(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0136340
(43) 공개일자 2021년11월17일

(51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) *G16H 50/20* (2018.01)
G16H 50/30 (2018.01)
(52) CPC특허분류
G16B 40/00 (2019.02)
G16H 50/20 (2018.01)
(21) 출원번호 10-2020-0054441
(22) 출원일자 2020년05월07일
심사청구일자 2020년05월07일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
박희남
서울특별시 영등포구 국제금융로7길 20 대교아파트 1-902
권오석
울산광역시 울주군 온양읍 망양길 50 온양e편한세상 107동 301호
(74) 대리인
나강은, 강현모, 김경용

전체 청구항 수 : 총 16 항

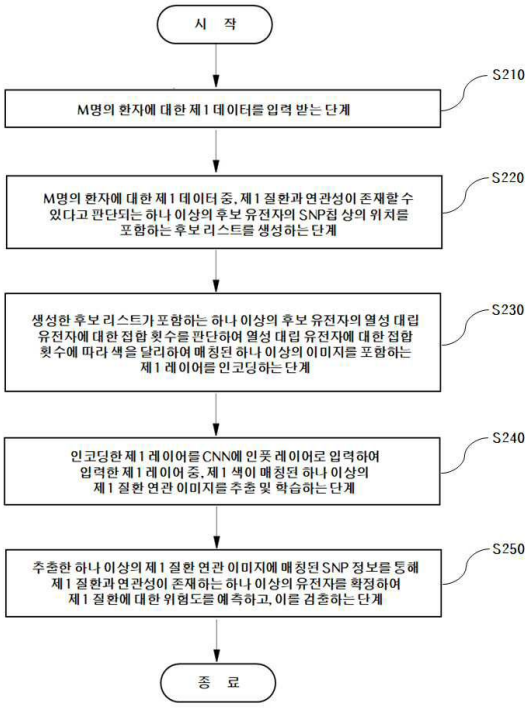
(54) 발명의 명칭 **유전 정보를 활용한 질환 위험도 예측 장치 및 예측 방법**

(57) 요 약

본 발명의 일 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법은 (a) 질환 위험도 예측 장치가 M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계, (b) 상기 질환 위험도 예측 장치가 상기 입력 받은 M

(뒷면에 계속)

대 표 도 - 도2



명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계, (c) 상기 질환 위험도 예측 장치가 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계, (d) 상기 질환 위험도 예측 장치가 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계 및 (e) 상기 질환 위험도 예측 장치가 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계를 포함한다.

(52) CPC특허분류
G16H 50/30 (2018.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1465030845
과제번호 HI19C0114010020
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 의료기기기술개발(R&D)
연구과제명 가상시술 시뮬레이션을 활용한 심방세동 고주파 전극도자 절제술의 임상적 유용성에 대한 전향적 무작위 배정 연구 (가상 로터 매핑에 대한 전극도자 절제술) (연세의대)
기 여 율 1/3
과제수행기관명 연세대학교 산학협력단
연구기간 2020.02.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1465030556
과제번호 HI18C0070010020
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 연구자주도질병극복연구(R&D)
연구과제명 심전도 적용 심장 부정맥 가상분석 시스템 솔루션 개발 및 임상검증
기 여 율 1/3
과제수행기관명 연세대학교 산학협력단
연구기간 2020.01.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711112470
과제번호 2020R1A2B5B01001695
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견연구자지원사업
연구과제명 유전형과 상환현상을 반영한 심방세동 치료제 효과평가 시뮬레이션
기 여 율 1/3
과제수행기관명 연세대학교
연구기간 2020.03.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

(a) 질환 위험도 예측 장치가 M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계;

(b) 상기 질환 위험도 예측 장치가 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계;

(c) 상기 질환 위험도 예측 장치가 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계;

(d) 상기 질환 위험도 예측 장치가 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출하는 단계; 및

(e) 상기 질환 위험도 예측 장치가 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계;

를 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 2

제1항에 있어서,

상기 (a) 단계는,

(a´) 제2 데이터를 입력 받되, 상기 제2 데이터는 상기 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인, 단계;

를 더 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 3

제2항에 있어서,

상기 후보 리스트는,

상기 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인 제2 데이터를 더 포함하는,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 4

제1항에 있어서,

상기 (b) 단계는,

(b-1) 상기 M명의 환자에 대한 제1 데이터 중, N(N은 양의 정수, N≤M)명의 환자에 대한 제1 데이터를 랜덤(Random)으로 추출하는 단계;

를 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 5

제4항에 있어서,

상기 (b-1) 단계 이후에,

(b-2) 상기 랜덤으로 추출한 N명의 환자에 대한 제1 데이터를 GWAS(Genome Wide Association Study, 전장 유전체 연관성 연구)에 적용하여 P-value 절단 임계 기준을 통과하는지 여부에 대한 연관성 분석을 수행하는 단계; 및

(b-3) 상기 연관성 분석 수행 결과 P-value 절단 임계 기준을 통과하여 상기 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 상기 후보 리스트에 기록하는 단계;

를 더 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 6

제5항에 있어서,

상기 (b-3) 단계 이후에,

(b-4) 상기 (b-1) 단계로 회귀하여 상기 (b-1) 단계 내지 (b-3) 단계를 K(K는 양의 정수)회 반복하는 단계;

를 더 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 7

제6항에 있어서,

상기 (b-3) 단계와 (b-4) 단계 사이에,

(b-3´) 상기 후보 리스트에 기록한 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 상기 제1 데이터에서 삭제하는 단계;

를 더 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 8

제6항에 있어서,

상기 K는,

80 내지 120 중 어느 하나인,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 9

제5항에 있어서,

상기 P-value 절단 임계 기준은,

$5*10^{-8}$ 내지 $1*10^{-2}$ 중 어느 하나인,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 10

제1항에 있어서,

상기 열성 대립 유전자에 대한 접합 횟수는,

0, 1 및 2 중 어느 하나인.

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 11

제10항에 있어서,

상기 열성 대립 유전자에 대한 접합 횟수가 2인 경우,

상기 하나 이상의 이미지에 상기 제1 색이 매칭되며,

상기 열성 대립 유전자에 대한 접합 횟수가 1인 경우,

상기 하나 이상의 이미지에 제2 색이 매칭되고,

상기 열성 대립 유전자에 대한 접합 횟수가 0인 경우,

상기 하나 이상의 이미지에 제3 색이 매칭되는,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 12

제1항에 있어서,

상기 CNN은,

풀링 레이어(Pooling Layer)를 미포함하는,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 13

제1항에 있어서,

상기 제1 질환은,

심방 세동인,

유전 정보를 활용한 질환 위험도 예측 방법.

청구항 14

제1항에 있어서,

(f) 상기 제1 질환에 대한 위험도 예측의 근거를 분석하는 단계;

를 더 포함하는 유전 정보를 활용한 질환 위험도 예측 방법.

청구항 15

하나 이상의 프로세서;

네트워크 인터페이스;

상기 프로세서에 의해 수행되는 컴퓨터 프로그램을 로드(Load)하는 메모리; 및

대용량 네트워크 데이터 및 상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하되,

상기 컴퓨터 프로그램은 상기 하나 이상의 프로세서에 의해,

(A) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 오퍼레이션;

(B) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 오퍼레이션;

(C) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 오퍼레이션;

(D) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입

력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 오퍼레이션; 및

(E) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 예측하고, 이를 검출하는 오퍼레이션;

을 실행하는 유전 정보를 활용한 질환 위험도 예측 장치.

청구항 16

컴퓨팅 장치와 결합하여,

(AA) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계;

(BB) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계;

(CC) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계;

(DD) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계; 및

(EE) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 예측하고, 이를 검출하는 단계;

를 실행시키기 위하여,

매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 유전 정보를 활용한 질환 위험도 예측 장치 및 예측 방법에 관한 것이다. 보다 자세하게는 개인의 유전자에 대한 정보를 포함하는 SNP 정보를 활용하여 질환 발생 가능성을 사전에 예측하는 유전 정보를 활용한 질환 위험도 예측 장치 및 예측 방법에 관한 것이다.

배 경 기 술

[0002] 부정맥(Arrhythmia)이란 심장에서 전기 자극이 잘 만들어지지 못하거나 자극의 전달이 제대로 이루어지지 않음으로 인해 규칙적인 수축이 계속되지 못하여 심장 박동이 비정상적으로 빨라지거나 늦어지거나 혹은 불규칙해지는 증상을 의미하며, 심방 세동(Atrial Fibrillation)이 주된 원인으로서 심한 경우 급사나 뇌졸중까지 초래할 수 있다.

[0003] 부정맥의 치료방법으로는 고주파 전극 도자 절제 시술과 같이 심장 조직을 소작함으로써 심장의 전기적 전도를 차단하여 부정맥을 막을 수 있는 수술 요법이 있으나, 이는 심방 세동이 이미 발생하여 부정맥으로까지 번진 경우에 해당하는 치료법이며, 심방 세동의 발생 가능성을 사전에 차단하는 예방 전략에 해당하지는 않는다.

