(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년12월08일
(11) 등록번호 10-2336311
(24) 등록일자 2021년12월02일

(51) 국제특허분류(Int. Cl.)
G16H 50/50 (2018.01) *G16H 10/60* (2018.01)
G16H 20/00 (2018.01) *G16H 50/30* (2018.01)
(52) CPC특허분류
G16H 50/50 (2018.01)
G16H 10/60 (2021.08)
(21) 출원번호 10-2019-0146627
(22) 출원일자 2019년11월15일
심사청구일자 2019년11월15일
(65) 공개번호 10-2021-0059325
(43) 공개일자 2021년05월25일
(56) 선행기술조사문헌
KR1020160072842 A*
KR1020190021471 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
한국과학기술원
대전광역시 유성구 대학로 291(구성동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
최정균
대전광역시 유성구 대학로 291(구성동)
배민균
대전광역시 유성구 대학로 291(구성동)
(뒷면에 계속)
(74) 대리인
특허법인다나

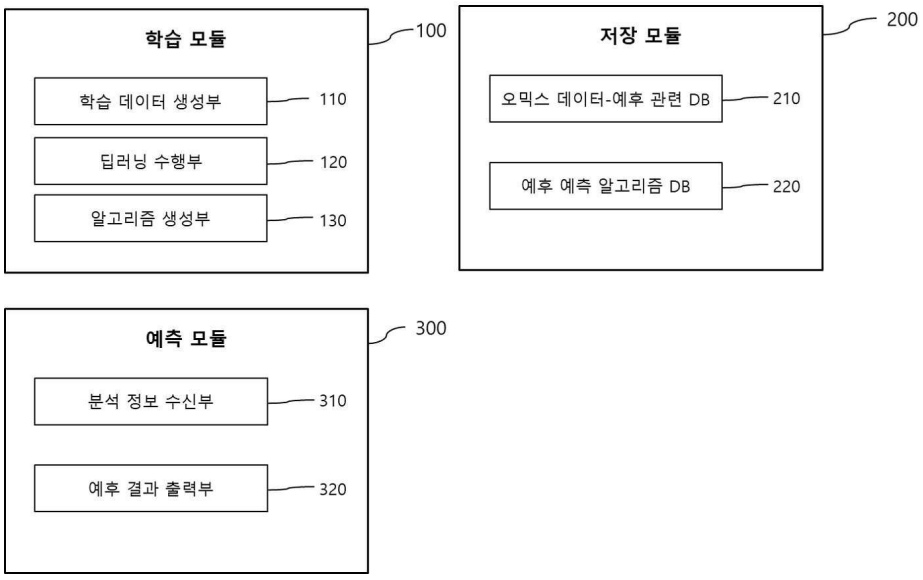
전체 청구항 수 : 총 6 항 심사관 : 백양규

(54) 발명의 명칭 **딥러닝을 이용한 암의 예후 예측 모델**

(57) 요 약

본 발명은 암 환자의 오믹스 데이터로부터 암의 예후를 예측할 수 있는 예후 예측 시스템에 관한 것으로, 상기 시스템에 따르면 복수의 오믹스 데이터를 사용함으로써 분자생물학적 다양성을 통합적으로 고려하여 암 환자의 예후를 예측할 수 있고, 단일 유전자가 암의 예후 예측에 미치는 중요도를 판단할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G16H 20/00 (2021.08)
G16H 50/30 (2018.01)
(72) 발명자
김영준
서울특별시 강남구 논현로160길 31, 202호 (신사동)
김다원
서울특별시 강남구 헌릉로590길 100, 106동 701호 (세곡동, 세곡리엔파크1단지)
최은지
서울특별시 강남구 언주로 105, 200동 407호 (개포동, 현대2차아파트)
이정우
서울특별시 서대문구 연세로 50, 연세대학교 첨단관 112호 (신촌동)

이 발명을 지원한 국가연구개발사업
과제고유번호 2019001617
부처명 과학기술정보통신부
과제관리(전문)기관명 연세대학교 산학협력단
연구사업명 원천기술개발사업
연구과제명 통합유전체 마커의 확장성과 정확도 향상을 위한 noncoding 분석과 딥러닝 방법 개발(2019)
기 여 율 1/2
과제수행기관명 한국과학기술원
연구기간 2019.01.01 ~ 2019.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711081659
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 바이오의료기술개발사업
연구과제명 대장암 특이적 정밀진단 마커 개발을 위한 다중유전체 데이터 연계 확장 분석(3/3, 1단계/총괄: 대장암 다중유전체 통합분석 기반의 정밀의료 원천기술 개발)
기 여 율 1/2
과제수행기관명 연세대학교 산학협력단
연구기간 2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

(1) 암 환자로부터 수집된 3개 이상의 오믹스 데이터와 암의 예후와의 상관관계를 학습하는 학습 모듈;

(2) 상기 학습 모듈에서 생성된 암의 예후 예측 알고리즘을 저장하는 저장 모듈; 및

(3) 분석 정보를 수신하고, 상기 저장 모듈에 저장된 암의 예후 예측 알고리즘을 사용하여 예후를 모르는 암 환자의 예후를 산출하는 예측 모듈;을 포함하고,

상기 학습 모듈은

(a) 예후를 알고 있는 암 환자의 오믹스 데이터를 개별 유전자에 대해 매트릭스 형태로 정렬하여 딥러닝용 학습 데이터를 생성하는 학습 데이터 생성부;

(b) 생성된 학습 데이터로부터 오믹스 데이터와 암의 예후와의 상관 관계를 학습하는 딥러닝 수행부; 및

(c) 상기 딥러닝 수행부의 결과로부터 암의 예후를 예측하는 알고리즘을 생성하는 알고리즘 생성부;를 포함하는 것인, 암의 예후 예측용 시스템.

