International Journal of
Molecular Sciences

MDPI

Article

A Study on the Prediction of Cancer Using Whole-Genome Data and Deep Learning

Young-Ji Lee [1], Jun-Hyung Park [2] and Seung-Ho Lee [1,*]

[1] Department of Electronic Engineering, Hanbat National University, Daejeon 34158, Korea
[2] 3BIGS Co., Ltd., 156, Gwanggyo-ro, Yeongtong-gu, Suwon-si 16506, Korea
[*] Correspondence: shlee@cad.hanbat.ac.kr

Abstract: The number of patients diagnosed with cancer continues to increasingly rise, and has nearly doubled in 20 years. Therefore, predicting cancer occurrence has a significant impact on reducing medical costs, and preventing cancer early can increase survival rates. In the data preprocessing step, since individual genome data are used as input data, they are classified as individual genome data. Subsequently, data embedding is performed in character units, so that it can be used in deep learning. In the deep learning network schema, using preprocessed data, a character-based deep learning network learns the correlation between individual feature data and predicts cancer occurrence. To evaluate the objective reliability of the method proposed in this study, various networks published in other studies were compared and evaluated using the TCGA dataset. As a result of comparing various networks published in other studies using the same data, excellent results were obtained in terms of accuracy, sensitivity, and specificity. Thus, the superiority of the effectiveness of deep learning networks in predicting cancer occurrence using individual whole-genome data was demonstrated. From the results of the confusion matrix, the validity of the model for predicting the cancer using an individual's whole-genome data and the deep learning proposed in this study was proven. In addition, the AUC, which is the area under the ROC curve, which judges the efficiency of diagnosis as a performance evaluation index of the model, was found to be 90% or more, good classification results were derived. The objectives of this study were to use individual genome data for 12 cancers as input data to analyze the whole genome pattern, and to not separately use reference genome sequence data of normal individuals. In addition, several mutation types, including SNV, DEL, and INS, were applied.

Keywords: personal genome data; genomic variation; cancers; TCGA dataset; data classification; data screening; data transformation; k-max pooling; deep learning; shortcut connection

check for updates

Citation: Lee, Y.-J.; Park, J.-H.; Lee, S.-H. A Study on the Prediction of Cancer Using Whole-Genome Data and Deep Learning. *Int. J. Mol. Sci.* **2022**, *23*, 10396. https://doi.org/10.3390/ijms231810396

Academic Editor: Alvaro Galli

Received: 14 July 2022
Accepted: 5 September 2022
Published: 8 September 2022

1. Introduction

Currently, the number of patients diagnosed with cancer continues to increase, and the WHO announced that the total number of people diagnosed with cancer has nearly doubled from 10 million in 2000 to 19.3 million in 2020 [1]. In 2020, there were about 10 million deaths from cancer [2]. In addition, due to the increase in the aged population, the proportion of cancer deaths among the total deaths is continuously increasing. Predicting cancer occurrence plays a significant role in reducing healthcare costs, and preventing cancer early also increases survival rates. Therefore, the bio-healthcare market, which can check an individual's health status, has expanded worldwide, and its recent aim is to improve the quality of medical services and reduce cost [3,4]. On the other hand, although there are many causes of cancer, it is known that the probability of occurrence of cancer is remarkably high when genetic and chromosomal mutations occur. Genes are vital for the growth and development of the human body and maintaining life. However, when a small part of a gene is deleted or abnormal, it leads to various abnormalities, diseases, and death in some cases [5,6]. With the advancement of technology, the study of which genes are related to which diseases has yielded insightful results, and research on mutated genes and genetic mutations related to specific diseases continues to evolve [7–9]. Most cancers are caused by gene mutations; gene mutations caused by external factors (smoking, UV rays, environmental changes, etc.) are the most common cancer development pattern [4,10]. Representative mutations were revealed by research include mutated genes, such as BRCA1 and BRCA2, that can initiate the development of breast cancer [11,12]. If a mutation occurs in these genes, women have 60–80% increased risk of breast cancer and 40% increased risk of ovarian cancer [13]. In the case of men, it is reported that the risk of prostate cancer and male breast cancer is increased, and it is also associated with pancreatic cancer and colorectal cancer. Thus, research on the relationship between genes and diseases is important, and is steadily progressing. Since mutations have unique characteristics depending on the cause of their occurrence, not only the pattern analysis of specific genes, but also the pattern analysis of the whole genome is needed. However, there are many difficulties in analyzing the patterns of thousands of genes. Recently, research into extracting results from various and complex data using deep learning techniques [14–20] is in progress. Um et al. [21] in 2017 measured the accuracy of hand tremors by applying wearable data sensors on the hands of people with Parkinson's disease. Similarly, Wei Jiao et al. [22] in 2020 predicted cancer using gene frequency data and gene mutation data, and Youngji Lee et al. [23] in 2020, predicted disease through a deep learning network by performing multiple data augmentations on individual microbiome data. A study by Sun et al. [24] predicted cancer through a deep learning network for 12 types of cancer based on reference genome sequence data [25] and specific genome mutation data for normal individuals. Due to the difficulty in using large training datasets, the learning procedure was limited to specific genes related to specific diseases. However, since the genome as well as mutations are unique for each person, it is possible to contract a disease even if the specific gene is not mutated. For this reason, it is necessary to analyze all gene patterns possessed by each cancer patient, rather than analyzing the gene pattern targeting only a specific gene related to a specific cancer.

Therefore, in this study, we propose a method to predict cancer using an individual's whole-genome data (whole-exome sequencing and whole-genome sequencing) and deep learning. The objectives of the proposed study were as follows: First is the use of individual genome data for 12 cancers as input data to analyze the whole genome pattern without analyzing the genome pattern related to a specific cancer. Secondly, since the whole genome pattern of an individual is analyzed, the reference genome sequence data of the normal person were not used separately because the reference genome sequence of the normal person is not compared. Thirdly, to apply not only gene mutation types, such as SNV, DEL, and INS of point mutations, but also general genetic mutation types. Fourth, to achieve learning based on the character convolution network [26–28], so it can be easily applied to various diseases and rare diseases. The final objective was to analyze not only specific genomic mutations related to specific diseases, but also overall genomic patterns.

2. Results

2.1. Experiment Environment

To evaluate the objective performance of individual genomic data using the deep learning techniques suggested in this study and predicting cancer outbreaks, we use public genomic data from the Cancer Genome Atlas (TCGA) project conducted by the U.S. National Institutes of Health (NIH) [29]. The hardware used in the experiments in this study is the Intel(R) Core(TM) i7-9700K 3.60 GHz CPU, it consists of 16 GB of RAM, Nvidia GeForce RTX 2080 Ti (V-RAM 11 GB) GPU, and development tools on the Window 10 Pro 64-bit operating system used JetBrains PyCharm Community Edition 2021.1.1. In addition, torch 1.2.0, CUDA 10.0, cuDNN 7.6.5 libraries were used.

2.2. Experimental Procedure

After preprocessing data by performing data classification, data screening, and data transformation, training sets were used to train the deep learning network. In the learning process, backpropagation was repeatedly performed until the loss was very small to correct the weights to improve the accuracy. Next, based on the trained deep learning model, cancer prediction was performed with the untrained test set as the input.

