



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0107535
(43) 공개일자 2024년07월09일

(51) 국제특허분류(Int. Cl.)
G16H 50/20 (2018.01) G06N 20/00 (2019.01)
G06N 3/08 (2023.01) G16H 50/50 (2018.01)
(52) CPC특허분류
G16H 50/20 (2018.01)
G06N 20/00 (2021.08)
(21) 출원번호 10-2022-0190278
(22) 출원일자 2022년12월30일
심사청구일자 2022년12월30일

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 갯벌로 27, 인천대학교 이노베이션센터(송도동)
(72) 발명자
전광길
인천광역시 연수구 해송로 70, 209동 805호(송도동, 송도웹카운티2단지)
(74) 대리인
특허법인충정

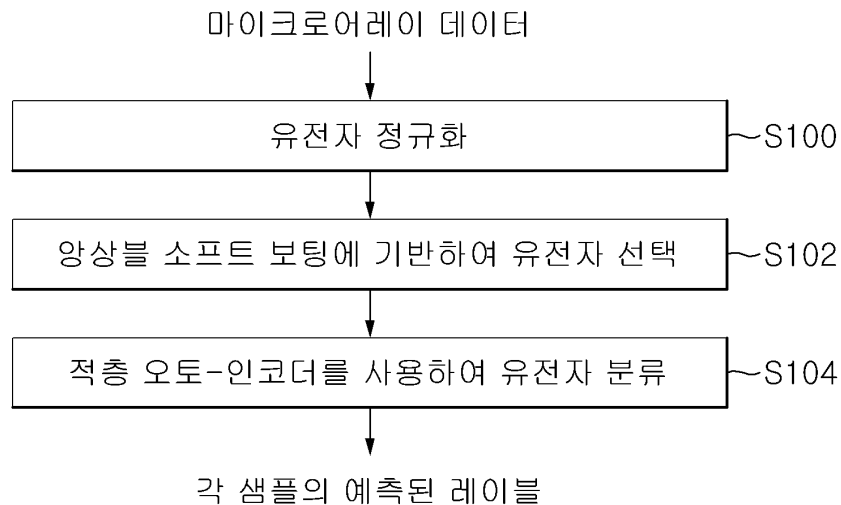
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 인공지능 기반 의료 진단 및 분류 방법

(57) 요약

본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은, 적어도 하나의 프로세서를 포함하는 인공지능 기반 마이크로어레이 암 분류 장치에 의해 수행되는 방법으로서, (A) 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계; 및 (B) 심층 신경망 기반의 적층 오토인코더(SAE: Stacked Auto-Encoder)를 사용하여 상기 선택된 유전자들을 분류하는 단계를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/08 (2023.01)

G16H 50/50 (2018.01)

명세서

청구범위

청구항 1

적어도 하나의 프로세서를 포함하는 인공지능 기반 의료 진단 및 분류 장치에 의해 수행되는 방법으로서,

(A) 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계; 및

(B) 심층 신경망 기반의 적층 오토인코더(SAE: Stacked Auto-Encoder)를 사용하여 상기 선택된 유전자들을 분류하는 단계를 포함하는, 인공지능 기반 의료 진단 및 분류 방법.

청구항 2

청구항 1에 있어서,

상기 단계 (A) 이전에,

(C-1) 상기 앙상블 소프트 보팅에 기반하여 훈련용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하여 분류하는 단계; 및

(C-2) 상기 분류된 결과에 기반하여 상기 앙상블 소프트 보팅에 사용된 각각의 알고리즘과 관련된 가중치를 결정하는 단계를 더 포함하는, 인공지능 기반 의료 진단 및 분류 방법.

청구항 3

청구항 2에 있어서,

상기 단계 (C-1) 및 단계 (C-2)에서, 상기 앙상블 소프트 보팅에 사용된 알고리즘들은, 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 윌콕슨(Wilcoxon) 방법 및 신호 대 잡음비(SNR: Signal to Noise) 테스트를 포함하는, 인공지능 기반 의료 진단 및 분류 방법.

청구항 4

청구항 3에 있어서,

상기 단계 (C-2)는,

상기 분류된 결과에 기반하여 상기 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 상기 윌콕슨(Wilcoxon) 방법 및 상기 신호 대 잡음비(SNR: Signal to Noise) 테스트 각각과 관련된 가중치를 결정하는 단계를 포함하고,

상기 단계 (A)는

(A-1) 상기 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계;

(A-2) 상기 윌콕슨(Wilcoxon) 방법에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계;

(A-3) 상기 신호 대 잡음비(SNR: Signal to Noise) 테스트에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계; 및

(A-4) 상기 선택된 각각의 유전자에 대한 가중 평균을 계산하고, 상기 가중 평균에 기반하여 모든 가중된 유전자들 중에서 암과 관련된 유전자들을 선택하는 단계를 포함하는, 인공지능 기반 의료 진단 및 분류 방법.

청구항 5

청구항 1에 있어서,

상기 단계 (A) 이전에,

상기 마이크로어레이 데이터의 유전자를 수학적 식 1에 기반하여 정규화하는 단계를 더 포함하고,

[수학적 식 1]

$$Y_{\text{norm}} = (Y_s - Y_{s_{\min}}) \times (Y_{s_{\max}} - Y_{s_{\min}})^{-1},$$

Y_s 는 현재 유전자의 발현 수준, $Y_{s_{\min}}$ 는 유전자의 최소값, $Y_{s_{\max}}$ 는 유전자의 최대값, 그리고 Y_{norm} 은 유전자의 정규화된 값을 나타내는, 인공지능 기반 의료 진단 및 분류 방법.

청구항 6

청구항 2에 있어서,

상기 (C-2)는, 상기 분류된 결과에 기반하여 k-최근접 이웃(k-NN: k-Nearest Neighbors) 알고리즘을 사용하여 상기 앙상블 소프트 보팅에 사용된 각각의 알고리즘과 관련된 가중치를 결정하는 단계를 포함하는, 인공지능 기반 의료 진단 및 분류 방법.

발명의 설명

기술 분야

[0001] 본 발명은 인공지능 기반 의료 진단 및 분류 방법에 관한 것이다.

배경 기술

[0002] 마이크로어레이 데이터 구조는 다양한 유형의 악성 조직 및 질병을 진단하고 분류하는 데 중요하다. 그러나 유전자의 높은 차원과 제한된 수의 샘플은 유전자 발현 및 분류 문제를 극복하는 데 효과적이다. 소수의 유전자만이 질병을 생물학적으로 정확하게 특성화하기 때문에 질병 유발 유전자를 해석하는 것은 어렵고 시간이 많이 걸린다.

[0003] 암과 같은 질병을 조기에 발견하면 효과적인 약물을 개발할 수 있다. 더욱이 실험실에서 효율적인 유전자를 발견하는 것은 어렵고 시간과 비용이 많이 든다. 마이크로어레이 데이터에서 유전자 분리를 자동화함으로써 분류 오류를 줄일 수 있을 뿐만 아니라 처리를 완료하는 데 필요한 시간 요소를 원하는 수준으로 줄일 수 있다.

[0004] 기계 학습에서 특징 선택은 여전히 최고 수준의 인식 및 분류를 달성하면서 문제 공간 특징의 가능한 가장 작은 하위 집합을 얻는 것을 목표로 한다. 또한 필터 모델, 래퍼(wrapper) 모델 및 내장 프로시저와 같은 특징 선택을 위한 많은 접근 방식에 대해 최적화 전략은 허용 가능한 결과를 제공할 수 있다. 그럼에도 불구하고 시간이 많이 걸리고 범용 최적화 가능성이 낮다. 필터 모델의 계산 비용은 래퍼 및 통합 모델의 계산 비용보다 낮다. 필터링 및 래퍼는 유전자 선택에 일반적으로 사용되는 전략이다. 필터링 방법에 따라 각 특징에는 클래스 레이블 및 단일 변수 점수 기준과의 연결을 기반으로 값이 할당된다.

[0005] 결과적으로 순위가 가장 높은 유전자가 선택되고 분류된다. 대조적으로 래퍼 기술은 순위 지정 과정에서 각 유전자의 성능을 평가하기 위해 분류 모음이 필요하다. 따라서, 유전자의 최적 하위 범주는 발견된 모든 하위 범주의 성능 순위/점수를 기반으로 결정된다. 필터링 방법이 개념 관계를 정량화할 수는 없지만 래퍼 전략은 높은 처리 비용으로 인해 제한될 수 있다.

[0006] 따라서, 적은 수의 마이크로어레이 데이터 샘플에도 불구하고 분류 오류를 최소화할 수 있고 분류 정확도를 높일 수 있는 마이크로어레이 암 분류 방법이 요구된다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) KR 10-1935094 B1

발명의 내용

해결하려는 과제

[0008] 본 발명의 목적은 적은 수의 마이크로어레이 데이터 샘플에도 불구하고 분류 오류를 최소화할 수 있고 분류 정확도를 높일 수 있는 인공지능 기반 의료 진단 및 분류 방법을 제공하는 것이다.