청구항 2

제1항에 있어서, 상기 암의 예후는 암 치료 후 생존 기간인 것인, 암의 예후 예측용 시스템.

청구항 3

제1항에 있어서, 상기 오믹스 데이터는 유전체, 전사체, 단백체, 대사체, 후성유전체, 지질체 및 메틸체로 이루어진 군에서 선택되는 것인, 암의 예후 예측용 시스템.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

제1항에 있어서, 상기 딥러닝 수행부는

(b-1) 정렬된 오믹스 데이터로부터 개별 유전자에 대한 오믹스 데이터의 공통 특징을 학습하는 단계;

(b-2) 학습된 개별 유전자에 대한 오믹스 데이터의 공통 특징에서 중요하지 않은 정보를 제거하는 단계;

(b-3) 상기 (b-2)의 결과로부터 개별 유전자당 가장 높은 값을 암의 예후 예측을 위한 대푯값으로 선별하는 단계;

(b-4) 상기 (b-3)에서 선별한 개별 유전자당 대푯값을 조합하여 암의 예후 예측의 정확도를 확인하는 단계; 및

(b-5) 상기 (b-4)의 결과에서 정확도가 가장 높은 대푯값의 조합을 암의 예후 예측용 마커로 선별하는 단계;를 포함하는 암의 예후 예측용 시스템.

청구항 8

제7항에 있어서, 상기 (b-4) 단계는 콕스 비례 위험 모델(Cox proportional hazards model)을 사용하여 개별 유전자당 대푯값과 암 환자의 예후와의 음의 로그 부분 유도(negative log partial likelihood)를 계산하는 것인, 암의 예후 예측용 시스템.

청구항 9

제1항 내지 제3항, 제7항 및 제8항 중 어느 한 항에 따른 암의 예후 예측용 시스템을 사용한 암의 예후 예측 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 암 환자의 멀티 오믹스 데이터를 이용하여 딥러닝 기반으로 암의 예후에 영향을 미치는 요인을 파악함으로써 효과적으로 암의 예후를 예측하는 시스템에 관한 것이다.

배 경 기 술

[0003] 암의 예후를 예측하는 것, 즉 암 환자의 수술 후 생존기간을 예측하는 것은 암 환자들에게 매우 중요한 요소이다. 과거의 예후 예측 모델은 환자의 나이, 암의 단계(stage), 성별 등 임상정보만을 이용하여 오직 선형적인 관계로 예후를 예측하였다. 그러나 이러한 방법은 같은 암이라도 개별 환자마다 암의 특성이 다르다는 점을 반영하지 못하므로 예측 성능이 떨어지는 한계점이 있었다.

[0004] 이러한 문제를 해결하기 위해 최근 암의 다양성에 대한 연구가 진행되고 있다. 2018년 종결된 암 유전체 지도(The Cancer Genome Atlas, TCGA) 프로젝트는 개별 환자의 암 특성을 확인할 수 있는 다양한 오믹스(omics) 데이터를 생산하였다. 최근에는 이런 오믹스 데이터를 암의 예후 예측에 적용함으로써 기존 예후 예측 모델의 한계점을 극복하려는 연구가 진행되고 있다.

[0005] 그러나 지금까지 연구된 암의 예후 예측 모델은 일부 한계점이 있다. 첫째는 현재 가장 많이 사용하는 예후 예측 방법은 선형적인 방식이라는 점이다. 이 방법은 직관적으로 예후에 미치는 영향을 확인할 수 있다는 장점이 있지만 생물학적 복잡성을 단순한 선형 모델로 설명하기 어렵다는 한계점이 있고 그로 인해 예후 예측 성능이 조금 떨어진다는 단점이 있다.

[0006] 둘째로 대부분의 암 예후 예측은 단일 오믹스만을 사용한다는 점이다. 최근까지 암의 특징을 확인할 수 있는 다양한 오믹스 방법들이 많이 개발되어 분석에 사용되고 있지만 예후 예측에는 RNA 서열분석과 같이 대부분 유전자의 발현을 확인하는 오믹스 방법만이 사용되고 있다. 하지만 암은 유전자의 발현뿐만 아니라 유전자의 변이 등 복잡한 원인으로 인해 발생하고 예후에 영향을 미친다. 따라서 단일 오믹스만으로는 암의 다양성을 설명하기 힘든 문제가 있다.

[0007] 셋째는 선행 기술에서는 단일 유전자당 예후에 미치는 중요도 확인을 통해 최종 마커 유전자를 선별하기 어렵다는 것이다. 선행 기술에서는 하나의 유전자당 여러 멀티 오믹스 특징들에 대해 독립적으로 할당한다. 하지만 이 방식으로는 각 오믹스 특징마다 중요한 유전자가 서로 다르기 때문에 최종적으로 어떤 유전자가 예후 예측의 마커로 사용될 수 있을지 선별하기 어렵다는 단점이 존재한다.

[0008] 본 발명자들은 상기 한계점을 해결하기 위해 비선형모델로 생물학적 다양성을 학습하고, 멀티 오믹스를 이용해 암의 분자생물학적 다양성을 통합적으로 고려하며, 단일 유전자당 중요도를 판단할 수 있는 암의 예후 예측용 시스템을 개발하여 본 발명을 완성하였다.

