(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0008273
(43) 공개일자 2021년01월21일

(51) 국제특허분류(Int. Cl.)
G16H 10/40 (2018.01) *G16B 25/10* (2019.01)
G16H 10/20 (2018.01)
(52) CPC특허분류
G16H 10/40 (2018.01)
G16B 25/10 (2019.02)
(21) 출원번호 10-2019-0084709
(22) 출원일자 2019년07월12일
심사청구일자 2019년07월12일

(71) 출원인
주식회사 디앤피바이오텍
서울특별시 서초구 서초대로49길 18, 304호 (서초동, 상림빌딩)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
황도식
서울특별시 서대문구 연세로 50 제3공학관 C618호 (신촌동)
박두현
서울특별시 서대문구 연세로 50 제3공학관 C516호 (신촌동)
(뒷면에 계속)
(74) 대리인
특허법인(유한) 다래

전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **임상 정보와 유전자 다형성 정보를 이용한 폐암 환자의 수술 후 예후 예측 방법**

(57) 요 약

임상 정보와 유전자 다형성 정보를 이용한 폐암 환자의 수술 후 예후 예측 방법이 제공된다. 본 발명의 일 실시예에 따른 예후 예측 방법은 복수의 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신하는 단계, 수신된 유전자 다형성 정보 및 임상 정보를 제1 정보 그룹과 제2 정보 그룹으로 분류하는 단계, 복수의 폐암 환자 중 일부 환자를 선택하고, 선택된 일부 환자 각각에 대해 유전자 다형성 정보 및 임상 정보 중 일부를 선택하는 단계, 선택하는 단계를 반복하여 복수의 선택 정보 세트를 구성하는 단계 및 구성된 복수의 선택 정보 세트를 기초로 예후 예측 모델을 생성하는 단계를 포함할 수 있다.

대 표 도 - 도2

(52) CPC특허분류
G16H 10/20 (2018.01)
(72) 발명자
이명훈
서울특별시 송파구 송파대로 345, 102동 2401호(가락동,헬리오시티)
전선곤
대전광역시 유성구 엑스포로 448 503동 1002호 (전민동,엑스포아파트)
이현철
대구시 북구 구암로21길 38 칠곡공작한양1차 아파트 105-701
박승연
대구시 북구 옥산로 87 태왕아너스로뎀플러스 308호
김민소
대구광역시 동구 경안로 각산 푸르지오 824 104-2505
오대중
대구광역시 동구 동부로22길 14(신천동) 부띠크시티오피스텔 1동 2116호

명 세 서

청구범위

청구항 1

임상 정보와 유전자 다형성 정보를 이용한 폐암 환자의 수술 후 예후 예측 방법에 있어서,

복수의 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신하는 단계;

상기 수신된 유전자 다형성 정보 및 임상 정보를 제1 정보 그룹과 제2 정보 그룹으로 분류하는 단계;

상기 복수의 폐암 환자 중 일부 환자를 선택하고, 상기 선택된 일부 환자 각각에 대해 유전자 다형성 정보 및 임상 정보 중 일부를 선택하는 단계;

상기 선택하는 단계를 반복하여 복수의 선택 정보 세트를 구성하는 단계; 및

상기 구성된 복수의 선택 정보 세트를 기초로 예후 예측 모델을 생성하는 단계;를 포함하는 예후 예측 방법.

청구항 2

제1항에 있어서,

상기 분류하는 단계는,

상기 유전자 다형성 정보 및 임상 정보 각각이 상기 예후 예측 모델의 정확도에 기여하는 정도를 나타내는 중요도 값을 기초로, 제1 정보 그룹과 제2 정보 그룹으로 분류하는 것을 특징으로 하는 예후 예측 방법.

청구항 3

제1항에 있어서,

상기 분류하는 단계는,

상기 유전자 다형성 정보 및 임상 정보 각각에 대한 사건 발생률 차이 값을 기초로, 제1 정보 그룹과 제2 정보 그룹으로 분류하는 것을 특징으로 하는 예후 예측 방법.

청구항 4

제1항에 있어서,

상기 분류하는 단계는,

상기 유전자 다형성 정보 및 임상 정보 각각에 대한 단변량 선형 회귀 모델을 생성하고, 상기 생성된 단변량 선형 회귀 모델의 AUC(Area Under Curve) 값을 도출하며, 상기 도출된 AUC 값을 기초로 제1 정보 그룹과 제2 정보 그룹으로 분류하는 것을 특징으로 하는 예후 예측 방법.

청구항 5

제2항 내지 제4항 중 어느 한 항에 있어서,

상기 수신된 복수의 폐암 환자 중 일부 환자를 랜덤하게 선택하여 복수의 서브 세트를 구성하는 단계;를 더 포함하고,

상기 분류하는 단계는,

상기 복수의 서브 세트 각각에 속한 일부 환자의 유전자 다형성 정보 및 임상 정보를 기초로, 상기 복수의 서브 세트 별로 중요도 값, 사건 발생률 차이 값 및 AUC 값 중 하나를 도출하여 합산하고, 상기 복수의 서브 세트의 수로 상기 합산된 값을 나누어 평균 값을 도출하며, 상기 도출된 평균 값을 기초로 제1 정보 그룹과 제2 정보 그룹으로 분류하는 것을 특징으로 하는 예후 예측 방법.

청구항 6

제1항에 있어서,

상기 분류하는 단계는,

상기 임상 정보 중 병리학적 종양 병기, 조직학적 유형, 나이, 성별 및 흡연량 정보 중 적어도 하나를 상기 제1 정보 그룹으로 분류하고, 유전자 다형성 정보 및 상기 제1 정보 그룹으로 분류되지 않은 임상 정보를 상기 제2 정보 그룹으로 분류하는 것을 특징으로 하는 예후 예측 방법.

청구항 7

제1항에 있어서,

상기 선택하는 단계는,

상기 제1 정보 그룹으로 분류된 정보에 가중치를 부여하고, 상기 부여된 가중치를 반영하여 유전자 다형성 정보 및 임상 정보 중 일부를 선택하는 것을 특징으로 하는 예후 예측 방법.

청구항 8

제1항에 있어서,

상기 선택하는 단계는,

상기 제1 정보 그룹으로 분류된 정보는 모두 선택하고, 상기 제2 정보 그룹으로 분류된 정보 중 일부를 랜덤하게 선택하여, 유전자 다형성 정보 및 임상 정보 중 일부를 선택하는 것을 특징으로 하는 예후 예측 방법.

청구항 9

제1항에 있어서,

상기 유전자 다형성 정보는,

CD3EAP, TNFRSF10B, AKT1, C3, HOMER2, GNB2L1, CD3D 및 ADAMTSL3 유전자의 다형성 정보 중 적어도 하나 이상의 다형성 정보인 예후 예측 방법.

청구항 10

제1항에 있어서,

상기 임상정보는,

폐암 환자의 병리학적 종양 병기, 조직학적 유형, 나이, 성별 및 흡연량 정보 중 적어도 하나인 예후 예측 방법.

청구항 11

제1항에 있어서,

상기 생성하는 단계는,

상기 복수의 선택 정보 세트 각각을 기초로 의사결정나무에 기반한 복수의 개별 모델을 생성하고, 상기 생성된 복수의 개별 모델을 종합하여 앙상블 예후 예측 모델을 생성하는 것을 특징으로 하는 예후 예측 방법.

청구항 12

제1항에 있어서,

상기 생성하는 단계는,

상기 복수의 선택 정보 세트 각각을 기초로 의사결정나무에 기반한 복수의 개별 모델을 생성하고, 상기 복수의 선택 정보 세트 각각을 구성하는 정보 중 상기 제1 정보 그룹에 속하는 정보의 수에 비례하게 상기 생성된 복수의 개별 모델 각각에 가중치를 부여하며, 상기 부여된 가중치를 반영하여 앙상블 예후 예측 모델을 생성하는 것을 특징으로 하는 예후 예측 방법.

청구항 13

제1항에 있어서,

예후를 예측할 폐암 환자의 유전자 다형성 정보 및 임상 정보를 입력하는 단계; 및

상기 입력된 유전자 다형성 정보 및 임상 정보를 기초로, 상기 생성된 예후 예측 모델을 이용하여 예측된 예후 결과를 출력하는 단계;를 더 포함하는 예후 예측 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 임상 정보와 유전자 다형성 정보를 이용한 폐암 환자의 수술 후 예후 예측 방법에 관한 것으로, 보다 상세하게는 유효성이 검증된 주요 특징을 반영하여 예후 예측 모델의 성능을 향상시킬 수 있는 예후 예측 방법에 관한 것이다.