2.3. Dataset

In this study, the National Cancer Institution Genomic Data Commons (NCI GDC) dataset used in this paper consisted of TCGA pan-cancer somatic mutation data [30]. The whole genome pattern was analyzed using individual genome data from 12 cancers such as BLCA, BRCA, COAD, GBM, KIRC, LGG, LUSC, OV, PRAD, SKCM, THCA, and UCEC as inputs. In addition, SNVs (SNV, DEL, and INS) were applied, and gender information was also included. Data from a total of 5554 patients were used, of which data from 4438 patients were used as the training set. In total, 892 patients were used as the validation set to determine whether learning progressed well, and 224 patients were used as the test set to confirm the learning results. Table 1 shows the names and numbers of datasets of 12 cancer types. Figure 1 shows an example of the test set used for evaluation.

Table 1. Name and number of dataset for 12 cancers.

Label	Cancer Name	Cancer	Training Set	Validation Set	Test Set	Total
1	Bladder urothelial carcinoma	BLCA	314	63	16	393
2	Breast invasive carcinoma	BRCA	776	157	38	971
3	Colon adenocarcinoma	COAD	258	52	13	323
4	Glioblastoma multiforme	GBM	304	61	16	381
5	Skin cutaneous melanoma	SKCM	285	58	14	357
6	Kidney renal clear cell carcinoma	KIRC	268	53	14	335
7	Brain lower-grade glioma	LGG	405	81	21	507
8	Lung squamous cell carcinoma	LUSC	379	76	19	474
9	Ovarian serous cystadenocarcinoma	OV	344	68	18	430
10	Prostate adenocarcinoma	PRAD	394	79	20	493
11	Thyroid carcinoma	THCA	393	79	20	492
12	Uterine corpus endometrial carcinoma	UCEC	318	65	15	398
Total	12	12	4438	892	224	5554

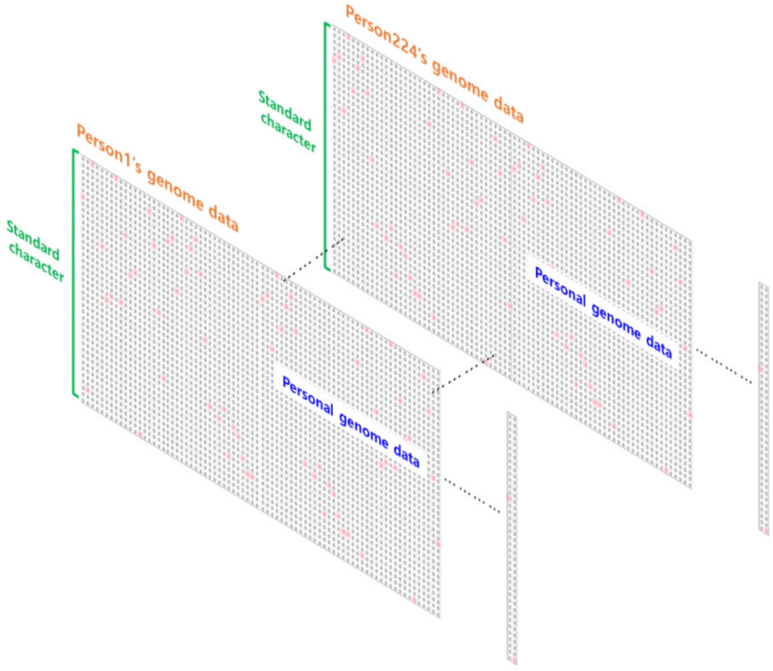


Figure 1. Example of a sample dataset used for evaluation.

2.4. Reliability Evaluation of Cancer Prediction Results

In this study, it took approximately 11 s to load the training model and process the 224 sources of test data. To evaluate the objective reliability of the prediction result of cancer using individual's whole-genome data and the deep learning technique proposed in this study, the accuracy, precision, specificity, sensitivity, F-score, etc., were comparatively evaluated. In a study by Sun et al. in 2019, based on the normal gene sequence, the authors generated data by writing different sections as 1 and the same sections as 0 in the base sequence of a specific gene, and then used deep learning to identify 12 cancers. Therefore, in this study, cancer was predicted using the genetic data for 12 cancers, same as in the study by Sun et al. [24]. The 12 predicted cancers were BLCA, BRCA, COAD, GBM, KIRC, LGG, LUSC, OV, PRAD, SKCM, THCA, and UCEC. Figure 2 illustrates how the performance of the model was evaluated. In Figure 2, TP (True Positive) is when the actual value is 1 and the model predicts it as 1, and TN (True Negative) is when the actual value is 0 and the model predicts it as 0. In addition, FP (False Positive) is a case in which the actual value is 0 but the model predicts 1, and FN (False Negative) is a case in which the actual value is 1 but the model predicts 0.

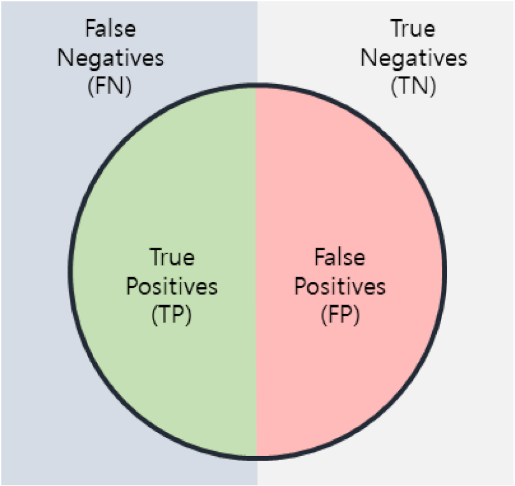


Figure 2. Performance evaluation index of the model.

Accuracy is a representative indicator used and indicates the degree to which the value to be measured is correctly predicted, and is obtained by Equation (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}} \tag{1}$$

Sensitivity represents a case in which a predictive result is positive among people with a disease, and is obtained by Equation (2).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{2}$$

Specificity indicates a case in which the prediction result is negative among people without disease, and is obtained by Equation (3).

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{3}$$

Precision represents a case in which a person with an actual disease is represented among the cases in which the prediction result is judged to be positive, and is obtained by Equation (4).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

F-score is an index that combines precision and sensitivity, and is obtained by Equation (5).

$$\text{F–score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (5)$$

The performance evaluation of this study had an accuracy of 74.11%. As shown in Table 2, the mean and standard deviation of precision were 75.7% and 14.53% respectively. The mean and standard deviation of sensitivity were 73.84 and 11.22% respectively. The mean and standard deviation of specificity were 97.61% and 1.86% respectively. The mean and standard deviation of F-score were 74.1% and 10.27% respectively. The reasons that the sensitivity was relatively lower than the specificity are as follows: First, there was little difference in the frequency of base mutations between cancers. Second, although the types of cancers were diverse, they were poorly derived because they exhibit common genetic mutations at the molecular level.

Table 2. Results of precision, sensitivity, specificity, and F-score for 12 cancers.