[0009] 본 발명의 목적은 유전자 발현 모델을 생성하기 위한 일반화 가능한 접근법으로서 합리적인 양의 오류로 분류할 수 있고, 소프트 검색 전략을 사용하여 마이크로어레이 데이터 세트에서 효과적인 유전자를 식별할 수 있는 인공지능 기반 의료 진단 및 분류 방법을 제공하는 것이다.

과제의 해결 수단

[0010] 상기 과제를 해결하기 위한 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은, 적어도 하나의 프로세서를 포함하는 인공지능 기반 의료 진단 및 분류 장치에 의해 수행되는 방법으로서,

[0011] (A) 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계; 및

[0012] (B) 심층 신경망 기반의 적층 오토인코더(SAE: Stacked Auto-Encoder)를 사용하여 상기 선택된 유전자들을 분류하는 단계를 포함할 수 있다.

[0013] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은, 상기 단계 (A) 이전에,

[0014] (C-1) 상기 앙상블 소프트 보팅에 기반하여 훈련용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하여 분류하는 단계; 및

[0015] (C-2) 상기 분류된 결과에 기반하여 상기 앙상블 소프트 보팅에 사용된 각각의 알고리즘과 관련된 가중치를 결정하는 단계를 더 포함할 수 있다.

[0016] 또한, 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 있어서, 상기 단계 (C-1) 및 단계 (C-2)에서, 상기 앙상블 소프트 보팅에 사용된 알고리즘들은, 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 윌콕슨(Wilcoxon) 방법 및 신호 대 잡음비(SNR: Signal to Noise) 테스트를 포함할 수 있다.

[0017] 또한, 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 있어서, 상기 단계 (C-2)는,

[0018] 상기 분류된 결과에 기반하여 상기 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 상기 윌콕슨(Wilcoxon) 방법 및 상기 신호 대 잡음비(SNR: Signal to Noise) 테스트 각각과 관련된 가중치를 결정하는 단계를 포함하고,

[0019] 상기 단계 (A)는

[0020] (A-1) 상기 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계;

[0021] (A-2) 상기 윌콕슨(Wilcoxon) 방법에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계;

[0022] (A-3) 상기 신호 대 잡음비(SNR: Signal to Noise) 테스트에 기반하여 상기 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용하는 단계; 및

[0023] (A-4) 상기 선택된 각각의 유전자에 대한 가중 평균을 계산하고, 상기 가중 평균에 기반하여 모든 가중된 유전자들 중에서 암과 관련된 유전자들을 선택하는 단계를 포함할 수 있다.

[0024] 또한, 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은, 상기 단계 (A) 이전에, 상기 마이크로어레이 데이터의 유전자를 수학적 1에 기반하여 정규화하는 단계를 더 포함하고,

[0025] [수학식 1]

$$Y_{\text{norm}} = (Y_s - Y_{s_{\min}}) \times (Y_{s_{\max}} - Y_{s_{\min}})^{-1},$$

[0027] Y_s 는 현재 유전자의 발현 수준, $Y_{s_{\min}}$ 는 유전자의 최소값, $Y_{s_{\max}}$ 는 유전자의 최대값, 그리고 Y_{norm} 은 유전자의 정규화된 값을 나타낼 수 있다.

[0028] 또한, 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 있어서, 상기 (C-2)는, 상기 분류된 결과에 기반하여 k-최근접 이웃(k-NN: k-Nearest Neighbors) 알고리즘을 사용하여 상기 앙상블 소프트 보팅에 사용된 각각의 알고리즘과 관련된 가중치를 결정할 수 있다.

발명의 효과

[0029] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 의하면, 적은 수의 마이크로어레이 데이터 샘플에도 불구하고 분류 오류를 최소화할 수 있고 분류 정확도를 높일 수 있다.

[0030] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 의하면, 유전자의 가중치를 평가하기 위해 K-NN 분류기를 사용한 지도 학습을 사용하여 유전자 선택에 대한 수정 접근법을 제안할 수 있고, 개선된 유전자 선택기와 적층 오토인코더(SAE: Stacked Auto-Encoder) 분류기를 사용하는 객관적인 순위 기술을 사용하여 관련 유전자의 통계적 결과를 결합하여 잠재적으로 부적절하고 편향된 전문 지식의 필요성을 제거할 수 있다.

[0031] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 의하면, 가혹한 테스트 조건에서 정확도를 유지하면서 낮은 계산량으로 강력한 분류를 달성할 수 있다.

[0032] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 의하면, 마이크로어레이 데이터를 사용하여 유전자를 연구하기 위한 일반화 가능하고 정확한 방법을 구축할 수 있다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법의 흐름도.

도 2는 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법이 적용될 수 있는 장치의 블록도.

도 3은 도 1에 도시된 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계의 상세 흐름도로서, 도 3a는 훈련 데이터를 이용하여 가중치들을 결정하는 것을 나타내고, 도 3b는 훈련 단계에서 결정된 가중치들을 이용하여 암과 관련된 유전자들을 선택하는 것을 나타낸 도면.

도 4는 적층 오토인코더의 구성을 도시한 도면.

도 5는 미만성 거대 B 세포 림프종(DLBCL), 백혈병 및 데이터에서 특징을 선택하고 이중 반복과 비교하는 다양한 방법의 박스 플롯.

도 6은 유전자 선택의 영향을 평가하기 위해 정확도 기준으로 5~45개의 최적 유전자를 평가한 결과를 도시한 도면.

도 7은 작고 둥근 청색 세포 종양(SRBCT)에 대한 분류 결과(퍼센트)를 도시한 도면.

도 8은 다발성 경화증에 대한 분류 결과(퍼센트)를 도시한 도면.

발명을 실시하기 위한 구체적인 내용

[0034] 이하, 첨부된 도면을 참조하여 본 명세서에 개시된 실시예를 상세히 설명하되, 도면 부호에 관계없이 동일하거나 유사한 구성요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다. 즉, 본 발명에서 사용되는 '부'라는 용어는 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, '부'는 어떤 역할들을 수행한다. 그렇지만 '부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일

예로서 '부'는 소프트웨어 구성요소들, 개체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 '부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '부'들로 결합되거나 추가적인 구성요소들과 '부'들로 더 분리될 수 있다.

- [0035] 또한, 본 명세서에 개시된 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다. 또한, 첨부된 도면은 본 명세서에 개시된 실시예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않으며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0036] 이하, 첨부한 도면들을 참조하여, 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에 대해 설명하기로 한다.
- [0037] 다양한 유전적 기원의 악성 종양 및 질병을 마이크로어레이 데이터로 진단하고 분류할 수 있다. 유전자의 큰 크기와 마이크로어레이의 샘플 수가 적기 때문에 극복해야 할 많은 장애물이 있다.
- [0038] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은 소프트 앙상블을 통해 가장 효과적인 유전자를 식별하는 단계 및 새로운 심층 신경망으로 분류하는 단계를 포함하는, 다양한 질병에서 유전자 발현을 위한 조합 전략이다.
- [0039] 특징 선택 접근 방식은 세 가지 전략을 결합하여 래퍼(wrapper) 유전자를 선택하고 k-최근접 이웃(K-NN) 알고리즘에 따라 순위를 지정하여, 오류 수준이 낮은 매우 일반화 가능한 모델을 생성한다.
- [0040] 소프트 앙상블을 사용하여 미만성 대세포 림프종(DLCL: diffuse large cell lymphoma), 백혈병 및 전립선암의 세 가지 마이크로어레이 데이터 세트에서 가장 효과적인 유전자 하위 집합을 식별할 수 있다.
- [0041] 누적 심층 신경망을 사용하여 3개의 데이터 세트를 모두 분류하여 각각 평균 97.51%, 99.6% 및 96.34%의 정확도를 달성했다. 또한, 이전에 보고되지 않은 작고 둥근 청색 세포 종양(SRBCT) 및 다발성 경화증 관련 뇌 조직 병변의 두 데이터 세트를 조사하여 모델 방법의 일반화 가능성을 보여주었다.
- [0042] 래퍼 방법을 빠른 분류 모델과 결합하면 마이크로어레이 데이터가 많은 수의 유전자와 연관되어 있는 경우에도 특징들의 순위를 매길 수 있다. 적절하게 수정된 분류 매개변수는 원하는 정확도와 정밀도를 달성하기 위한 기반을 설정하며, 이것이 다양한 연구에서 k-최근접 이웃(k-NN) 브레이싱(bracing) 조정을 조사한 이유이다.
- [0043] 마이크로어레이 데이터의 특징 선택과 함께 효율적인 분류는 정확성과 효율성을 향상시킨다. 실제로 인공 신경망(ANNs: Artificial Neural Networks)은 문제 도메인의 복잡성에 관계없이 적절한 조정 및 학습을 통해 최적의 결과를 생성할 수 있는 효율적인 알고리즘 클래스이다. 최대 학습 잠재력 신경망, 경계 클래스 회귀 분류, 로지스틱 회귀, 무작위 결정 트리 및 베이지안 이론이 있다. 유방암, 대장암, 림프절암, 폐암 모두 유전자 발현 분석을 통해 진단할 수 있다. 폐암 검출을 위해 기계 학습 기술과 유전자 발현의 특징 선택이 권장된다.
- [0044] 기계 학습(특히 딥 러닝(DL))은 많은 응용 분야에서 분류 연구에서 인기 있는 도구가 되었다. 최근 DL은 마이크로어레이 데이터 처리에 상당한 영향을 미쳤다. 이러한 데이터 세트의 최근 가용성으로 인해 DL 접근 방식을 사용하여 데이터 분석 속도를 높이고 암 진단, 예후 및 치료 반응의 정확도를 향상시킬 수 있다. 마이크로어레이 데이터를 사용하여 유전자 발현을 자동으로 모델링하기 위한 정확하고 빠른 접근법이 절실히 필요한 것 같다. 고정 방법과 빠른 바인딩 클래스를 결합하여 동시에 시간을 늘리고 선택한 유전자의 순위를 매길 수 있다. 또한 효율적인 DL 모델을 구현함으로써 유전자 발현의 정확도를 훨씬 더 쉽게 향상시킬 수 있다.
- [0045] 소프트 군집화 방법과 세 가지 래퍼 전략을 사용하여 본 발명은 다양한 질병에 대한 유전자 발현의 성공적인 조합 기술을 제공한다. 효과적인 유전자 부분 집합을 선택할 때 특징 선택 부분에서 k-NN 방법을 의사 결정에 사용할 수 있다.
- [0046] 몇 가지 접근법은 특징 선택 및 분류 절차를 사용하여 마이크로어레이를 분류한다. 유전자 발현 분석은 유전자 선택이 중요하기 때문에 이러한 근본적인 문제에 대해 다양한 측면에서 많은 연구가 진행되어 왔다.
- [0047] 마이크로어레이 데이터에는 중복되거나 예측 설계에 적합하지 않은 상당수의 유전자가 포함될 수 있다. 그러나 마이크로어레이 데이터의 데이터 양이 주어지면 적은 수의 관련 유전자가 학습에 더 많은 이점을 제공할 수 있