배 경 기 술

[0002] 최근 데이터의 중요성이 강조되면서, 의료 빅데이터를 이용한 고성능의 예후 예측 모델을 확보하고자 하는 요구가 증대하고 있다. 특정 질병에 걸린 환자의 생존을 예측하는데 있어서, 생존과 관련된 여러 인자가 존재하는 경우 종래에는 각 예측 인자간의 가중합을 통해 고위험군과 저위험군을 나누는 선형 모델을 이용하였다. Cox Proportional Hazard Model이 선형 모델의 대표적인 예이다. 선형 모델은 각 인자간의 상호작용을 고려하지 않는 단순한 방식이어서 성능 향상에 한계가 존재한다.

[0003] 이에 따라 선형 모델의 한계를 극복하기 위해 다양한 기계학습 방식을 이용하고자 하는 시도들이 존재하였다. 그러나 학습 모델이 복잡할수록 과적합 문제가 발생하여 충분한 성능이 나오지 못하는 문제가 발생하였는데, 특히 의료 분야 연구의 경우 데이터 수의 부족으로 과적합 문제 발생 확률이 높다는 점에서 기계학습 적용에 어려움이 있었다.

[0004] 의료 데이터를 이용한 예측 모델에서는 랜덤포레스트와 같은 기법을 사용하여 단점을 극복하는 시도가 있었다. 랜덤포레스트는 특징의 부분집합을 선택하는 방식으로 개별 모델의 다양성을 확보할 수 있기 때문에, 데이터 수의 부족으로 인한 문제는 극복할 수 있었다. 하지만 개별 모델의 다양성이 확보됨으로 인해 발생하는 장점보다 개별 모델의 성능이 떨어지는 단점이 더 커서 여전히 예후 예측 모델의 성능 향상에 한계가 있다는 문제점이 존재한다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 상술한 문제점을 해결하기 위한 것으로, 유전자 다형성 정보 및 임상정보 중 예후 예측과의 연관성이 보다 검증된 정보를 학습 모델 생성에 높은 빈도로 반영시킬 수 있는 예후 예측 방법을 제공함을 목적으로 한다.

과제의 해결 수단

[0006] 상기 목적을 달성하기 위한 본 발명의 일 실시 예에 따른 임상 정보와 유전자 다형성 정보를 이용한 폐암 환자의 수술 후 예후 예측 방법은, 복수의 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신하는 단계, 상기 수신된 유전자 다형성 정보 및 임상 정보를 제1 정보 그룹과 제2 정보 그룹으로 분류하는 단계, 상기 복수의 폐암 환자 중 일부 환자를 선택하고, 상기 선택된 일부 환자 각각에 대해 유전자 다형성 정보 및 임상 정보 중 일부를 선택하는 단계, 상기 선택하는 단계를 반복하여 복수의 선택 정보 세트를 구성하는 단계 및 상기 구성된 복수의 선택 정보 세트를 기초로 예후 예측 모델을 생성하는 단계를 포함할 수 있다.

[0007] 그리고 상기 분류하는 단계는, 상기 유전자 다형성 정보 및 임상 정보 각각이 상기 예후 예측 모델의 정확도에 기여하는 정도를 나타내는 중요도 값을 기초로, 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다.

[0008] 또한, 상기 분류하는 단계는, 상기 유전자 다형성 정보 및 임상 정보 각각에 대한 사건 발생률 차이 값을 기초로, 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다.

[0009] 그리고 상기 분류하는 단계는, 상기 유전자 다형성 정보 및 임상 정보 각각에 대한 단변량 선형 회귀 모델을 생성하고, 상기 생성된 단변량 선형 회귀 모델의 AUC(Area Under Curve) 값을 도출하며, 상기 도출된 AUC 값을 기초로 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다.

[0010] 또한, 상기 수신된 복수의 폐암 환자 중 일부 환자를 랜덤하게 선택하여 복수의 서브 세트를 구성하는 단계를 더 포함하고, 상기 분류하는 단계는, 상기 복수의 서브 세트 각각에 속한 일부 환자의 유전자 다형성 정보 및 임상 정보를 기초로, 상기 복수의 서브 세트 별로 중요도 값, 사건 발생률 차이 값 및 AUC 값 중 하나를 도출하여 합산하고, 상기 복수의 서브 세트의 수로 상기 합산된 값을 나누어 평균 값을 도출하며, 상기 도출된 평균 값을 기초로 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다.

[0011] 그리고 상기 분류하는 단계는, 상기 임상 정보 중 병리학적 종양 병기, 조직학적 유형, 나이, 성별 및 흡연량 정보 중 적어도 하나를 상기 제1 정보 그룹으로 분류하고, 유전자 다형성 정보 및 상기 제1 정보 그룹으로 분류되지 않은 임상 정보를 상기 제2 정보 그룹으로 분류할 수 있다.

[0012] 또한, 상기 선택하는 단계는, 상기 제1 정보 그룹으로 분류된 정보에 가중치를 부여하고, 상기 부여된 가중치를 반영하여 유전자 다형성 정보 및 임상 정보 중 일부를 선택할 수 있다.

[0013] 그리고 상기 선택하는 단계는, 상기 제1 정보 그룹으로 분류된 정보는 모두 선택하고, 상기 제2 정보 그룹으로 분류된 정보 중 일부를 랜덤하게 선택하여, 유전자 다형성 정보 및 임상 정보 중 일부를 선택할 수 있다.

[0014] 또한, 상기 유전자 다형성 정보는, CD3EAP, TNFRSF10B, AKT1, C3, HOMER2, GNB2L1, CD3D 및 ADAMTSL3 유전자의 다형성 정보 중 적어도 하나 이상의 다형성 정보일 수 있다.

[0015] 그리고 상기 임상정보는, 폐암 환자의 병리학적 종양 병기, 조직학적 유형, 나이, 성별 및 흡연량 정보 중 적어도 하나일 수 있다.

[0016] 또한, 상기 생성하는 단계는, 상기 복수의 선택 정보 세트 각각을 기초로 의사결정나무에 기반한 복수의 개별 모델을 생성하고, 상기 생성된 복수의 개별 모델을 종합하여 앙상블 예후 예측 모델을 생성할 수 있다.

[0017] 그리고 상기 생성하는 단계는, 상기 복수의 선택 정보 세트 각각을 기초로 의사결정나무에 기반한 복수의 개별 모델을 생성하고, 상기 복수의 선택 정보 세트 각각을 구성하는 정보 중 상기 제1 정보 그룹에 속하는 정보의 수에 비례하게 상기 생성된 복수의 개별 모델 각각에 가중치를 부여하며, 상기 부여된 가중치를 반영하여 앙상블 예후 예측 모델을 생성할 수 있다.

[0018] 또한, 예후를 예측할 폐암 환자의 유전자 다형성 정보 및 임상 정보를 입력하는 단계 및 상기 입력된 유전자 다형성 정보 및 임상 정보를 기초로, 상기 생성된 예후 예측 모델을 이용하여 예측된 예후 결과를 출력하는 단계를 더 포함할 수 있다.

발명의 효과

[0019] 이상과 같은 본 발명의 다양한 실시 예에 따르면, 유전자 다형성 정보 및 임상정보 중 예후 예측과의 연관성이 보다 검증된 정보가 포함된 개별 모델을 다양하게 확보할 수 있어, 데이터 수의 부족으로 인한 과적합 문제를 해결함과 동시에 예후 예측 모델 성능 향상의 한계를 뛰어넘을 수 있는 효과가 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시 예에 따른 예후 예측 방법을 수행할 수 있는 전자 장치의 구성을 도시한 블럭도,

도 2는 본 발명의 일 실시 예에 따른 예후 예측 방법을 설명하기 위한 흐름도,

도 3은 종래의 랜덤포레스트 방식으로 학습 데이터를 선별한 예를 도시한 도면,

도 4는 본 발명의 일 실시 예에 따라 학습 데이터를 선별한 예를 도시한 도면,

도 5는 본 발명의 일 실시 예에 따른 예후 예측 방법의 성능을 검증하는데 사용된 환자 데이터의 통계값을 도시한 도면, 그리고,

도 6은 예측 방법 성능 비교를 위한 Kaplan-Meier 생존분석 결과 값을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하, 본 문서의 다양한 실시 예가 첨부된 도면을 참조하여 기재된다. 그러나, 이는 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 문서의 실시 예의 다양한 변경(modifications), 균등물(equivalents), 및/또는 대체물(alternatives)을 포함하는 것으로 이해되어야 한다. 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.

[0022] 본 문서에서, "가진다," "가질 수 있다," "포함한다," 또는 "포함할 수 있다" 등의 표현은 해당 특징(예: 수치, 기능, 동작, 또는 부품 등의 구성요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[0023] 본 문서에서, "A 또는 B," "A 또는/및 B 중 적어도 하나," 또는 "A 또는/및 B 중 하나 또는 그 이상"등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다. 예를 들면, "A 또는 B," "A 및 B 중 적어도 하나," 또는 "A 또는 B 중 적어도 하나"는, (1) 적어도 하나의 A를 포함, (2) 적어도 하나의 B를 포함, 또는 (3) 적어도 하나의 A 및 적어도 하나의 B 모두를 포함하는 경우를 모두 지칭할 수 있다.본 문서에서 사용된 "제1," "제2," "첫째," 또는 "둘째,"등의 표현들은 다양한 구성요소들을, 순서 및/또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.

