ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 22-03-23
162
j.inst.Korean.electr.electron.eng.Vol.26,No.3,494~499,September 2022
https://doi.org/10.7471/ikeee.2022.26.3.494

피부섬유모세포 전사체 정보를 활용한 구간 선택 기반 연령 예측

Age Prediction based on the Transcriptome of Human Dermal Fibroblasts through Interval Selection

석 호 식*★

Ho-Sik Seok*★

Abstract

It is reported that genome-wide RNA-seq profiles has potential as biomarkers of aging. A number of researches achieved promising prediction performance based on gene expression profiles. We develop an age prediction method based on the transcriptome of human dermal fibroblasts by selecting a proper age interval. The proposed method executes multiple rules in a sequential manner and a rule utilizes a classifier and a regression model to determine whether a given test sample belongs to the target age interval of the rule. If a given test sample satisfies the selection condition of a rule, age is predicted from the associated target age interval. Our method predicts age to a mean absolute error of 5.7 years. Our method outperforms prior best performance of mean absolute error of 7.7 years achieved by an ensemble based prediction method. We observe that it is possible to predict age based on genome-wide RNA-seq profiles but prediction performance is not stable but varying with age.

요 약

본 논문에서는 인간의 피부섬유모세포(Human dermal fibroblasts)로부터 확보한 전사체 정보를 활용하여 나이를 예측하는 방법을 소개한다. 제안 방법에서는 훈련을 통해 확보한 분류기 및 회귀 모델을 이용하여 샘플이 속한 적합한 연령 그룹을 선택한 후, 선택된 연령 그룹에 속하는 훈련 데이터의 관측값을 활용하여 구체적인 연령을 예측한다. 연령을 예측하려는 샘플이 입력되면 복수개의 판별 규칙이 순서대로 실행되는데, 개별 판별 규칙에서는 분류기와 회귀 모델을 동시에 실행하여 해당 판별 규칙에 대한 선택 조건이 만족되는지 여부를 확인한다. 선택 조건이 만족될 경우 판별 규칙의 타겟 연령 그룹에 속하는 데이터를 이용하여 훈련된 회귀 모델로 연령을 예측하며, 선택 조건이 만족되지 않으면 후속 판별 규칙을 실행한다. 공개 데이터에 대하여 실험한 결과 기존 연구에서 달성한 7.7년의 평균 예측 오차보다 우수한 5.7년이라는 평균 예측 오차를 달성함을 확인하였다.

Key words : age prediction, genotype-phenotype association, transcriptome, dermal fibroblast, machine learning

Ⅰ. 서론

측정 기술의 발전과 함께 다양한 표현형에 대한 바이오마커를 발견하고, 유전 정보 및 표현형의 관계를 이해하기 위한 연구에 관심이 집중되고 있다[1]-[3]. 최근 인공지능/기계학습 기술들이 고도화되면서 기계학습을 이용하여 유전정보와 표현형의 관계를 탐색하려는 시도들이 있었는데[4]-[6], Fleischer의 그룹에서는 인간의 피부섬유모세포(Human dermal fibroblasts)로부터 확보

* Dept. of Artificial Intelligence and Data Science, Korea Military Academy
★ Corresponding author
E-mail : hosik.seok@gmail.com, Tel : +82-2-2197-2873
Manuscript received Sep. 20, 2022; revised Sep. 23, 2022; accepted Sep. 25, 2022.

한 RNA-seq 데이터 셋에 분류기의 앙상블을 적용하여 나이를 예측하는 방법을 소개하였다[7]. 해당 연구가 발표되기 전부터 바이오마커를 활용하여 나이를 예측하려는 연구가 활발하게 진행되어 왔으나[8], [9], 이전 연구와 비교했을 때 Fleischer 그룹의 연구는 (1) 피부섬유모세포의 유전자 발현 데이터를 활용하였고 (2) 상대적으로 적은 숫자의 샘플을 활용하였다는 점에서 많은 흥미를 끌었다. 피부섬유모세포는 지속적으로 발생하는 손상이 축적되고 연령과 관계된 변화를 보인다는 점에서 노화에 대한 바이오마커로써 잠재력을 가지고 있음이 보고되어왔는데[10], [11], Fleischer 그룹의 연구를 통해 피부섬유모세포를 바이오마커로 활용하여 나이를 예측하는 것이 가능하다는 것을 알 수 있었다.