[0004] 한편, 최근 유전자를 분석하여 인간이 건강한 삶을 영위할 수 있도록 이바지하는 연구가 활발하게 진행되고 있는바, 유전자는 개개인의 생리학적 특성이 반영된 생체 지도이기에 유전자를 분석함으로써 특정 질환과 연관성이 존재하는 유전자를 예측한다면 이에 걸맞는 예방 전략을 선택적으로 적용하여 해당 질환이 발현되지 않도록 조절할 수 있기 때문이다.

[0005] 그러나 인간의 유전자 수는 수 만개를 초과하며, 개인별로 그 특성이 상이하기 때문에 특정 질환에 대하여 보편적으로 연관성이 존재하는 유전자를 예측함에 있어서 정확도가 다소 결여되는 것이 현재까지의 연구 결과이며, 예측의 정확도를 높이기 위해 다방면의 노력을 기울이고 있는 실정이다.

[0006] 본 발명은 이러한 사항들을 반영하여 심방 세동과 연관성이 존재하는 유전자를 질환 발생 이전에 높은 정확도로 예측함으로써 심방 세동 발생 가능성의 사전 차단이 가능한 예방 전략을 효과적으로 적용할 수 있도록 이바지하

는 새롭고 획기적인 기술에 관한 것이다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제10-2016-0008040호(2016.07.26)

발명의 내용

해결하려는 과제

[0008] 본 발명이 해결하고자 하는 기술적 과제는 심방 세동과 연관성이 존재하는 유전자를 질환 발생 이전에 높은 정확도로 예측함으로써 심방 세동 발생 가능성의 사전 차단이 가능한 예방 전략을 효과적으로 적용할 수 있도록 이바지하는 유전 정보를 활용한 질환 위험도 예측 장치 및 예측 방법을 제공하는 것이다.

[0009] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법은 (a) 질환 위험도 예측 장치가 M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계, (b) 상기 질환 위험도 예측 장치가 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계, (c) 상기 질환 위험도 예측 장치가 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계, (d) 상기 질환 위험도 예측 장치가 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계 및 (e) 상기 질환 위험도 예측 장치가 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계를 포함한다.

[0011] 일 실시 예에 따르면, 상기 (a) 단계는, (a´) 제2 데이터를 입력 받되, 상기 제2 데이터는 상기 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인, 단계를 더 포함할 수 있다.

[0012] 일 실시 예에 따르면, 상기 후보 리스트는, 상기 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인 제2 데이터를 더 포함할 수 있다.

[0013] 일 실시 예에 따르면, 상기 (b) 단계는, (b-1) 상기 M명의 환자에 대한 제1 데이터 중, N(N은 양의 정수, N<M)명의 환자에 대한 제1 데이터를 랜덤(Random)으로 추출하는 단계를 포함할 수 있다.

[0014] 일 실시 예에 따르면, 상기 (b-1) 단계 이후에, (b-2) 상기 랜덤으로 추출한 N명의 환자에 대한 제1 데이터를 GWAS(Genome Wide Association Study, 전장 유전체 연관성 연구)에 적용하여 P-value절단 임계 기준을 통과하는지 여부에 대한 연관성 분석을 수행하는 단계 및 (b-3) 상기 연관성 분석 수행 결과 P-value 절단 임계 기준을 통과하여 상기 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 상기 후보 리스트에 기록하는 단계를 더 포함할 수 있다.

[0015] 일 실시 예에 따르면, 상기 (b-3) 단계 이후에, (b-4) 상기 (b-1) 단계로 회귀하여 상기 (b-1) 단계 내지 (b-3) 단계를 K(K는 양의 정수)회 반복하는 단계를 더 포함할 수 있다.

[0016] 일 실시 예에 따르면, 상기 (b-3) 단계와 (b-4) 단계 사이에, (b-3´) 상기 후보 리스트에 기록한 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 상기 제1 데이터에서 삭제하는 단계를 더 포함할 수 있다.

[0017] 일 실시 예에 따르면, 상기 K는, 80 내지 120 중 어느 하나일 수 있다.

[0018] 일 실시 예에 따르면, 상기 P-value 절단 임계 기준은, 5*10-8 내지 1*10-2 중 어느 하나일 수 있다.

[0019] 일 실시 예에 따르면, 상기 열성 대립 유전자에 대한 접합 횟수는, 0, 1 및 2 중 어느 하나일 수 있다.

[0020] 일 실시 예에 따르면, 상기 열성 대립 유전자에 대한 접합 횟수가 2인 경우, 상기 하나 이상의 이미지에 상기 제1 색이 매칭되며, 상기 열성 대립 유전자에 대한 접합 횟수가 1인 경우, 상기 하나 이상의 이미지에 제2 색이 매칭되고, 상기 열성 대립 유전자에 대한 접합 횟수가 0인 경우, 상기 하나 이상의 이미지에 제3 색이 매칭될 수 있다.

[0021] 일 실시 예에 따르면, 상기 CNN은, 풀링 레이어(Pooling Layer)를 미포함할 수 있다.

[0022] 일 실시 예에 따르면, 상기 제1 질환은, 심방 세동일 수 있다.

[0023] 일 실시 예에 따르면, (f) 상기 제1 질환에 대한 위험도 예측의 근거를 분석하는 단계를 더 포함할 수 있다.

[0024] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치는 하나 이상의 프로세서, 네트워크 인터페이스, 상기 프로세서에 의해 수행되는 컴퓨터 프로그램을 로드(Load)하는 메모리 및 대용량 네트워크 데이터 및 상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하되, 상기 컴퓨터 프로그램은 상기 하나 이상의 프로세서에 의해 (A) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 오퍼레이션, (B) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 오퍼레이션, (C) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 오퍼레이션, (D) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 오퍼레이션 및 (E) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 오퍼레이션을 실행한다.

[0025] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합하여, (AA) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계, (BB) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계, (CC) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계, (DD) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계 및 (EE) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계를 실행시킨다.

발명의 효과

[0026] 상기와 같은 본 발명에 따르면, 개개인의 생리학적 특성을 반영하는 SNP 정보를 분석의 기초 데이터로 이용하기에 개개인에 최적화된 의료 서비스를 제공할 수 있다는 효과가 있다.

[0027] 또한, 환자 별 SNP 정보인 제1 데이터와 더불어 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP칩 상의 특정 위치인 제2 데이터를 입력 받고, 제2 데이터를 기본적으로 포함하는 후보 리스트를 생성하기에 제1 질환과 연관성이 존재하는 SNP 정보임에도 불구하고 P-Value을 통과하지 못하여 해당 SNP 정보가 누락되는 상황 자체를 방지할 수 있는바, 질환 위험도를 예측함에 있어서의 정확도, 더 나아가 본 발명 자체의 신뢰성을 향상시킬 수 있다는 효과가 있다.

[0028] 또한 M 명의 환자에 대한 제1 데이터 중에서 N명의 환자에 대한 제1 데이터를 랜덤으로 추출하기 때문에 어느 한 명에 대한 제1 데이터에 치우치지 않는 균일한 제1 데이터의 처리가 가능해질 수 있다는 효과가 있다.

[0029] 또한 GWAS를 적용함에 있어 병렬 프로세싱을 수행하며, 후보 리스트에 기록한 하나 이상의 후보 유전자의 SNP칩

상의 위치를 제1 데이터에서 삭제하며 GWAS를 K회 반복 수행하기 때문에 전체 처리 시간이 획기적으로 단축될 수 있다는 효과가 있다.

[0030] 또한, 제1 질환에 대한 위험도를 예측함에 있어서 CNN을 이용하기에, 지속적인 사용을 통해 위험도 예측의 정확도가 비약적으로 향상될 수 있다는 효과가 있다.

[0031] 또한, 제1 질환과 연관성이 존재하는 유전자가 특정되기에, 환자는 자신의 유전자 정보 중에서 제1 질환과 연관성이 존재하는 유전자가 포함되어 있는지를 확인하여 포함되어 있다면 제1 질환 발생 가능성의 사전 차단이 가능한 예방 전략을 적용할 수 있다는 효과가 있다.

[0032] 또한 제1 질환에 대한 위험도 예측의 근거를 분석하여 제공할 수 있는바, 본 발명 자체의 신뢰성을 향상시킬 수 있다.

[0033] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해 될 수 있을 것이다.

도면의 간단한 설명

[0034] 도 1은 본 발명의 제1 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치가 포함하는 전체 구성을 나타낸 도면이다.

도 2는 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법의 대표적인 단계를 도시한 순서도이다.

도 3은 도 2에 도시된 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법의 순서도에서 S210´ 단계를 더 포함한 순서도이다.

도 4는 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치가 2개의 입력, 보다 구체적으로 제1 데이터와 제2 데이터를 입력 받는 모습을 도시한 도면이다.