다. 이전 연구에서는 유의미한 클래스 상관 관계를 기반으로 서로 관련된 유전자를 식별하려고 했다.

[0048] 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은 도 1과 같다. 전처리(단계 S100), 특징(유전자) 선택(단계 S102), 및 분류(단계 S104)는 제안하는 본 발명의 방법의 주요 단계이다.

[0049] 도 1은 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법의 흐름도이고, 도 2는 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법이 적용될 수 있는 장치의 블록도이다.

[0050] 도 1에 도시된 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법은, 마이크로어레이 데이터의 유전자를 정규화하는 단계(단계 S100), 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계(단계 S102) 및 심층 신경망 기반의 적층 오토인코더(SAE: Stacked Auto-Encoder)를 사용하여 상기 선택된 유전자들을 분류하는 단계(단계 S104)를 포함할 수 있다.

[0051] 한편, 도 2에 도시된 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법이 적용될 수 있는 장치(200)는, 프로세서(202), 메모리(204), 입출력 인터페이스(208) 및 네트워크 통신 인터페이스(206)를 포함할 수 있다.

[0052] 프로세서(202)는, 마이크로어레이 데이터의 유전자를 정규화하는 동작, 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 동작 및 심층 신경망 기반의 적층 오토인코더(SAE: Stacked Auto-Encoder)를 사용하여 상기 선택된 유전자들을 분류하는 동작을 수행할 수 있다.

[0053] 메모리(204)는 프로세서(202)가 상기 동작들을 수행하게 하는 명령어들을 저장하고, 입출력 인터페이스(208)는 입력 및 출력 기능을 제공하는 입출력 장치에 연결될 수 있고, 네트워크 통신 인터페이스(206)는 네트워크와 통신할 수 있는 기능을 제공할 수 있다.

[0055] 전처리

[0056] 단계 S100에서, 마이크로어레이 데이터의 유전자를 정규화한다.

[0057] 전처리 중에, 자동 분류 모델이 과도하게 훈련되지 않도록 데이터를 섞었다. 샘플이 무작위로 혼합되어 있지만 레이블 위치는 분석된 각 샘플이 나타내는 클래스에 따라 다를 수 있다. 검증 단계에서 샘플의 조합과 샘플의 위치에 해당하는 레이블을 조정하면 학습 알고리즘이 한 클래스의 유전자 또는 그 하위 집합에만 조정할 수 있다. 이 방법은 과적합을 방지하여 분류 정확도를 향상시킨다.

[0058] 해당 샘플의 위치가 무작위로 이동하고 레이블도 이동된다. 각 데이터셋의 각 유전자는 자연적으로 고도로 분산되어 있고 다양한 유전자 집합을 활용하므로 전처리 단계에서 정규화되어야 한다. 유전자 정규화는 처리 비용을 줄이고 유전자 발현을 최적화한다. 적절한 정규화 범위는 최소-최대 정규화를 기반으로 (0-1)이다. 따라서 다음과 같이 정규화를 위해 최소-최대 방법이 적용된다.

수학식 1

$$Y_{\text{norm}} = (Y_s - Y_{s_{\min}}) \times (Y_{s_{\max}} - Y_{s_{\min}})^{-1}$$

[0059]

[0060] 이는 모든 숫자가 0과 1 사이에서 조정됨을 나타낸다. Y_s 는 현재 유전자의 발현 수준, $Y_{s_{\min}}$ 는 유전자의 최소값, $Y_{s_{\max}}$ 는 유전자의 최대값, 그리고 Y_{norm} 은 유전자의 정규화된 값을 나타낼 수 있다.

[0062] 유전자 선택

[0063] 단계 S102에서, 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택한다.

[0064] 특징을 선택하는 데 사용되는 세 가지 방법은 신호 대 잡음비(SNR: Signal to Noise) 테스트, 윌콕슨(Wilcoxon)

방법 및 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석이다. ROC 곡선은 첫 번째 접근 방식에서 참양성률 대 거짓양성률을 나타낸다. 곡선 아래 면적(AUC: area under the curve)은 수학적 2를 사용하여 결정될 수 있다.

수학적 2

$$AUC = \int_0^1 T_1(T_2^{-1}(t)) dt \Rightarrow t \in (0, 1)$$

T 는 F_i 에 대한 보완(complement)/보충(supplement)이며, 여기서 $F_1(x)$ 및 $F_2(x)$ 는 두 개의 서로 다른 통계 커뮤니티 $F_i(x)$ 에서 x 의 분포 함수이다. AUC 값이 높은 유전자는 샘플내의 유전자를 포함하는 레이블들의 중첩 정도가 낮다는 것을 나타내기 때문에, AUC가 높은 유전자만 유전자 선택 벡터로 구성된다.

두 모집단이 통계적으로 비교 가능한지 결정하는 Mann-Whitney 테스트는 Wilcoxon 테스트와 유사하다. 귀무 가설은 두 모집단의 분포 함수가 동일하다고 주장하는 반면, 대립 가설은 두 모집단의 분포 함수가 동일하지 않다고 주장한다. 이 테스트는 두 샘플 간의 차이에 대한 기본 가정이 필요하지 않기 때문에, 많은 실제 시나리오에서 선호된다. Wilcoxon 테스트를 수행하는 방법은 다음과 같다.

(1) 두 그룹의 모든 관측치를 중첩하고 총 수를 기준으로 오름차순으로 정렬한다.

(2) Wilcoxon의 통계는 더 작은 그룹에 포함된 모든 순위의 합계를 기반으로 한다.

(3) 전체 분포표를 기반으로 p-값을 계산하여 가정에 사용한다.

유전자 선택의 경우, Wilcoxon 계수보다 Wilcoxon 테스트에서 표준화된 Wilcoxon 통계의 절대값이 더 효과적이고 유익하다.

수학적 3의 연결의 관점에서 표현된 유전자 레이블 간의 차이를 측정하는 SNR 방법론은 사용된 세 번째 방법이다.

수학적 3

$$SNR(f_i, c) = (\mu_1 - \mu_2) \times (\sigma_1 + \sigma_2)^{-1}$$

여기서 μ_1 및 μ_2 와 σ_1 및 σ_2 는 각각 샘플의 평균 및 표준 편차를 나타낸다. 변수 c 및 f_i 는 각각 유전자 레이블 및 i 번째 특징 벡터를 포함하는 벡터로 간주된다. SNR은 유전자 선택 문제를 해결하고 마이크로어레이 데이터를 분류하기 위해 각 유전자에서 특징을 선택할 것으로 예상된다.

마지막으로 훈련 단계에서 세 개의 관련 클래스 중에서 가장 많은 표를 얻은 유전자가 가장 관련성이 높은 특성으로 선택된다. 훈련 데이터는 K-fold 방법(CV = 10)을 사용하여 신규 및 검증 및 훈련 세트로 구분하고 가장 자주 발생하는 특징을 사용하여 선택된 유전자를 나타낸다.

