International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Potential of Disease Prediction using Deep Learning Algorithms

Jian Oh
Ewha Womans University (Class of 2022)
52, Ewhayeodae-gil, Seodaemun-gu, Seoul, Republic of Korea, 03760
Major in Life Science

Abstract— *When big data is applied to the medical field, the first thing to consider is the algorithm. Numerous treatment data are being generated every day in general hospitals. Algorithms are essential in classifying these data and constructing a model that is useful for treatment. The binary algorithm is the beginning of all algorithms. This study aims to implement a process for predicting cancer-based on an algorithm used in the early stages of Big Data extraction through machine learning.*

Keywords— *Artificial Neural Networks, Deep Learning, Artificial Intelligence, Cancer Prediction.*

I. Introduction

Cancer is the largest cause of death in humans. Thanks to advances in modern medical technology, the 5-year survival rate of cancer patients are increasing. However, there is still a long way to go for medical technology to overcome cancer. Preventing cancer in advance rather than treating it after cancer is a critical factor in determining life quality. The era of finding mutations in cancer genes by Next Generation Sequencing(NGS) technologies has arrived. Using Deep Learning, which collects data on mutations in actual cancer patients, it is possible to determine the driver gene that caused cancer. For predicting cancer on genetic mutation, it is necessary to construct an optimal Deep Learning model. The basic principle applied to Deep Learning models is a binary algorithm. Here, we will use a cancer prediction program based on ANN (Artificial Neural Networks), an early model of Deep Learning.

II. Artificial Neural Networks(ANN)

Artificial Neural Networks are inspired by biological neural networks' structure and function [1]. They are a class of pattern matching that is commonly used for regression and classification problems. Still, They are an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types [1]. Deep Learning is one of the classically popular methods. Deep Learning methods are a modern update to Artificial Neural Networks that exploit abundant cheap computation [1]. They are concerned with building much larger and more complex neural networks. As commented above, many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data [2].

Deep Learning aims to create computer systems that can better understand complex data such as pictures, voices, and handwriting. It is closely related to neural networks and machine learning, two buzzwords in the tech industry. A neural network is a type of computer architecture consisting of network nodes that exchange messages [2]. Upon receiving signals, each node conducts an independent task and exchanges information with other node peers to achieve the same goal or goals [2]. Let's say, for instance, to tell a cat apart from a dog. Deep learning is an advanced type of machine learning with algorithms and multiple layers of network nodes [2]. It can reinforce the recognition capabilities of neural network systems through self-training, for instance, watching millions of pictures of cats and dogs. Deep Learning is a class of neural network models [3]. That is a model with an input layer, an output layer, and an arbitrary number of hidden layers [3]. These layers consist of neurons or neural units as below Figure 1.

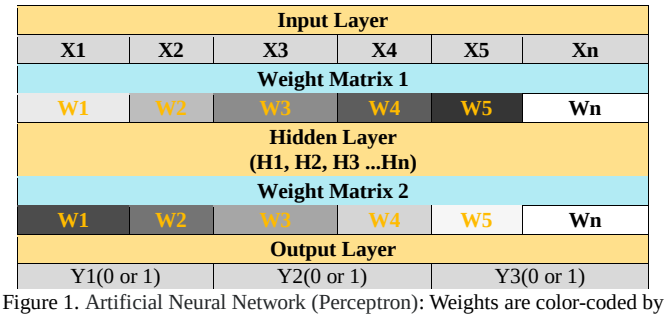

Figure 1. Artificial Neural Network (Perceptron): Weights are color-coded by sign (black +, grey -).

A neuron is a function that maps an input vector $\{x1,...,xn\}$ to a scalar output y via a weight vector $\{w1,...,wn\}$ and a nonlinear function f [4]. The function f is called the link function, which provides the nonlinearity between the input and output. The hidden layer consists of a vector of N neurons H={H1,...,Hn}[4]. Finally, there is an output layer with one neuron for every element of the output vector Y={Y1,...,Yn}[4]. Every component of the input layer is connected to every neuron in the hidden layer, with $\boldsymbol{w}$ indicating the weight associated with the connection between the input element and the hidden neuron [4]. It is needed to construct update equations for both sets of weights - the input-to-hidden layer weights $\boldsymbol{w}$ and the hidden-to-output weights $\boldsymbol{w'}$ [4]. For this works, it is necessary to compute the Gradient descent algorithm [5].