본 논문에서는 피부섬유모세포에서 확보한 전사체(transcriptome) 정보에 기반하여 나이를 예측하는 방법을 소개한다. 제안 방법은 분류기와 회귀 모델을 결합하여 샘플의 연령을 예측하는 방법을 활용한다. 제안 방법에서는 먼저 입력된 샘플에 해당하는 연령 구간을 추정한 후, 추정된 연령 구간의 데이터에 대한 회귀 분석을 통해 입력된 샘플의 연령을 추정한다. 연령 구간의 추정은 훈련 데이터에서 추출된 21개의 규칙을 활용한다. 21개의 규칙을 통해 샘플에 해당하는 연령 구간이 추정되며 동시에 선택된 연령 구간에 해당하는 훈련 데이터들이 추려진다. 적절한 규칙의 선택을 위하여 각 규칙의 선택 기준이 명시되는데, 분류기 및 회귀 모델의 실행 결과를 활용하여 적절한 연령 구간이 선택된다. 공개 데이터에 대한 실험을 통해 제안 방법이 기존 방법을 능가하는 예측 성능을 보유함을 확인할 수 있었으며 피부섬유모세포의 전사체를 활용하여 연령을 예측하는 것이 가능함을 확인하였다.

II. 제안 방법

2.1. 실험 데이터 및 예측 성능 측정

본 논문에서는 Fleischer 그룹이 생성한 데이터를 활용한다. Fleischer 그룹은 피부섬유모세포로부터 RNA-seq 데이터셋을 확보하였으며 해당 데이터셋은 Gene Expression Omnibus(접근번호 : GSE113957)에서 획득할 수 있다[7]. 해당 데이터셋은 133명을 대상으로 27,142개의 유전자 발현 정보를 측정하여 생성된 것이며, 1~9세 연령대에서 12개, 10~19세 연령대에서 14개, 20~29세 연령대에서 17개, 30~39세 연령대에서 14개, 40~49세 연령대에서 14개, 50~59세 연령대에서 6개, 60~69세 연령대에서 19개, 70~79세 연령대에서 4개, 80~89세 연령대에서 26개, 90세 이상 연령대에서 7개의 샘플이 측정되었다. 예측 성능은 LOO(Leave-one-out) 교차 검증을 통해 측정되었다. [7]에서 설명된 절차와 동일한 절차를 통해 데이터를 전처리하였는데, 유전자 발현값 차이 정도(5배 이상 차이) 및 5 FPKM(Fragments per killobase of transcript per million)보다 큰 발현값을 갖는 샘플의 존재 여부를 활용하여 훈련 과정에서 고려될 유전자 집합을 선택하였다(자세한 데이터전처리 절차는 [7]의 설명 참고). 전처리 결과 최소 4,755개, 최대 4,861개, 평균 4,852개의 유전자 발현 정보가 예측 모델 구축에 활용되었다.

본 연구에서 우리는 구간 선택을 통해 예측을 수행하고자 하였다. 주어진 샘플이 속하는 적절한 연령 구간을 먼저 선택한 후 연령 예측을 수행하여 예측 성능을 높이려고 계획하였으며, 예측 성능은 각 교차 검증 폴드에 해당하는 샘플의 실제 연령(y_i)과 예측 연령($\hat{y}_i$)의 차이의 절대값에 기반하여 측정되었다.