Label	Cancer Name	Cancer	Test Set	Precision	Sensitivity	Specificity	F-Score
1	Bladder urothelial carcinoma	BLCA	16	58.8	62.5	96.63	60.6
2	Breast invasive carcinoma	BRCA	38	70.7	76.3	93.55	73.4
3	Colon adenocarcinoma	COAD	13	100.0	76.9	100.0	87.0
4	Glioblastoma multiforme	GBM	16	58.8	62.5	96.63	60.6
5	Skin cutaneous melanoma	SKCM	14	68.8	78.6	97.62	73.3
6	Kidney renal clear cell carcinoma	KIRC	14	68.4	92.9	97.14	78.8
7	Brain lower-grade glioma	LGG	21	93.8	71.4	99.51	81.1
8	Lung squamous cell carcinoma	LUSC	19	94.4	89.5	99.51	91.9
9	Ovarian serous cystadenocarcinoma	OV	18	76.5	72.2	98.06	74.3
10	Prostate adenocarcinoma	PRAD	20	64	80.0	95.59	71.1
11	Thyroid carcinoma	THCA	20	87.5	70.0	99.02	77.8
12	Uterine corpus endometrial carcinoma	UCEC	15	66.7	53.3	98.09	59.3
Average				75.7	73.84	97.61	74.1

The confusion matrix [31] is an index used to evaluate the performance of a specific classification model, and expresses the relationship between the actual value and the value predicted by the model at a glance. In Figure 3, the vertical axis represents the actual cancer number, and the horizontal axis represents the predicted cancer number. In Figure 3, BRCA can be judged to be a cancer associated with UCEC and OV, since UCEC and OV were predicted as well. Since the prediction results of other cancers are also expressed as the related cancers, the adequacy of the model for predicting cancer using an individual's whole-genome data and deep learning proposed in this study is proven.

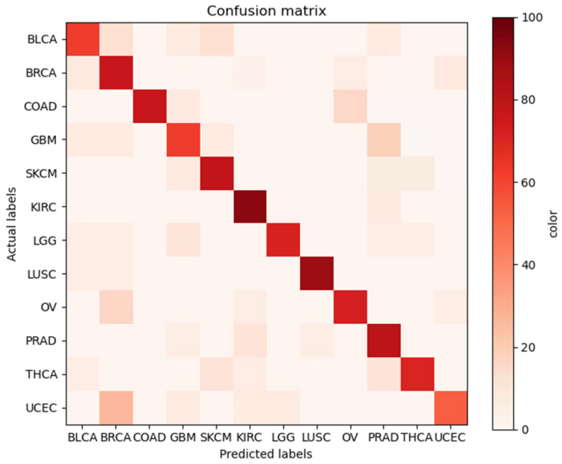


Figure 3. Confusion matrix of 12 cancers.

In addition, Receiver Operating Characteristic (ROC) curves judge the efficiency of diagnosis as a performance evaluation index of the model [32]. The ROC curve is not only widely used for diagnostic purposes in the medical field, but is also a commonly used index for algorithm performance evaluation. The ROC curve is a graph expressing the relationship between specificity and sensitivity according to a standard, and the horizontal axis shows specificity, and the vertical axis indicates sensitivity. As the ROC curve rises to the top left, and thus the Area Under the Curve (AUC) increases, it can be judged that the prediction was better. Figure 4 presents the ROC curves for the 12 cancers predicted in this study. If the AUC of the ROC curve is more than 90%, it is generally recognized that the classification result is excellent. From the analysis, it is shown that the AUC is over 90% for all 12 cancers predicted in this study (Figure 4), indicating good classification results.

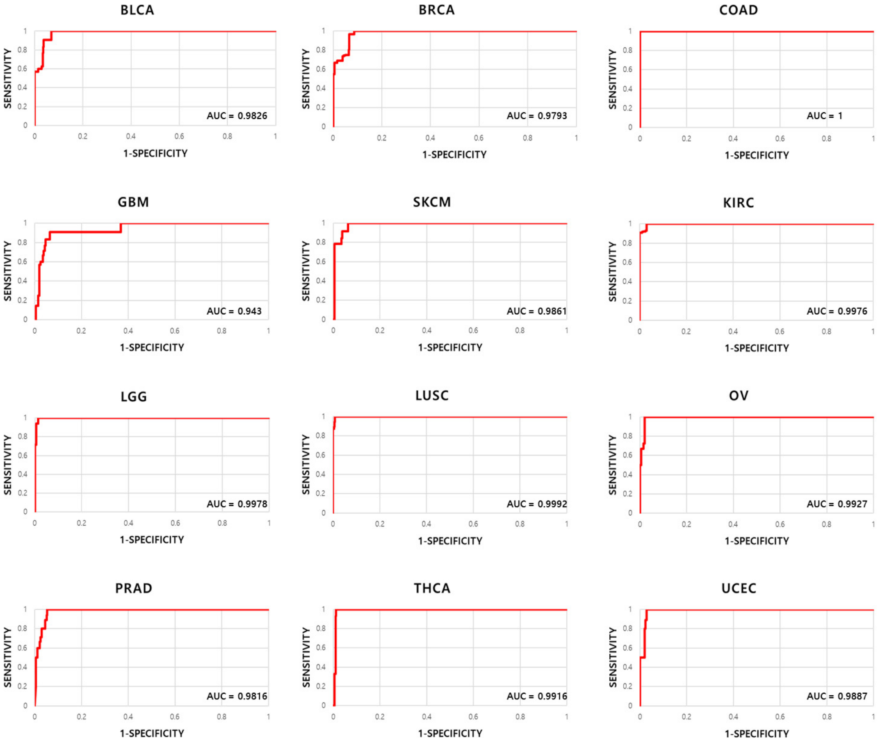


Figure 4. ROC curves of 12 cancers.

We compared our results with multiple networks using data for 12 cancers (BLCA, BRCA, COAD, GBM, KIRC, LGG, LUSC, OV, PRAD, SKCM, THCA, UCEC). As shown in Table 3, the method proposed in this paper had higher accuracy, sensitivity, and specificity than other networks. As a result of learning the same data using networks such as AlexNet, ResNet18, and ResNet34, poorer results were obtained than the network proposed in this paper [33,34]. It is shown that the network proposed in this paper produces better results than other existing networks because it learns by extracting features from the entire section by applying k-max pooling. Therefore, the superiority of the deep learning network proposed in this study to predict using individual whole-genome data and cancer was determined.

Table 3. Results of proposed method and other networks.

Method	Accuracy	Sensitivity	Specificity
AlexNet [33]	56.68	56.89	95.62
ResNet18 [34]	66.07	64.95	96.84
ResNet34 [34]	67.69	66.21	96.99
The Proposed Method	74.11	73.84	97.61

Furthermore, our results are compared with Sun et al. [24] in Table 4. Sun et al. [24] used WES tumor germline variants and somatic mutation data. Our paper used only somatic mutation data. Although the data of the papers are different, we compared them because it is a method for predicting multiple cancers with one network.

Table 4. Results of the proposed method and those of Sun et al. [24].