선행기술문헌

비특허문헌

[0010] (비특허문헌 0001) 1. SCIENTIFIC REPORTS | 7: 16954

(비특허문헌 0002) 2. Clin Cancer Res; 24(6) March 15, 2018

발명의 내용

해결하려는 과제

[0011] 본 발명의 목적은 멀티 오믹스 데이터에 기반하여 암의 분자생물학적 다양성을 통합적으로 고려함으로써 암의 예후를 효과적으로 예측할 수 있는 암의 예후 예측용 시스템을 제공하는 것이다.

과제의 해결 수단

[0013] 상기 목적을 달성하기 위하여, 본 발명의 일 양상은 하기 단계를 포함하는 딥러닝을 이용한 암의 예후 예측용 시스템을 제공한다:

[0014] (1) 암 환자로부터 수집된 오믹스 데이터와 암의 예후와의 상관관계를 학습하는 학습 모듈;

[0015] (2) 상기 학습 모듈에서 학습된 암의 예후 예측 알고리즘을 저장하는 저장 모듈; 및

[0016] (3) 분석 정보를 수신하고, 상기 저장 모듈에 저장된 암의 예후 예측 알고리즘을 사용하여 예후를 모르는 암 환자의 예후를 산출하는 예측 모듈.

[0017] 본 발명의 일 구체예에 있어서, 상기 암의 예후 예측용 시스템은 암 환자의 치료 후 생존 기간, 또는 외과적 수술 후 생존 기간 예측에 사용될 수 있다.

[0018] 본 명세서에 사용된 용어, "오믹스(omics)"는 유전체, 전사체, 단백체, 대사체 등 다양한 분자 수준에서 생성된 여러 데이터들을 의미하며, 초고속, 대량분석이 가능한 분자생물학적 기술의 발전, 컴퓨터 산업의 발전에 따른 정보 처리 능력의 비약적 발달에 따라 멀티 오믹스 데이터가 생성되고 있다.

[0019] 본 발명의 일 구체예에 있어서, 상기 오믹스 데이터는 유전체(genome), 전사체(transcriptome), 단백체(proteome), 대사체(metabolome), 후성유전체 (epigenome), 지질체(lipodome) 및 메틸체(methylome)로 이루어진 군에서 선택될 수 있으나, 이에 제한되지 않는다. 본 발명에서, 상기 오믹스 데이터는 복수개가 사용될 수 있고, 바람직하게는 최소 3개의 오믹스 데이터가 사용될 수 있으며, 예를 들어 유전체, 전사체 및 메틸체 데이터가 사용될 수 있다.

[0020] 본 발명의 일 구체예에 있어서, 상기 학습 모듈은 하기 단계를 포함할 수 있다.

[0021] (a) 예후를 알고 있는 암 환자의 오믹스 데이터로부터 딥러닝용 학습 데이터를 생성하는 학습 데이터 생성부;

[0022] (b) 생성된 학습 데이터로부터 오믹스 데이터와 암의 예후와의 상관 관계를 학습하는 딥러닝 수행부; 및

[0023] (c) 상기 딥러닝 수행부의 결과로부터 암의 예후를 예측하는 알고리즘을 생성하는 알고리즘 생성부.

[0024] 본 명세서에 사용된 용어, "딥러닝(deep learning)"은 심층 학습으로도 불리우며, 최근 기계학습 분야에서 대두되고 있는 기술 중 하나로 복수개의 은닉 계층(hidden layer)와 이들에 포함되는 복수 개의 유닛(hidden unit)으로 구성되는 신경망(neural network)이다.

[0025] 딥러닝 모델에 기본 특성(low level feature)들을 입력하는 경우, 이러한 기본 특성들이 복수 개의 은닉 계층을 통과하면서 예측하고자 하는 문제를 보다 잘 설명할 수 있는 상위 레벨 특성(high level feature)로 변형된다. 이러한 과정에서 전문가의 사전 지식 또는 직관이 요구되지 않기 때문에 특성 추출에서의 주관적 요인을 제거할 수 있으며, 보다 높은 일반화 능력을 갖는 모델을 개발할 수 있게 된다.

[0026] 또한, 딥러닝의 경우 특징 추출과 모델 구축이 하나의 세트로 구성되어 있기 때문에 기존의 기계학습 이론 대비 보다 단순한 과정을 통하여 최종 모델을 형성할 수 있는 장점이 있다.

[0027] 본 발명에서, 상기 딥러닝은 심층 신경망(Deep Neural Network, DNN)에 기반하여 수행될 수 있으며, 예를 들어 구글 오픈소스인 텐서플로우(TensorFlow)로 수행될 수 있다.

[0028] 본 발명에서, 상기 학습 데이터 생성부는 예후를 알고 있는 암 환자의 오믹스 데이터를 개별 유전자에 대해 매트릭스 형태로 정렬할 수 있다. 매트릭스 형태로 정렬함으로써 개별 유전자, 단일 오믹스에 대한 정보만으로 암의 예후를 예측하는 것이 아니라, 개별 유전자에 대한 오믹스 데이터들의 공통 특징을 학습하여 비선형적인 방법으로 암의 예후를 예측할 수 있다. 사용되는 유전자는 따로 선별하는 것이 아니라, 오믹스 데이터에 포함된 전체 유전자를 사용할 수 있다.