2.2. 제안방법

주어진 샘플에 대응하는 연령은 해당 샘플에 대한 연령 구간을 먼저 선택하여 예측된다. 연령 구간은 복수 개의 규칙을 순차적으로 확인하여 결정되는데, 개별 규칙은 타겟 연령 구간(T), 분류 연령 구간(C), 회귀 연령 구간(R), 선택 조건으로 구성된다. 타겟 연령 구간(T)은 주어진 샘플의 연령을 예측하기 위하여 활용되는 연령 구간으로, T에 해당하는 연령의 훈련 데이터로부터 회귀 모델을 획득한 후 훈련된 회귀 모델을 이용하여 연령을 예측한다. T가 주어진 샘플의 연령을 예측하기에 적합한 연령 구간인지 여부를 결정하기 위하여 분류기 및 회귀 모델이 활용된다. 분류기는 분류 연령 구간(C)를 이용하여 훈련되는데, C는 여러 개의 연령 구간으로 구성되어 개별 T에 대하여 고유한 분류기가 훈련된다. 회귀모델은 회귀 연령 구간(R)에 속하는 훈련 데이터에 기반하여 획득되며, 테이블 1의 선택 조건에서 “regV”가 회귀 모델의 출력값을 의미한다. 분류기의 분류 결과와 회귀 모델의 출력값(regV)를 활용하여 타겟 연령 구간(T)이 주어진 샘플에 적절한 연령 구간인지 여부를 판별한다.

제안 모델에서는 21개의 연령 구간을 활용하며 개별 연령 구간의 구체적인 선택 규칙은 테이블 1의 R_A(규칙 A)에서 R_U(규칙 U)로 표시되어 있다. 연령을 알 수 없는 샘플이 주어질 경우 R_A에서부터 R_B, R_C의 순서로 R_U까지 차례대로 규칙이 확인된다. 만약 해당 규칙의 선택 조

Table 1. A set of rules for age prediction.
표 1. 연령 예측 규칙 집합

	Target age interval (T)	Classification age intervals (C)	Regression age interval (R)		Target age interval (T)	Classification age intervals (C)	Regression age interval (R)
	Selection condition				Selection condition		
R_A	[1,3]	[1,3],[11,81]	[1, 96]	R_B	[4,8]	[2,8],[13,30]	[4,96]
	Its classifier selects C_0 and $regV \leq 6$				Its classifier selects C_0 and $regV \leq 24.65$		
R_C	[9,11]	[1,11], [13,96]	[9,96]	R_D	[86,96]	[86,96],[6,80]	[12,96]
	Its classifier selects C_0 and $regV \geq 23$				Its classifier selects C_0 and $regV \geq 62$		
R_E	[84,85]	[84,96],[12,70],[71,80]	[12,85]	R_F	[12,18]	[6,18],[51,65]	[12,85]
	Its classifier selects C_0 and $regV \geq 45$				Its classifier selects C_0 and $regV \leq 31$		
R_G	[19,20]	[10,20], [31,40], [51,61], [71,79]	[19,83]	R_H	[21,22]	[12,22],[41,50],[51,54],[61,64], [65,68],[69,75],[77,83]	[21,83]
	Its classifier selects C_0 and $regV \leq 42.2$				Its classifier selects C_0 and $50.4 \leq regV \leq 53.9$		
R_I	[23,24]	[19,24],[41,50],[51,54], [61,64],[65,68],[69,75],[77,83]	[23,83]	R_J	[25,26]	[21,26],[41,50],[51,54], [61,64],[65,68],[69,75],[77,83]	[25,83]
	Its classifier selects C_0 and $regV \leq 32.5$				Its classifier selects C_0 and $43.8 \leq regV \leq 44.8$		
R_K	[27,30]	[25,30],[86,96],[1,20]	[27,83]	R_L	[82,83]	[82,90],[11,16],[64,68]	[31,83]
	Its classifier selects C_0 and $regV \leq 55.3$				Its classifier selects C_0 and $regV \geq 70.0$		
R_M	[80,82]	[80,86],[91,96], [26,29],[41,45],[47,56]	[31, 83]	R_N	[78, 79]	[17,21],[27,30],[31,35], [37,40],[46,49],[51,56], [57,60],[63,69],[70,75]	[31, 96]
	Its classifier selects C_0 and $44.5 \leq regV \leq 46.2$				Its classifier selects C_0 and $67.9 \leq regV \leq 69$		
R_O	[25,31]	[25,31],[1,11],[46,50], [51,56],[57,60],[61,65]	[1, 96]	R_P	[32, 34]	[24,34],[1,11],[46,50], [51,56],[57,61],[65,69]	[1,96]
	Its classifier selects C_0 and $63 \leq regV \leq 63.5$				Its classifier selects C_0 and $10 \leq regV \leq 34.6$		
R_Q	[75,78]	[75,85],[1,11],[86,96]	[1,96]	R_R	[68,75]	[68,85],[1,11], [30,40],[41,48]	[1,96]
	Its classifier selects C_0 and $45 \leq regV \leq 47$				Its classifier selects C_0 and $regV \geq 63.7$		
R_S	[35,40]	[21,40],[1,11],[68, 85]	[1,96]	R_T	[60,67]	[60,85],[1,12],[21,40], [41,50],[51,54]	[1,96]
	Its classifier selects C_0 and $10 \leq regV \leq 49$				Its classifier selects C_0 and $50.85 \leq regV \leq 62.2$		
R_U	[40, 59]						