Method	Accuracy	Sensitivity	Specificity
Sun et al. [24] (using the WES tumor germline variants and somatic mutation data)	70.4	65.92	96.27
The Proposed Method (using only somatic mutation data)	74.11	73.84	97.61

3. Discussion

Current causes of cancer include environmentally induced mutations (somatic mutation) and genetic mutation. As the number of cancers are increasing, the seriousness of somatic mutation is rising. Analyzing the mutation pattern in genes is required. There are many difficulties in analyzing thousands of gene patterns. Thus, research on various and complex data using deep learning is currently being conducted. However, since the genome and genome mutations are different for each person, even if you do not possess a specific gene, you can contract a disease. Rather than predicting with data limited to a specific genome, by learning and predicting with the whole-genome data of an individual, a more comprehensive prediction can be performed. For this reason, it is necessary to analyze all gene patterns possessed by each cancer patient, rather than analyzing the gene pattern targeting only a specific gene related to a specific cancer.

To evaluate the objective reliability of the method proposed in this study, the accuracy, precision, specificity, and F-score, were comparatively evaluated. We preformed this study using the whole-genome data for 12 cancers. The accuracy, precision, sensitivity, specificity, and F-score were 74.11%, 75.7%, 73.84%, 97.61%, and 74.1%, respectively. These results are higher than those produce by other networks. Thus, the effectiveness of our network is proven. In this study, since the reference genome sequence data for normal individuals were not needed, only individual genome data for 12 cancers were used as training data to analyze an individual's whole-genome mutation pattern. The training data includes several mutation types includes SNV, DEL, and INS. It also includes characteristic information about chromosome number, position, and type of genetic variation. The whole-genome data used in this paper consisted of genome sequencing data and exome sequencing data, and information on genomic mutations occurring between the start and end positions of genome mutations is specified. Therefore, it was judged that the information that affects genomic variation, such as missense, silent, frame shift, etc., is necessary for learning. As learning is performed based on the character convolution network, it can be easily applied to various diseases, including rare diseases. Therefore, it is possible to analyze not only specific genome mutations, but also the whole genome pattern. The preprocessed character-based one-hot encoding vector extracts feature information through a convolution layer and a pooling layer. Therefore, overfitting does not occur because there is a greater possibility of extracting feature information that affects the result rather than feature information on repeated characters. The proposed method uses individual genome data for 12 cancers as input data to analyze the whole genome pattern, and does not separately use reference genome sequence data for normal individuals.

On the other hand, our results were expressed as a confusion matrix to evaluate the performance between cancers related to the prediction results of a specific cancer predicted by the model. In addition, the classification results were shown as the AUC, the area under the ROC curve, which is an index for evaluating the performance of algorithms, and these results were more than 90%; therefore, the ROC curve is shaped well for the 12 predicted cancers.

As a result of comparing the method proposed in this study using the same data for various networks published in other studies, excellent results were obtained in terms of accuracy, sensitivity, and specificity. Thus, the superiority of deep learning networks in predicting cancer using individual whole-genome data is demonstrated.

In summary, the evaluation of the prediction of cancer using whole-genome data and deep learning proposed in this study is as follows. First, it is considered as a basis for predicting various types of cancer because it does not learn specific genetic data related to specific cancers, but analyzes all genome patterns of each cancer patient. Second, the reason that the results for uterine cancer were lower than those of other cancers is that BRCA1 and BRCA2 mutant genes are related to both breast and uterine cancer. Third, this study did not use reference sequence information for normal people and predicts cancer using only individual genomic data information. Given the flexibility of this approach, it can be extended to other diseases as well. As a future research plan, it is considered necessary to apply genomic data not only to cancer, but also other diseases, including rare diseases.

4. Materials and Methods

The proposed network predicts 12 types of cancer through a deep learning network using individual whole-genome data. Figure 5 presents the overall schema of the prediction of cancer using whole-genome data and deep learning. The data used in the proposed method are individual whole-genome data. Since genomic data has various mutation information for each chromosome, the type of word increases exponentially as the quantity of data, including new genomic mutation information, increases. On the other hand, since the number of characters is specified for a character, the number of characters does not change, even if the quantity of data increases. Therefore, after embedding data in character units, the results are derived through a deep learning network.

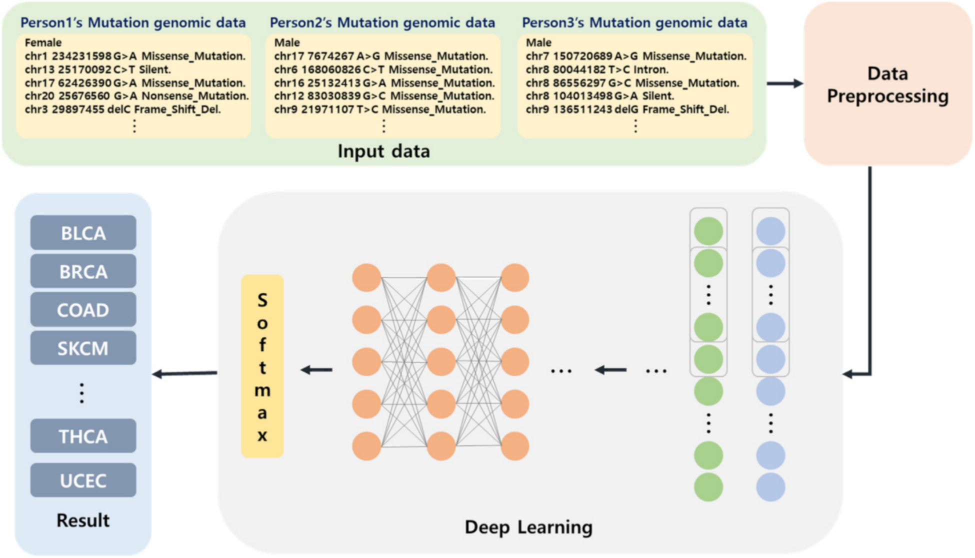

Person1's Mutation genomic data
Female
chr1 234231598 G>A Missense_Mutation.
chr13 25170092 C>T Silent.
chr17 62426390 G>A Missense_Mutation.
chr20 25676560 G>A Nonsense_Mutation.
chr3 29897455 delC Frame_Shift_Del.
⋮
Person2's Mutation genomic data
Male
chr17 7674267 A>G Missense_Mutation.
chr6 168060826 C>T Missense_Mutation.
chr16 25132413 G>A Missense_Mutation.
chr12 83030839 G>C Missense_Mutation.
chr9 21971107 T>C Missense_Mutation.
⋮
Person3's Mutation genomic data
Male
chr7 150720689 A>G Missense_Mutation.
chr8 80044182 T>C Intron.
chr8 86556297 G>C Missense_Mutation.
chr8 104013498 G>A Silent.
chr9 136511243 delG Frame_Shift_Del.
⋮
Input data
Data Preprocessing
BLCA
BRCA
COAD
SKCM
⋮
THCA
UCEC
Result
Softmax
Deep Learning

Figure 5. The overall schema for the prediction of cancer using whole-genome data and deep learning.