1차 유전자는 분류 클래스를 예측하고, k 는 가장 가까운 선택된 이웃으로, 빠르고 거대한 데이터 유전자 세트에 대해 좋은 응답을 제공한다. 본 발명은 k -NN의 성능을 향상시키기 위해 사용되었으며, 쿼리와 k -d 알고리즘을 활용하여 계산을 최소화하고 클래스 성능을 향상시켰다. 훈련 단계에서 얻은 정밀도를 고려하는 유전자 선택의 상세 흐름도가 도 3에 도시된다.

도 3은 도 1에 도시된 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하는 단계의 상세 흐름도로서, 도 3a는 훈련 데이터를 이용하여 가중치들을 결정하는 것을 나타내고, 도 3b는 훈련 단계에서 결정된 가중치들을 이용하여 암과 관련된 유전자들을 선택하는 것을 나타낸 도면이다.

- [0078] 도 3a를 참조하면, 앙상블 소프트 보팅에 기반하여 훈련용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하여 분류하고(단계 S300, S302 및 S304), 상기 분류된 결과에 기반하여 k-최근접 이웃(k-NN: k-Nearest Neighbors) 알고리즘을 사용하여 상기 앙상블 소프트 보팅에 사용된 각각의 알고리즘과 관련된 가중치를 결정한다(단계 S306).
- [0079] 상기 앙상블 소프트 보팅에 사용된 알고리즘들은, 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 윌콕슨(Wilcoxon) 방법 및 신호 대 잡음비(SNR: Signal to Noise) 테스트를 포함할 수 있다.
- [0080] 상기 단계 S306에서는, 상기 분류된 결과에 기반하여 상기 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석, 상기 윌콕슨(Wilcoxon) 방법 및 상기 신호 대 잡음비(SNR: Signal to Noise) 테스트 각각과 관련된 가중치를 결정할 수 있다.
- [0081] 도 3b를 참조하면, 단계 S102에서, 앙상블 소프트 보팅(ensemble soft voting) 학습에 기반하여 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택한다.
- [0082] 이를 상세히 설명하면, 단계 S308에서, 수신자 조작 특성(ROC: Receiver Operating Characteristic) 분석에 기반하여 테스트용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용한다.
- [0083] 단계 S310에서, 윌콕슨(Wilcoxon) 방법에 기반하여 테스트용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용한다.
- [0084] 단계 S312에서, 신호 대 잡음비(SNR: Signal to Noise) 테스트에 기반하여 테스트용 마이크로어레이 데이터에서 암과 관련된 유전자들을 선택하고, 선택된 각각의 유전자에 대응하는 가중치를 적용한다.
- [0085] 단계 S314에서, 상기 선택된 각각의 유전자에 대한 가중 평균을 계산하고, 단계 S316에서 상기 가중 평균에 기반하여 유전자들을 소팅(sorting)한 후 모든 가중된 유전자들 중에서 암과 관련된 유전자들을 선택한다.
- [0087] **적층 오토인코더(SAE: Stacked Auto-encoders)**
- [0088] 적층 오토인코더(SAE: Stacked Auto-encoders)는 오토인코더를 사용하여 구성된다. 각 오토인코더의 숨겨진 레이어는 신경망에서 다음 오토인코더의 숨겨진 레이어에 연결된다. 학습 단계 전반에 걸쳐 이전의 오토인코더의 숨겨진 계층을 후속 오토인코더의 입력으로 사용해야 한다.
- [0089] 도 4는 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법에서 사용된 SAE 아키텍처를 보여준다. SAE를 사용하면 기존 추상화 위에 계층화하여 새 추상화를 만들 수 있다. 재구성할 때 최종 숨겨진 계층 출력에는 유전자 데이터의 높은 수준 속성이 포함된다. 개체 필드에서 개체의 속성은 전도율 분포를 정의한다. 로지스틱 회귀 레이어를 사용하여 전도도 분포를 결정했다.
- [0090] 정규화된 유전자 선택을 DNN에 공급한다. 그것은 기호 $U = \{G^{(1)}, G^{(2)}, \dots, G^{(M)}\}$ 에 의해 표기된다. 여기서, M 은 훈련 세트의 수이고, $G^{(k)} \in [0,1]^m$ 는 정규화된 유전자들의 수를 나타낸다. m 은 무작위로 선택된 유전자 서열 모음에 알려지지 않은 수의 유전자 값들이 있음을 나타낸다.
- [0091] 내부 전도도 분포 샘플은 $U = \{\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(M)}\}$ 이다. 여기서 $\sigma^{(k)} \in [0,1]^n$ 이고, n 은 각 클래스의 확률을 나타낸다. 가중치, 편향 행렬 및 벡터를 시드(seed)하기 위해 비지도 계층별 학습이 사용된다.
- [0092] DNN은 값이 $U = \{G^{(1)}, G^{(2)}, \dots, G^{(M)}\}$ 인 유전자를 처리하는 임무를 맡는다. 전체 기술은 아래에 요약되어 있다. 이전 레이어의 출력을 사용하여 첫 번째 히든 레이어를 먼저 훈련시키는 것이 중요하다. 히든 레이어가 모두 학습될 때까지 동일한 절차가 반복된다. 지도 미세 조정 단계에서 전체 DNN을 초기화하기 위해 DNN의 최종 히든 레이어에서 사전 훈련된 네트워크 매개변수는 로지스틱 회귀 모델에 입력된다.
- [0093] 네트워크는 실제 전도도 분포의 대표 샘플에서 이름을 따왔다. 네트워크 매개변수를 미세 조정하기 위한 하향식 접근 방식은 기울기 최적화 기술을 기반으로 하는 역전파 알고리즘을 기반으로 한다. "드롭아웃(dropout)"을 사용하여 과적합을 줄임으로써 보다 일반화 가능한 모델을 향상시킬 수 있다. 각 훈련 세션 동안 네트워크의 히

든 유닛 중 0.5%가 네트워크에서 무작위로 삭제된다.

[0094] 뉴런 공동 적응을 단순화하면 보다 탄력적인 네트워크를 구축할 수 있다. 드롭아웃 레이어는 거대한 데이터 세트에서 훈련될 때 훌륭하게 수행된다. 일반적인 오토인코더에서 볼 수 있듯이 드롭아웃은 수학적 4와 수학적 5에 영향을 미친다.

수학적 4

$$y_i = f \left(\sum_{j=1}^m w_{ij} \text{Bernoulli}(p) * x_j + b_i \right)$$

[0095]

수학적 5

$$z_j = f \left(\sum_{i=1}^m w_{ij}^T \text{Bernoulli}(p) * y_i + b'_j \right)$$

[0096]

[0097] $\text{Bernoulli}()$ 는 확률 p 가 0.50인 0 또는 1의 랜덤 벡터를 생성하는 함수로 정의된다.

[0099] 결과

[0100] 이 섹션에서는 유전자 선택의 결과 및 분류 체계를 설명하고, 개선된 본 발명의 DL 모델의 결과를 설명한다.

[0102] 데이터 세트

[0103] 이 연구는 주로 유전자 설명과 같은 도서관-도서관 데이터베이스에서 데이터를 수집하는 기술-분석 조사이다. 이러한 기록은 동적 검색 및 찾아보기 특징을 갖춘 무료 온라인 의료 검색 포털을 통해 자주 액세스할 수 있다.

[0104] 미세 수술용 칩 제작, 암 및 건강한 검체를 포함한 다양한 샘플의 샘플 수집, RNA 추출 및 DNA 보충과 같은 실험실 작업을 통해 유전자 발현 데이터가 수집되었다. 실험은 미만성 거대 B 세포 림프종(DLBCL), 백혈병, 전립선암의 세 가지 데이터 세트를 사용하여 수행되었다. 데이터 세트의 각 데이터 샘플에는 샘플링 기술 및 기기의 신뢰성에 대한 정보가 포함되어 있으며, DLBCL 및 여포성 림프종(FL)과 관련하여 분류된 두 가지 유형의 양성 종양이 있다.

[0105] DLBCL 데이터 세트는 77개 샘플의 7070개 유전자를 포함하며, 그 중 58개는 DLBCL에 대해 양성이고 나머지 샘플은 FL에 대해 양성이었다. 두 림프종을 구별하기 위해 유전자 발현 데이터를 활용하여 분류 모델을 개발했다. 백혈병 데이터 세트에는 급성 림프구성 백혈병(ALL) 및 급성 골수성 백혈병(AML) 환자의 골수 및 혈액 샘플이 포함되어 있다. 이 데이터 수집에는 72개의 샘플(47개의 ALL 및 25개의 AML)이 포함되어 있으며 각 샘플에는 7129개의 유전자 샘플이 있다. 세 번째 데이터 세트에는 102개의 샘플(천연 소스에서 50개, 전립선 조직에서 52개)이 포함되어 있으며 각 샘플에는 125,333개의 유전자가 포함되어 있다. 단면에서 모든 유전자 데이터는 정규화된다.

