossRef]
18. Oliveira, D.V.; Prahm, K.P.; Christensen, I.J.; Hansen, A.; Høgdall, C.K.; Høgdall, E.V. Gene expression profile association with poor prognosis in epithelial ovarian cancer patients. *Sci. Rep.* **2021**, *11*, 5438. [CrossRef]
19. Van De Vijver, M.J.; He, Y.D.; Van't Veer, L.J.; Dai, H.; Hart, A.A.; Voskuil, D.W.; Schreiber, G.J.; Peterse, J.L.; Roberts, C.; Marton, M.J. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **2002**, *347*, 1999–2009. [CrossRef]
20. Zeng, D.; Zhou, R.; Yu, Y.; Luo, Y.; Zhang, J.; Sun, H.; Bin, J.; Liao, Y.; Rao, J.; Zhang, Y. Gene expression profiles for a prognostic immunoscore in gastric cancer. *J. Br. Surg.* **2018**, *105*, 1338–1348. [CrossRef]
21. Wong, K.K.; Rostomily, R.; Wong, S.T. Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers* **2019**, *11*, 53. [CrossRef]
22. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **2018**, *24*, 1248–1259. [CrossRef]
23. 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. [CrossRef]
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
25. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
26. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
27. 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. Process. Syst.* **2017**, *30*, 1–11.
28. Tay, Y.; Bahri, D.; Metzler, D.; Juan, D.-C.; Zhao, Z.; Zheng, C. Synthesizer: Rethinking self-attention for transformer models. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10183–10192.
29. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–13.
30. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 234–250.
31. Li, W.; Liu, K.; Zhang, L.; Cheng, F. Object detection based on an adaptive attention mechanism. *Sci. Rep.* **2020**, *10*, 11307. [CrossRef]
32. Ye, Y.; Dai, Q.; Qi, H. A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer. *Cell Death Discov.* **2021**, *7*, 71. [CrossRef]
33. Shen, S.; Wang, G.; Zhang, R.; Zhao, Y.; Yu, H.; Wei, Y.; Chen, F. Development and validation of an immune gene-set based Prognostic signature in ovarian cancer. *eBioMedicine* **2019**, *40*, 318–326. [CrossRef]
34. Li, S.; Xuan, Y.; Gao, B.; Sun, X.; Miao, S.; Lu, T.; Wang, Y.; Jiao, W. Identification of an eight-gene prognostic signature for lung adenocarcinoma. *Cancer Manag. Res.* **2018**, *10*, 3383. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
37. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
38. Claus, E.B.; Walsh, K.M.; Wiencke, J.K.; Molinaro, A.M.; Wiemels, J.L.; Schildkraut, J.M.; Bondy, M.L.; Berger, M.; Jenkins, R.; Wrensch, M. Survival and low-grade glioma: The emergence of genetic information. *Neurosurg. Focus* **2015**, *38*, E6. [CrossRef] [PubMed]
39. Lee, M.; Han, S.W.; Seok, J. Prediction of survival risks with adjusted gene expression through risk-gene networks. *Bioinformatics* **2019**, *35*, 4898–4906. [CrossRef] [PubMed]
40. Shrestha, A.; Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access* **2019**, *7*, 53040–53065. [CrossRef]

41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
42. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
43. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
44. Zhou, T.; Lu, H.; Yang, Z.; Qiu, S.; Huo, B.; Dong, Y. The ensemble deep learning model for novel COVID-19 on CT images. *Appl. Soft Comput.* **2021**, *98*, 106885. [CrossRef]
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. Pölsterl, S.; Navab, N.; Katouzian, A. Fast training of support vector machines for survival analysis. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Porto, Portugal, 7–11 September 2015; pp. 243–259.
47. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [CrossRef]
48. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **2011**, *39*, 1. [CrossRef]
49. Lee, S.; Lim, H. Review of statistical methods for survival analysis using genomic data. *Genom. Inform.* **2019**, *17*, e41. [CrossRef]
50. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]
51. Wang, Y.; Liu, X.; Guan, G.; Xiao, Z.; Zhao, W.; Zhuang, M. Identification of a five-pseudogene signature for predicting survival and its ceRNA network in glioma. *Front. Oncol.* **2019**, *9*, 1059. [CrossRef]
52. Xiao-Jie, L.; Ai-Mei, G.; Li-Juan, J.; Jiang, X. Pseudogene in cancer: Real functions and promising signature. *J. Med. Genet.* **2015**, *52*, 17–24. [CrossRef]
53. Liu, B.; Liu, J.; Liu, K.; Huang, H.; Li, Y.; Hu, X.; Wang, K.; Cao, H.; Cheng, Q. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. *Biomed. Pharmacother.* **2019**, *117*, 109116. [CrossRef]
54. Xu, H.; Yu, S.; Yuan, X.; Xiong, J.; Kuang, D.; Pestell, R.G.; Wu, K. DACH1 suppresses breast cancer as a negative regulator of CD44. *Sci. Rep.* **2017**, *7*, 4361. [CrossRef]
55. Jetten, A.M. GLIS1–3 transcription factors: Critical roles in the regulation of multiple physiological processes and diseases. *Cell. Mol. Life Sci.* **2018**, *75*, 3473–3494. [CrossRef]
56. Stasiak, M.; Kolenda, T.; Kozłowska-Masłoń, J.; Sobocińska, J.; Poter, P.; Guglas, K.; Paszkowska, A.; Bliźniak, R.; Teresiak, A.; Kazimierczak, U. The World of Pseudogenes: New Diagnostic and Therapeutic Targets in Cancers or Still Mystery Molecules? *Life* **2021**, *11*, 1354. [CrossRef]
57. Wang, S.; Jin, F.; Fan, W.; Liu, F.; Zou, Y.; Hu, X.; Xu, H.; Han, P. Gene expression meta-analysis in diffuse low-grade glioma and the corresponding histological subtypes. *Sci. Rep.* **2017**, *7*, 11741. [CrossRef]
58. Hsu, J.B.-K.; Chang, T.-H.; Lee, G.A.; Lee, T.-Y.; Chen, C.-Y. Identification of potential biomarkers related to glioma survival by gene expression profile analysis. *BMC Med. Genom.* **2019**, *11*, 34. [CrossRef]
59. Sanson, M.; Marie, Y.; Paris, S.; Idbaih, A.; Laffaire, J.; Ducray, F.; El Hallani, S.; Boisselier, B.; Mokhtari, K.; Hoang-Xuan, K. Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. *J. Clin. Oncol.* **2009**, *27*, 4150–4154. [CrossRef] [PubMed]
60. Labussiere, M.; Idbaih, A.; Wang, X.-W.; Marie, Y.; Boisselier, B.; Falet, C.; Paris, S.; Laffaire, J.; Carpentier, C.; Criniere, E. All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. *Neurology* **2010**, *74*, 1886–1890. [CrossRef] [PubMed]