Jian Oh, "Potential of Disease Prediction using Deep Learning Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 4, pp. 283-286, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Gradient descent is an optimization algorithm used to find the parameters (coefficients) of a function *f* that minimizes an error function [5]. Using the Gradient descent optimization algorithm, the weights are updated incrementally after each epoch (pass over the training dataset) [6]. The magnitude and direction of the weight update are computed by taking a step in the opposite direction of the cost gradient [7]. The weights that connect variables in a neural network are partially analogous to parameter coefficients in a standard regression model and can describe relationships between variables [7]. The weights dictate the relative influence of information processed in the network. The weights suppress input variables that are not relevant in their correlation with a response variable [7]. The opposite effect is seen for weights assigned to explanatory variables with strong, positive associations with a response variable [7]. There are multiple algorithms that all lead to the same final weight. No algorithm can ensure proper learning of perceptron because perceptron is used to approximate output.

In perceptron learning, the weights are adjusted only when a pattern is misclassified; this corresponds to the error values' gradient descent [4]. The backpropagation algorithm was developed for training multilayer perceptron networks [4]. The idea is to train a network by propagating the output errors backward through the layers [4]. The errors serve to evaluate the error function derivatives concerning the weights, which can then be adjusted [4]. This process is known as backpropagation because it is based on the neuron's final output error [4]. This error gets propagated backward throughout the network to update the weights. The difference between neurons in hidden layers and neurons in output layers is only in the way the error is computed (if using backpropagation); For the output layers, the error is calculated directly by measuring the difference between the required output and the output the network currently yields given a sample on its input [4]. The error is computed for the hidden layers by computing the neurons' contributions in the next layer (the output layer in case you have only one hidden layer) [4]. A neural network element adds a linear combination of its input signals and applies a sigmoid function to the result(output).

Activation functions transform the weighted sum of inputs that goes into the artificial neurons [8]. These functions should be nonlinear to encode intricate patterns of the data. The most popular activation functions are Sigmoid [5]. The activation function is used to transform the activation level (weighted sum of inputs) to an output signal. The output is a specific value, A_1, if the input sum is above a certain threshold and A_0 if the input sum is below a certain threshold [5]. The values used by the Perceptron were $A_1 = 1$ and $A_0 = 0$[5].

The activation function's common choice is the sigmoid function $\boldsymbol{\sigma}$ since it takes a real-valued input (the signal strength after the sum) and squashes it to a range between 0 and 1 [5]. The sigmoid function has the mathematical form $\sigma(x)=1/(1+e-x)$, and produces a curve with an "S" shape, and is shown in below Figure 2 [5].

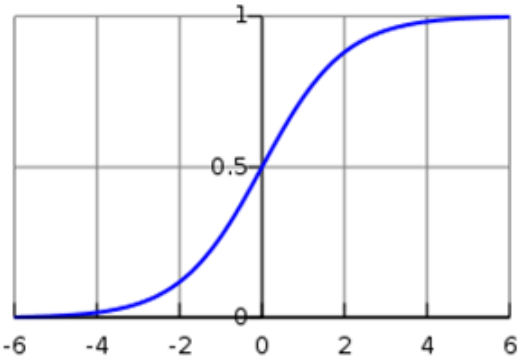

Figure 2. The sigmoid function [5]

In particular, large negative numbers become 0, and large positive numbers become 1. The sigmoid function has seen frequent use historically since it has an excellent interpretation of a neuron's firing rate: from not firing at all (0) to fully-saturated firing at an assumed maximum frequency (1) [5].

III. Application

The bio-health care market using artificial intelligence is growing significantly, centering on precision medicine, disease prediction, and pharmacogenomics [9]. In particular, the influence of precision medicine on diagnosis and treatment based on big cancer data accumulated over the past decades is increasing. Here, we will design a cancer prediction program by applying an algorithm based on Deep Learning. The process of performing machine learning begins with specifying the problem and preparing the data [9].

To know which cancers are predicted based on genetic variation, data on cancer risk factors is needed. Cancer risk factors include gene mutation, copy number variant (CNV), gene expression, DNA methylation, micro RNA, and reverse-phase protein array (RPPA) [10]. Among each factor, weights are assigned by classifying them into positive, neutral, and negative according to the degree related to carcinogenesis. Table 1 shows the risk factors and the binary weights for positive, neutral, and negative factors.