[0024] 본 문서에서 사용된 표현 "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, "~에 적합한(suitable for)," "~하는 능력을 가지는(having the capacity to)," "~하도록 설계된(designed to)," "~하도록 변경된(adapted to)," "~하도록 만들어진(made to)," 또는 "~를 할 수 있는(capable of)"과 바꾸어 사용될 수 있다. 용어 "~하도록 구성된(또는 설정된)"은 하드웨어적으로 "특별히 설계된(specifically designed to)" 것만을 반드시 의미하지 않을 수 있다. 대신, 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다. 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 부프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(generic-purpose processor)(예: CPU 또는 application processor)를 의미할 수 있다.

[0025] 이하의 설명에서 '예후'란 폐암과 같은 신생물 질환의 발병, 재발, 전이성 확산, 약물 내성, 폐암 기인성 사망, 폐암 기인성 사망으로의 진행 가능성, 병의 경과, 완치 여부를 포함하는 용어이다. 이하의 설명에서 '폐암'은 편평상피암, 편평세포암, 선암, 소세포암(small cell carcinoma)를 포함하는 용어이다. 이하의 설명에서 '다형성(polymorphism)'이란 유전적으로 결정된 집단 내에서 2 이상의 대체적 서열 또는 대립형질의 발생을 의미하는 용어이다.

[0026] 이하에서는 도면을 참조하여 본 발명에 대해 상세히 설명하기로 한다.

[0027] 도 1은 본 발명의 일 실시 예에 따른 예후 예측 방법을 수행할 수 있는 전자 장치(100)의 구성을 설명하기 위한 블록도이다. 도 1을 참조하면 전자 장치(100)는 입력부(110), 통신부(120), 메모리(130), 출력부(140), 프로세서(150)를 포함할 수 있다. 다만 상술한 모든 구성요소가 본 발명의 일 실시 예에 따른 예후 예측 방법을 수행할 때 필수적으로 필요한 것은 아니며, 상술한 구성요소 이외에도 다양한 구성요소들이 추가적으로 포함될 수도 있다.

[0028] 입력부(110)는 폐암 환자의 유전자 다형성 정보 및 임상 정보를 입력 받을 수 있다. 예를 들어, 입력부(110)는 키보드, 터치 스크린 등으로 구현되어 텍스트 형태로 사용자가 입력한 문장을 수신할 수 있다.

[0029] 통신부(120)는 외부 서버(200)와 통신을 수행할 수 있다. 통신부(120)는 외부 서버(200)로부터 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신할 수 있다. 통신부(120)는 다양한 유무선 통신 모듈을 포함할 수 있다. 예를 들어, 통신부(120)는 유선 LAN, 블루투스(Bluetooth), 지그비(Zigbee), WiFi, WiFi direct와 같은 방식으로 외부 네트워크에 연결되어 통신을 수행할 수 있다. 통신부(120)는 이 밖에 3G(3rd Generation), 3GPP(3rd Generation Partnership Project), LTE(Long Term Evoloution), LTE-A(LTE Advanced), 5G(5^{th} Generation) 등과 같은 다양한 이동 통신 규격에 따라 이동 통신망에 접속하여 통신을 수행하는 이동 통신 모듈을 더 포함할 수도 있다.

[0030] 입력부(110) 또는 통신부(120)를 통해 수신한 폐암 환자의 유전자 다형성 정보 및 임상 정보는 예후 예측 모델을 생성하는데 학습 데이터로서 이용될 수 있다. 또한 학습을 통해 예후 예측 모델이 생성된 이후에, 전자 장치(100)는 실제 예후 예측을 하고자 하는 환자의 유전자 다형성 정보 및 임상 정보를 입력부(110)나 통신부(120)를 통해 수신할 수 있다.

[0031] 메모리(130)는 전자 장치(100)를 구동하기 위한 다양한 모듈, 소프트웨어, 데이터를 저장할 수 있다. 예를

들어, 메모리(130)에는 환자의 유전자 다형성 정보, 임상 정보, 학습이 완료된 예후 예측 모델 등이 저장될 수 있다. 메모리(130)는 전자 장치(100)를 동작시키기 위해 필요한 각종 프로그램 등이 저장되는 저장매체로서, 플래쉬 메모리, HDD(Hard Disk Drive), SSD (Solid State Drive) 등의 형태로 구현 가능하다. 예를 들어, 메모리(130)는 전자 장치(100)의 동작 수행을 위한 프로그램을 저장하기 위한 ROM, 전자 장치(100)의 동작 수행에 따른 데이터를 일시적으로 저장하기 위한 RAM을 구비할 수 있다.

[0032] 출력부(140)는 생성된 예후 예측 모델을 이용하여 예측된 예후 결과를 출력할 수 있다. 출력부(140)는 디스플레이, 프린터, 스피커 등 다양한 형태로 구현될 수 있다. 예를 들어, 출력부(140)는 액정 표시 장치(Liquid Crystal Display, LCD), 유기 전기 발광 다이오드(Organic Light Emitting Display, OLED) 또는 플라즈마 표시 패널(Plasma Display Panel, PDP) 등으로 구현되어, 전자 장치(100)를 통해 제공되는 다양한 화면을 표시할 수 있다.

[0033] 프로세서(150)는 전자 장치(100)의 상술한 구성들을 제어할 수 있다. 예를 들어, 프로세서(150)는 학습에 사용될 복수의 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신하도록 통신부(120)를 제어할 수 있다. 프로세서(150)는 하나 또는 복수의 하드웨어 칩 형태로 제작되어 전자 장치(100)에 탑재될 수 있다. 예를 들어, 프로세서(150)는 인공지능을 위한 전용 하드웨어 칩 형태로 제작될 수도 있고, 기존의 범용 프로세서(예를 들어, CPU 또는 application processor)로 제작될 수도 있다.

[0034] 프로세서(150)는 입력되거나 수신된 유전자 다형성 정보와 임상 정보를 두 종류의 정보 그룹으로 분류할 수 있다. 프로세서(150)는 수신된 폐암 환자의 정보를 둘로 나누고, 그 중 하나의 그룹에 포함된 정보가 예후 예측 모델 생성을 위한 학습에 보다 많이 이용되도록 할 수 있다.

[0035] 임상 정보는 폐암 환자의 병리학적 종양 병기, 조직학적 유형, 나이, 성별, 흡연량 정보 중 적어도 하나일 수 있다. 유전자 다형성 정보는 CD3EAP, TNFRSF10B, AKT1, C3, HOMER2, GNB2L1, CD3D 및 ADAMTSL3 유전자의 다형성 정보 중 적어도 하나일 수 있다. 이 8개 유전자의 다형성 정보는 폐암 예후 예측과 유의미한 상관관계를 갖는다.

[0036] CD3EAP 유전자는 Genebank accession No. NT_011109.16으로 공지되어 있다. Genebank accession No. NT_011109.16 서열 중 18178152번째 염기(CD3EAP 유전자의 전사시작점으로부터 +468번째 염기)가 G인지 A인지에 대한 정보가 CD3EAP 유전자의 다형성 정보이다. 이는 rs967591 G/A로 명명된다.

[0037] TNFRSF10B 유전자는 Genebank accession No. NG_012145.1로 공지되어 있다. Genebank accession No. NG_012145.1 서열 중 31000번째 염기(TNFRSF10B 유전자의 전사시작점으로부터 +26000번째 염기)가 C인지 T인지에 대한 정보가 TNFRSF10B 유전자의 다형성 정보이다. 이는 rs1047266 C/T로 명명된다.

[0038] AKT1 유전자는 Genebank accession No. AL590327.3으로 공지되어 있다. Genebank accession No. AL590327.3 서열 중 20697번째 염기(AKT1 유전자의 전사시작점으로부터 -7699번째 염기)가 A인지 G인지에 대한 정보가 AKT1 유전자의 다형성 정보이다. 이는 rs3803300 A/G로 명명된다.

[0039] C3 유전자는 Genebank accession No. AY513239.1로 공지되어 있다. Genebank accession No. AY513239.1 서열 중 26076번째 염기(C3 유전자의 엑손 22의 전사종결점으로부터 +7번째 염기)가 T인지 C인지에 대한 정보가 C3 유전자의 다형성 정보이다. 이는 rs2287845 T/C로 명명된다.

[0040] HOMER2 유전자는 Genebank accession No. AC022558.9로 공지되어 있다. Genebank accession No. AC022558.9 서열 중 169850번째 염기(HOMER2 유전자의 전사시작점으로부터 +99659번째 염기, 엑손 7의 전사시작점으로부터 -814번째 염기)가 A인지 G인지에 대한 정보가 HOMER2 유전자의 다형성 정보이다. 이는 rs1256428 A/G로 명명된다.