건이 만족되면 해당 규칙의 타겟 연령 구간(T)를 활용하여 주어진 샘플의 연령을 예측한다. 만약 주어진 샘플이 현재 규칙의 선택 조건을 만족하지 못하면 후속 규칙을 적용한다. R_A의 경우 타겟 연령 구간은 [1, 3]으로 1세부터 3세를 타겟 연령으로 한다. 분류기는 [1, 3] 연령 구간(클래스 C_0)과 [11, 81] 연령 구간(클래스 C_1)에 속하는 훈련 데이터를 이용하여 훈련되며 회귀 모델은 1세부터 96세에 속하는 훈련 데이터를 이용하여 훈련된다. R_A의 선택 조건은 분류기가 주어진 샘플이 C_0, 즉 [1, 3] 연령 구간에 속한다고 분류하며 회귀 모델의 출력값(regV)이 6보다 작거나 같은 경우로, 두 조건이 모두 만족되면 R_A의 타겟 연령 구간인 1세부터 3세에 속하는 훈련 데이터를 활용하여 획득된 회귀 모델로 연령을 예측한다. 만약 R_A의 선택 조건이 만족되지 않으면 R_B가 적용된다. R_B의 타겟 연령 구간은 [4, 8]이며 분류기는 [2, 8] (클래스 C_0), [13, 30] (클래스 C_1)에 속하는 훈련 데이터로 학습된다. 회귀 모델은 [4, 96] 연령 구간에 속하는 훈련 데이터로 획득되며, R_B의 선택 조건은 분류기가 주어진 입력의 소속 클래스로 C_0를 선택하고 회귀 모델의 출력 값이 24.65 이하인 경우이다. 이와 같은 절차가 R_A부터 차례대로 실행되는데, 만약 R_T의 선택 조건마저 만족하지 못하는 경우 R_U가 적용되어 연령을 예측한다.