[0029] 본 발명의 일 구체예에 있어서, 상기 딥러닝 수행부는 하기 단계를 포함할 수 있다:

[0030] (b-1) 정렬된 오믹스 데이터로부터 개별 유전자에 대한 오믹스 데이터의 공통 특징을 학습하는 단계;

[0031] (b-2) 학습된 개별 유전자에 대한 오믹스 데이터의 공통 특징에서 중요하지 않은 정보를 제거하는 단계;

[0032] (b-3) 상기 (b-2)의 결과로부터 개별 유전자당 가장 높은 값을 암의 예후 예측을 위한 대푯값으로 선별하는 단계;

[0033] (b-4) 상기 (b-3)에서 선별한 개별 유전자당 대푯값을 조합하여 암의 예후 예측의 정확도를 확인하는 단계; 및

[0034] (b-5) 상기 (b-4)의 결과에서 정확도가 가장 높은 대푯값의 조합을 선별하는 단계.

[0035] 본 발명의 일 구체예에 있어서, 상기 (b-1) 단계는 정렬된 오믹스 데이터로 컨볼루션(convolution)을 진행하여 수행될 수 있으며, (b-2) 단계는 학습된 개별 유전자에 대한 오믹스 데이터의 공통 특징에 ReLU (rectified linear unit) 함수를 적용하여 수행될 수 있다.

[0036] 본 발명의 일 구체예에 있어서, 상기 (b-4) 단계는 콕스 비례 위험 모델(Cox proportional hazards model)을 사용하여 개별 유전자당 대푯값과 암 환자의 예후와의 음의 로그 부분 유도(negative log partial likelihood)를 계산하는 것일 수 있다.

[0038] 본 발명의 다른 양상은 상기 암의 예후 예측 시스템을 사용한 암의 예후 예측 방법을 제공한다. 상기 암의 예후 예측 시스템을 사용하면 예후를 모르는 암 환자에서 수집된 멀티 오믹스 데이터로부터 암의 예후에 가장 큰 영향을 미치는 개별 유전자를 선별할 수 있으므로 예후가 불분명한 암 환자의 예후를 효과적으로 예측할 수 있다. 또한, 암의 예후를 예측함으로써 예후가 나쁠 것으로 예상되면 공격적 치료를 수행한다는 등의 방식으로 개인별 맞춤 의료를 제공할 수 있다.

발명의 효과

[0040] 본 발명의 암의 예후 예측용 시스템에 따르면 복수의 오믹스 데이터를 사용함으로써 암의 분자생물학적 다양성을 통합적으로 고려하여 암 환자의 예후를 예측할 수 있고, 단일 유전자가 암의 예후 예측에 미치는 중요도를 판단할 수 있다.

도면의 간단한 설명

[0042] 도 1은 본 발명에 따른 암의 예후 예측 시스템의 구성을 설명하기 위한 블록도이다.

도 2는 본 발명에 따른 암의 예후 예측 시스템에서 학습 모듈의 학습 데이터 생성부의 일 예를 나타낸다.

도 3은 본 발명에 따른 암의 예후 예측 시스템에서 학습 모듈의 딥러닝 수행부의 일 예를 나타낸다.

도 4는 본 발명의 일 실시예에 따른 암의 예후 예측 시스템과 기존의 암 예후 예측 모델의 예후 예측 성능을 비교한 결과를 나타낸다.

도 5는 본 발명의 일 실시예에 따른 암의 예후 예측 시스템과 암의 예후 예측 마커로 알려진 유전자의 예후 예측 성능을 비교한 결과를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0043] 이하에서는 첨부한 도면을 참조하여 본 발명을 설명하기로 한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 따라서 여기에서 설명하는 실시예로 한정되는 것은 아니다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0044] 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 추가로 구비할 수 있다는 것을 의미한다.

[0046] 이하, 첨부된 도면을 참고하여 본 발명을 더욱 상세히 설명한다.

[0047] 도 1은 본 발명의 암의 예후 예측 시스템의 구성을 예시적으로 나타낸다.

[0048] 상기 암의 예후 예측 시스템은

[0049] (a) 암 환자로부터 수집된 오믹스 데이터와 암의 예후와의 상관관계를 학습하는 학습 모듈(100);

[0050] (b) 상기 학습 모듈에서 학습된 암의 예후 예측 알고리즘을 저장하는 저장 모듈(200); 및

[0051] (c) 분석 정보를 수신하고, 상기 저장 모듈에 저장된 암의 예후 예측 알고리즘을 사용하여 예후를 모르는 암 환자의 예후를 산출하는 예측 모듈(300)을 포함한다.

[0052] 상기 (a)의 학습 모듈(100)은 암의 예후, 즉 치료 후 생존 기간을 이미 알고 있는 암 환자의 오믹스 데이터와 암의 예후와의 상관관계를 딥러닝에 의해 학습하는 부분이며, 학습 데이터 생성부(110), 딥러닝 수행부(120) 및 알고리즘 생성부(130)를 포함한다.

[0053] 상기 오믹스 데이터는 암 유전체 지도(The Cancer Genome Atlas, TCGA)와 같은 공개 데이터베이스로부터 수집할 수 있으며, 이외에 NCBI (National Center for Biotechnology Information) 등으로부터도 수집할 수 있다. 상기 오믹스 데이터는 암의 분자생물학적 다양성을 통합적으로 위해 복수개가 사용될 수 있으며, 바람직하게는 3개 이상의 오믹스 데이터가 사용될 수 있다.