[0041] GNB2L1 유전자는 Genebank accession No. NT_023133으로 공지되어 있다. Genebank accession No. NT_023133 서열 중 234232번째 염기(OGNB2L1 유전자의 전사시작점으로부터 -123번째 염기)가 T인지 G인지에 대한 정보가 GNB2L1 유전자의 다형성 정보이다. 이는 rs3756585 T/G로 명명된다.

[0042] CD3D 유전자는 Genebank accession No. NG_009891.1로 공지되어 있다. Genebank accession No. NG_009891.1 서열 중 4393번째 염기(CD3D 유전자의 전사시작점으로부터 -610번째 염기)가 C인지 T인지에 대한 정보가 CD3D 유전자의 다형성 정보이다. 이는 rs3181259 C/T로 명명된다.

[0043] ADAMTSL3 유전자는 Genebank accession No. NT_077661.3으로 공지되어 있다. Genebank accession No. NT_077661.3 서열 중 1686899번째 염기(ADAMTSL3 유전자의 전사시작점으로부터 +243707번째, 엑손 14의 전사시

작점으로부터 -66번째 염기)가 C인지 T인지에 대한 정보가 ADAMTSL3 유전자의 다형성 정보이다. 이는 rs11259927 C/T로 명명된다.

[0044] 본 발명의 일 실시 예에 따르면, 프로세서(150)는 중요도를 고려하여 임상정보 및 유전자 다형성 정보를 두 개의 정보 그룹으로 분류할 수 있다. 예를 들어, 프로세서(150)는 임상 정보 중 적어도 하나를 제1 정보 그룹으로 분류하고, 유전자 다형성 정보 및 제1 정보 그룹으로 분류되지 않은 임상 정보를 제2 정보 그룹으로 분류할 수 있다. 임상정보 및 유전자 다형성 정보를 두 개의 정보 그룹으로 분류하는 방법에 대해서는 이하에서 다시 상세히 설명하기로 한다.

[0045] 프로세서(150)는 수신되고 분류된 폐암 환자의 정보를 이용하여, 의사결정나무에 기반한 복수의 개별 모델을 생성할 수 있다. 구체적으로, 프로세서(150)는 폐암 환자의 정보 중 선택된 일부 정보만을 이용하여 각각의 개별 모델을 생성할 수 있다. 이와 같이 개별 모델의 다양성을 확보함으로써 데이터 부족으로 인한 단점을 극복할 수 있다. 그리고 프로세서(150)는 생성된 복수의 개별 모델을 종합하여, 최종적으로 앙상블 예후 예측 모델을 생성할 수 있다.

[0046] 복수의 개별 모델을 생성하기 위하여, 프로세서(150)는 각각의 개별 모델을 생성하는데 이용할 선택 정보 세트를 구성하여야 한다. 메모리(140)에는 수신된 복수의 환자에 대한 데이터가 저장되어 있고, 각각의 환자에 대한 데이터는 유전자 다형성 정보와 임상 정보로 구성된다. 프로세서(140)는 복수의 환자 중 일부 환자를 선택하고, 선택된 일부 환자 각각에 대해 유전자 다형성 정보 및 임상 정보 중 일부를 선택하여 선택 정보 세트를 구성할 수 있다.

[0047] 본 발명의 일 실시 예에 따르면, 프로세서(150)가 복수의 환자 중 일부 환자를 선택하는 방법에 대해서는 한정하지 않는다. 프로세서(150)는 단일 환자의 데이터(즉, 단일 환자의 유전자 다형성 정보 및 임상 정보) 중에서 학습에 이용할 정보를 선택할 때, 상술한 제1 정보 그룹과 제2 정보 그룹으로의 분류를 이용할 수 있다. 개별 모델의 다양성을 확보하면서도 개별 모델 생성에 주요 정보가 반영되어, 최종 앙상블 예후 예측 모델의 성능을 향상시키기 위함이다.

[0048] 예를 들어, 프로세서(150)는 중요도(importance) 값을 측정하여 유전자 다형성 정보 및 임상 정보 중 일부를 제1 정보 그룹으로, 나머지를 제2 정보 그룹으로 분류할 수 있다. 설명의 편의를 위해 유전자 다형성 정보 및 임상 정보를 바이오마커로 통칭하기로 한다. 중요도 값이란 특정 바이오마커가 예후 예측을 위한 모델 학습에 얼마나 관여하는지를 나타내는 지표를 말한다. 특정 바이오마커 A에 대한 중요도 값을 계산하는 방법은 다음과 같다. 우선 랜덤포레스트 모델을 만든 후 모델을 만들 때 사용하지 않았던 검증용 샘플을 이용해 모델의 정확도를 측정한다. 다음으로 검증용 샘플에서 바이오마커 A가 가지는 값을 무작위로 교환한 후, 교환된 데이터를 이용해 모델의 정확도를 다시 측정한다. 만일 바이오마커 A가 가지는 값을 교환한 후의 정확도가 크게 하락한다면, 이는 바이오마커 A가 사건 발생 판별에 중요하다는 의미가 된다. 반대로 교환 후의 정확도에 변화가 거의 발생하지 않거나 오히려 정확도가 상승한다면, 바이오마커 A는 모델 학습에 중요한 정보가 아니라는 의미가 된다. 따라서 특정 바이오마커 값을 교환하기 전과 후의 정확도 차이를 기준으로 중요도 값을 계산하여, 프로세서(150)는 계산된 중요도 값에 비례하게 바이오마커 선택에 가중치를 부여할 수 있다. 마찬가지로 프로세서(150)는 임계 값을 설정하고, 임계 값 이상의 중요도 값을 보이는 바이오마커는 제1 정보 그룹으로 분류하고, 임계 값 미만의 중요도 값을 보이는 바이오마커는 제2 정보 그룹으로 분류할 수 있다.

[0049] 다른 예로, 프로세서(150)는 사건 발생률 차이 값을 측정하여 유전자 다형성 정보 및 임상 정보 중 일부를 제1 정보 그룹으로, 나머지를 제2 정보 그룹으로 분류할 수 있다. 사건 발생률 차이 값에 대해서는 아래의 표 1을 참조하여 설명하기로 한다. 표 1은 3개의 클래스(class)를 갖는 바이오마커 A 및 2개의 클래스(class)를 갖는 바이오마커 B에 대하여, 사건 발생률 차이 값을 계산하는 예시를 보여준다.

표 1

[0050]

바이오마커	사건 발생률(%)	사건 발생률 차이 값
A1	a	$\frac{\lvert a-b\rvert+\lvert a-c\rvert+\lvert b-c\rvert}{3}$
A2	b	
A3	c	
B1	d	$\lvert d-e\rvert$
B2	e	

[0052] 바이오마커 A의 값은 1, 2, 3 중 하나일 수 있고, 바이오마커 B의 값은 1, 2 중 하나인 경우를 가정한다. 표 1의 구성을 설명하면, 바이오마커 A의 값이 1인 집단(A1)의 N년 후의 사건 발생률이 a%, 바이오마커 A의 값이 2인 집단(A2)의 N년 후 사건 발생률이 b%라는 의미이다. A3, B1, B2도 마찬가지로 해석하면 된다.

[0053] 학습에 유리한 바이오마커이기 위해서는 A1, A2, A3의 사건 발생률 차이가 커야 한다. 마찬가지로 B1, B2의 사건 발생률 차이가 크다면 바이오마커 B는 학습에 유리한 바이오마커라는 의미이다. 사건 발생률 차이 값은 표 1의 우측 열에 주어진 수식과 같이 계산될 수 있다. 바이오마커 A와 같이 클래스가 3개 이상인 경우에는 모든 조합에 대하여 차이 값을 평균하여 사건 발생률 차이 값을 구할 수 있다.

[0054] 만일 특정 바이오마커가 연속된 값을 갖는 경우에는 기설정된 값을 기준으로 2진화(binarize)하여 사건 발생률 차이 값을 도출할 수 있다. 예를 들어 기설정된 값은 중위 값으로 하고, 중위 값 이상인 경우를 클래스 1, 중위 값 미만인 경우를 클래스 2로 구분할 수 있다. 각각의 클래스에 따라 사건 발생률이 도출될 것이고, 이를 통해 사건 발생률 차이 값을 구할 수 있다.

[0055] 표 1의 예시에서는 바이오마커 A 및 B에 대해서 기술하였으나, 실제 적용에 있어서는 유전자 다형성 정보 및 임상 정보 모두에 대해 사건 발생률 차이 값을 구할 수 있다. 프로세서(150)는 이렇게 도출된 사건 발생률 차이 값은 각각의 바이오마커에 대한 가중치로 이용할 수 있다. 사건 발생률 차이 값의 순위를 가중치로 이용할 수도 있으며, 정규화된 값을 가중치로 이용할 수도 있다. 또한 프로세서(150)는 임계 값을 설정하고, 임계 값 이상의 사건 발생률 차이 값을 보이는 바이오마커는 제1 정보 그룹으로 분류하고, 임계 값 미만의 사건 발생률 차이 값을 보이는 바이오마커는 제2 정보 그룹으로 분류할 수 있다.