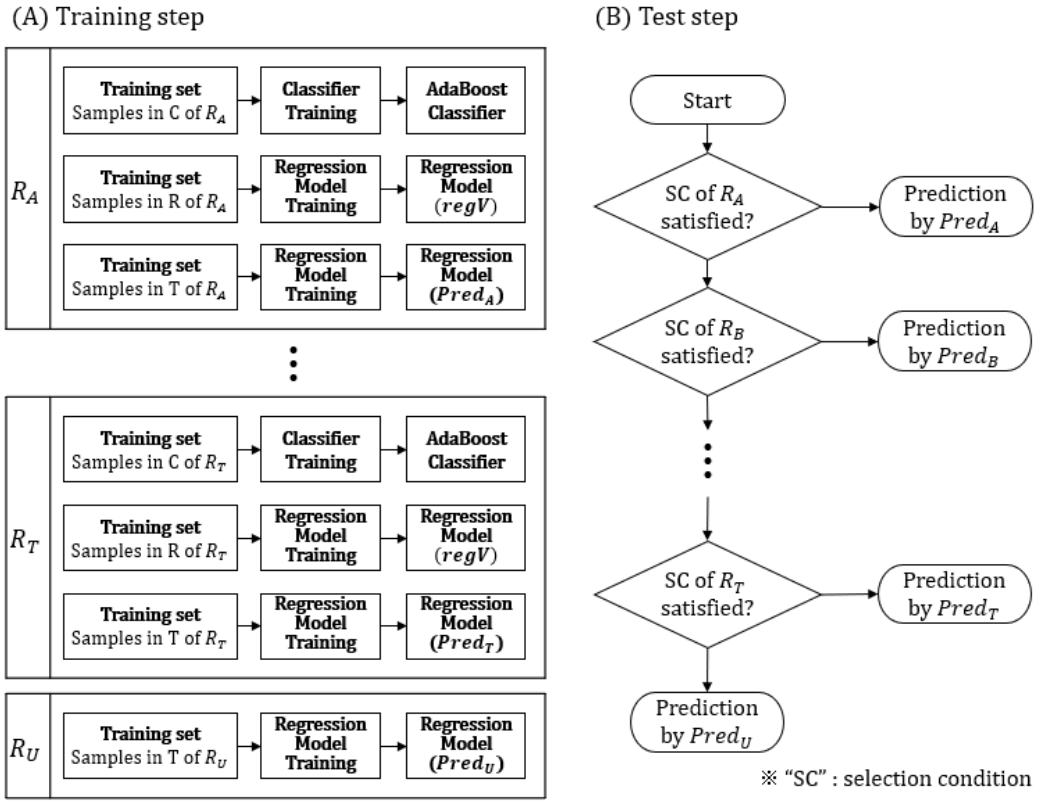

Fig. 1. A schematic of training and test procedure.
그림 1. 훈련 및 테스트 절차

테이블 1의 규칙을 살펴보면 분류 연령 구간을 구성하는 연령 구간이 세 개 이상인 규칙들이 존재하는데, 이 때는 테이블 1의 분류 연령 구간을 구성하는 구간들에 왼쪽부터 C_0, C_1, C_2...의 순서로 클래스 레이블을 부여하며 선택 조건에서는 C_0가 선택되었는지의 여부를 확인하게 된다. 또한 타겟 연령 구간에 속하는 훈련 데이터들을 이용하여 훈련된 회귀 모델의 예측 값이 T의 최소값보다 작거나 T의 최대값보다 클 경우 T의 최소값 혹은 T의 최대값을 예측 결과로 선택한다(그림 1).

제안 방법에서 분류기는 AdaBoost 분류기(estimator의 갯수: 3)[12]가 활용되었으며 차수가 6인 다항 커널(Polynomial kernel)에 기반한 커널 릿지 리그레션 모델(Kernel ridge regression model)이 회귀 모델로 사용되었다[13]. 분류기와 회귀 모델 모두 scikit-learn 라이브러리[14]에서 제공하는 모델을 활용하였다.

Ⅲ. 실험 결과 및 토의

133개의 샘플로 구성된 데이터셋에 대하여 LOO 교차검증을 통해 예측 성능을 확인한 결과 제안 모델은 5.7년의 예측 오류를 달성하였다. 동일한 데이터셋에 대하여 기존 예측 결과[7]를 비교하여 보면 제안 모델이 기존 접근법을 능가하는 예측 성능을 달성하였음을 알 수 있다(표 2).