도 5는 S220 단계가 포함하는 구체적인 단계를 도시한 순서도이다.

도 6은 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치가 복수 개의 프로세서를 포함하고, 각각의 프로세서가 병렬적으로 동작 가능하도록 구현한 모습을 도시한 도면이다.

도 7은 도 6에 도시된 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치에 있어서, S220-3´ 단계의 수행 결과를 서로 공유하는 모습을 도시한 도면이다.

도 8은 Minor Allele Encoding 단계인 S230 단계를 도식화하여 도시한 도면이다

도 9는 S230 단계, 보다 구체적으로 S230-1 단계 및 S230-2 단계를 수행함으로써 인코딩된 제1 레이어를 별도로 분리하여 도시한 도면이다.

도 10은 도 9에 도시된 제1 레이어를 CNN에 인풋 레이어로 입력하여 제1 질환 연관 이미지를 추출 및 학습하는 모습을 예시적으로 도시한 도면이다.

도 11은 도 10에서 제1 질환에 대한 위험도 예측의 근거를 분석하는 모습을 추가적으로 도시한 도면이다.

도 12는 S260-1 단계 내지 S260-2 단계를 거친 예측의 근거를 분석한 결과이다.

도 13은 일반적인 GWAS 적용 결과이다.

도 14는 P-Value 절단 임계 기준에 따른 ROC 커브를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0035] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시 예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 게시되는 실시 예에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시 예들은 본 발명의 게시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0036] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다. 본 명세서에서 사용된 용어는 실시 예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다.

[0037] 명세서에서 사용되는 "포함한다 (comprises)" 및/또는 "포함하는 (comprising)"은 언급된 구성 요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성 요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0038] 도 1은 본 발명의 제1 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치(100)가 포함하는 전체 구성을 나타낸 도면이다.

[0039] 그러나 이는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.

[0040] 본 발명의 제1 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치(100)는 프로세서(10), 네트워크 인터페이스(20), 메모리(30), 스토리지(40) 및 이들을 연결하는 데이터 버스(50)를 포함할 수 있다.

[0041] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processer Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(10)는 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0042] 네트워크 인터페이스(20)는 본 발명의 제1 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치(100)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(20)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.

[0043] 메모리(30)는 각종 데이터, 명령 및/또는 정보를 저장하며, 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법을 수행하기 위해 스토리지(40)로부터 하나 이상의 컴퓨터 프로그램(41)을 로드할 수 있다. 도 1에서는 메모리(30)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(30)로 이용할 수 있음은 물론이다.

[0044] 스토리지(40)는 하나 이상의 컴퓨터 프로그램(41) 및 대용량 네트워크 데이터(42)를 비임시적으로 저장할 수 있다. 이러한 스토리지(40)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크, 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.

[0045] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해 (A) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 오퍼레이션, (B) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 오퍼레이션, (C) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 오퍼레이션, (D) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출하는 오퍼레이션 및 (E) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 예측하고, 이를 학습하는 오퍼레이션을 실행할 수 있다.

[0046] 지금까지 간단하게 언급한 컴퓨터 프로그램(41)이 수행하는 오퍼레이션은 컴퓨터 프로그램(41)의 일 기능으로 볼 수 있으며, 보다 자세한 설명은 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법에 대한 설명에서 후술하도록 한다.

[0047] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의

명령 및/또는 정보의 이동 경로가 된다.

[0048] 이하, 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법에 대하여 도 2 내지 도 14를 참조하여 설명하도록 한다.

[0049] 도 2는 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법의 대표적인 단계를 도시한 순서도이다.

[0050] 이는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가되거나 삭제될 수 있고, 더 나아가 어느 한 단계가 다른 단계에 포함될 수도 있음은 물론이다.

[0051] 한편, 모든 단계는 본 발명의 제1 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 장치(100, 이하 "질환 위험도 예측 장치"라 한다.)에 의해 수행됨을 전제로 한다.

[0052] 우선, 질환 위험도 예측 장치 (100)가 M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받는다(S210).

[0053] 여기서 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩(미도시)에 포함된 환자 별 SNP 정보이다. 휴먼 게놈 프로젝트에 따라 인간의 유전자(DNA)를 분석한 결과 모든 인간들은 거의 대부분인 99.9%의 염기 서열이 동일하되 0.1% 정도만이 미세한 차이가 있으며, 모든 인간들의 유전자 차이는 특정 지점마다 되풀이해서 1개의 염기의 차이로 나타나는바, 이때 나타나는 염기를 단일 염기라 하며, 이러한 단일 염기가 다형성을 가짐을 의미하는 것이 SNP이다. 이러한 SNP의 유전적 차이 때문에 생리학적 특성이 달라지게 되는바, SNP 정보를 효과적으로 이용한다면 개개인에 최적화된 의료 서비스를 제공할 수 있으며, 본 발명 역시 이러한 측면을 적극적으로 이용하고자 한다.

[0054] 제1 데이터는 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)와 연결된 SNP칩(미도시), 또는 SNP칩(미도시)과 연결된 다른 디바이스(미도시)를 통해 입력 받을 수 있는바, 여기서 입력은 수신의 의미도 포함하는 광의의 개념이라 할 것이며, SNP칩(미도시)이 포함하는 SNP정보를 특정 경로와 무관하게 획득할 수 있는 경우라면 전부 포함된다 할 것이다.

[0055] 한편, M은 양의 정수로서 그 수치에 제한은 없다. 보다 구체적으로 M이 1이라면 해당 한 명의 환자에 대한 질환 위험도 예측과 이에 대한 학습이 가능해지기에 개개인에 최적화된 의료 서비스 제공이 가능해지며, M이 1을 초과한다면 복수의 환자에 대한 질환 위험도 예측과 이에 대한 학습이 가능해지기에 특정 질환에 대한 보편적인 예방 전략을 수립할 수 있게 되는바, 일석이조(一石二鳥)의 효과가 있다 할 것이다.

[0056] 이러한 S210 단계는 도 3에 도시된 바와 같이 제1 데이터뿐만 아니라 제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인 제2 데이터를 입력 받는 S210´ 단계를 더 포함할 수 있다.

[0057] 도 4를 참조하면 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)가 2개의 입력, 보다 구체적으로 제1 데이터와 제2 데이터를 입력 받는 것을 확인할 수 있는바, 여기서 제2 데이터는 SNP 정보의 형태로서 분석 이전의 로우 데이터 성격을 갖는 제1 데이터와 상이하게 이미 분석이 완료되어 통계적으로 입증된 공지된 데이터이며, 제1 데이터와 더불어 제2 데이터를 함께 입력 받기에 제1 질환과 연관성이 존재하는 SNP 정보임에도 불구하고 후술할 P-Value 절단 임계 기준을 통과하지 못하여 해당 SNP 정보가 누락되는 상황 자체를 방지할 수 있는바, 질환 위험도를 예측함에 있어서의 정확도, 더 나아가 본 발명 자체의 신뢰성을 향상시킬 수 있다는 장점이 있다.

[0058] 한편, 여기서 제1 질환은 특정 질환을 의미하는 단어로서 특정 유전자와 연관성이 존재할 수 있는 어떠한 질환이라도 제1 질환이 될 수 있으며 그 종류에 제한은 없다 할 것이나, 이하의 설명에선 제1 질환을 심방 세동으로 특정하여 설명을 이어가도록 한다.

[0059] M명의 환자에 대한 제1 데이터를 입력 받았다면, 질환 위험도 예측 장치(100)가 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 SNP칩 상의 위치를 포함하는 후보 리스트를 생성한다(S220).

[0060] 이러한 S220 단계의 핵심적인 내용은 후보 리스트를 생성하는 것인바, 그에 따라 다음과 같은 단계를 포함할 수 있다. 이하 설명하도록 한다.

[0061] 도 5는 S220 단계가 포함하는 구체적인 단계를 도시한 순서도이다.

[0062] 이는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가되거나 삭제

될 수 있고, 더 나아가 어느 한 단계가 다른 단계에 포함될 수도 있음은 물론이다.

[0063] 우선, M명의 환자에 대한 제1 데이터 중, N(N은 양의 정수, N≤M)명의 환자에 대한 제1 데이터를 랜덤(Random)으로 추출한다(S220-1).

[0064] 여기서 N은 양의 정수로서 앞서 S210 단계에서 설명한 M 이하의 수인바, 이는 M 명의 환자에 대한 제1 데이터 중에서 선택하는 것이기 때문이며, M 명의 환자에 대한 제1 데이터 중에서 N명의 환자에 대한 제1 데이터를 랜덤으로 추출하기 때문에 어느 한 명에 대한 제1 데이터에 치우치지 않는 균일한 제1 데이터의 처리가 가능해진다는 장점이 있다.