Table 1. List of cancer risk factors and Weights for each data [10]

Data	Positive Factor	Neutral Factor	Negative Factor
Gene mutation	100	010	001
Copy Number Variant	100	010	001
Gene expression	100	010	001
DNA Methylation	100	010	001
Micro RNA	100	010	001
RPPA	100	010	001

In particular, mutation, CNV, amount of gene expression, and DNA methylation of oncogenes and tumor suppressor genes in Table 2 can directly induce carcinogenesis [10]. For example, the bcl2 gene is on chromosome 18 and is an oncogene that causes B cell leukemia and lymphoma when translocated and overexpressed. As cancer genome research becomes active, new oncogenes and tumor suppressor genes are continually being revealed, so periodic updates are

Jian Oh, "Potential of Disease Prediction using Deep Learning Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 4, pp. 283-286, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

required.

Table 2. List of Oncogenes and Tumor suppressors [10]

AKT 1	BCL2	CDH1	EGFR	FGFR 1	KRA S	PDGFR B	RB1
AKT 2	BRAF	CDK4	EPHB 4	FGFR 2	NRA S	PIK3C A	SM O
AKT 3	BRCA 1	CDK6	ERBB 2	FGFR 3	HRA S	PIK3R1	TOP 1
APC	BRCA 2	CSF1R	ERBB 3	FLT3	MDM 2	PTCH1	TP5 3
ATM	BAP1	CDKN1 B	ERBB 4	FAS	MDM 4	PTCH2	HER 2

Table 3 and Table 4 are examples of the presence or absence of gene mutation and methylation at a specific location in binary format. For instance, if a point mutation is found in a gene and the rest of the gene variants are not observed, only the point mutation is marked as 1, and all others are marked as 0. Also, methylation on the CpG island of the DNA promoter silences the gene. If the gene is a tumor suppressor gene, cancer can occur. Methylation of histone protein affects the expression of downstream genes. In particular, methylation of lysine in histone H3 subunit is found at a high frequency in cancer patients.

Table 3. Patterns of Gene Aberrations [10]

Point Mutation	Insertion (CNV)	Deletion (CNV)	Duplication (CNV)	Fusion	Translocation
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
1	0	0	1	0	0

Table 4. Location of DNA Methylation [11]

DNA CpG	H3K3(Lysine 3)	H3K9(Lysine 9)	H3K29(Lysine 29)
1	0	0	0
1	1	0	0
1	0	1	0
1	0	0	1

Table 5 shows each base in a binary notation and applying it to the point mutation of the KRAS gene. KRAS Gene is an oncogene involved in cell signaling, maturation, and division. Mutations of the KRAS gene mainly occur at codons. G12A indicated in the table means that the amino acid Glycine (G) was converted to Alanine (A) due to the mutation at codon 12. In this way, we could prepare data on point mutations of other oncogenes or tumor suppressor genes.

However, after training, an overfitting problem may occur. Overfitting refers to a case where high accuracy is shown when the model is driven with a training data set, whereas low accuracy is shown in an actual test data set. It is mainly caused by high complexity caused by hidden nodes due to too many input variables or a problem in the training data set. Therefore, we should pay attention to the feature selection and data arrangement before training.

Table 5. Sequence Variations on KRAS GENE Point Mutation [12]

Nucleotide	A	C	G	T
Binary notation	00	01	10	11

Wild Type : **-GGTGGC-**	[1] Mutant Type : (G12A) -GCTGGC	[2] Mutant Type : (G13D) -GGTGAC-
-101011101001-	**-100111101001-**	**-101011100001-**
[3] Mutant Type : (G12S) -AGTGGC-	[4] Mutant Type: (G12V) -GTTGGC-	[5] Mutant Type: (G13A) -GGTGCC-
-001011101001-	**-101111101001-**	**-101011100101-**
[6] Mutant Type : (G12R) -CGTGGC-	[7] Mutant Type : (G13S) -GGTAGC-	[8] Mutant Type : (G13V) -GGTGTC-
-011011101001-	**-101011001001-**	**-101011101101-**
[9] Mutant Type : (G12C) -TGTGGC-	[10] Mutant Type : (G13R) -GGTCGC-	[11] Mutant Type : (G12S2) -TCTGGC-
-111011101001-	**-101011011001-**	**-110111101001-**

Table 6 is an example of an algorithm implemented for cancer prediction.

A threshold is determined from big data on the Effect of mutations in the gene on carcinogenesis. When the gene mutation, methylation, and CNV gene mutation found in the genetic test result of cancer patients are input through the input layer, the corresponding weight is given, and when the threshold is exceeded, it is transferred to the hidden layer (1). Suppose the gene delivered to the hidden layer (1) is a driver gene that directly affects carcinogenesis by referring to the big data of genetic cancer diagnosis. In that case, appropriate weight is assigned to it, and when this value exceeds the threshold, it is transferred to the hidden layer (2).