[0056] 또 다른 예로, 프로세서(150)는 AUC(Area Under Curve) 값을 이용하여 유전자 다형성 정보 및 임상 정보 중 일부를 제1 정보 그룹으로, 나머지를 제2 정보 그룹으로 분류할 수 있다. 프로세서(150)는 학습에 이용될 수 있는 유전자 다형성 정보 및 임상 정보 각각에 대해 개별적으로 단변량 선형 회귀 모델을 만들어 예후 예측 모델을 만들 수 있다. 학습이 완료된 모델을 바탕으로 학습 데이터에 대한 예측 유무를 이용하여 수신자 조작 특성(Receiver Operating Characteristic, ROC) 곡선을 그리고 AUC 값을 획득할 수 있다. AUC 값은 선형 모델을 통해 데이터가 사건 발생 유무에 대해 얼마나 잘 분리되는지를 나타내는 지표이다 AUC 값은 0.5~1 사이의 값을 갖는다. 값이 1에 가까울수록 모델이 사건 발생 유무를 잘 구분한다는 의미로, AUC 값이 높은 바이오마커가 학습에 보다 유의미한 바이오마커임을 알 수 있다. 프로세서(150)는 AUC 값에 비례하게 바이오마커에 가중치를 부여할 수 있다. 또한 프로세서(150)는 임계 값을 설정하고, 임계 값 이상의 AUC 값을 보이는 바이오마커는 제1 정보 그룹으로 분류하고, 임계 값 미만의 AUC 값을 보이는 바이오마커는 제2 정보 그룹으로 분류할 수 있다.

[0057] 또 다른 예로, 프로세서(150)는 윌콕슨 순위 합 검증을 이용하여 유전자 다형성 정보 및 임상 정보 중 일부를 제1 정보 그룹으로, 나머지를 제2 정보 그룹으로 분류하거나, 가중치를 부여할 수 있다. 윌콕스 순위 합 검증이란 사건이 발생한 환자 집단과 사건이 발생하지 않은 환자 집단에 대하여 각 집단 특징의 중앙값이 동일하다는 가설을 검증하는 방법이다. 두 집단의 분포 차이를 계산하여 가설에 대한 유의 확률(p-value)을 얻을 수 있고, 유의 확률이 작은 값일수록 두 집단의 중앙값이 다르다는 것을 의미한다. 즉 이는 사건 발생 집단과 미발생 집단의 분포의 차이 정도를 나타내는 지표로, 유의 확률이 낮을수록 생존 예후 예측 모델을 생성할 때 보다 효과적인 바이오마커로 작용할 수 있다는 의미이다. 프로세서(150)는 사용하는 모든 바이오마커에 대해 윌콕슨 순위 합 검증에 대한 유의 확률을 계산할 수 있다. 그리고 프로세서(150)는 계산된 유의 확률과 반비례한 형태로 각각의 바이오마커에 가중치를 부여할 수 있다. 또한 프로세서(150)는 임계 값을 설정하고, 임계 값 이하의 유의 확률 값을 보이는 바이오마커는 제1 정보 그룹으로 분류하고, 임계 값을 초과하는 유의 확률 값을 보이는 바이오마커는 제2 정보 그룹으로 분류할 수 있다.

[0058] 상술한 윌콕슨 순위 합 검증은 특징 선택 방법의 하나로, 윌콕슨 순위 합 검증 이외에 Fisher Score, Relief, Chi-square, Joint Mutual Information(JMI), Conditional Infomax Feature Extraction(CIFE), Double Input Symmetric Relevance(DISR), Mutual Information Maximization(MIM), Conditional Mutual Information Maximization(CMIM), Interaction Capping, T-test Score, Minimum Redundancy Maximum Relevance(MRMR), Mutual Information Feature Selection(MIFS), Least Absolute Shirinkage and Selection Operator(LASSO)와 같은 다른 방법도 이용할 수 있다.

[0059] 상술한 다양한 정보 분류 방법은 전체 샘플(모든 환자의 바이오마커)을 대상으로 하여 각각의 바이오마커에 대

한 가중치를 획득하거나, 획득된 가중치 값을 이용하여 각각의 바이오마커가 속하는 정보 그룹을 결정할 수 있다는 내용을 설명한 것이다. 하지만 본 발명의 실시 예는 전체 샘플을 이용하는 경우로만 한정되지 않는다. 본 발명의 일 실시 예에 따르면, 프로세서(150)는 전체 샘플 중 일부를 랜덤하게 추출하여 얻은 서브 샘플 세트를 구성할 수 있다. 그리고 프로세서(150)는 구성된 복수의 서브 샘플 각각에 대해 상술한 다양한 정보 분류 방법을 이용해 가중치를 획득할 수 있다. 예를 들어, 프로세서(150)는 전체 N 명의 환자 데이터 중 (N-x) 명에 대한 데이터로 구성된 서브 샘플 세트를 M개 구성할 수 있다. 그리고 프로세서(150)는 각각의 서브 샘플 세트에 대해 바이오마커 별로 사건 발생률 차이 값(이는 예시이며 사건 발생률 차이 값 대신에 중요도 값, AUC 값, 유의 확률 값을 이용하는 것도 당연히 가능하다)을 도출하여 가중치를 결정할 수 있다. 프로세서(150)는 M개의 결정된 가중치 값을 평균하여 해당 바이오마커에 대한 최종적인 가중치를 도출할 수 있다. 서브 샘플 세트를 구성하고 그에 대한 가중치를 구하는 과정을 반복함으로써, 전체 샘플(학습 데이터)에 과적합(overfitting)되는 것을 방지할 수 있는 효과가 있다.

[0060] 프로세서(150)는 상술한 다양한 정보 분류 방법 중 적어도 하나를 이용하여 유전자 다형성 정보 및 임상 정보를 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다. 이어서 프로세서(150)는 분류된 정보 그룹을 이용하여 학습에 이용할 정보를 선택할 수 있다.

[0061] 예를 들어, 프로세서(150)는 제1 정보 그룹으로 분류된 정보는 모두 학습에 이용할 정보로 선택하고, 제2 정보 그룹으로 분류된 정보 중 일부를 랜덤하게 학습에 이용할 정보로 선택할 수 있다.

[0062] 다른 예로, 프로세서(150)는 제1 정보 그룹으로 분류된 정보에 가중치를 부여하고, 부여된 가중치를 반영하여 유전자 다형성 정보 및 임상 정보 중 일부를 학습에 이용할 정보로 선택할 수 있다. 가중치는 상술한 정보 분류 방법에서 도출된 파라미터들을 이용할 수 있음은 물론이다.

[0063] 또 다른 예로는 프로세서(150)가 정보 선택 과정에서는 랜덤하게 학습에 이용할 정보를 선택하되, 생성된 개별 모델 각각에 대해 다르게 가중치를 부여하여 앙상블 예후 예측 모델을 생성할 수도 있다. 구체적으로, 프로세서(150)는 개별 모델 각각을 생성할 때 이용한 정보 중 제1 정보 그룹에 속하는 정보의 수에 비례하게 개별 모델에 대해 가중치를 부여할 수 있다. 즉, 정보의 종류에 따라 선택될 가중치를 부여하는 것이 아닌, 생성된 개별 모델이 앙상블 모델에 기여하는 정도에 대한 가중치를 부여하는 것이다.

[0064] 최종적으로 예후 예측 모델이 생성된 이후에, 프로세서(150)는 학습 완료 모델을 이용하여 폐암 환자의 예후 예측을 수행할 수 있다. 프로세서(150)는 입력부(110)나 통신부(120)를 통해 예후 예측을 수행할 환자의 데이터(임상 정보 및 유전자 다형성 정보)를 입력 받을 수 있다. 그리고 입력된 환자의 데이터를 기초로, 프로세서(150)는 학습 완료 모델을 이용하여 예측된 예후 결과를 출력하도록 출력부(140)를 제어할 수 있다.

[0065] 도 2는 본 발명의 일 실시 예에 따른 폐암 예후 예측 방법을 설명하기 위한 흐름도이다. 이하에서는 도 2를 참조하여 예후 예측 모델 생성을 중심으로 본 발명을 설명하기로 한다.

[0066] 우선 전자 장치(100)는 복수의 폐암 환자의 유전자 다형성 정보 및 임상 정보를 수신할 수 있다(S210). S210 단계에서 수신하는 정보는 예측 모델의 학습에 사용되는 정보에 해당한다.