4.1. Overall Flow Chart of This Study

The overall flow chart of this study consists of a data preprocessing stage that processes data classification, data screening, data transformation, and the deep learning network, which learns using training sets among the preprocessed data, as shown in Figure 6. After data preprocessing is performed, genome data are learned using the deep learning network based on character convolution. In the execution process, cancer prediction is performed using test sets based on the learned model.

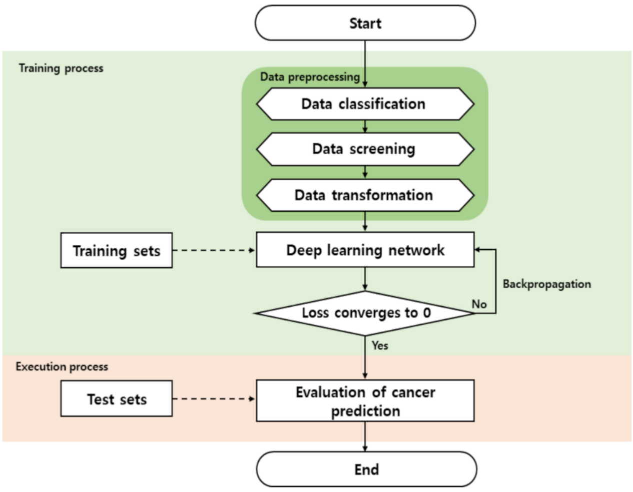


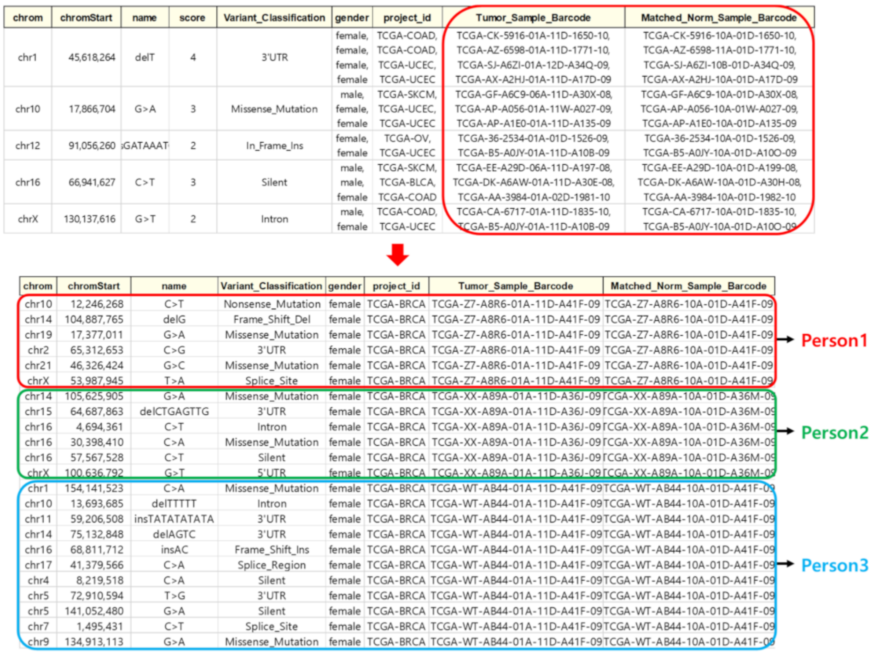
Figure 6. The overall flow chart of this study.

4.2. Data Preprocessing

The data for 12 cancers, BLCA, BRCA, COAD, GBM, KIRC, LGG, LUSC, OV, PRAD, SKCM, THCA, and UCEC, among the published genomic data of the TCGA project conducted by the National Institutes of Health [35], were included in this study. In total, 80% of the total 5554 data were in the training set, and the rest were divided to a validation set and test set. The data preprocessing stage consists of a data classification process that classifies individual data, a data screening process that reduces training time for the network results, and a data transformation process that converts into a learning-enabled form.

4.2.1. Data Classification

Individual whole-genome data were used as input data. However, since TCGA data were not classified by individual, they were classified as individual whole-genome data using Sample Barcode. The Sample Barcode is a unique key for classifying each biological sample data in the TCGA database. Figure 7 shows the classification of TCGA data into individual whole-genome data using Sample Barcode.



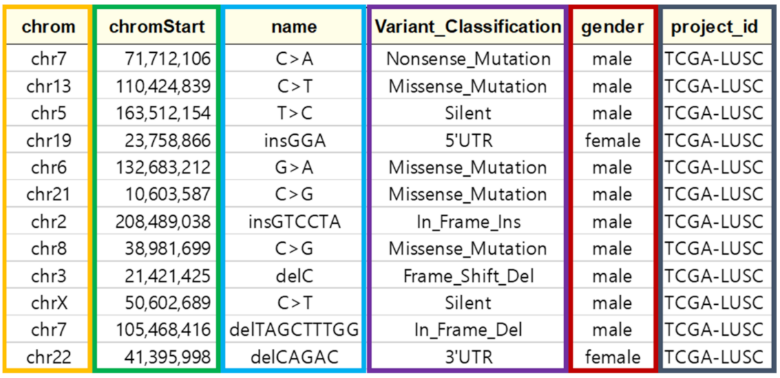
chrom	chromStart	name	score	Variant_Classification	gender	project_id	Tumor_Sample_Barcode	Matched_Norm_Sample_Barcode
chr1	45,618,264	delT	4	3'UTR	female, female, female, female	TCGA-COAD, TCGA-COAD, TCGA-UCEC, TCGA-UCEC	TCGA-CK-5916-01A-11D-1650-10, TCGA-AZ-6598-01A-11D-1771-10, TCGA-SJ-A6ZI-01A-12D-A34Q-09, TCGA-AX-A2HJ-01A-11D-A17D-09	TCGA-CK-5916-10A-01D-1650-10, TCGA-AZ-6598-11A-01D-1771-10, TCGA-SJ-A6ZI-10B-01D-A34Q-09, TCGA-AX-A2HJ-10A-01D-A17D-09
chr10	17,866,704	G>A	3	Missense_Mutation	male, female, female	TCGA-SKCM, TCGA-UCEC, TCGA-UCEC	TCGA-GF-A6C9-06A-11D-A30X-08, TCGA-AP-A056-01A-11W-A027-09, TCGA-AP-A1E0-01A-11D-A135-09	TCGA-GF-A6C9-10A-01D-A30X-08, TCGA-AP-A056-10A-01W-A027-09, TCGA-AP-A1E0-10A-01D-A135-09
chr12	91,056,260	GATAAAT	2	In_Frame_Ins	female, female	TCGA-OV, TCGA-UCEC	TCGA-36-2534-01A-01D-1526-09, TCGA-B5-A0JY-01A-11D-A10B-09	TCGA-36-2534-10A-01D-1526-09, TCGA-B5-A0JY-10A-01D-A10O-09
chr16	66,941,627	C>T	3	Silent	male, male, female	TCGA-SKCM, TCGA-BLCA, TCGA-COAD	TCGA-EE-A29D-06A-11D-A197-08, TCGA-DK-A6AW-01A-11D-A30E-08, TCGA-AA-3984-01A-02D-1981-10	TCGA-EE-A29D-10A-01D-A199-08, TCGA-DK-A6AW-10A-01D-A30H-08, TCGA-AA-3984-10A-01D-1982-10
chrX	130,137,616	G>T	2	Intron	male, female	TCGA-COAD, TCGA-UCEC	TCGA-CA-6717-01A-11D-1835-10, TCGA-B5-A0JY-01A-11D-A10B-09	TCGA-CA-6717-10A-01D-1835-10, TCGA-B5-A0JY-10A-01D-A10O-09