[0107] 설정

[0108] 데이터 분석은 2022 MATLAB 프로그래밍 환경에서 이루어졌으며 통계 테스트에는 소프트웨어 모델의 시뮬레이션을 따르는 정량적 및 정성적 결과가 포함된다. 모델링에 사용되는 시스템은 8기가바이트 RAM과 64비트 운영 체제를 갖춘 Intel(R), Core(TM) 및 Core i7 프로세서의 속성을 가지고 있다.

- [0109] 특징 선택 단계에서는 다양한 설정을 고려하여 초기 데이터 집합의 처리 비용을 줄이기 위해 초기 특징을 변경하지 않고 훨씬 적은 수의 초기 특징을 선택하여 적절한 집합을 선택했다.
- [0110] 본 발명의 제안된 특징 선택 모델과 다른 특징 선택 방법을 비교하기 위해 정확도를 평가할 때마다 특정 수의 특징(5~45개 유전자)을 k-NN 분류에 부여한다. 과적합을 방지하기 위해 알고리즘을 활용하여 (테스트된 모든 구조 중에서) 여러 번의 반복을 통해 최소 평균 제곱 오차(MSE)를 찾았다. 샘플 데이터에 대한 성능은 평균 오차 제곱 계수가 지정된 값(0.05)보다 작으면, 해당 네트워크가 기본 네트워크로 선택되고, 그렇지 않으면 최소 MSE 및 해당 구조가 선택된다.
- [0111] DNN 모델의 히든 레이어의 수는 모델의 학습 능력에 상당한 영향을 미친다. 일반적으로 모델을 미세 조정하기 위해 각 계층의 학습 속도와 반복 횟수를 조정한다. 3~5개의 레이어 신경망이 구축되었다. 각 레이어에는 입력 레이어에 208개의 노드, 출력 레이어에 812개, 히든 레이어에 150개의 노드가 포함되어 있다.
- [0112] 0.1에서 0.01 사이에서, 학습 속도는 훈련 전에서 훈련 중으로 다양했다. 미세 조정 방법은 1~500회 반복으로 구성되었다. 또한 다음 네트워크 구성이 사용되었다. 사전 훈련 세션은 각각 10명씩 10개의 그룹으로 나뉘었고, 각 그룹은 100-epoch 훈련 epoch를 받았다. 이 시점에서 배치 크기는 40으로 낮아졌다. MSE는 네트워크의 성능을 평가하는 데 사용되었다.
- [0114] 평가
- [0115] 표 1 내지 표 3에는 DLBCL 데이터, 백혈병 및 전립선암 각각에 대해 5개의 폴드(각 폴드는 10개의 분할을 포함함)에 대한 시뮬레이션 결과가 나타나 있으며, 실험 데이터에 대한 출력이 계산된다.

표 1

Data dividing	No. Layers	Feature selection by wrapper method			Feature selection by ensemble model		
		Max	Mean	Min	Max	Mean	Min
10-fold (1)	Low	0.90 ± (0.045)	0.87 ± (0.078)	0.85 ± (0.118)	0.92 ± (0.021)	0.90 ± (0.056)	0.89 ± (0.072)
	Med	0.95 ± (0.022)	0.91 ± (0.034)	0.89 ± (0.044)	100	0.97 ± (0.011)	0.95 ± (0.016)
	High	0.93 ± (0.034)	0.90 ± (0.042)	0.87 ± (0.078)	0.95 ± (0.034)	0.94 ± (0.032)	0.93 ± (0.036)
10-fold (2)	Low	0.92 ± (0.036)	0.88 ± (0.064)	0.86 ± (0.084)	0.93 ± (0.029)	0.92 ± (0.043)	0.91 ± (0.065)
	Med	0.94 ± (0.032)	0.89 ± (0.051)	0.88 ± (0.064)	100	0.98 ± (0.017)	0.96 ± (0.026)
	High	0.94 ± (0.029)	0.92 ± (0.038)	0.89 ± (0.060)	0.96 ± (0.027)	0.98 ± (0.042)	0.94 ± (0.051)
10-fold (3)	Low	0.91 ± (0.045)	0.85 ± (0.084)	0.83 ± (0.096)	0.94 ± (0.018)	0.92 ± (0.033)	0.90 ± (0.065)
	Med	0.95 ± (0.025)	0.89 ± (0.041)	0.87 ± (0.085)	0.99 ± (0.007)	0.96 ± (0.014)	0.94 ± (0.022)
	High	0.92 ± (0.037)	0.87 ± (0.055)	0.84 ± (0.086)	0.95 ± (0.026)	0.93 ± (0.035)	0.92 ± (0.044)
10-fold (4)	Low	0.89 ± (0.073)	0.85 ± (0.089)	0.82 ± (0.124)	0.96 ± (0.028)	0.94 ± (0.054)	0.91 ± (0.073)
	Med	0.94 ± (0.031)	0.90 ± (0.046)	0.88 ± (0.067)	100	0.98 ± (0.013)	0.96 ± (0.024)
	High	0.93 ± (0.048)	0.88 ± (0.076)	0.84 ± (0.109)	0.95 ± (0.031)	0.93 ± (0.044)	0.91 ± (0.056)
10-fold (5)	Low	0.90 ± (0.045)	0.87 ± (0.078)	0.85 ± (0.118)	0.94 ± (0.034)	0.92 ± (0.051)	0.91 ± (0.078)
	Med	0.95 ± (0.043)	0.93 ± (0.068)	0.90 ± (0.074)	0.98 ± (0.022)	0.97 ± (0.030)	0.95 ± (0.046)
	High	0.94 ± (0.051)	0.90 ± (0.077)	0.88 ± (0.092)	0.96 ± (0.028)	0.94 ± (0.048)	0.93 ± (0.053)

[0116]

표 2

Data dividing	No. Layers	Feature selection by wrapper method			Feature selection by ensemble model		
		Max	Mean	Min	Max	Mean	Min
10-fold (1)	Low	0.92 ± (0.025)	0.91 ± (0.032)	0.89 ± (0.056)	0.96 ± (0.014)	0.94 ± (0.025)	0.95 ± (0.048)
	Med	0.96 ± (0.021)	0.94 ± (0.027)	0.92 ± (0.35)	100	0.99 ± (0.009)	0.98 ± (0.018)
	High	0.92 ± (0.031)	0.91 ± (0.044)	0.89 ± (0.076)	0.97 ± (0.026)	0.96 ± (0.027)	0.94 ± (0.024)
10-fold (2)	Low	0.94 ± (0.024)	0.93 ± (0.036)	0.91 ± (0.051)	0.97 ± (0.013)	0.95 ± (0.026)	0.94 ± (0.028)
	Med	0.96 ± (0.021)	0.94 ± (0.024)	0.92 ± (0.036)	100	0.98 ± (0.011)	0.97 ± (0.021)
	High	0.93 ± (0.034)	0.91 ± (0.054)	0.89 ± (0.067)	0.97 ± (0.016)	0.96 ± (0.023)	0.95 ± (0.041)
10-fold (3)	Low	0.92 ± (0.028)	0.91 ± (0.035)	0.88 ± (0.070)	0.98 ± (0.011)	0.96 ± (0.017)	0.94 ± (0.034)
	Med	0.97 ± (0.018)	0.94 ± (0.022)	0.92 ± (0.044)	100	0.99 ± (0.008)	0.98 ± (0.010)
	High	0.95 ± (0.035)	0.93 ± (0.037)	0.92 ± (0.056)	0.98 ± (0.018)	0.97 ± (0.018)	0.95 ± (0.037)
10-fold (4)	Low	0.95 ± (0.037)	0.93 ± (0.041)	0.90 ± (0.065)	0.98 ± (0.018)	0.97 ± (0.026)	0.95 ± (0.032)
	Med	0.98 ± (0.020)	0.94 ± (0.027)	0.93 ± (0.039)	100	0.99 ± (0.006)	0.98 ± (0.011)
	High	0.94 ± (0.029)	0.92 ± (0.039)	0.90 ± (0.041)	0.98 ± (0.023)	0.97 ± (0.026)	0.95 ± (0.041)
10-fold (5)	Low	0.94 ± (0.031)	0.92 ± (0.036)	0.88 ± (0.086)	0.98 ± (0.014)	0.96 ± (0.019)	0.95 ± (0.039)
	Med	0.97 ± (0.023)	0.96 ± (0.031)	0.93 ± (0.053)	100	0.98 ± (0.012)	0.95 ± (0.039)
	High	0.96 ± (0.032)	0.94 ± (0.042)	0.92 ± (0.078)	0.98 ± (0.015)	0.97 ± (0.025)	0.96 ± (0.028)

[0117]