[0067] 그리고 전자 장치(100)는 수신된 유전자 다형성 정보 및 임상 정보를 제1 정보 그룹과 제2 정보 그룹으로 분류할 수 있다(S220). 예를 들어, 제1 정보 그룹은 통계적이나 임상적으로 폐암의 예후에 자주 사용되는 정보, 상관관계의 유효성이 높다고 판단되는 정보일 수 있다. 즉 제1 정보 그룹에 속하는 정보는 예후 예측 모델을 학습할 때 더 잦은 빈도로 선택되는 것이 바람직한 정보일 수 있다. 이러한 분류는 중요도 값, 사건 발생률 차이 값, AUC 값, 유의 확률 값 등에 기초하여 결정될 수 있다.

[0068] 다만 제2 정보 그룹에 속하는 정보가 폐암 예후 예측과 통계적, 임상적으로 검증되지 않았다는 의미가 아니라는 점에 유의해야 할 것이다. 제2 정보 그룹에 속하는 정보 역시 폐암 예후 예측과의 상관관계가 인정되는 정보이나, 제1 정보 그룹에 속하는 정보보다 사용 빈도나 통계적 유의성이 낮을 뿐이다.

[0069] 이어서 전자 장치(100)는 복수의 폐암 환자 중 일부 환자를 선택하고, 선택된 일부 환자 각각에 대해 유전자 다형성 정보 및 임상 정보 중 일부를 선택할 수 있다(S230). 예를 들어, 전자 장치(100)는 제1 정보 그룹으로 분류된 정보를 학습에 이용할 데이터로 반드시 선택하거나, 제1 정보 그룹으로 분류된 정보에 가중치를 부여하여 더 높은 빈도로 선택되도록 할 수 있다. 폐암 환자의 정보 중 일부 정보가 더 많이 학습 데이터로 선택되도록 해야 하는 이유는 다음과 같다.

[0070] 본 발명의 일 실시 예에 따르면 폐암 환자의 예후 예측을 위해 임상 정보인 폐암 환자의 병리학적 종양 병기,

조직학적 유형, 나이, 성별, 흡연량 정보와 유전자 다형성 정보인 CD3EAP, TNFRSF10B, AKT1, C3, HOMER2, GNB2L1, CD3D 및 ADAMTSL3 유전자의 다형성 정보 중 일부를 이용하여 예후 예측 모델을 학습시킬 수 있다. 본 발명이 적용될 수 있는 기계학습 모델의 종류는 한정되지 않으며, 다양한 기계학습/딥러닝 기반의 앙상블 학습 모델을 생성할 때 본 발명의 방법이 사용될 수 있다. 이하에서는 설명의 편의를 위해 앙상블 기법을 이용하는 랜덤포레스트 알고리즘을 기준으로 설명하기로 한다. 랜덤포레스트 알고리즘에서는 학습에 사용할 정보(특징)을 랜덤하게 선택하는 방식으로 다수의 개별 모델을 만들고, 이들 다수의 개별 모델을 앙상블 기법으로 통합하여 최종 예후 예측 모델을 생성한다.

[0071] 종래의 랜덤포레스트 방식에서는 학습을 통해 예후 예측 모델을 생성할 때, 모든 노드의 결정(decision, split, 분류, 의사결정) 과정에서 모든 환자 및 모든 정보 중 일부를 랜덤하게 선택한 후, 가장 분류에 적합하다고 판단되는 특징 하나를 이용하여 결정을 진행한다. 도 3은 종래의 랜덤포레스트 방식으로 한 번의 결정에 사용할 학습 데이터를 선별한 예시이다. 7명의 환자 각각에 대해 11가지 정보를 가진 데이터베이스에서 두 차례 랜덤한 선택을 통해 두 번의 결정에 사용될 데이터를 선택한 것이다. Dominant로 분류된 정보는 제1 정보 그룹에 속하는 정보로 볼 수 있고, Non-Dominant로 분류된 정보를 제2 정보 그룹에 속하는 정보로 볼 수 있다.

[0072] 도 3에서 제1 선택인 좌측 표에서는 7명의 환자 중 1, 2, 5, 6번째 환자를 선택하였고, 선택된 4명의 환자 데이터 중에서 2, 4, 8, 9번째 정보를 선택하였다. 그리고 제2 선택인 우측 표에서는 7명의 환자 중 3, 4, 5, 7번째 환자를 선택하였고, 선택된 4명의 환자 데이터 중에서 1, 5, 7, 10번째 정보를 선택하였다.

[0073] 이렇게 모든 정보 중 일부를 랜덤하게 선택하는 경우에는, 예후 예측에 상대적으로 크게 도움이 되는 유효성이 검증된 Dominant 특징을 사용하는 빈도가 줄어들게 된다. 특히 Non-Dominant 특징이 Dominant 특징의 수보다 많은 경우에 이러한 현상이 빈번하게 나타난다. 이에 따라 전체적인 예측 모델의 성능 향상에 한계가 있으며, 더 나아가 학습을 하면 할수록 성능이 저하될 가능성도 배제할 수 없다.

[0074] 이러한 문제점을 개선하기 위하여 본 발명의 일 실시 예에 따른 예후 예측 모델을 학습할 때는 Dominant 특징이 높은 빈도로 선택될 수 있도록 하였다. 도 4는 본 발명의 다양한 실시 예 중에서 제1 정보 그룹에 속하는 정보는 반드시 선택되도록 하고, 제2 정보 그룹에 속하는 정보는 일부를 랜덤하게 선택되게 하는 방식을 적용한 예시이다.

[0075] 도 4 역시 도 3과 마찬가지로 7명의 환자 각각에 대해 11가지 정보를 가진 데이터베이스에서 개별 모델 생성 과정 중 두 번의 결정 과정에서 사용될 데이터를 선택한 것이다. 제1 선택인 좌측 표에서는 7명의 환자 중 1, 2, 5, 6 번째 환자를 선택하였고, 선택된 4명의 환자 데이터 중에서 Dominant 특징에 해당하는 1, 2, 3번째 정보를 선택하고, Non-Dominant 특징에 해당하는 정보 중에서 랜덤하게 4, 8, 9번째 정보를 선택하였다. 그리고 제2 선택인 우측 표에서는 7명의 환자 중 3, 4, 5, 7번째 환자를 선택하였고, 선택된 4명의 환자 데이터 중에서 Dominant 특징에 해당하는 1, 2, 3번째 정보를 선택하고, Non-Dominant 특징에 해당하는 정보 중에서 랜덤하게 5, 7, 10번째 정보를 선택하였다. 즉 본 발명의 일 실시 예에 따를 경우에는 예후 예측에 상대적으로 크게 도움이 되는 유효성이 검증된 Dominant 특징을 사용하여 학습한 개별 모델이 많아지고, 학습을 하면 할수록 성능이 향상될 것이라는 점을 보장할 수 있게 된다.

[0076] 전자 장치(100)는 이렇게 선택하는 단계를 반복하여 복수의 선택 정보 세트를 구성할 수 있다(S240). 앙상블 학습 모델을 이용하기 위해서는 복수의 개별 모델이 필요하고, 복수의 개별 모델을 생성하기 위해서 복수의 선택 정보 세트가 필요하기 때문이다.

[0077] 이어서 전자 장치(100)는 구성된 복수의 선택 정보 세트를 기초로 예후 예측 모델을 생성할 수 있다(S250). 구체적으로, 전자 장치(100)는 복수의 선택 정보 세트 각각을 기초로 의사결정나무에 기반한 복수의 개별 모델을 생성할 수 있다. 각각의 선택 정보 세트는 의사결정나무의 각 결정 과정에 사용된다. 그리고 전자 장치(100)는 생성된 복수의 개별 모델을 종합하여 앙상블 예후 예측 모델을 생성할 수 있다.

[0078] 도 2에 도시하지는 않았으나, 상술한 과정을 통해 예후 예측 모델이 생성되면, 전자 장치(100)는 이러한 학습 완료 모델을 이용하여 폐암 환자의 예후를 예측할 수 있다. 전자 장치(100)는 예후를 예측할 폐암 환자의 유전자 다형성 정보 및 임상 정보를 입력 받을 수 있다. 그리고 전자 장치(100)는 입력된 유전자 다형성 정보 및 임상 정보를 기초로, 생성된 예후 예측 모델을 이용하여 예측된 예후 결과를 도출할 수 있다. 이어서 전자 장치(100)는 도출된 예후 결과를 다양한 방식으로 출력하여 사용자에게 제공할 수 있다.

[0079] 이하에서는 본 발명의 일 실시 예에 따른 폐암 예후 예측 모델을 생성하고, 생성된 예측 모델을 기반으로 향상된 효과를 검증한 연구 내용을 설명한다.