[0065] 한편, 이하 설명할 GWAS(Genome Wide Association Study, 전장 유전체 연관성 연구)에 랜덤으로 추출한 N명의 환자에 대한 제1 데이터를 적용함에 있어서 M명의 환자에 대한 제1 데이터 중, N 명의 환자에 대한 제1 데이터를 랜덤으로 추출하는 S220-1 단계가 포함되기에 이를 CRR(Computational Randomized Replication)로 명명할 수 있다 할 것이다.

[0066] N명의 환자에 대한 제1 데이터를 랜덤으로 추출했다면, 랜덤으로 추출한 N명의 환자에 대한 제1 데이터를 GWAS에 적용하여 P-value 절단 임계 기준을 통과하는지 여부에 대한 연관성 분석을 수행한다(S220-2).

[0067] GWAS란 Genome Wide Association Study 의 약자로서, 전장 유전체 연관성 연구를 의미하는바, 인간의 생리학적 특성과 게놈 전체의 유전자의 열성 대립 유전자의 발생 빈도 사이의 관련성을 조사하는 공지된 분석 기법 중 하나이다.

[0068] 이러한 GWAS에는 특정 질환과 연관성이 존재하는지 여부를 판단하는 기준인 P-value 절단 임계 기준이 요구되는바, 제1 질환이 심방 세동인 경우 P-value 절단 임계 기준은 $5*10^{-8}$ 내지 $1*10^{-2}$ 중 어느 하나일 수 있다.

[0069] 여기서 P-value 절단 임계 기준의 값은 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)의 관리자 또는 사용자가 자유롭게 설정할 수 있는바, P-value 절단 임계 기준이 작다면 P-value 절단 임계 기준을 통과함에 있어 엄격한 기준이, P-value 절단 임계 기준이 크다면 P-value 절단 임계 기준을 통과함에 있어 유연한 기준이 적용될 것이며, P-value 절단 임계 기준이 0에 가까울수록 명확한(Specific)한 분석이 이루어질 수 있다.

[0070] 연관성 분석을 수행했다면, 연관성 분석 수행 결과 P-value 절단 임계 기준을 통과하여 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 SNP 칩 상의 위치를 후보 리스트에 기록한다(S220-3).

[0071] P-value 절단 임계 기준을 통과한 유전자는 제1 질환과 연관성이 존재할 수 있다는 1차적인 심증이 형성된 유전자인바, 이를 기초로 질환 위험도를 예측하기 위해서는 2차적인 검증이 추가적으로 요구되나, 이에 대해서는 후술하도록 하며, S220-3 단계는 2차적인 검증의 대상이 되는 후보 리스트를 생성하는 과정으로 이해하면 충분하다 할 것이다.

[0072] 한편, 앞서 S220´ 단계에서 제1 데이터와 더불어 제2 데이터를 입력 받은 경우, S220-3단계에서 기록한 후보 리스트에는 제2 데이터가 포함되어 있을 것이며, S220-3 단계에 의해 후보 리스트에 기록한 SNP칩 상의 위치와 제2 데이터가 중복되는 경우, 이에 대한 사항을 별도로 표시하거나, 어느 하나만을 선택하여 중복 자체가 되지 않도록 기록할 수도 있을 것이다.

[0073] 후보 리스트에 기록했다면, S220-1 단계로 회귀하여 S220-1 단계 내지 S220-3 단계를 K(K는 양의 정수)회 반복한다(S220-4).

[0074] 여기서 K는 80 내지 120 중 어느 하나일 수 있으며, S220-1 단계 내지 S220-3 단계가 K회 반복됨으로써 M명의 환자에 대한 제1 데이터 중 N명의 환자에 대한 제1 데이터가 어느 한 명에 대한 제1 데이터에 치우치지 않고 균일하게 추출될 수 있을 것이고, 그에 따른 GWAS 적용도 가능해질 것인바, M명이 환자에 대한 제1 데이터를 최대한으로 이용할 수 있으므로 질환 위험도를 예측함에 있어서의 정확도, 더 나아가 본 발명 자체의 신뢰성을 향상시킬 수 있다는 장점이 있다.

[0075] 그러나 K가 증가함에 따라 S220-1 단계 내지 S220-3 단계를 그만큼 여러 번 반복해야 하므로 처리 시간이 오래 소요될 수 있는바, 이는 병렬 프로세싱을 통해 해결할 수 있다. 예를 들어, 도 6에 도시된 바와 같이 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)가 복수 개의 프로세서(10)를 포함하고, 각각의 프로세서가 병렬적으로 동작 가능하도록 구현한다면 어느 하나의 프로세서에서 S220-1 단계 내지 S220-3 단계를 수행함과 동시에 또 다른 프로세서에서 S220-1 단계 내지 S220-3 단계를 수행할 수 있을 것이므로 전체 처리 시간이 획기적으로 단축될 수 있을 것이다. 이는 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)가 하나의 프로세

서(10)를 포함하고, 해당 프로세서(10)에 따른 내부 처리를 복수 개로 분리한 경우 역시 마찬가지일 것이나, 하나의 프로세서(10)를 분리하여 개별적으로 처리하는 것이므로 처리 속도가 떨어질 가능성이 있는바, 복수 개의 프로세서(10)를 포함하는 상태에서 병렬 프로세싱으로 구현하는 것이 바람직하나, 이에 반드시 한정하는 것은 아니라 할 것이다.

[0076] 한편, S220-3 단계와 S220-4 단계 사이에 후보 리스트에 기록한 하나 이상의 후보 유전자의 SNP칩 상의 위치를 제1 데이터에서 삭제하는 단계(S220-3´)를 더 포함할 수 있는바, 후보 리스트에 기록한 하나 이상의 후보 유전자의 SNP칩 상의 위치는 이미 P-value 절단 임계 기준을 통과한 것이기에, S220-1 단계로 회귀하여 재차 추출될 가능성을 사전에 방지하기 위함이며, 병렬 프로세싱으로 구현하는 경우 도 7에 도시된 바와 같이 S220-3´ 단계의 수행 결과를 서로 공유한다면 프로세싱이 진행될수록 후보 리스트에 기록된 만큼 M명의 환자에 대한 제1 데이터가 줄어들 것이기에 전체 처리 시간이 단축된다는 부수적인 효과까지 얻을 수 있을 것이다.

[0077] 다시 도 2에 대한 설명으로 돌아가도록 한다.

[0078] 후보 리스트를 생성했다면, 질환 위험도 예측 장치(100)가 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩한다(S230).

[0079] S220 단계까지 수행하며 생성한 후보 리스트는 SNP 정보를 포함하는 리스트 형태이기 때문에 질환 위험도를 예측하고 이를 학습하기 위해서는 컴퓨팅 장치가 인식 가능한 형태로 인코딩해야 하며, 본 발명은 CNN(Convolution Neural Network)을 이용하기에 입력 가능한 이미지 형태로 인코딩해야 하고, 해당 단계가 S230 단계, 즉 Minor Allele Encoding 단계에 해당한다.

[0080] 도 8에는 이를 도식화하여 예시적으로 도시한바, 본 발명의 일 실시 예에 따른 질환 위험도 예측 장치(100)는 후보 리스트가 포함하는 하나 이상의 후보 유전자의 SNP칩 상의 위치를 기초로 해당 후보 유전자의 SNP 정보(후보 유전자는 복수 개의 SNP로 구성되며, 우(A), 열(a)서엥 따라 AA, Aa, aA, aa의 경우의 수를 가짐)를 로딩해올 수 있으며, Allele 자체가 대립형질의 유전자이기 때문에 한 쌍의 기준 유전자와 후보 리스트가 포함하는 하나 이상의 후보 유전자의 SNP 정보(한 쌍)의 열성 대립 유전자에 대한 접합 여부를 1차적으로 판단하여 후보 리스트가 포함하는 하나 이상의 후보 유전자의 SNP 정보에 열성 대립 유전자에 대한 접합 여부에 따라 색이 부여된 이미지를 개별적으로 매칭한다(S230-1).

[0081] 여기서 열성 대립 유전자에 대한 접합 여부는 접합 자체가 없는 경우와 접합이 존재하는 경우(동형 접합 또는 이형 접합) 2가지로 나뉘어질 수 있는바, 열성 대립 유전자에 대한 접합 자체가 없는 경우 부여된 색이 연한 회색, 열성 대립 유전자에 대한 접합이 존재하는 경우에 부여된 색이 진한 회색이라는 전제하에 도 8에 도시된 Area 1을 참조하면 기준 유전자를 기준으로 SNP1은 빨간색이 부여된 이미지가 1개, 연한 회색이 부여된 이미지가 1개 매칭되어 있으며, SNP2는 연한 회색이 부여된 이미지가 2개, SNP3은 연한 회색이 부여된 이미지가 1개, 빨간색이 부여된 이미지가 1개, SNP5 내지 SNP6은 빨간색이 부여된 이미지가 2개 매칭되어 있음을 확인할 수 있다.