표 3

Data dividing	No. Layers	Feature selection by wrapper method			Feature selection by ensemble model		
		Max	Mean	Min	Max	Mean	Min
10-fold (1)	Low	0.89 ± (0.046)	0.86 ± (0.071)	0.85 ± (0.141)	0.91 ± (0.032)	0.88 ± (0.048)	0.86 ± (0.132)
	Med	0.94 ± (0.033)	0.90 ± (0.060)	0.88 ± (0.073)	0.98 ± (0.026)	0.96 ± (0.031)	0.93 ± (0.056)
	High	0.90 ± (0.057)	0.86 ± (0.058)	0.85 ± (0.129)	0.94 ± (0.044)	0.91 ± (0.058)	0.89 ± (0.108)
10-fold (2)	Low	0.90 ± (0.068)	0.86 ± (0.092)	0.83 ± (0.121)	0.93 ± (0.045)	0.90 ± (0.063)	0.87 ± (0.119)
	Med	0.94 ± (0.053)	0.92 ± (0.064)	0.90 ± (0.070)	100	0.97 ± (0.024)	0.94 ± (0.051)
	High	0.91 ± (0.073)	0.89 ± (0.084)	0.87 ± (0.113)	0.95 ± (0.034)	0.93 ± (0.055)	0.90 ± (0.095)
10-fold (3)	Low	0.90 ± (0.056)	0.88 ± (0.094)	0.86 ± (0.118)	0.92 ± (0.035)	0.90 ± (0.044)	0.87 ± (0.093)
	Med	0.94 ± (0.034)	0.90 ± (0.056)	0.88 ± (0.096)	0.98 ± (0.025)	0.95 ± (0.035)	0.92 ± (0.048)
	High	0.91 ± (0.076)	0.89 ± (0.073)	0.87 ± (0.108)	0.94 ± (0.042)	0.91 ± (0.057)	0.88 ± (0.084)
10-fold (4)	Low	0.89 ± (0.067)	0.87 ± (0.075)	0.86 ± (0.123)	0.93 ± (0.035)	0.91 ± (0.067)	0.87 ± (0.083)
	Med	0.95 ± (0.048)	0.93 ± (0.041)	0.91 ± (0.087)	100	0.97 ± (0.024)	0.94 ± (0.041)
	High	0.89 ± (0.074)	0.87 ± (0.085)	0.86 ± (0.106)	0.94 ± (0.041)	0.91 ± (0.051)	0.89 ± (0.072)
10-fold (5)	Low	0.89 ± (0.056)	0.85 ± (0.074)	0.84 ± (0.143)	0.92 ± (0.033)	0.88 ± (0.061)	0.85 ± (0.128)
	Med	0.94 ± (0.041)	0.92 ± (0.060)	0.90 ± (0.082)	0.98 ± (0.020)	0.95 ± (0.049)	0.91 ± (0.086)
	High	0.91 ± (0.050)	0.89 ± (0.071)	0.88 ± (0.113)	0.96 ± (0.039)	0.93 ± (0.064)	0.90 ± (0.096)

[0118]

[0119]

도 5에서, 특징 선택 방법의 박스 플롯과 중복에 대한 유추는 세 가지 데이터 모델 모두에서 설명된다. 유전자 발현 데이터의 세 가지 모델 모두에서 소프트 유전자의 최적 하위 집합을 찾기 위한 최소 및 최대 거리는 더 작은 폭을 갖는다. 이 표에서 레이어 수의 변화와 뉴런 수의 변화가 있는 경우 네트워크의 역학을 가정하고 서로 다른 해당 정확도를 추정한다. 대부분의 경우 낮은 뉴런과 높은 뉴런을 기준으로 조정하면 정확도가 좋아지고 정확도의 분산이 적다.

[0120]

낮은 계층의 수는 네트워크에 대해 1~3개의 히든 계층으로 가정하고 많은 계층은 4~8개까지 다양하다. 뉴런의 개수가 적은 경우에는 5~15개, 많은 경우에는 20~30개까지 다양하며, 증분 추정법에 따라 레이어 및 뉴런 안정화 과정을 수행하였다. 입력 및 출력 레이어 노드의 수는 각각 선택된 유전자 수와 유전자 발현과 관련된 클래스 수에 따라 조정되었다.

[0121]

특징 선택의 성능을 비교하기 위해 현 단계에서 본 발명의 제안하는 방법과 단일 래퍼 방법을 비교한다. 정확도 사이의 분산을 줄이는 것 외에도 출력도 크게 향상된다. 소프트 균질화 방법을 적용한 결과 평균적으로 4~6%의 거리에서 최적화가 이루어졌으며, 세 가지 유전자 데이터 모델 모두에서 100% 정확도를 보였다. 백혈병 유전자

발현의 가장 낮은 분포 수준과 가장 높은 분포가 전립선암 데이터에 대해 추정된다.

[0122] 또한 유효 유전자를 선별함에 있어서 제안하는 특징을 선별하는 방법은 신뢰 구간보다 작은 정확도 값을 가진다. 각 샘플에서 유전자 발현에 대한 전문가의 의견과 제안된 알고리즘으로 얻은 레이블 간에는 유의미한 차이가 없다($p\text{-Value} < 0.02$). 그리고 출력의 표준 편차는 무시할 수 있었다.

[0124] 논의

[0125] 다양한 분류 주제와 다양한 데이터에서 제안하는 방법은 기능적 포괄성을 가져온다. 유전자 발현 분야에서 제안된 솔루션 중 다수는 제한된 반복, 고도로 분산된 분류 결과 및 불내성 불확실성 범위에서 수많은 시간을 소비하여 얻을 수 있다. 본 발명의 제안된 방법을 통해 유전자 선택의 영향을 평가하기 위해 도 6에 제시된 정확도 기준으로 5-45개의 최적 유전자를 평가했다. 세 가지 형식 모두 알고리즘의 재현성과 견고성을 나타낸다.

[0126] 도 6에는, 3개의 마이크로어레이 데이터 세트 각각에 대해 자동 유전자 선택 및 분류에 대한 결과(퍼센트)가 1행, 2행 및 3행에서 각각의 첫 번째, 두 번째 및 세 번째 반복에서 유효 유전자의 수가 가장 적은 소프트 앙상블 및 SAE 방법을 통해 표시된다.

[0127] 또한 그것은 유전자의 최소 정확도(10개에서 20개 유전자)를 정확하게 결정하고 분류할 수 있음을 나타낸다. 이전 알고리즘이 이미 유전자(특징)를 선택했지만, 최소한의 개수의 유효 유전자의 분석을 건너뛰었다는 점은 흥미롭다.

[0128] 본 발명의 제안하는 방법에서는 교차 검증 방법과 변수 K값을 5, 10, 15, 20으로 채택한 K-fold를 이용하여 기준을 추정하였다. 일반적으로, 선택된 유전자의 수가 너무 많지도 너무 적지도 않은 경우 출력이 최대값이다. 반면 DLBCL, 전립선암 등의 데이터는 평균적인 개수의 유전자를 적용해 알고리즘에 정확하게 구현했다.

[0129] 그럼에도 불구하고 본질적으로 만족스러운 백혈병과 같은 데이터의 경우 소수의 선택된 유전자를 적용하더라도 출력은 허용 가능하다. 대부분의 유사한 방법에서 선택된 최소량의 유전자는 버려진다. 이것은 처리량 감소에 직접적인 영향을 미치고 최적의 출력을 달성하는 속도를 높일 수 있기 때문에 매우 중요하다.

[0130] 이전 방법론과 달리, 마이크로어레이 데이터에서 유전자 발현 모델의 일반화 가능성은 덜 주목을 받았다. 실제로, 모델이 이전에 본 적이 없는 데이터를 분류할 때 필요한 견고성과 신뢰성을 입증하는 경우 일반화 가능성을 주장할 수 있다. 일반화 가능성을 평가하기 위해 초기 마이크로어레이 데이터세트와 다른 기능을 가진 두 개의 추가 데이터세트를 사용했다.

[0131] 첫 번째 마이크로어레이는 어린이에게서 얻은 SRBCT 모음이다. 이들 종양은 신경모세포종(NB), 횡문근육종(RMS), 비립프종 호지킨병(NHL) 및 유잉 계열 악성종양(EWS)의 4가지 뚜렷한 아형을 나타낸다. 각 클래스에는 23개의 샘플(EWS), 20개의 샘플(RMS), 12개의 샘플(NB) 및 8개의 샘플(NHL)이 있으며, 총 2308개의 유전자를 포함하는 63개의 샘플이 있다. 이 데이터와 이전에 설명한 데이터의 주요 차이점은 이전에 알려지지 않은 마이크로어레이 데이터에는 4개의 개별 유전자 발현 클래스(4 클래스 모드)가 있는 반면, 이전에 논의된 데이터에는 두 클래스(질병의 유무)만 포함된다는 것이다.

[0132] 두 번째 데이터는 27,336개의 유전자가 있는 44개의 마이크로어레이 샘플을 포함하는 다발성 경화증의 전사 프로파일 모음이다. 44개 샘플 중 26개 섹션은 질병의 진행 단계가 있는 다발성 경화증(MS)을 보였고, 18개 샘플은 신경학적 합병증이 없었다. 마이크로어레이 분자 사슬을 생성하기 위한 전사 과정의 차이와 함께 솔루션의 낮은 분류 정확도로 인해 이러한 유형의 데이터에 대한 연구는 거의 수행되지 않았다. 도 7과 도 8에, 분류 정확도가 SRBCT와 MS 보이지 않는 데이터 모두에서 선택된 유효 특징 또는 유전자의 수로 표현된다.