제안 방법의 예측 성능을 보다 자세히 알아보기 위하여 타겟 연령 구간 별로 예측 성능을 분석한 결과, 연령 구간 별로 예측 성능에 큰 변화가 발생함을 확인할 수 있었다(표 3). 예를 들어 타겟 연령 구간 [9, 11]과 [84, 85]의 경우 각각 0.9와 0.6의 MAE를 달성하였다. 타겟 연령 구간 [1, 3], [21, 22], [23, 24], [25, 26], [27, 30], [82, 83]에 대해서도 MAE 1.1, 1.5, 1.6, 1.8, 1.6, 1.0이라는 양호한 예측 성능을 달성하였다. 그러나 [35, 40], [60, 67], [75, 78]의 타겟 연령 구간에 대해서는 MAE 12.0, 10.3, 9.3로 예측 성능이 악화되었음을 확인하였다. 표 3의 예측 성능을 보면 1세부터 11세까지의 연령층과 80세부터 96세까지의 연령층에 대해서는 매우 우수한 예측 성능(전자의 MAE = 1.7, 후자의 MAE = 1.9)을 달성하였음을 확인할 수 있으나 12세부터 79세의 연령 그룹 중 일부 연령 그룹에 대해서는 개선의 여지가 존재함을 알 수 있다.

12세부터 79세까지의 연령 그룹에서 발생한 예측 성능 저하는 해당 그룹의 샘플에서 확보한 유전자 발현 정

Table 2. Prediction performance.
표 2. 예측 성능

Method	Parameters	Mean absolute error	Ref.
Our method	AdaBoost classifier (Number of estimator = 3) Kernel ridge regression (Polynomial kernel, degree = 6)	5.7	-
LDA ensemble	Age bin width = 20	7.7	[7]
Gaussian näive Bayes ensemble	Age bin width = 30	15.7	
k-nearest neighbors ensemble	Age bin width = 20	19.7	
Random forest ensemble	Age bin width = 20	11.8	
Linear regression	-	12.1	
Elastic net regression	Alpha = 0.1 L1/L2 = 0	12.0	
Support vector regression	Second order polynomial kernel	11.9	

Table 3. Prediction performance per target age interval.
표 3. 타겟 연령 구간별 예측 성능

Target age interval (C)	Prediction error	Target age interval (C)	Prediction error	Target age interval (C)	Prediction error
[1, 3]	1.1	[4, 8]	2.4	[9, 11]	0.9
[12, 18]	6.1	[19, 20]	10.5	[21, 22]	1.5
[23, 24]	1.6	[25, 26]	1.8	[27, 30]	1.6
[30, 31]	2.4	[32, 34]	2.4	[35, 40]	12.0
[40, 59]	6.4	[60, 67]	10.3	[68, 75]	2.9
[75, 78]	9.3	[78, 79]	5.0	[80, 82]	2.0
[82, 83]	1.0	[84, 85]	0.6	[86, 96]	2.5

보에 큰 변화가 존재하기 때문인것으로 추정된다. 그러나 주어진 데이터 집합의 크기와 복잡도(데이터 전처리 후 평균 4,850개의 유전자 발현 값을 133명을 대상으로 측정, 1세부터 96세의 연령 그룹을 대상으로 예측 시도)를 고려했을 때, 연령 예측 매개체로서의 피부섬유모세포의 잠재력을 단정하기는 어렵다.

피부섬유모세포를 활용한 나이 예측의 선행연구[7]에서는 LDA(Linear discriminant analysis) 분류기의 앙상블을 이용하여 전사체 정보에 기반한 나이 예측 방법을 소개하였다. 본 논문에서는 분류기와 회구 모델의 결합을 통해 예측 성능을 향상시킴으로써 모델링을 통한 예측 성능 여지가 존재함을 확인하였다.

Ⅳ. 결론

피부섬유모세포의 유전자 발현값에 기반하여 연령을 예측할 수 있도록 먼저 적합한 연령 구간을 선택한 후 회귀 모델을 적용하는 방법을 소개하였다. 예측 성능을 향상시키기 위하여 다양한 타겟 연령 구간에 대하여 연령을 예측하는 방법을 개발하였으며, 분류기와 회귀 모델의 조합을 통해 이미 소개된 연구 결과를 능가하는 예측 성능을 달성하였다. 실험 결과 피부섬유모세포의 전사체 정보를 활용하여 연령을 예측하는 것이 어느 정도 가능하나, 연령 예측을 위한 피부섬유모세포의 잠재력을 결정하기 위해서는 예측 성능 차이를 발생시키는 정확한 원인에 대한 추가 연구가 필요함을 확인할 수 있었다. 특히 본 연구에서 분석한 피부섬유모세포와 같은 문제들은 관측된 샘플의 수는 상대적으로 적지만, 해당 도메인을 여러 생체 현상 관점에서 관측한 지식이 많이 축적된 문제이다. 이런 문제들은 예측 모델 구축뿐만 아니라 다양한 소스에서 관측된 데이터의 융합 차원에서 후속 연구를 진행할 필요가 있다.