[0082] 이는 SNP1에서 열성 대립 유전자에 대한 접합이 1회, SNP2에서 열성 대립 유전자에 대한 접합이 0회, SNP3에서 열성 대립 유전자에 대한 접합이 1회, SNP5 내지 SNP6에서 열성 대립 유전자에 대한 접합이 2회라는 것을 의미하는바, 후보 리스트가 포함하는 하나 이상의 후보 유전자의 SNP 정보에서 열성 대립 유전자에 대한 접합 횟수는 0, 1 및 2 중 어느 하나일 수 있다.

[0083] 열성 대립 유전자에 대한 접합 여부에 따라 색이 부여된 이미지를 매칭했다면, 열성 대립 유전자에 대한 접합 횟수가 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어를 인코딩한다(S230-2).

[0084] 열성 대립 유전자에 대한 접합 횟수가 0, 1 및 2 중 어느 하나일 수 있다고 한바, 열성 대립 유전자에 대한 접합 횟수가 2인 경우(동형 접합), 하나 이상의 이미지에 제1색이, 열성 대립 유전자에 대한 접합 횟수가 1인 경우(이형 접합), 하나 이상의 이미지에 제2 색이, 열성 대립 유전자에 대한 접합 횟수가 0인 경우(접합 자체가 없음), 하나 이상의 이미지에 제3색이 매칭될 수 있으며, 앞서 SNP1에서 열성 대립 유전자에 대한 접합이 1회, SNP2에서 열성 대립 유전자에 대한 접합이 0회, SNP3에서 열성 대립 유전자에 대한 접합이 1회, SNP5 내지 SNP6에서 열성 대립 유전자에 대한 접합이 2회라 하였기에 도 8에 도시된 Area 2를 참조하면 SNP1은 제2색(하늘색)이, SNP2는 제3 색(연한 회색)이, SNP3은 제2 색(하늘색)이, SNP5 내지 SNP6은 제1 색(남색)이 매칭된 것을 확인할 수 있다.

[0085] 도 9에는 S230 단계, 보다 구체적으로 S230-1 단계 및 S230-2 단계를 수행함으로써 인코딩된 제1 레이어를 별도

로 분리하여 도시한바, 각각의 이미지는 SNP1 내지 SNPQ(Q는 양의 정수)의 열성 대립 유전자에 대한 접합 횟수를 나타내는 것이며, 제1 색 내지 제3 색 중 어느 하나가 매칭된 이미지의 수는 후보 리스트가 포함하는 후보 유전자의 수에 따라 상이해질 수 있을 것이다.

[0086] S230 단계까지 수행하면 CNN에 인풋 레이어로 입력할 이미지 형태의 제1 레이어가 인코딩되며, 제1 레이어를 인코딩했다면 질환 위험도 예측 장치(100)가 인코딩한 제1 레이어를 CNN에 인풋 레이어로 입력하여 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습한다(S240).

[0087] CNN은 대표적인 하나 이상의 컨볼루션 레이어를 통해 인풋 레이어에서 특징(Feature)을 추출해 결과값을 예측하는 딥러닝(Deep Learning) 기법으로서, 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)에 적용되는 CNN은 기존의 CNN과 상이하게 별도의 풀링 레이어(Pooling Layer)는 포함하지 않는바, SNP 정보에서 해당 환자의 유전 정보의 위치는 SNP칩 상에 고정되어 있음과 동시에 풀링 레이어에 의한 정보 손실을 방지하기 위함이며, 이는 CNN이 본 발명에 최적화되어 커스터마이징된 하나의 대표적인 모습으로 볼 수 있다.

[0088] 이와 더불어 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)에 적용되는 CNN의 또 다른 독자적인 특징을 설명하면 다음과 같다.

[0089] 우선, M이 1을 초과하는 경우 인코딩된 환자 개인의 SNP 정보는 1:1의 비율로 환자군과 대조군(Case/Control)으로 입력되는바, 통상적인 유전자 연구에서는 1:3 또는 그 이상으로 대조군을 상대적으로 크게 구성하여 분석하며, 이 경우 대조군에 편향이 발생하는바, 이를 방지하기 위함이다.

[0090] 또한, K-fold 기법을 적용하여 신경망에서 사용되는 Mini-batch를 대체하며, Cross-calidation 의 효과도 함께 유도한다. 예를 들어, 환자 100명/비환자 900명의 총 1,000명으로 구성된 모집단의 경우, K-fold(K=10, 10등분) 100명은 환자 50명/비환자 50명이 입력 데이터로 연산될 수 있다(Stochastic Gradient Decent)

[0091] 아울러, 컨볼루션 레이어의 출력은 Leaky ReLU를 채택하는바, 일반적으로 사용되는 ReLU는 {0, 1, 2}로 코딩된 SNP 정보 중 0 값에 의해 뉴런이 죽어버리는 영향이 발생하므로 적합하지 않기에 보호 유전자(Protective Gene) 효과를 반영하기 위함이다.

[0092] 도 10에는 도 9에 도시된 제1 레이어를 CNN에 인풋 레이어로 입력하여 제1 질환 연관 이미지를 추출 및 학습하는 모습을 예시적으로 도시한바(가로축은 환자 수, 세로축은 제1 레이어이다), 하나 이상의 컨볼루션 레이어(Con1, Con2, ConN)를 통해 인풋 레이어로부터 특징을 추출하고, 풀리 커넥티드 레이어(Fc1, FcN)를 거쳐 제1 질환 연관 이미지를 추출하는 모습을 확인할 수 있으며, 최종적으로 추출한 제1 질환 연관 이미지는 열성 대립 유전자에 대한 접합 횟수인 2가 제1 색으로 매칭된 이미지이다.

[0093] 한편, 열성 대립 유전자에 대한 접합 횟수가 높을수록 제1 질환에 대한 위험도가 높다고 볼 수 있으나, 이는 제1 레이어가 포함하는 하나 이상의 이미지 전체를 기준으로 판단을 수행하는 것이기에 열성 대립 유전자에 대한 접합 횟수가 2인 제1 색으로 매칭된 이미지만 제1 질환 연관 이미지로 추출되는 것은 아니며, 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)가 상대적으로 판단한다 할 것이다. 예를 들어, 제1 레이어가 포함하는 하나 이상이 이미지가 열성 대립 유전자에 대한 접합 횟수인 1이 제2 색으로 매칭된 이미지와 열성 대립 유전자에 대한 접합 횟수인 0이 제1 색으로 매칭된 이미지만을 포함한다면, 제1 질환 연관 이미지로 열성 대립 유전자에 대한 접합 횟수인 1이 제2 색으로 매칭된 이미지가 추출될 수도 있을 것이다.

[0094] 하나 이상의 제1 질환 연관 이미지를 추출 및 학습했다면 질환 위험도 예측 장치(100)가 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출한다(S250).

[0095] S240 단계에서 추출 및 학습한 하나 이상의 제1 질환 연관 이미지는 이미지의 형태이나, S230 단계에서 인코딩된 이미지이기 때문에 최초 제1 데이터가 포함하는 SNP 정보가 매칭되어 있으며, S250 단계에서는 제1 질환 연관 이미지에 매칭된 SNP정보를 통해 제1 질환과 연관성이 존재하는 유전자를 확정하여 위험도를 예측하고 이를 검출하는 것이다.

[0096] S250 단계까지 수행하면 제1 질환과 연관성이 존재하는 유전자가 특정되기에, 환자는 자신의 유전자 정보 중에서 제1 질환과 연관성이 존재하는 유전자가 포함되어 있는지를 확인하여 포함되어 있다면 제1 질환 발생 가능성의 사전 차단이 가능한 예방 전략을 적용할 수 있다.

[0097] 또한, 제1 질환에 대한 위험도는 0 내지 1의 확률 표현형으로 출력될 수 있는바, 1에 가까울수록 해당 환자가 제1 질환이 발생할 확률이 높으며 0에 가까울수록 확률이 낮기에, 이를 기초로 제1 질환 발생 가능성의 사전 차

단이 가능한 예방 전략을 적용할 수 있다.

[0098] 추가적으로, 도 11 하단에 도시된 바와 같이 제1 질환에 대한 위험도 예측의 근거를 분석하는 단계(S260)가 더 수행될 수 있는바, Explainable AI 기술 중 하나인 Gradient Class Activation Map 기술을 사용하며, 이는 신경망의 예측 근거를 설명하기 위해 블랙박스로 알려진 CNN의 가중치를 유도하여 출력할 수 있는 기술이다. 이하 설명하도록 한다.

[0099] 우선, 수학식 1을 이용하여 각 표본이 가지는 모든 SNP 정보에 대해 제1 질환을 예측하기 위해 어떠한 영향을 미치는지 점수화한다(S260-1).

[0100] 수학식 1: $S^{c}_{Grad\text{-}CAM} = \sum_{k} w^{c}_{k} Conv^{k}$

[0101] 여기서 $S^{c}_{Grad\text{-}CAM}$은 벡터, w^{c}_{k}는 각 클래스(c, 군, 그룹)의 k번째 Feature Map에 대한 역전파법에 의해 유도된 가중치이며, $Conv^{k}$는 k번째 Feature Map의 컨볼루션 가중치이다.