[0133] 효과적인 유전자를 찾기 위해 필터링 방법 중 소프트 앙상블을 사용하는 본 발명에서 제안한 연구 방법과 달리 실제로 여러 군집 지능 기반 학습 알고리즘은 대답 공간 검색을 기반으로 한다. 이러한 알고리즘은 일반적인 것을 검색하는 동안 로컬 최적값에 갇힐 가능성이 있다. 분류기는 일반적으로 다른 알고리즘을 사용하여 매개 변수를 학습하는 과정에서 더 적합하므로, 분류 인식을 향상시키는 것은 유전자 발현에 국한되어서는 안된다.

[0134] 특징 선택 및 분류를 위해 마이크로어레이 데이터의 5가지 유효 유전자에 순위를 지정하는 소프트 앙상블 및 적층 오토인코더를 기반으로 하는 새로운 하이브리드 접근 방식이 제안되었다. 소프트 래퍼(soft wrapper)의 세 가지 방법과 k-NN 알고리즘을 사용한 분류를 결합하여 최종 분류에 필요한 유전자 수가 최소화되었다. 낮은 정확도와 느린 데이터 처리와 관련된 필터링 방법 및 최적화 알고리즘과 비교할 때, 세 가지 래핑 방법을 기반으

로 하는 조합 방법은 높은 수준의 정확도를 유지하면서 적합한 하위 집합을 선택하는 프로세스를 가속화한다.

[0135] 상대적으로 적은 수의 샘플에도 불구하고 새로운 적층 오토인코더는 높은 분류 오류로 이어질 수 있는 과적합의 복잡성을 거의 완전히 피했다. 본 발명의 제안된 방법의 견고성과 단순성 외에도 모델의 일반화 가능성은 분류 오류를 최소화하면서 정확도를 높이기 위해 조정할 수 있는 또 다른 중요한 측면이다.

[0136] 전술한 본 발명은, 프로그램이 기록된 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 매체는, 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다. 따라서, 상기의 상세한 설명은 모든 면에서 제한적으로 해석되어서는 아니되고 예시적인 것으로 고려되어야 한다. 본 발명의 범위는 첨부된 청구항의 합리적 해석에 의해 결정되어야 하고, 본 발명의 등가적 범위 내에서의 모든 변경은 본 발명의 범위에 포함된다.

[0137] 본 발명은 전술한 실시예 및 첨부된 도면에 의해 한정되는 것이 아니다. 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 있어, 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 본 발명에 따른 구성요소를 치환, 변형 및 변경할 수 있다는 것이 명백할 것이다.

부호의 설명

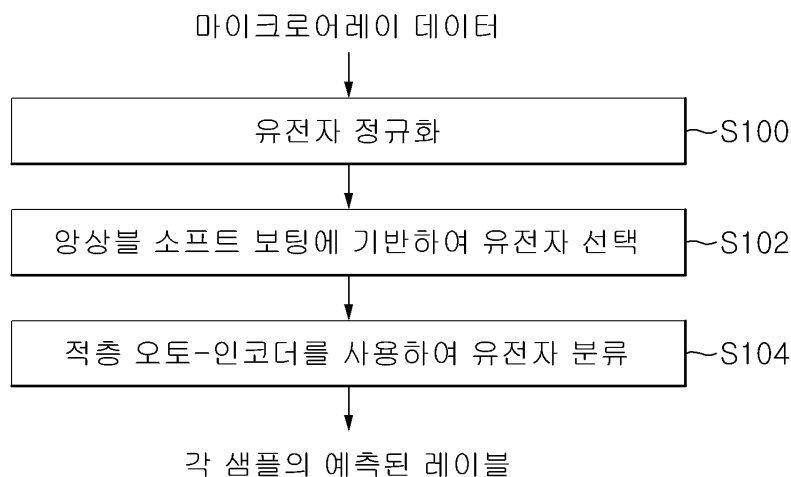
[0138] 200 : 본 발명의 일 실시예에 의한 인공지능 기반 의료 진단 및 분류 방법이 적용될 수 있는 장치