[0080] <1-1> 연구대상의 선정

[0081] 본 연구에서는 암과 관련된 유전자의 다형성과 폐암 수술 후 예후와의 관계를 다기관 임상코호트를 통해 평가하였다. 서울대학교 분당병원에서 62례, 전남대학교 병원에서 245례, 경북대학교 병원에서 126례, 계명대학교 병원에서 33례, 아산병원에서 42례를 합하여 전체 508명의 환자를 대상으로 하였다. 그 중 세계보건기구 분류에 따라 조직학적 유형이 선암(adenocarcinomas, ACs)인 환자와 편평세포암(squamous cell carcinomas, SQs)인 환자의 데이터를 이용하였다. 또한 폐암 병기에 대한 국제 시스템(International System for Staging Lung Cancer)에 따라 종양의 병리학적 병기가 2기 이하인 환자의 데이터를 이용하였다. 또한 5년 후 생존 여부가 확인 가능한 환자의 데이터를 이용하였다.

[0082] 도 5는 본 연구에서 사용한 데이터에 대한 통계치를 나타낸다. 본 연구에서는 5년을 기준으로 한 예후 예측을 진행하였다.

[0083] <1-2> 예측 모델 생성

[0084] 본 연구에서는 예후 예측 모델의 성능 차이를 검증하기 위하여 3가지 종류의 예측 모델을 생성하였다. (1)학습 데이터로 임상정보만을 사용하였으며 종래의 랜덤포레스트 방식으로 생성한 예측 모델, (2) 학습 데이터로 임상정보 및 8개의 유전자 다형성 정보를 사용하였으며 종래의 랜덤포레스트 방식으로 생성한 예측 모델, (3)본 발명의 일 실시 예에 따라 생성한 예측 모델이 그것이다. (3)본 발명의 일 실시 예에 따라 생성한 예측 모델은 학습 데이터로 임상정보 및 8개의 유전자 다형성 정보를 사용하였으며, 임상정보 중 병리학적 종양 병기, 조직학적 유형, 나이를 제1 정보 그룹으로 설정하여 학습하였다.

[0085] <1-3> 통계분석

[0086] 앞서 설명한 바와 같이, 본 연구에서 예측 모델 학습에 사용한 정보는 8개 유전자(CD3EAP, TNFRSF10B, AKT1, C3, HOMER2, GNB2L1, CD3D, ADAMTSL3)의 다형성 정보 및 5개 임상정보(병리학적 종양 병기, 조직학적 유형, 나이, 성별, 흡연량)이다. 예측 모델 생성에 이어 본 연구에서는 예측 모델 성능 평가를 진행하였다.

[0087] 성능 평가에는 학습에 사용된 트레이닝 코호트(training cohort)와 평가를 위한 검증 코호트(validation cohort)를 통해 평가하였다. 구체적으로 트레이닝 코호트는 서울대학교 분당병원에서 62례, 경북대학교 병원에서 126례, 계명대학교 병원에서 33례, 아산병원에서 42례를 합하여 263명의 환자의 데이터로 구성하였다. 그리고 검증 코호트는 전남대학교 병원로부터 획득한 245례의 환자 데이터로 구성하였다.

[0088] 예측 모델에 대한 성능은 Area Under the Curve(AUC) 값을 기준으로 평가하였다. 전체 생존(overall survival, OS)은 수술을 한 날부터 사망일 혹은 최종 확인일로 정의하였다. 이번 성능 평가에서는 본 발명의 일 실시 예에 따른 예측 모델과 종래의 랜덤포레스트 방식으로 생성한 예측 모델을 비교하였다. 비교 결과 본 발명의 일 실시 예에 따른 예측 모델의 성능이 개선된 것을 확인할 수 있었다.

[0089] 트레이닝 코호트를 구성하는 263명의 환자 중 사망자는 144명으로 생존율은 45.25%이다. 검증 코호트를 구성하는 245명의 환자 중 사망자는 111명으로 생존율은 54.69%이다. 아래의 표 2는 (1)임상정보만을 이용하여 종래의 랜덤포레스트 방식으로 생성한 예측 모델을 이용한 경우, (2)임상정보 및 8개의 유전자 정보를 이용하여 종래의 랜덤포레스트 방식으로 생성한 예측 모델을 이용한 경우, (3)본 발명의 일 실시 예에 따른 예측 모델을 이용한 경우 각각에 대한 성능 검증 표이다. AUC 결과 수치는 동일 실험을 301번 반복한 통계 값이며, AUC 값에는 반복한 결과에 대한 표준편차를 기입하였다. AUC는 값이 높을수록 예측 모델이 보다 정확함을 나타내는 지표이다.

표 2

[0090]

예측 모델 (예측인자)	data	예측 성능 AUC(표준편차)
종래 랜덤포레스트 (임상정보)	트레이닝 코호트	0.709(0.016)
	검증 코호트	0.743(0.003)
종래 랜덤포레스트 (임상정보 및 유전자 정보)	트레이닝 코호트	0.711(0.014)
	검증 코호트	0.749(0.005)
본 발명의 예측 모델 (임상정보 및 유전자 정보)	트레이닝 코호트	0.718(0.012)
	검증 코호트	0.758(0.005)

[0092] 임상정보만을 이용하여 종래의 랜덤포레스트 방법으로 학습한 결과 트레이닝 코호트에 대해서는 AUC 0.709(표준편차 0.016), 검증 코호트에 대해서는 AUC 0.743(표준편차 0.003)을 얻었다. 종래의 랜덤포레스트 방법을 이용하는 경우에도 8개의 유전자 정보를 함께 이용하였을 때 트레이닝 코호트와 검증 코호트 모두에 대해 AUC가 상승하는 결과를 얻었다.

[0093] 특히 본 발명의 일 실시 예에 따른 예측 모델을 이용한 경우에 트레이닝 코호트와 검증 코호트 모두에서 AUC가 더욱 상승하는 결과를 얻었다. 본 발명의 일 실시 예에 따른 예측 모델을 사용할 경우 추가적인 성능 향상이 가능함이 검증된 것이다.

[0094] AUC 결과 예측을 통한 검증에 더하여, 본 발명의 일 실시 예에 따른 예측 모델이 고위험군과 저위험군을 구분하는 성능도 향상됨을 검증하였다. 고위험군과 저위험군 구분 성능 검증은 다음의 방법으로 수행하였다.

[0095] 우선 상술한 검증 코호트를 구성하는 245명의 환자를 저위험군과 고위험군으로 분류하였다. 301번 반복한 실험에서 예측 모델의 각각의 트리는 생존 혹은 사망에 대한 최종 결과를 나타낸다. 301개 트리의 예측 결과에 대해 150개 이하의 트리가 사망했다고 판단한 환자는 저위험군으로, 151개 이상의 트리가 사망했다고 판단한 환자는 고위험군으로 설정하였다. 예측 모델이 정확하다면 고위험군으로 분류된 환자의 생존율은 기간이 지남에 따라 보다 빠르게 줄어들 것이고, 저위험군으로 분류된 환자의 생존율은 보다 서서히 줄어들 것이다. 따라서 기간이 지남에 따라 고위험군과 저위험군의 생존율 격차가 클수록 예측 모델의 정확도가 높다고 볼 수 있다.

[0096] 이러한 기준에 따라 본 검증 실험에서는 (1)임상정보만을 이용하여 종래의 랜덤포레스트 방식으로 생성한 예측 모델을 이용한 경우와, (3)본 발명의 일 실시 예에 따른 예측 모델을 이용한 경우 각각에 대해 저위험군과 고위험군으로 분류를 진행하였다. 그리고 각 예측 모델 및 각 군에 대한 Kaplan-Meier 생존 곡선을 도 6에 도시하였다.

[0097] Kaplan-Meier 생존분석은 생존율을 산출하는 대표적인 방법으로 누적한계추정법으로도 불린다. Kaplan-Meier 생존분석은 사건(사망)이 발생한 시점마다 구간 생존율을 구하고, 이들의 누적으로 누적생존율을 추정한다. 도 6은 각 예측 모델 및 각 군에 대해 Kaplan-Meier 생존분석을 한 결과 값을 도시한 것이다. 이에 따라 도 6에는 4개의 선(저위험군-모델(1), 저위험군-모델(3), 고위험군-모델(1), 고위험군-모델(3))이 표시되었다.

[0098] 도 6에서 파란 색 선은 (1)종래의 방식으로 생성한 예측 모델에 대한 결과를 나타내고, 빨간 색 선은 (3)본 발명의 일 실시 예에 따른 예측 모델에 대한 결과를 나타낸다. 앞서 설명한 바와 같이 각각의 예측 모델에서 도출한 고위험군의 생존율과 저위험군의 생존율의 차가 클수록, 예측 모델의 위험도 구분 성능이 뛰어나다고 볼 수 있다. 도 6를 참조하면 본 발명의 일 실시 예에 따른 예측 모델이 종래의 방식으로 생성한 예측 모델보다 생존율 차이가 큰 것으로 예측하고 있기 때문에 저위험군과 고위험군을 보다 잘 구분하고 있음을 확인할 수 있다.