References

[1] P. Benfey and T. Mitchell-Olds, “From Genotype

to Phenotype: Systems Biology meets Natural Variation," *Science*, vol.320, pp.495-497, 2008. DOI: 10.1126/science.1153716
[2] A. Drouin, et al., "Predictive Computational Phenotyping and Biomarker Discovery using Reference-free Genome Comparison," *BMC Genom.*, vol.17, pp.754, 2016.
[3] A. Young, et al, "Deconstructing the Sources of Genotype-phenotype Associations in Humans," *Science*, vol.365, pp.1396-1400, 2019. DOI: 10.1126/science.aax3710
[4] A. Drouin, et al., "Interpretable Genotype-to-phenotype Classifiers with Performance Guarantees," *Sci. Rep.*, vol.9, p.4071, 2019.
[5] A. Smith, et al., "Standard Machine Learning Approaches outperform Deep Representation Learning on Phenotype Prediction from "Transcriptomics Data," *BMC Bioinform.*, vol.21, pp.119, 2020. DOI: 10.1186/s12859-020-3427-8
[6] Z. Tang, et al., "Deep Learning of Imaging Phenotype and Genotype for Predicting Overall Survival Time of Glioblastoma Patients," *IEEE Trans. Med. Imaging*, vol.39, pp.2100-2109, 2020. DOI: 10.1109/TMI.2020.2964310
[7] J. Fleischer,et al., "Predicting Age from The Transcriptome of Human Dermal Fibroblasts," *Genome Biol.*, vol.19, p.221, 2018. DOI: 10.1186/s13059-018-1599-6
[8] S. Horvath, "DNA Methylation Age of Human Tissues and Cell Types," *Genome Biol.*, vol.14, p.3156, 2013. DOI: 10.1186/gb-2013-14-10-r115
[9] R. Zbieć-Piekarska, et al., "Development of a Forensically Useful Age Prediction Method based on DNA Methylation Analysis," *Forensic Sci. Int. Genet.*, vol.17, pp.173-179, 2015. DOI: 10.1016/j.fsigen.2015.05.001.
[10] J. Tigges, et al, "The Hallmarks of Fibroblast Ageing," *Mech. Ageing Dev.*, vol.138, pp.26-44, 2014. DOI: 10.1016/j.mad.2014.03.004.
[11] J. Phillip, et al., "Biophysical and Biomolecular Determination of Cellular Age in Humans," *Nat. Biomed. Eng*, vol.1, 2017.
[12] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, pp.119-139, 1997, DOI: 10.1006/jcss.1997.1504
[13] K. P. Murphy, *Machine learning: a probabilistic perspective*, The MIT Press, Cambridge, pp.492-493, 2012.
[14] F. Pedregosa, et al., "Scikit-Learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol.12, pp.2825-2830, 2011.

BIOGRAPHY

Ho-Sik Seok (Member)

1999 : BS degree in Computer Engineering, Seoul National University.
2001 : MS degree in Electrical Engineering and Computer Science, Seoul National University.
2012 : PhD degree in Electrical Engineering and Computer Science, Seoul National University.
2016~2020.2 : Assistant professor, Dept. of Computer and Communications Engineering, Kangwon National University.
2020.3~2022.1 : Assistant professor, Dept. of Computer Science and Engineering, Kangwon National University.
2022.2~present : Assistant professor, Dept. of Artificial Intelligence and Data Science, Korea Military Academy.