[0054] 상기 학습 데이터 생성부(110)는 암 환자의 오믹스 데이터를 딥러닝용 학습 데이터로 가공하는 부분이며, 도 2에 도시된 바와 같이 각각의 오믹스 데이터를 개별 유전자에 대해 매트릭스 형태로 정렬한다. 매트릭스 형태로 정렬하는 것은 유전자 발현 데이터만을 이용해 암의 예후를 예측했던 기존의 방법과는 달리 유전자의 발현, 유전자 변이 등 암의 다양성을 통합적으로 고려하는 예후 예측 시스템을 구축하기 위함이다.

[0055] 상기 딥러닝 수행부(120)는 학습 데이터로부터 서로 다른 딥러닝 함수를 사용하여 개별 유전자에 대해 정렬된 오믹스 데이터와 암의 예후와의 상관 관계를 학습하는 부분이며, 도 3에 도시된 바와 같이 하기 단계를 포함할 수 있다:

[0056] (b-1) 정렬된 오믹스 데이터로부터 개별 유전자에 대한 오믹스 데이터의 공통 특징을 학습하는 단계;

[0057] (b-2) 학습된 개별 유전자에 대한 오믹스 데이터의 공통 특징에서 중요하지 않은 정보를 제거하는 단계;

[0058] (b-3) 상기 (b-2)의 결과로부터 개별 유전자당 가장 높은 값을 암의 예후 예측을 위한 대푯값으로 선별하는 단계;

[0059] (b-4) 상기 (b-3)에서 선별한 개별 유전자당 대푯값을 조합하여 암의 예후 예측의 정확도를 확인하는 단계; 및

[0060] (b-5) 상기 (b-4)의 결과에서 정확도가 가장 높은 대푯값의 조합을 암의 예후 예측용 마커로 선별하는 단계.

[0061] 상기 (b-1) 단계는 매트릭스 형태로 정렬된 오믹스 데이터로 컨볼루션 (convolution)을 진행하여 수행될 수 있다. 구체적으로 하기 수학식 1에 따라 n개의 유전자에 대해 k개의 커널로 컨볼루션(convolution)을 진행하며, 예를 들어 k=3이라면 k=1, 2, 3 각각에 대해 컨볼루션을 진행하게 된다. 이 과정을 통해 각 커널마다 개별 유전자에 대한 오믹스 데이터들의 공통 특징이 학습된다. 딥러닝이 완료된 후 하기 수학식 1에서 w의 크기는 해당 오믹스 데이터의 예후 예측에 있어서의 중요도를 나타내며, k는 하이퍼파라미터(hyperperparameter) 튜닝(최적화) 과정에서 높은 성능을 나타내는 값을 사용한다. 하이퍼파라미터는 딥러닝을 수행하기 위해 사전에 설정해야 하는 값을 의미한다.

[0062] [수학식 1]

[0063] $$c_n^k = \sum_{m=1}^{M} w_m^k X_{mn}$$

[0064] (m=오믹스, n=유전자, *Xmn*=유전자와 오믹스 특징들로 이루어진 매트릭스, w= 개별 유전자의 멀티 오믹스 데이터와 곱해지는 웨이트(weight), k=컨볼루션을 진행하는 커널(kernel), 및 c_n^k는 커널 k로 컨볼루션을 진행한 유전자 n에 대한 값을 의미함)

[0066] 상기 (b-2) 단계는 학습된 개별 유전자에 대한 오믹스 데이터들의 공통 특징에 하기 수학식 2를 적용하여 중요하지 않은 정보를 제거하는 단계이며, 이 과정에서 중요도가 높은 유전자에 대한 결괏값(h_n^k)만 남게 된다.

[0067] [수학식 2]

$$h_n^k = ReLU(c_n^k + b_k)$$

[0068] $$ReLU(x) = \begin{cases} x \ if \ x \geq 0 \\ 0 \ if \ x < 0 \end{cases}$$

[0069] 상기 수학식 2에 사용된 ReLU(rectified linear unit) 함수는 딥러닝 과정에서 사용되는 활성화 함수(activation fuction)로 상기 수학식 2에 기재된 바와 같이 x가 0보다 작은 값일 때는 0을 사용하고, 반대로 0보다 큰 값일 때는 해당 값을 그대로 사용하는 함수이다. 또한, 상기 b는 바이어스(bias)를 의미하며, 딥러닝 과정에서 자동적으로 정해진다.

[0070] 상기 (b-3) 단계는 상기 (b-2)에서 나온 중요도가 높은 유전자에 대한 결괏값에 대하여 수학식 3에 따른 맥스 풀링(max pooling)을 진행하여 개별 유전자당 가장 높은 값을 암의 예후 예측을 위한 대푯값(h_n)으로 선별하는 단계이다.

[0071] [수학식 3]

[0072] $$pooling(h_n) = \max(\{h_n^1, h_n^2, h_n^3, \ldots, h_n^k\})$$

[0073] 상기 (b-4) 단계는 (b-3)에서 선별한 개별 유전자당 대푯값을 조합하여 암의 예후 예측의 정확도를 확인하는 단계이며, 구체적으로 개별 유전자당 대푯값의 조합으로 예측된 암의 예후와 이미 알고 있는 예후를 비교하여 정확도를 확인한다. 이때 하기 수학식 4에 따른 콕스 비례 위험 모델(Cox proportional hazards model)을 사용하여 개별 유전자당 대푯값과 암 환자의 예후와의 음의 로그 부분 유도(negative log partial likelihood)를 계산하여 학습을 진행하였다.