[0102] 앞서 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)에 적용되는 CNN은 보호 유전자 효과를 반영하도록 구현했다고 했던바, 종래의 Grad-CAM과는 달리 ReLU가 적용되지 않는다.

[0103] 점수화했다면, 모든 표본에 대하여 산정한 Grad-CAM score = $S^{c}_{Grad\text{-}CAM}$에서 제1 질환 클래스(c, 군, 그룹)를 대상으로, 제1 질환 환자군 예측에 영향을 주는 평균 Grad-CAM score를 수학식 2를 이용하여 산정한다(S260-2).

[0104] 수학식 2: $Y^{c}_{n} = \sum_{m} S^{c}_{m}$

[0105] 여기서 Y^{c}_{n}는 제1 질환 클래스(c, 군, 그룹)에 대해서 n번째 SNP 정보의 평균 점수이고, m은 환자 표본의 수, S^{c}_{m}은 m번째 환자의 Grad-CAM score이다.

[0106] 도 12는 S260-1 단계 내지 S260-2 단계를 거친 예측의 근거를 분석한 결과이며, 도 13은 일반적인 GWAS 적용 결과이다. 도 12의 하단부를 참조하면, 제1 질환인 심방 세동 환자군의 각 SNP 정보에 대한 $S^{c}_{Grad\text{-}CAM}$점수가 하단의 Heat Map에 적층되어 쌓여 있음을 확인할 수 있으며, 이는 제1 질환인 심방 세동을 예측하기 위해 학습된 CNN의 가중치를 시각화한 결과이다. 한편, 도 12의 상단부는 하단부의 Heat Map의 각 열(Column)의 평균을 나타낸 것이다.

[0107] 이러한 도 12와 도 13을 비교하면 제1 질환인 심방 세동과 연관성이 존재하는 유전자로 알려진 6개의 유전자 PRRX1, PPFIA4, PITX2, HAND2, NEURL, ZFHX3이 S260-1 단계 내지 S260-2 단계를 거친 예측의 근거를 분석한 결과에서도 높은 점수로 기여하는 것을 확인할 수 있다.

[0108] 지금까지 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법에 대하여 설명하였다. 본 발명에 따르면, 개개인의 생리학적 특성을 반영하는 SNP 정보를 분석의 기초 데이터로 이용하기에 개개인에 최적화된 의료 서비스를 제공할 수 있다. 또한, 환자 별 SNP 정보인 제1 데이터와 더불어 제1 질환과 연관성이 존재한다고 임상적으로 입증된 유전자의 SNP칩 상의 특정 위치인 제2 데이터를 입력 받고, 제2 데이터를 기본적으로 포함하는 후보 리스트를 생성하기에 제1 질환과 연관성이 존재하는 SNP 정보임에도 불구하고 P-Value 절단 임계 기준을 통과하지 못하여 해당 SNP 정보가 누락되는 상황 자체를 방지할 수 있는바, 질환 위험도를 예측함에 있어서의 정확도, 더 나아가 본 발명 자체의 신뢰성을 향상시킬 수 있다. 또한 M 명의 환자에 대한 제1 데이터 중에서 N명의 환자에 대한 제1 데이터를 랜덤으로 추출하기 때문에 어느 한 명에 대한 제1 데이터에 치우치지 않는 균일한 제1 데이터의 처리가 가능해질 수 있다. 또한 GWAS를 적용함에 있어 병렬 프로세싱을 수행하며, 후보 리스트에 기록한 하나 이상의 후보 유전자의 SNP칩 상의 위치를 제1 데이터에서 삭제하며 GWAS를 K회 반복 수행하기 때문에 전체 처리 시간이 획기적으로 단축될 수 있다. 또한, 제1 질환에 대한 위험도를 예측함에 있어서 CNN을 이용하기에, 지속적인 사용을 통해 위험도 예측의 정확도가 비약적으로 향상될 수 있다. 또한, 제1 질환과 연관성이 존재하는 유전자가 특정되거나 제1 질환에 대한 위험도가 0 내지 1의 확률 표현형으로 출력되기에, 환자는 자신의 유전자 정보 중에서 제1 질환과 연관성이 존재하는 유전자가 포함되어 있는지를 확인하여 포

함되어 있거나, 제1 질환에 대한 위험도를 확인하여 1에 가깝다면 제1 질환 발생 가능성의 사전 차단이 가능한 예방 전략을 적용할 수 있다. 또한 제1 질환에 대한 위험도 예측의 근거를 분석하여 제공할 수 있는바, 본 발명 자체의 신뢰성을 향상시킬 수 있다.

[0109] 도 14는 P-Value 절단 임계 기준에 따른 ROC 커브를 도시한 도면인바, 예측의 정확도를 나타내는 AUC(Area Under the Curve)가 60 내지 94임을 확인할 수 있으며, 그에 따라 본 발명에 따른 최대 예측의 정확도는 P-Value 절단 임계 기준이 $1*10^{-2}$인 경우에 94%에 달함을 확인할 수 있다.

[0110] 한편, 중복 서술을 방지하기 위해 자세히 설명하지는 않았지만, 본 발명의 제1 실시 예에 따른 질환 위험도 예측 장치(100)와 본 발명의 제2 실시 예에 따른 유전 정보를 활용한 질환 위험도 예측 방법은 동일한 기술적 특징을 포함하는 본 발명의 제3 실시 예에 따른 매체에 저장된 컴퓨터 프로그램으로 구현할 수 있다. 이 경우 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합하여, (AA) M(M은 양의 정수)명의 환자에 대한 제1 데이터를 입력 받되, 상기 제1 데이터는 SNP(Single Nucleotide Polymorphism, 단일 염기 다형성) 칩에 포함된 환자 별 SNP 정보인, 단계, (BB) 상기 입력 받은 M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 상기 SNP 칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계, (CC) 상기 생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 상기 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어(Layer)를 인코딩하는 단계, (DD) 상기 인코딩한 제1 레이어를 CNN(Convolution Neural Network)에 인풋 레이어(Layer)로 입력하여 상기 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계 및 (EE) 상기 추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 상기 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계를 실행할 수 있을 것이다.