chrom	chromStart	name	Variant_Classification	gender	project_id	Tumor_Sample_Barcode	Matched_Norm_Sample_Barcode
chr10	12,246,268	C>T	Nonsense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr14	104,887,765	delG	Frame_Shift_Del	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr19	17,377,011	G>A	Missense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr2	65,312,653	C>G	3'UTR	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr21	46,326,424	G>C	Missense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chrX	53,987,945	T>A	Splice_Site	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr14	105,625,905	G>A	Missense_Mutation	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr15	64,687,863	delCTGAGTTG	3'UTR	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	4,694,361	C>T	Intron	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	30,398,410	C>A	Missense_Mutation	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	57,567,528	C>T	Silent	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chrX	100,636,792	G>T	5'UTR	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr1	154,141,523	C>A	Missense_Mutation	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr10	13,693,685	delTTTTT	Intron	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr11	59,206,508	insTATATATATA	3'UTR	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr14	75,132,848	delAGTC	3'UTR	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr16	68,811,712	insAC	Frame_Shift_Ins	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr17	41,379,566	C>A	Splice_Region	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr4	8,219,518	C>A	Silent	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr5	72,910,594	T>G	3'UTR	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr5	141,052,480	G>A	Silent	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr7	1,495,431	C>T	Splice_Site	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09
chr9	134,913,113	G>A	Missense_Mutation	female	TCGA-BRCA	TCGA-WT-AB44-01A-11D-A41F-09	TCGA-WT-AB44-10A-01D-A41F-09

Figure 7. Classified by individual genomic data.

4.2.2. Data Screening

After classifying into individual whole-genome data, data screening is performed. Data screening is an important process that increases accuracy and performance of the network result by excluding unclear and unrelated features. First, TCGA genetic data holds a variety of information, but environmental and acquired factors such as tobacco, weight, alcohol, etc., were excluded from variable information in this study because there was not enough data provided due to personal information protection. Six categories of the TCGA data variables are selected, and five categories for variables excluding cancer types are used as training data. Variables used in training data are chromosome number, chromosome location, genetic mutation, mutation type, and gender. The mutation type is an important variable in identifying genetic mutations because it contains information about whether the amino acid sequence changes when the sequence changes. Figure 8 shows the variable information to be used for learning data.

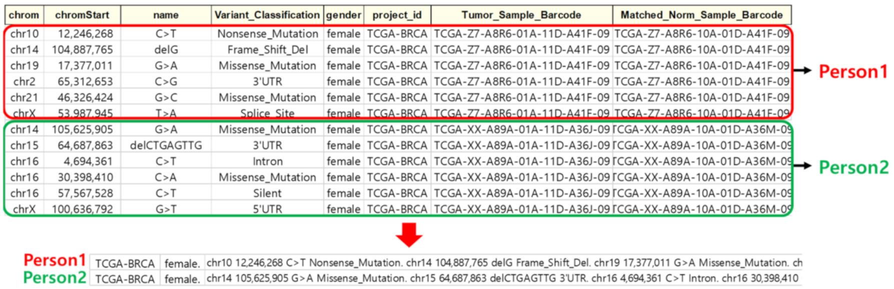


chrom	chromStart	name	Variant_Classification	gender	project_id
chr7	71,712,106	C>A	Nonsense_Mutation	male	TCGA-LUSC
chr13	110,424,839	C>T	Missense_Mutation	male	TCGA-LUSC
chr5	163,512,154	T>C	Silent	male	TCGA-LUSC
chr19	23,758,866	insGGA	5'UTR	female	TCGA-LUSC
chr6	132,683,212	G>A	Missense_Mutation	male	TCGA-LUSC
chr21	10,603,587	C>G	Missense_Mutation	male	TCGA-LUSC
chr2	208,489,038	insGTCCTA	In_Frame_Ins	male	TCGA-LUSC
chr8	38,981,699	C>G	Missense_Mutation	male	TCGA-LUSC
chr3	21,421,425	delC	Frame_Shift_Del	male	TCGA-LUSC
chrX	50,602,689	C>T	Silent	male	TCGA-LUSC
chr7	105,468,416	delTAGCTTTGG	In_Frame_Del	male	TCGA-LUSC
chr22	41,395,998	delCAGAC	3'UTR	female	TCGA-LUSC

Figure 8. Variable information to be used as training data.

4.2.3. Data Transformation

After the data selection, this process performs data transformation into a form that can be learned by the deep learning network. Various individual genomic data are converted into one sentence in order to apply to a character convolution-based deep learning network. As shown in Figure 9, each column represents different mutation feature information. Each row represents mutation feature information separated by ' '(spacing), and each row is separated by a '.'(dot).

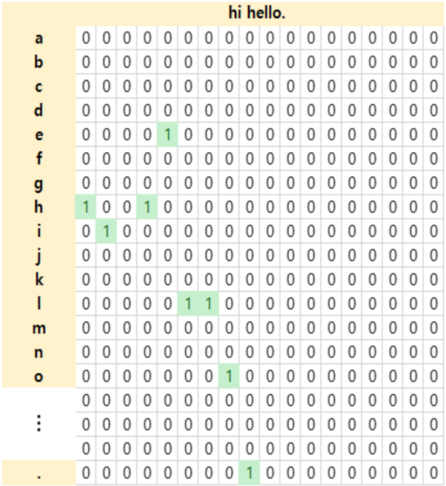


chrom	chromStart	name	Variant_Classification	gender	project_id	Tumor_Sample_Barcode	Matched_Norm_Sample_Barcode
chr10	12,246,268	C>T	Nonsense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr14	104,887,765	delG	Frame_Shift_Del	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr19	17,377,011	G>A	Missense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr2	65,312,653	C>G	3'UTR	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr21	46,326,424	G>C	Missense_Mutation	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chrX	53,987,945	T>A	Splice_Site	female	TCGA-BRCA	TCGA-Z7-A8R6-01A-11D-A41F-09	TCGA-Z7-A8R6-10A-01D-A41F-09
chr14	105,625,905	G>A	Missense_Mutation	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr15	64,687,863	delCTGAGTTG	3'UTR	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	4,694,361	C>T	Intron	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	30,398,410	C>A	Missense_Mutation	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chr16	57,567,528	C>T	Silent	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09
chrX	100,636,792	G>T	5'UTR	female	TCGA-BRCA	TCGA-XX-A89A-01A-11D-A36J-09	TCGA-XX-A89A-10A-01D-A36M-09

Person1	TCGA-BRCA	female.	chr10 12,246,268 C>T Nonsense_Mutation. chr14 104,887,765 delG Frame_Shift_Del. chr19 17,377,011 G>A Missense_Mutation. ch
Person2	TCGA-BRCA	female.	chr14 105,625,905 G>A Missense_Mutation. chr15 64,687,863 delCTGAGTTG 3'UTR. chr16 4,694,361 C>T Intron. chr16 30,398,410

Figure 9. The process of converting an individual's genome data into a single sentence.