202 : 프로세서 204 : 메모리

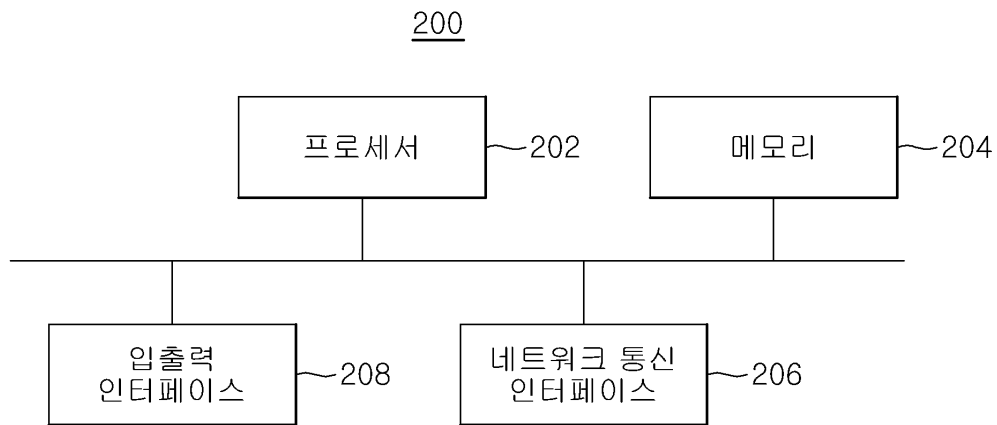
206 : 네트워크 통신 인터페이스 208 : 입출력 인터페이스

도면

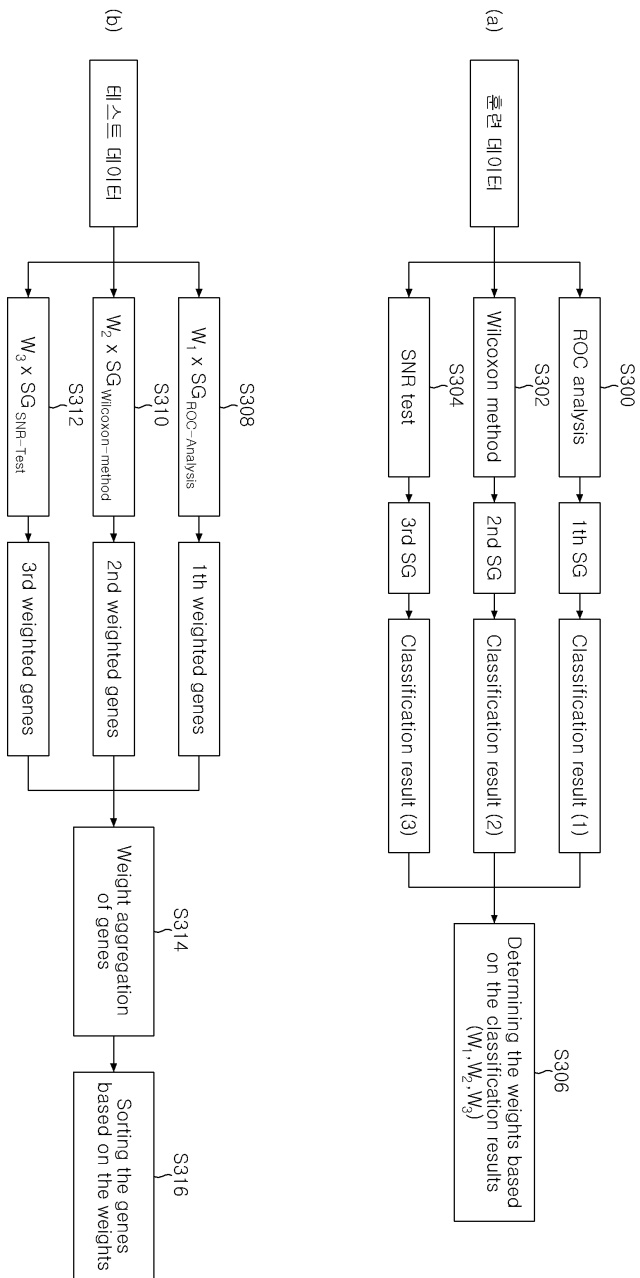
도면1



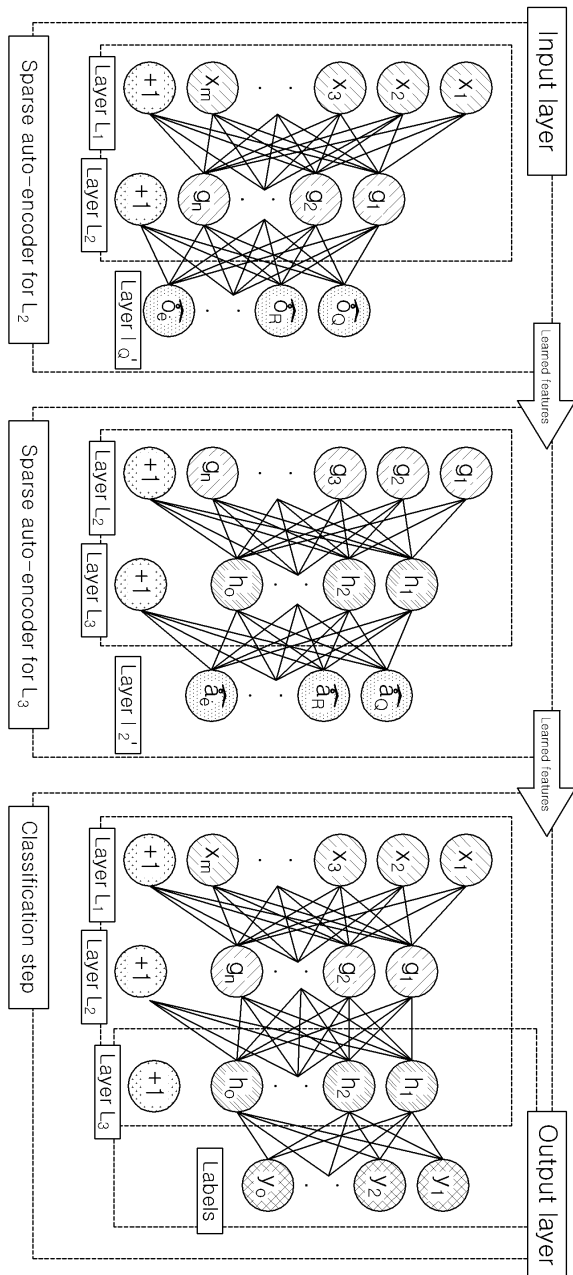
도면2



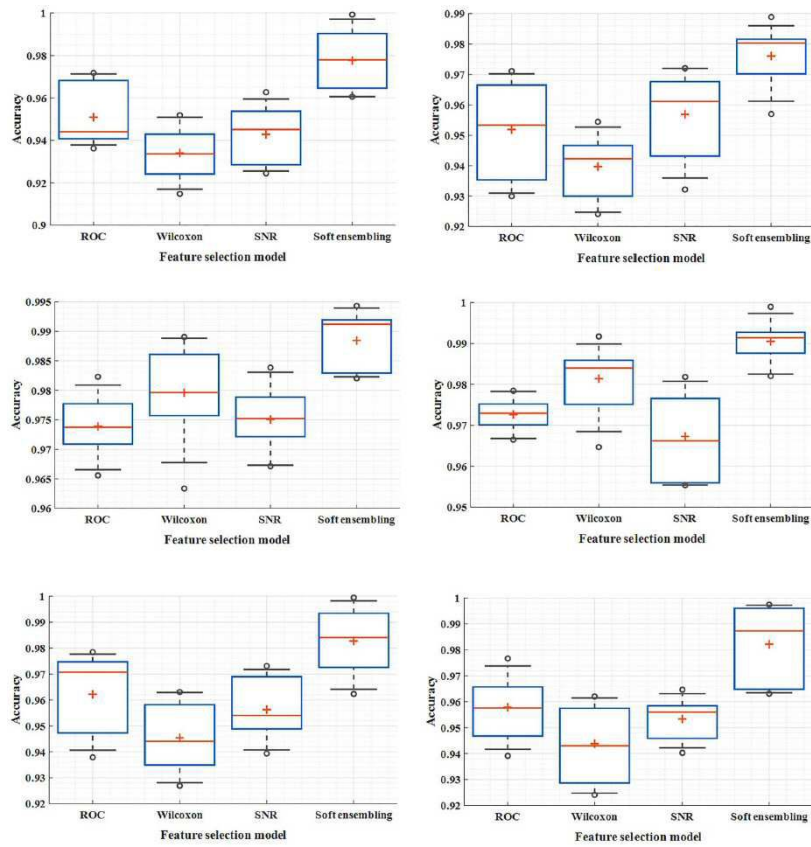
도면3



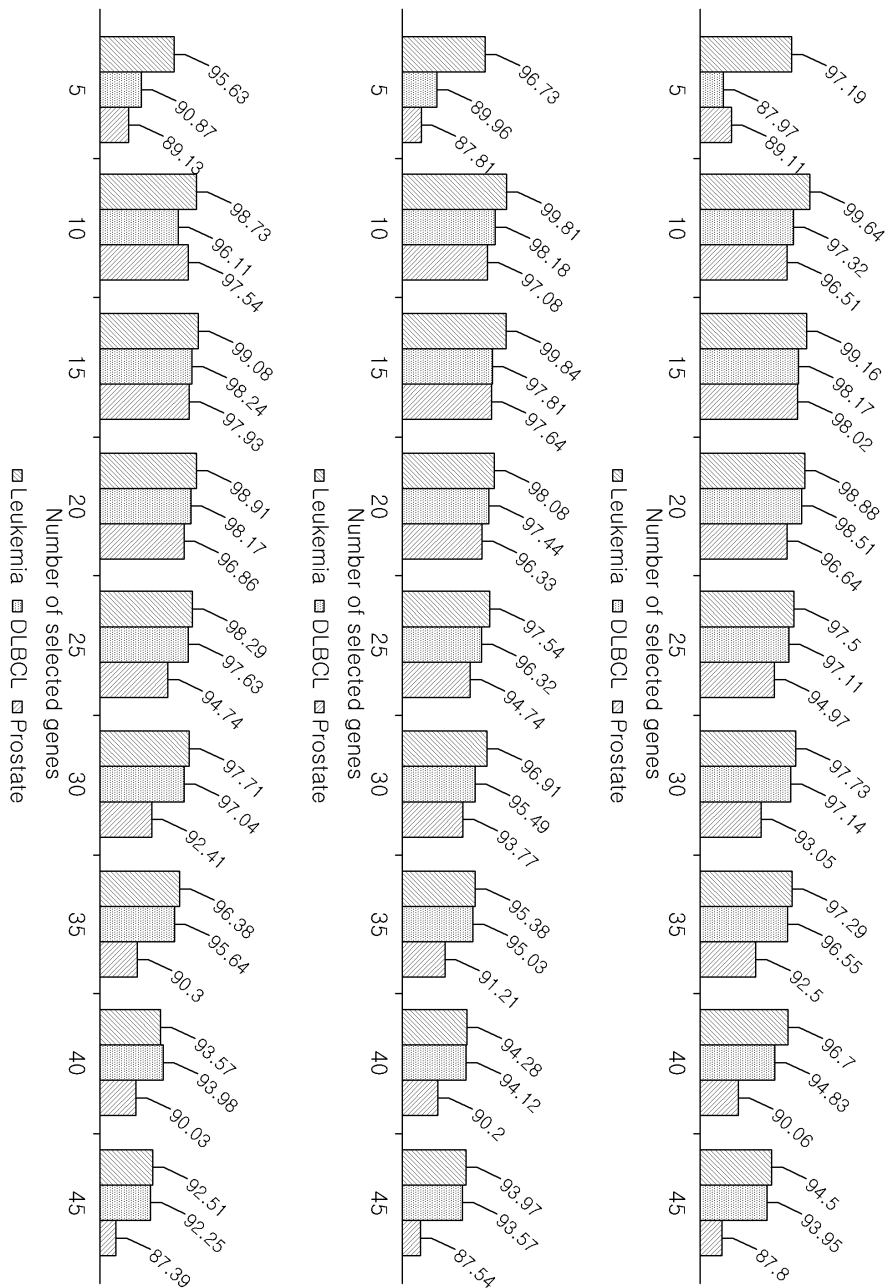
도면4



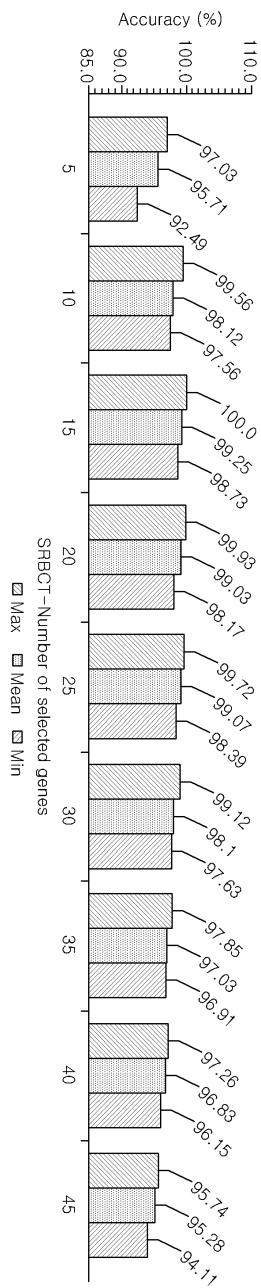
도면5



도면6



도면7



도면8