[0111] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0112] 10: 프로세서

20: 네트워크 인터페이스

30: 메모리

40: 스토리지

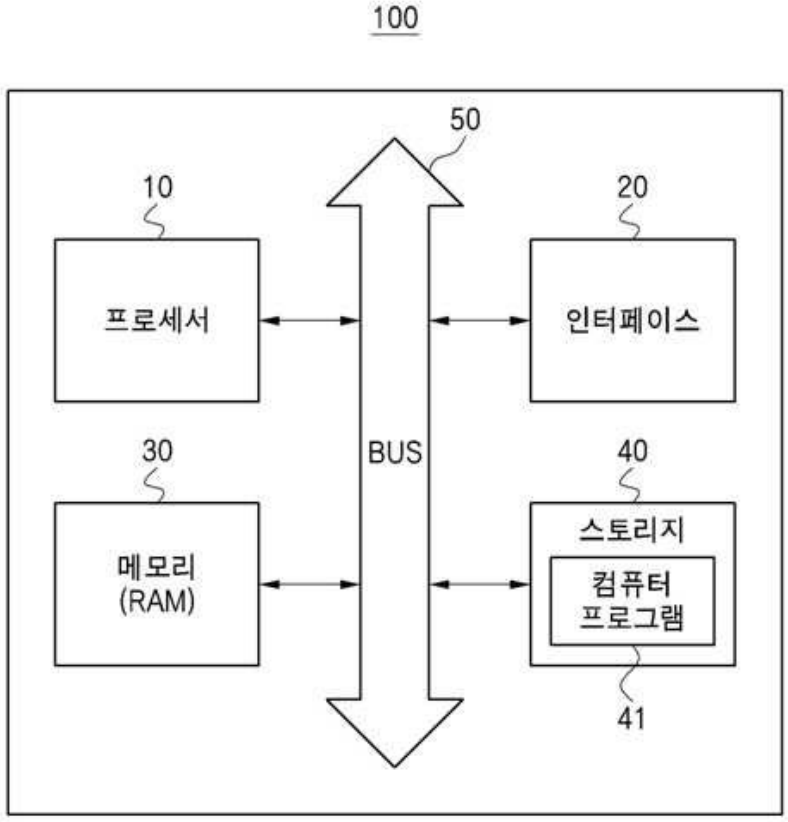
41: 컴퓨터 프로그램

50: 데이터 버스

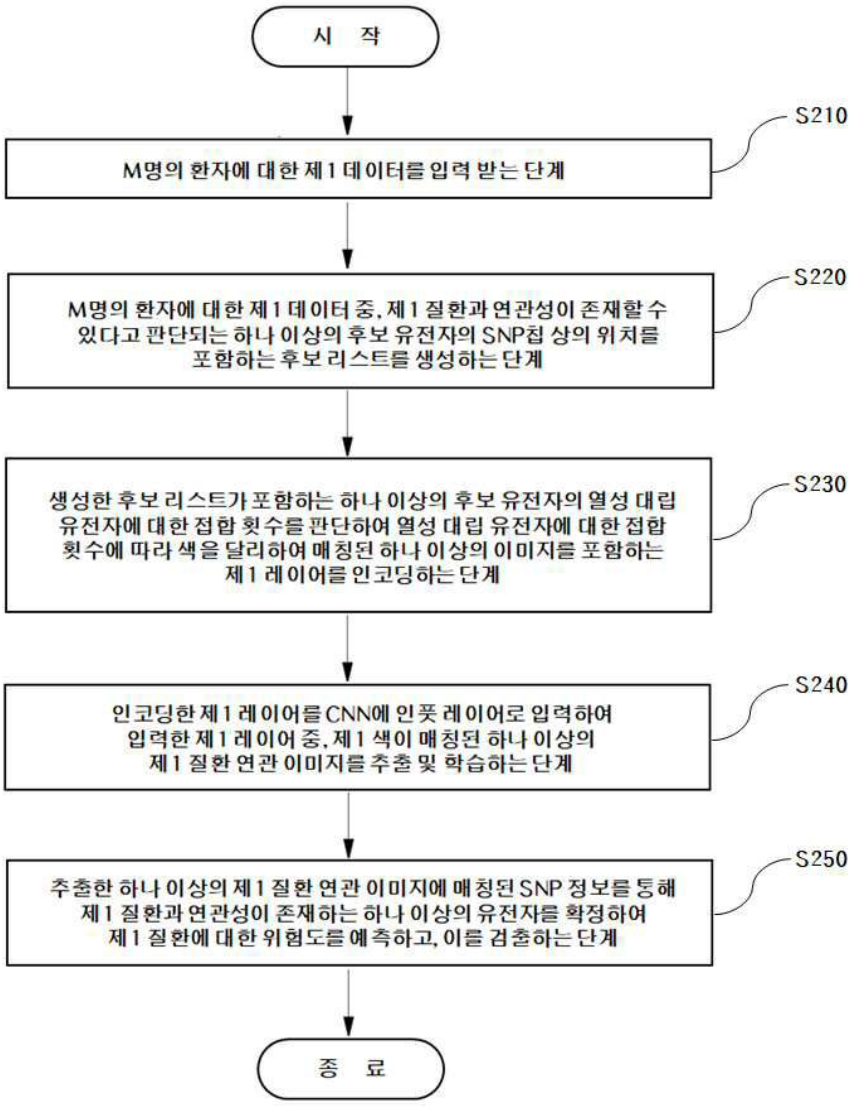
100: 유전 정보를 활용한 질환 위험도 예측 장치

도면

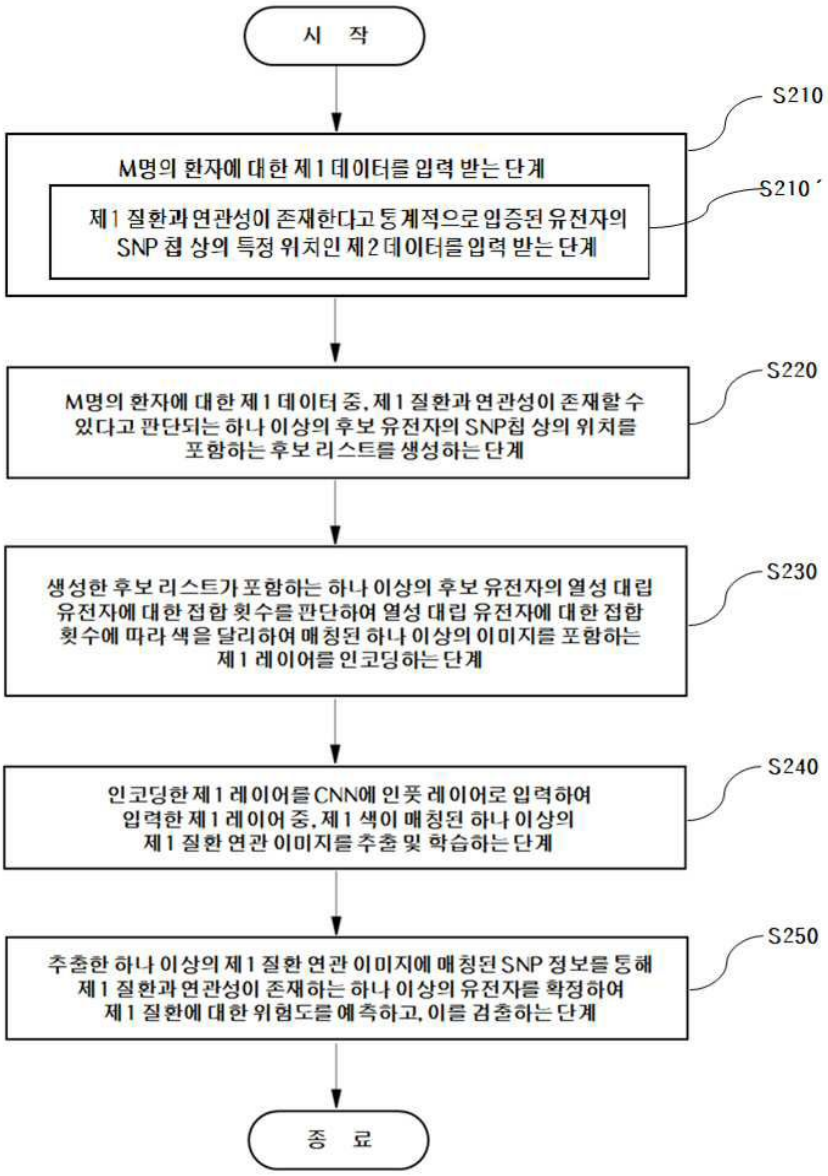
도면1



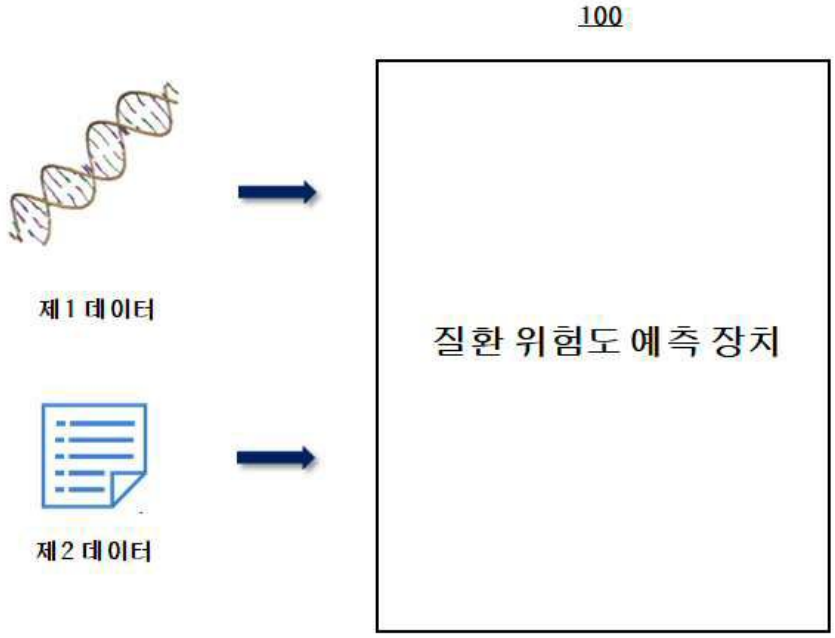
도면2


시 작
S210
M명의 환자에 대한 제1 데이터를 입력 받는 단계
S220
M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 SNP칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계
S230
생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어를 인코딩하는 단계
S240
인코딩한 제1 레이어를 CNN에 인풋 레이어로 입력하여 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계
S250
추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계
종 료