The oncogene and tumor suppressor genes in Table 2 are driver genes. If a passenger gene does not directly affect carcinogenesis, it is stopped here without moving to the next layer. In the hidden layer (2), if the probability that the abnormality of the gene transmitted from the hidden layer (1) leads to actual cancer is more than 50%, cancer caused by the gene mutation is specified and output through the output layer. Subsequently, from the cancer diagnosis of big data, the average value of the initial diagnosis ages of patients who have developed cancer due to a specific gene mutation is calculated. We can then inform the probability of developing cancer, considering the genetic testing sponsor's age. The volume of data on cancer genes and cancer patients is increasing rapidly over time. As data accumulates, we will discover a new driver gene, and the passenger gene will be converted into a driver gene. Therefore, when diagnosing genes, testing as many gene mutations as possible will help the sponsor's healthcare and improve algorithm-based cancer prediction accuracy.

Jian Oh, "Potential of Disease Prediction using Deep Learning Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 4, pp. 283-286, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Table 6. Algorithm Matrix Example for Cancer Prediction.

Input Layer	TH (0.8)		Hidden Layer(1)	TH (0.5)		Hidden Layer(2)	TH (0.5)		Out Layer
Gene -1	0.2								Colon
Gene -2	0.5		Gene 4	0.7					
Gene -3	0.4					Gene 4	0.4		
Gene -4	0.8								
Gene -5	0.9								Breast
Gene -6	0.4		Gene 5	0.3					
Gene -7	0.1								
Gene -8	0.6								
Gene -9	0.6								Lung
Gene-10	0.6								
Gene-11	0.5								
Gene-12	0.7								
Gene-13	0.7								Gastric
Gene-14	0.2		Gene 15	0.6					
Gene-15	0.8					Gene 15	0.6		
Gene-16	0.5								
Gene-17	0.7								Pancreatic
Gene-18	0.3		Gene 20	0.4					
Gene-19	0.4								
Gene-20	0.9								

TH : Threshold

IV. Conclusion

I tried a cancer prediction program using an ANN-based algorithm system. Weight calculation is a pivotal variable to ensure the accuracy of prediction. In the future, as big data on genetic mutations of cancer patients accumulate, the prediction accuracy will increase. Even with the same gene mutation, if it is a driver gene for lung cancer and a passenger gene for colon cancer, a weight is given to the prediction of lung cancer, and no weight is given to the prediction of colon cancer. It needs to develop an advanced algorithm model that can more accurately calculate weights.

Acknowledgments

He has answered any questions in detail and kindly in the course of this research. Thanks to Dr. James Oh, American Institute of Standards.

References

[1] Robert E. Schapire. The boosting approach to machine learning: An overview. In Nonlinear Estimation and Classification, Springer, 2003.
[2] Rebecca Fiebrink and Baptiste Caramiaux, The Machine Learning Algorithm as Creative Musical Tool, Oxford University Press, 2016.
[3] Jenna Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society, January–June 2016 : 1–12.
[4] Vladislav Skorpil and Jiri Stastny, Neural Networks And Back Propagation Algorithm, Electronics' 2006
[5] Joe Kilian and Hava T. Siegelmann, On the Power of Sigmoid Neural Networks, January 1993, Conference: Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, COLT 1993, Santa Cruz, CA, USA, July 26-28, 1993, DOI: 10.1145/168304.168321
[6] Jasper Snoek et al., Practical Bayesian Optimization of Machine Learning Algorithms, Supplimentary material, 2012.
[7] Matthew J. Maenner et al., Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder, PLoS ONE, 2016, 11(12): e0168224. doi:10.1371/journal.pone.0168224.
[8] R. Ryan Williams, Algorithms and Resource Requirements for Fundamental Problems, CMU-CS-07-147, 2007.
[9] Adnan Darwiche and Knot Pipatsrisawat, Complete Algorithms, Armin Biere, Marijn Heule, Hans van Maaren and Toby Walsh IOS Press, 2008.
[10] [10. Stefan Harmeling, Solving Satisfiability Problems with Genetic Algorithms, Stanford University, 2000.
[11] Junjeong Choi et al., CpG Island Methylation According to the Histologic Patterns of Early Gastric Adenocarcinoma, The Korean Journal of Pathology 2011; 45: 469-476 http://dx.doi.org/10.4132/KoreanJPathol.2011.45.5.469
[12] Sanford Burnham Prebys Medical Discovery Institute. "Conquering cancer's infamous KRAS mutation." ScienceDaily. ScienceDaily, 14 May 2019. <www.sciencedaily.com/releases/2019/05/190514081557.htm>.