[0074] [수학식 4]

$$l(\beta, X) = -\sum_{i \in U} (X_i \beta - log \sum_{j \in R_i} e^{X_j \beta})$$

[0075]

[0076] (X_i=매트릭스, β=콕스 비례 위험 모델의 파라미터, U=검열이 안된 샘플들(uncensored samples) 집단, R=검열된 샘플들(at-risk samples) 및 j=생존 추적(follow up) 기간으로 i보다 더 길 때를 의미함)

[0078] 상기 (b-5) 단계는 상기 (b-4)의 결과에서 정확도가 가장 높은 대푯값의 조합을 암의 예후 예측용 마커로 선별하는 단계이며, 이 과정에서 암의 예후 예측에 중요한 영향을 미치는 유전자 또는 유전자들의 조합이 선별된다.

[0079] 상기 저장 모듈(200)은 상기 학습모듈에 의해 학습된 암의 예후 예측 알고리즘을 저장하는 부분으로, 암의 예후 예측 알고리즘 데이터베이스(database, DB)(220)를 포함하여 구성되고, 학습된 개별 유전자당 오믹스 데이터와 암의 예후와의 관련성을 저장하기 위한 오믹스 데이터-예후 관련 DB(210)를 더 포함하여 구성될 수도 있다.

[0080] 상기 예측 모듈(300)은 분석할 정보를 수신하는 분석 정보 수신부(310) 및 저장 모듈에 저장된 예후 예측 알고리즘을 사용하여 분석 정보로부터 산출한 예후를 출력하는 예후 결과 출력부(320)를 포함한다.

[0081] 상기 분석 정보는 예후가 불확실한 암 환자의 오믹스 데이터일 수 있으며, 상기 예후 결과는 암 치료 후 생존 기간일 수 있다.

[0083] 이하, 실시예를 통해 본 발명을 더욱 상세히 기술한다.

[0085] **실시예: 암의 예후 예측 시스템 검증**

[0086] 본 발명의 암의 예후 예측 시스템(이하, 예후 예측 시스템으로 기재함)은 TCGA 간암 환자 데이터 중에서 유전제, 전사체 및 메틸체 데이터를 학습에 사용하였다. 구체적으로 TCGA 간암 환자 데이터의 80%는 예후 예측 시스템의 학습, 검증 및 하이퍼파라미터 최적화에 사용하였고, 나머지 20%는 예후 예측 시스템의 테스트에 사용

하였다.

[0087] 본 발명의 예후 예측 시스템이 학습에 사용된 데이터에 과적합(overfitting)된 것이 아니라 완전히 다른 데이터셋에서도 작동하는 것을 확인하기 위하여 TCGA 데이터뿐만 아니라 완전히 독립적인 데이터셋인 한국인 간암 샘플 데이터에서 성능 확인을 진행했다.

[0088] 구체적으로 TCGA 간암 환자 211명과 한국인 간암 환자 144명에 대해 유전자 발현, DNA 체세포 변이(somatic mutation), DNA 생식선 변이(germline mutation), 복수체 변이(copy number variation), 발현 조절(DNA methylation)에 해당하는 오믹스 특징들을 사용해 본 발명의 예후 예측 시스템으로 간암 환자의 예후를 예측하고, 선행기술인 Glmnet, RSF(random survival forest) 모델과 성능을 비교하였다.

[0089] 그 결과, 도 4에 나타낸 바와 같이 선형모델을 사용하는 Glmnet 모델은 두 데이터셋 모두에서 성능(concordance index, C-index)이 0.5 근처로 나타나 제일 낮았다. 상기 C-index는 0.0 내지 1.0 사이의 값을 가지며, 0.5에 가까울수록 무작위로 예측하는 것으로 평가하고, 1.0에 가까울수록 정확히 예측한다고 평가한다. 비선형 모델을 사용하는 RSF 모델은 학습에 사용되는 TCGA 간암 샘플에서는 높은 성능을 보였으나, 한국인 간암 샘플에서는 성능이 낮은 것으로 나타나 학습에 사용되는 데이터셋에 과적합되어 있는 현상을 확인할 수 있다. 반면 본 발명의 예후 예측 시스템은 모든 데이터셋에서 다른 모델들에 비해 높은 성능을 나타내었다.

[0091] 또한, 논문에서 간암의 예후를 예측하는 것으로 알려진 3개 유전자 (UPB1, SOCS2 및 RTN3), 11개 유전자 (ACSM3, CXCL14, INTS8, LCAT, MARCO, PAMR1, CRHBP, DNASE1L3, FCN2, MT1X 및 VIPR1) 및 종양 관련 섬유아세포(cancer associated fibroblasts, CAF)와 관련된 12종의 유전자(ACSM3, CXCL14, INTS8, LCAT, MARCO, PAMR1, CRHBP, DNASE1L3, FCN2, MT1X 및 VIPR1)를 확인하고, 상기 유전자들의 간암 예후 예측 성능을 평가하였다. 그 결과, 도 5에 나타낸 바와 같이 논문에서 제시하는 간암의 예후 예측 유전자들은 학습에 사용되는 TCGA 샘플에 과적합되어 있는 현상을 확인할 수 있었다.

[0092] 본 결과를 통해 본 발명의 예후 예측 시스템은 다양한 오믹스 데이터를 활용한 유전자 단위의 통합적인 학습으로 인해 다른 예후 예측 모델과 비교하여 간암의 예후 예측에 현저히 우수한 성능을 보이는 것을 확인할 수 있었다.