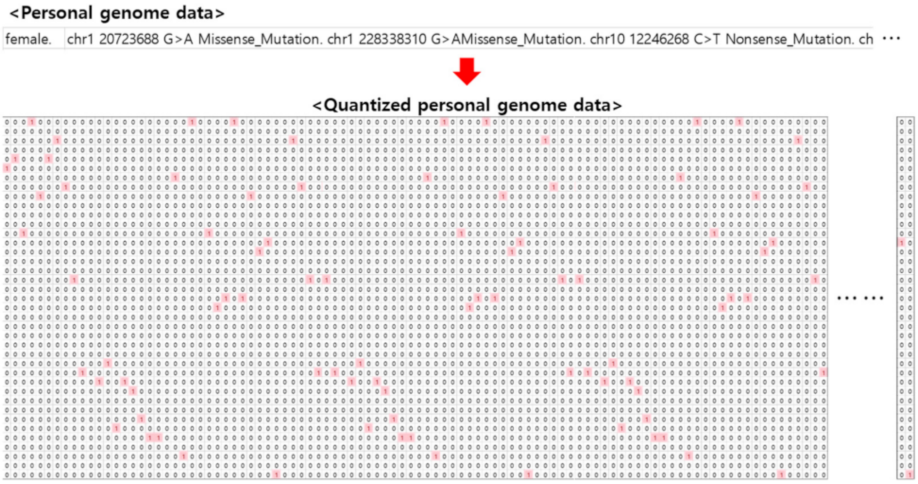
After converting multiple genomic data of an individual into a single sentence, uppercase letters are changed to lowercase letters and one-hot encoding is performed. Since the training data is the whole-genome data for each individual, there are several mutations in the genome of each individual. Therefore, even when new genome mutation information comes in, one-hot encoding is performed that converts data into character units that are easy to generate and learn data. Using a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, -, ', ;, etc.(70 characters) for characters convolution, the sentence of "hi hello" is performed one-hot encoding with 1 for the corresponding character and 0 for the non-corresponding character as in the example of Figure 10.



	hi hello.																	
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
⋮	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Figure 10. Example of one-hot encoding.

Based on the selected TCGA genome data variable information, one-hot encoding is per-formed to build 4438 training data as shown in Figure 11. The average number of variants for each patient was 217. The constructed training data is given as input data to deep learning network based on character convolution.



female. chr1 20723688 G>A Missense_Mutation. chr1 228338310 G>AMissense_Mutation. chr10 12246268 C>T Nonsense_Mutation. ch …

Figure 11. Construction of learning data after data conversion for personal genome data.

4.3. Deep Learning Network Architecture

Based on the selected TCGA genome data variable information, the constructed training data goes through the character convolution neural network shown in Figure 12, and the results of predicting cancer using genomic data are the output. It has been proven by previously published studies that when the layer of the deep learning network model is too deep, the performance of the model is lowered due to problems such as gradient vanishing and gradient exploding [36,37]. Gradient vanishing is a case in which the weights are not properly updated as the gradient gradually decreases toward the input layer in the backpropagation process during deep model training. Gradient exploding refers to a case in which the gradient increases and diverges to an abnormally large value. To solve this problem, the ResNet model for learning deep layers has been published [34]. In this study, we use shortcut connection, which is the basic schema of the ResNet model. Skipping several layers through a shortcut connection that skips more than one layer can solve the gradient vanishing problem and improve the learning speed. On the other hand, by using batch normalization that normalizes the distribution of data, the dependence of the initial weight value is reduced, and overfitting, which decreases the error for training data but increases the error for actual data, is reduced.

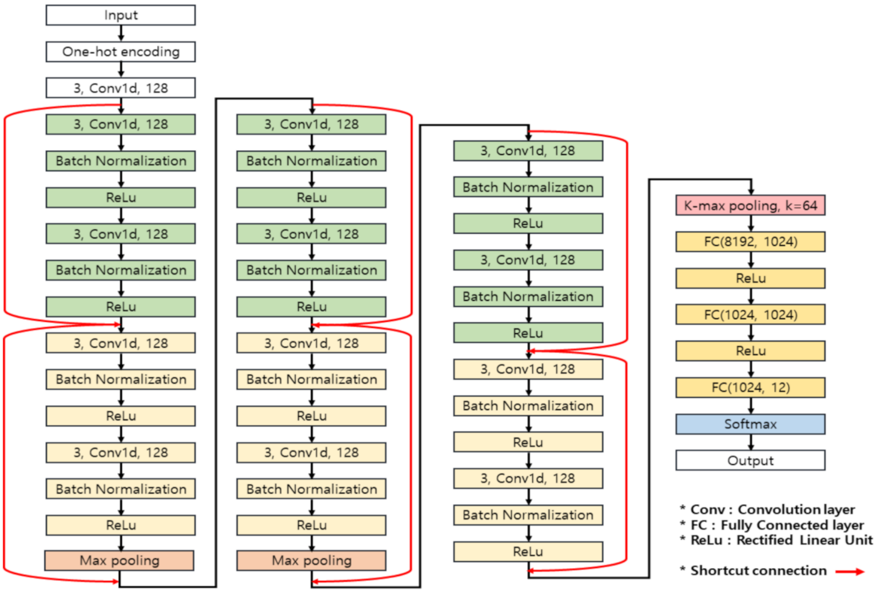


Figure 12. Structure of the character convolution neural network applied in this study.

Convolutional neural networks generally consist of a convolution layer and a pooling layer, and a fully connected layer [38] is applied at the last stage. The convolution layer extracts various features from the input data, and the pooling layer uses both max pooling and k-max pooling to prevent overfitting by reducing the size and parameters of the input data. Max pooling is down sampling by extracting the maximum value for each period of data. K-max pooling [39] performs down sampling by extracting k with the highest value from each row of learning data. Since the dataset used in this study is sequencing data, even though there are important data values in a certain section, there is a possibility of loss if it is not the maximum value. Therefore, k-max pooling is performed before the fully connected layer to extract features from the whole section. After the pooling layer, features are classified using a fully connected layer. Finally, the learning time is improved using the ReLu activation function [40], and the probability of cancer prediction results is estimated by performing softmax.

Backpropagation

In this study, backpropagation [41,42] is used to reduce the loss when the loss rises higher than certain level. The effect of improving the accuracy of cancer prediction results was obtained. After comparing the results obtained by the deep learning network and the results of the validation set, the backpropagation is repeated until the loss is very small to obtain the desired result for the output data. The weights are modified so that the data used for training is the genome data of a total of 4438 people, and the data used in the validation set are the genome data of 892 people. Epoch was set to 150 times, batch size was set to 8, and learning rate was set to 0.0001.