[0099] 상술한 바와 같은 본 발명의 다양한 실시 예에 따르면, 유전자 다형성 정보와 임상 정보를 선별적으로 이용함으로써 예후 예측 모델의 정확도를 향상시킬 수 있다. 정확도 향상은 실제 환자의 데이터를 이용한 실험을 통해 충분히 검증되었다. 본 발명의 다양한 실시 예에 따른 예측 모델을 통해 환자의 예후를 정확히 진단함으로써 개별 맞춤형 의료 방법을 제공하는 정밀 의료가 가능해진다. 종국적으로는 본 발명을 통해 폐암 환자의 생존율을 높이는데 기여할 수 있을 것이다.

[0100] 한편, 본 개시에서 사용된 용어 "부" 또는 "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구성된 유닛을 포함하며, 예를 들면, 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. "부" 또는 "모듈"은, 일체로 구성된 부품 또는 하나 또는 그 이상의 기능을 수행하는 최소 단위 또는 그 일부가 될 수 있다. 예를 들면, 모듈은 ASIC(application-specific integrated circuit)으로 구성될 수 있다.

[0101] 본 개시의 다양한 실시 예들은 기기(machine)(예: 컴퓨터)로 읽을 수 있는 저장 매체(machine-readable storage media에 저장된 명령어를 포함하는 소프트웨어로 구현될 수 있다. 기기는, 저장 매체로부터 저장된 명령어를 호출하고, 호출된 명령어에 따라 동작이 가능한 장치로서, 개시된 실시 예들에 따른 전자 장치(예: 전자 장치(100))를 포함할 수 있다. 상기 명령이 프로세서에 의해 실행될 경우, 프로세서가 직접, 또는 상기 프로세서의 제어 하에 다른 구성요소들을 이용하여 상기 명령에 해당하는 기능을 수행할 수 있다. 명령은 컴파일러 또는 인터프리터에 의해 생성 또는 실행되는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장매체가 신호(signal)를 포함하지 않으며 실재(tangible)한다는 것을 의미할 뿐 데이터가 저장매체에 반영구적 또는 임시적으로 저장됨을 구분하지 않는다.

[0102] 일 실시 예에 따르면, 본 문서에 개시된 다양한 실시 예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로, 또는 어플리케이션 스토어(예: 플레이 스토어TM)를 통해 온라인으로 배포될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

[0103] 다양한 실시 예들에 따른 구성 요소(예: 모듈 또는 프로그램) 각각은 단수 또는 복수의 개체로 구성될 수 있으며, 전술한 해당 서브 구성 요소들 중 일부 서브 구성 요소가 생략되거나, 또는 다른 서브 구성 요소가 다양한 실시 예에 더 포함될 수 있다. 대체적으로 또는 추가적으로, 일부 구성 요소들(예: 모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성 요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 다양한 실시 예들에 따른, 모듈, 프로그램 또는 다른 구성 요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이 추가될 수 있다.

부호의 설명

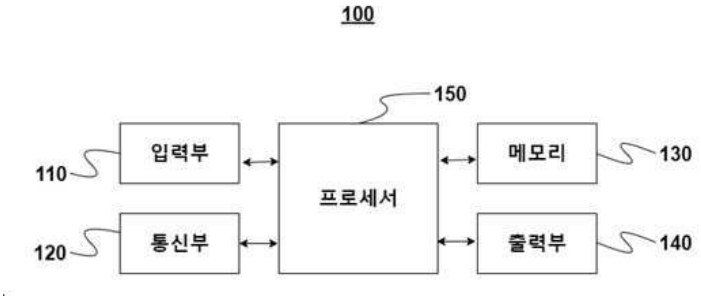
[0105] 100: 전자 장치 110: 입력부

120: 통신부 130: 메모리

140: 출력부 150: 프로세서

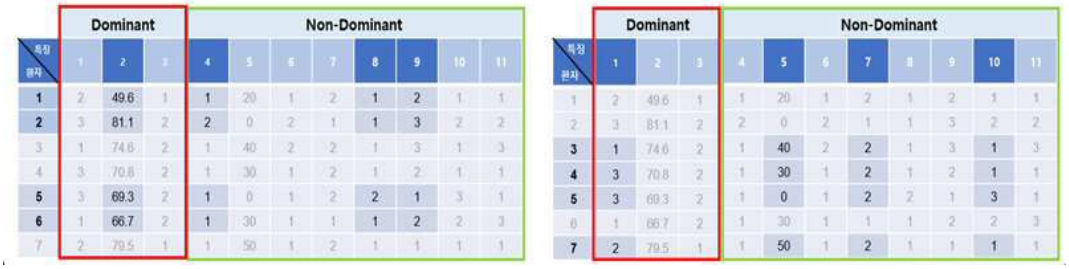
도면

도면1



도면2

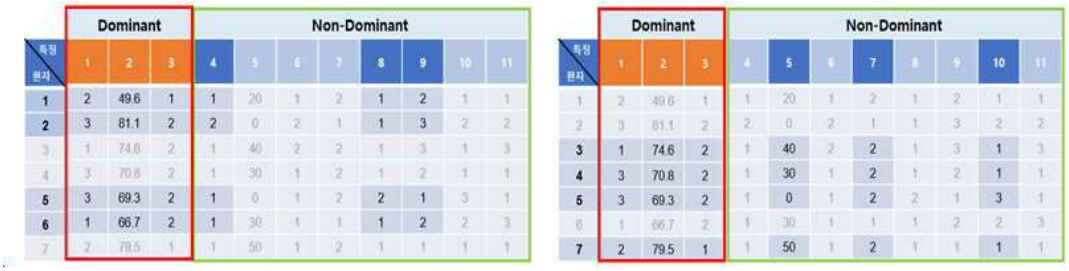
도면3



특징 / 환자	Dominant			Non-Dominant							
	1	2	3	4	5	6	7	8	9	10	11
1	2	49.6	1	1	20	1	2	1	2	1	1
2	3	81.1	2	2	0	2	1	1	3	2	2
3	1	74.6	2	1	40	2	2	1	3	1	3
4	3	70.8	2	1	30	1	2	1	2	1	1
5	3	69.3	2	1	0	1	2	2	1	3	1
6	1	66.7	2	1	30	1	1	1	2	2	3
7	2	79.5	1	1	50	1	2	1	1	1	1

특징 / 환자	Dominant			Non-Dominant							
	1	2	3	4	5	6	7	8	9	10	11
1	2	49.6	1	1	20	1	2	1	2	1	1
2	3	81.1	2	2	0	2	1	1	3	2	2
3	1	74.6	2	1	40	2	2	1	3	1	3
4	3	70.8	2	1	30	1	2	1	2	1	1
5	3	69.3	2	1	0	1	2	2	1	3	1
6	1	66.7	2	1	30	1	1	1	2	2	3
7	2	79.5	1	1	50	1	2	1	1	1	1

도면4



특징 / 환자	Dominant			Non-Dominant							
	1	2	3	4	5	6	7	8	9	10	11
1	2	49.6	1	1	20	1	2	1	2	1	1
2	3	81.1	2	2	0	2	1	1	3	2	2
3	1	74.6	2	1	40	2	2	1	3	1	3
4	3	70.8	2	1	30	1	2	1	2	1	1
5	3	69.3	2	1	0	1	2	2	1	3	1
6	1	66.7	2	1	30	1	1	1	2	2	3
7	2	79.5	1	1	50	1	2	1	1	1	1

특징 / 환자	Dominant			Non-Dominant							
	1	2	3	4	5	6	7	8	9	10	11
1	2	49.6	1	1	20	1	2	1	2	1	1
2	3	81.1	2	2	0	2	1	1	3	2	2
3	1	74.6	2	1	40	2	2	1	3	1	3
4	3	70.8	2	1	30	1	2	1	2	1	1
5	3	69.3	2	1	0	1	2	2	1	3	1
6	1	66.7	2	1	30	1	1	1	2	2	3
7	2	79.5	1	1	50	1	2	1	1	1	1

도면5

	5년 기준			
	환자수(비율)	생존율(%)	환자수(비율)	무병 생존율(%)
전체	508	49.80	544	31.62
나이				
64이하	245 (48.23)	65.31	274 (50.37)	39.42
64초과	263 (51.77)	35.36	270 (49.63)	23.70
성별				
남자	381 (75.00)	43.83	387 (71.14)	29.20
여자	127 (25.00)	67.72	157 (28.86)	37.58
흡연 상태				
비 흡연	137 (26.97)	62.04	169 (31.07)	33.73
흡연	371 (73.03)	45.28	375 (68.93)	30.67
흡연량				
30미만	230 (45.28)	58.70	267 (49.08)	33.33
30이상	278 (54.72)	42.45	277 (50.92)	29.96
Histological				
SCC	251 (49.41)	43.03	244 (44.85)	30.33
AC	257 (50.59)	56.42	300 (55.15)	32.67
Stage				
IA	148 (29.13)	67.57	152 (27.94)	50.66
IB	210 (41.34)	52.38	213 (39.15)	30.99
IIA	86 (16.93)	27.91	100 (18.38)	15.00
IIB	64 (12.60)	29.69	79 (14.52)	17.72

도면6