부호의 설명

[0093] 100: 학습 모듈

110: 학습 데이터 생성부

120: 딥러닝 수행부

130: 알고리즘 생성부

200: 저장 모듈

210: 오믹스 데이터-예후 관련 DB

220: 예후 예측 알고리즘 DB

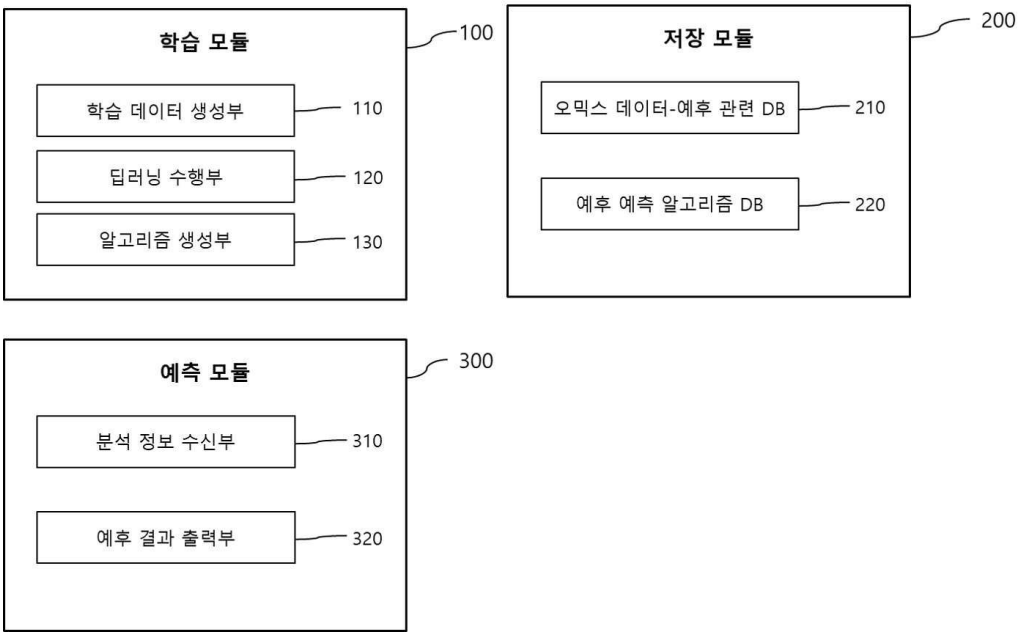
300: 예측 모듈

310: 분석 정보 수신부

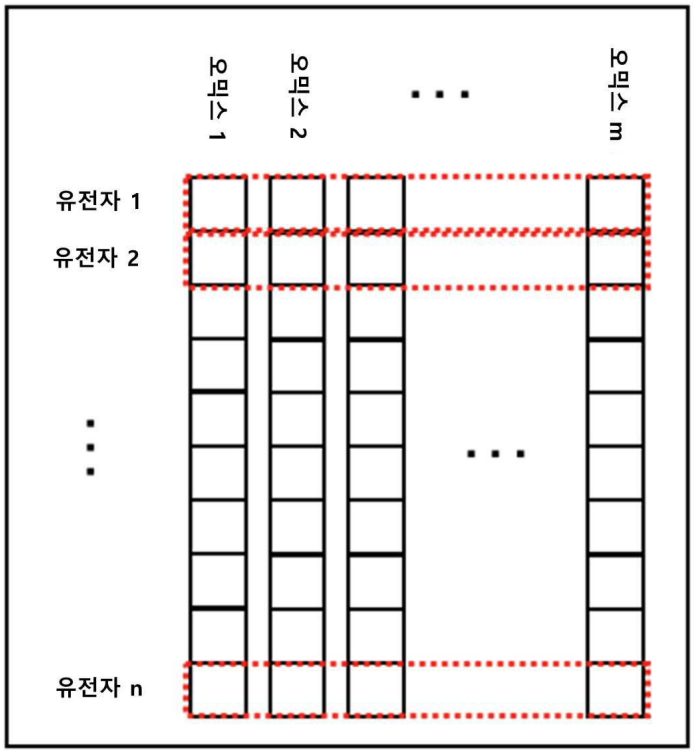
320: 예후 결과 출력부

도면

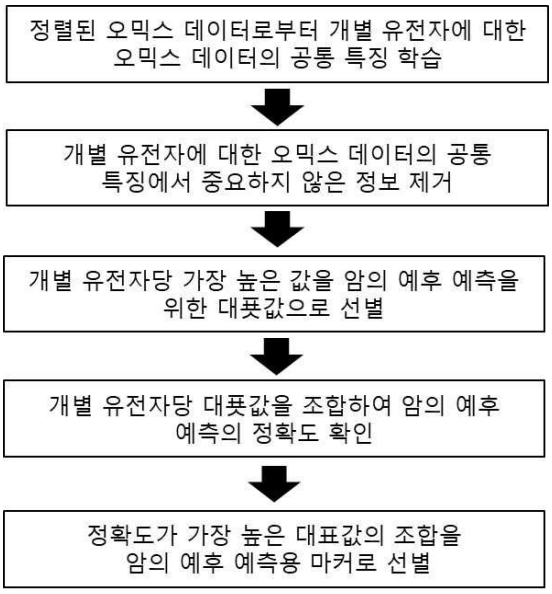
도면1


학습 모듈
100
학습 데이터 생성부
110
딥러닝 수행부
120
알고리즘 생성부
130
저장 모듈
200
오믹스 데이터-예후 관련 DB
210
예후 예측 알고리즘 DB
220
예측 모듈
300
분석 정보 수신부
310
예후 결과 출력부
320