Author Contributions: Conceptualization, Y.-J.L.; Data curation, Y.-J.L.; Formal analysis, Y.-J.L.; Investigation, Y.-J.L.; Methodology, Y.-J.L.; Project administration, Y.-J.L.; Software, Y.-J.L.; Supervision, J.-H.P., S.-H.L.; Validation, J.-H.P., S.-H.L.; Writing—original draft, Y.-J.L.; Writing—review & editing, J.-H.P., S.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Technology Development Program (S2788640), Ministry of SMEs and Startups, Republic of Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. NRF-2022R1F1A1066371). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (IITP-2022-RS-2022-00156212) super-vised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting the conclusions of this article is available for download at the https://tcga-data.nci.nih.gov/docs/publications/tcga/, accessed on 31 July 2022. All source code developed by this study is publicly available at: https://github.com/leeyoungji/Prediction-Cancer-using-Whole-Genome-Data, accessed on 31 July 2022.

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

References

1. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2021. *CA Cancer J. Clin.* **2021**, *71*, 7–33. [CrossRef] [PubMed]
2. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef] [PubMed]
3. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **2018**, *1*, 18. [CrossRef] [PubMed]
4. Greenman, C.; Stephens, P.; Smith, R.; Dalgliesh, G.L.; Hunter, C.; Bignell, G.; Davies, H.; Teague, J.; Butler, A.; Stevens, C.; et al. Patterns of somatic mutation in human cancer genomes. *Nature* **2007**, *446*, 153–158. [CrossRef] [PubMed]
5. Emilsson, V.; Thorleifsson, G.; Zhang, B.; Leonardson, A.S.; Zink, F.; Zhu, J.; Carlson, S.; Helgason, A.; Walters, G.B.; Gunnarsdottir, S.; et al. Genetics of gene expression and its effect on disease. *Nature* **2008**, *452*, 423–428. [CrossRef]
6. Chun, H.-W.; Tsuruoka, Y.; Kim, J.-D.; Shiba, R.; Nagata, N.; Hishiki, T.; Tsujii, J. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Biocomputing* **2006**, *2006*, 4–15.
7. Shuch, B.; Vourganti, S.; Ricketts, C.J.; Middleton, L.; Peterson, J.; Merino, M.J.; Metwalli, A.R.; Srinivasan, R.; Linehan, W.M. Defining early-onset kidney cancer: Implications for germline and somatic mutation testing and clinical management. *J. Clin. Oncol.* **2014**, *32*, 431. [CrossRef] [PubMed]
8. Gilissen, C.; Hoischen, A.; Brunner, H.G.; Veltman, J. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* **2012**, *20*, 490–497. [CrossRef]
9. Van Dam, S.; Vosa, U.; van der Graaf, A.; Franke, L.; de Magalhaes, J.P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* **2018**, *19*, 575–592. [CrossRef]
10. Martincorena, I.; Campbell, P.J. Somatic mutation in cancer and normal cells. *Science* **2015**, *349*, 1483–1489. [CrossRef]
11. Antoniou, A.C.; Pharoah, P.D.P.; McMullan, G.; Day, N.E.; Stratton, M.R.; Peto, J.; Ponder, B.J.; Easton, D.F. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **2002**, *86*, 76–83. [CrossRef] [PubMed]
12. Levy-Lahad, E.; Friedman, E. Cancer risks among BRCA1 and BRCA2 mutation carriers. *Br. J. Cancer* **2007**, *96*, 11–15. [CrossRef] [PubMed]
13. Petrucelli, N.; Daly, M.B.; Pal, T. *BRCA1-and BRCA2-Associated Hereditary Breast and Ovarian Cancer*; University of Washington: Seattle, WA, USA, 2016.
14. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
15. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
19. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603.
20. Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable object detection using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2147–2154.

21. Um, T.T.; Pfister, F.M.J.; Pichler, D.; Endo, S.; Lang, M.; Hirche, S.; Fietzek, U.; Kulić, D. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 216–220.
22. Jiao, W.; Atwal, G.; Polak, P.; Karlic, R.; Cuppen, E.; Danyi, A.; de Ridder, J.; van Herpen, C.; Lolkema, M.P.; Steeghs, N.; et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **2020**, *11*, 728. [CrossRef]
23. Lee, Y.-J.; Park, J.-H.; Chung, H.-Y.; Kim, K.-M.; Lee, S.-H. A Data Augmentation Methodology for Predicting the Association of Microbiome Community and Diseases Based on Artificial Intelligence. *J. Inst. Electron. Inf. Eng.* **2021**, *58*, 59–66.
24. Sun, Y.; Zhu, S.; Ma, K.; Liu, W.; Yue, Y.; Hu, G.; Lu, H.; Chen, W. Identification of 12 cancer types through genome deep learning. *Sci. Rep.* **2019**, *9*, 17256. [CrossRef]
25. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **2010**, *467*, 1061. [CrossRef]
26. Koomsubha, T.; Vateekul, P. A character-level convolutional neural network with dynamic input length for Thai text categorization. In Proceedings of the 2017 9th International Conference on Knowledge and Smart Technology (KST), Chon Buri, Thailand, 1–4 February 2017; IEEE: Piscataway, NJ, USA, 2017.
27. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
28. Conneau, A.; Schwenk, H.; Barrault, L.; Lecun, Y. Very deep convolutional networks for text classification. *arXiv* **2016**, arXiv:1606.01781.
29. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. Review the Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol./Współczesna Onkol.* **2015**, *2015*, 68–77. [CrossRef]
30. 33 TCGA Cancer Projects Summary. Available online: https://genome.ucsc.edu/cgi-bin/hgTables?db=hg38&hgta_group=phenDis&hgta_track=gdcCancer&hgta_table=allCancer&hgta_doSchema=describe+table+schema (accessed on 6 May 2019).
31. Visa, S.; Ramsay, B.; Ralescu, A.L.; van der Knaap, E. Confusion matrix-based feature selection. *MAICS* **2011**, *710*, 120–127.
32. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commmun. ACM* **2017**, *60*, 84–90.
34. 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, 26 June–1 July 2016.
35. Ding, L.; Bailey, M.H.; Porta-Pardo, E.; Thorsson, V.; Colaprico, A.; Bertrand, D.; Gibbs, D.L.; Weerasinghe, A.; Huang, K.-L.; Tokheim, C.; et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **2018**, *173*, 305–320. [CrossRef] [PubMed]
36. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
37. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **1998**, *6*, 107–116. [CrossRef]
38. Sun, D.; Wulff, J.; Sudderth, E.B.; Pfister, H.; Black, M.J. A fully-connected layered model of foreground and background flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
39. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. *arXiv* **2014**, arXiv:1404.2188.
40. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
41. Hecht-Nielsen, R. Theory of the backpropagation neural network. In *Neural Networks for Perception*; Academic Press: Cambridge, MA, USA, 1992; pp. 65–93.
42. Werbos, P.J. Backpropagation through time: What it does and how to do it. *Proc. IEEE* **1990**, *78*, 1550–1560. [CrossRef]