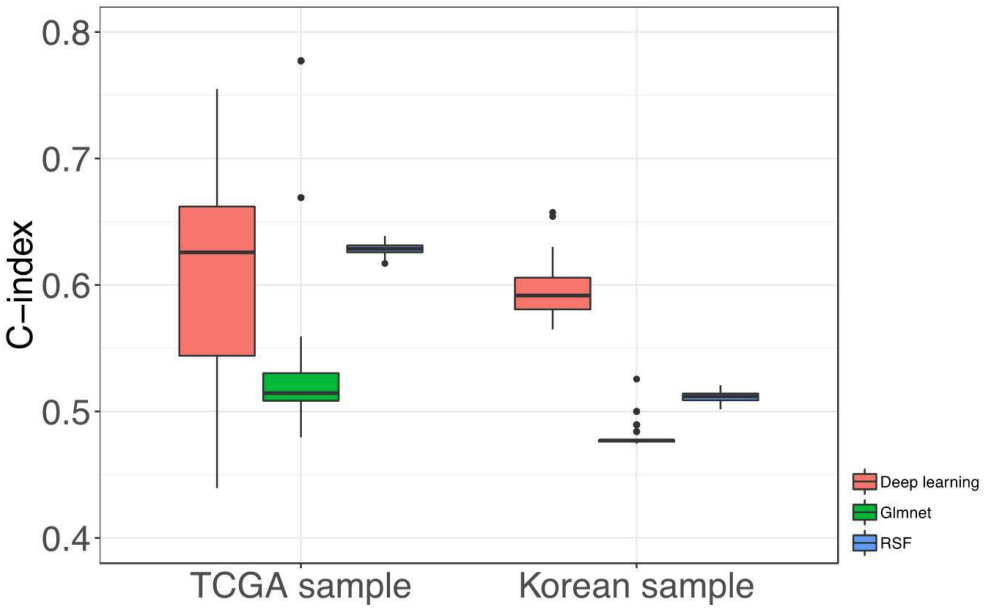
도면2


오믹스 1
오믹스 2
오믹스 m
유전자 1
유전자 2
유전자 n

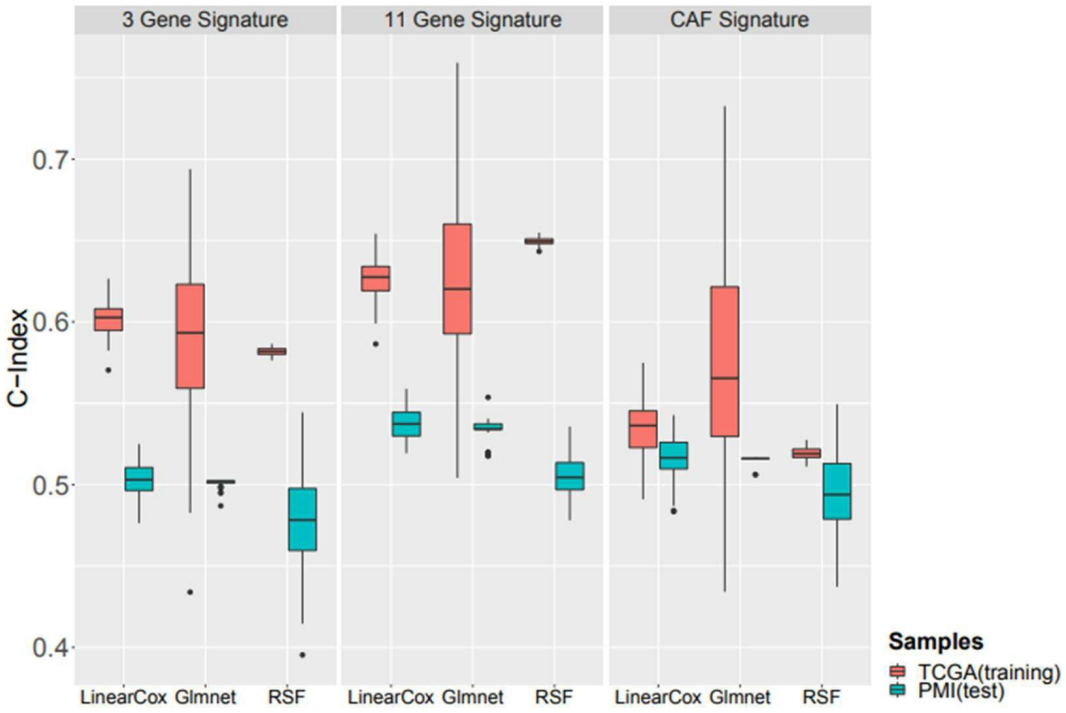
도면3


정렬된 오믹스 데이터로부터 개별 유전자에 대한 오믹스 데이터의 공통 특징 학습
개별 유전자에 대한 오믹스 데이터의 공통 특징에서 중요하지 않은 정보 제거
개별 유전자당 가장 높은 값을 암의 예후 예측을 위한 대푯값으로 선별
개별 유전자당 대푯값을 조합하여 암의 예후 예측의 정확도 확인
정확도가 가장 높은 대표값의 조합을 암의 예후 예측용 마커로 선별

도면4


C-index
0.8
0.7
0.6
0.5
0.4
TCGA sample
Korean sample
Deep learning
Glmnet
RSF

도면5


3 Gene Signature
11 Gene Signature
CAF Signature
C-Index
0.7
0.6
0.5
0.4
LinearCox
Glmnet
RSF
Samples
TCGA(training)
PMI(test)