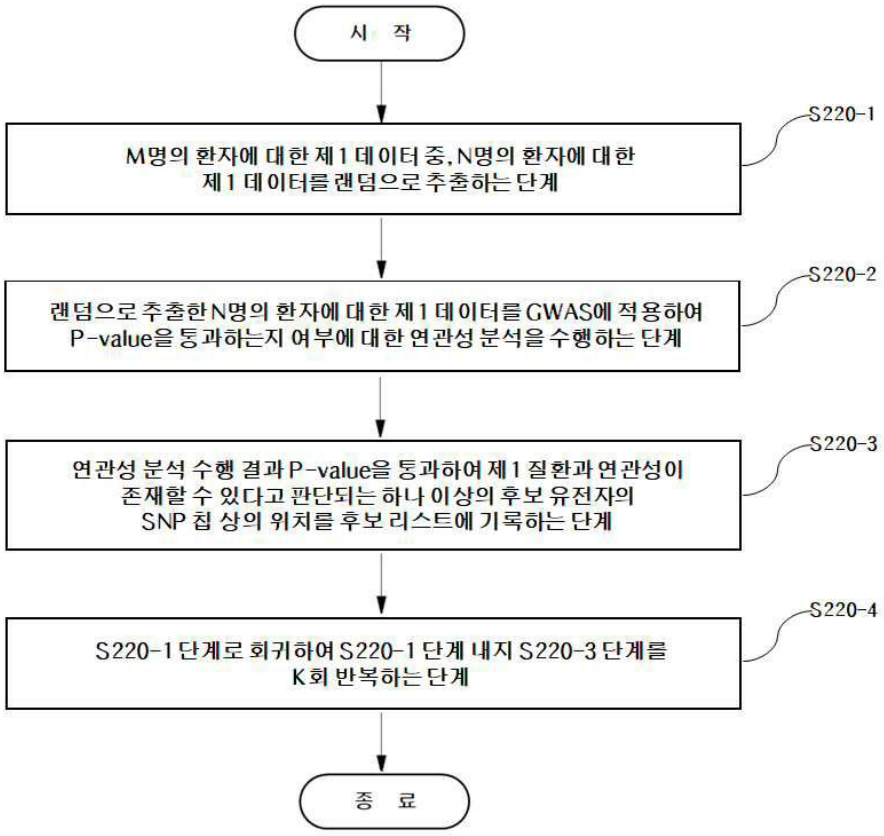
도면3


시 작
S210
M명의 환자에 대한 제1 데이터를 입력 받는 단계
S210'
제1 질환과 연관성이 존재한다고 통계적으로 입증된 유전자의 SNP 칩 상의 특정 위치인 제2 데이터를 입력 받는 단계
S220
M명의 환자에 대한 제1 데이터 중, 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 SNP칩 상의 위치를 포함하는 후보 리스트를 생성하는 단계
S230
생성한 후보 리스트가 포함하는 하나 이상의 후보 유전자의 열성 대립 유전자에 대한 접합 횟수를 판단하여 열성 대립 유전자에 대한 접합 횟수에 따라 색을 달리하여 매칭된 하나 이상의 이미지를 포함하는 제1 레이어를 인코딩하는 단계
S240
인코딩한 제1 레이어를 CNN에 인풋 레이어로 입력하여 입력한 제1 레이어 중, 제1 색이 매칭된 하나 이상의 제1 질환 연관 이미지를 추출 및 학습하는 단계
S250
추출한 하나 이상의 제1 질환 연관 이미지에 매칭된 SNP 정보를 통해 제1 질환과 연관성이 존재하는 하나 이상의 유전자를 확정하여 제1 질환에 대한 위험도를 예측하고, 이를 검출하는 단계
종 료

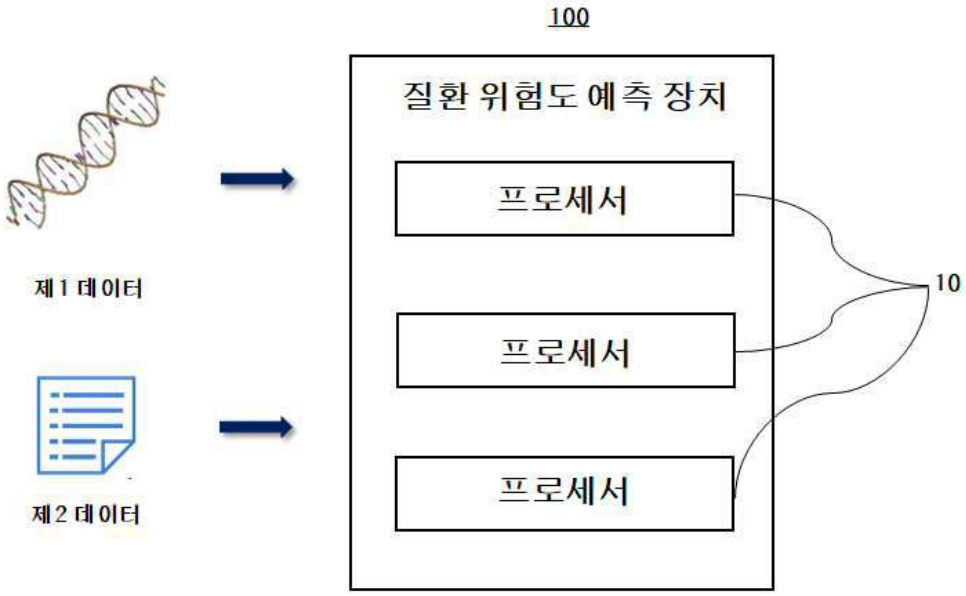
도면4


100
제1 데이터
제2 데이터
질환 위험도 예측 장치

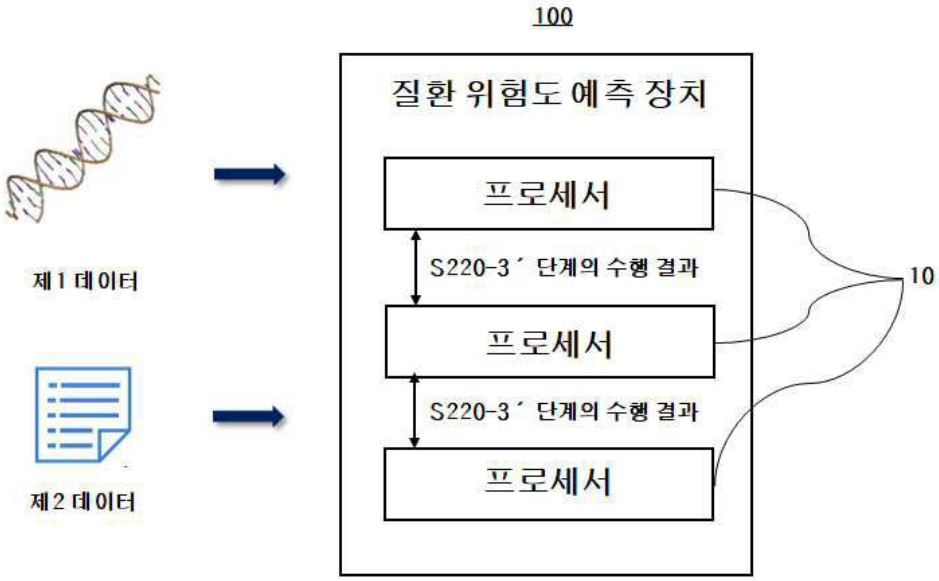
도면5


시 작
M명의 환자에 대한 제1 데이터 중, N명의 환자에 대한 제1 데이터를 랜덤으로 추출하는 단계
S220-1
랜덤으로 추출한 N명의 환자에 대한 제1 데이터를 GWAS에 적용하여 P-value을 통과하는지 여부에 대한 연관성 분석을 수행하는 단계
S220-2
연관성 분석 수행 결과 P-value을 통과하여 제1 질환과 연관성이 존재할 수 있다고 판단되는 하나 이상의 후보 유전자의 SNP 칩 상의 위치를 후보 리스트에 기록하는 단계
S220-3
S220-1 단계로 회귀하여 S220-1 단계 내지 S220-3 단계를 K회 반복하는 단계
S220-4
종 료

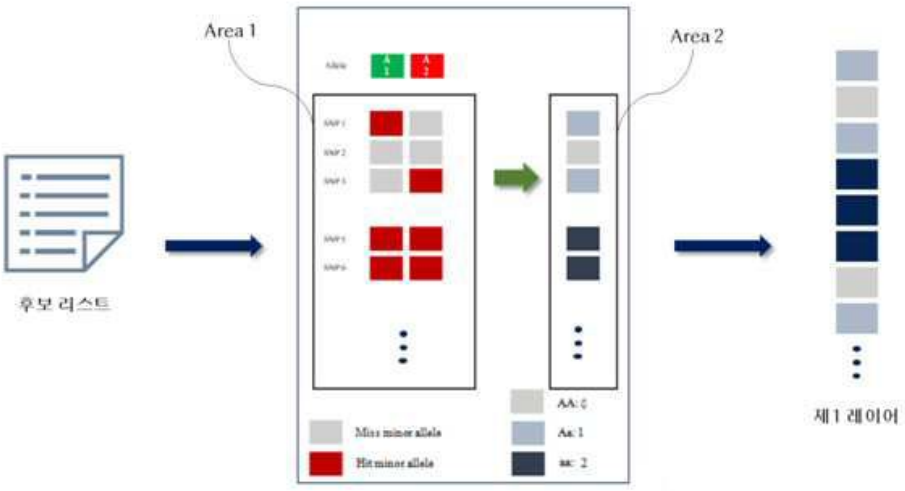
도면6


100
질환 위험도 예측 장치
프로세서
프로세서
프로세서
10
제1 데이터
제2 데이터

도면7

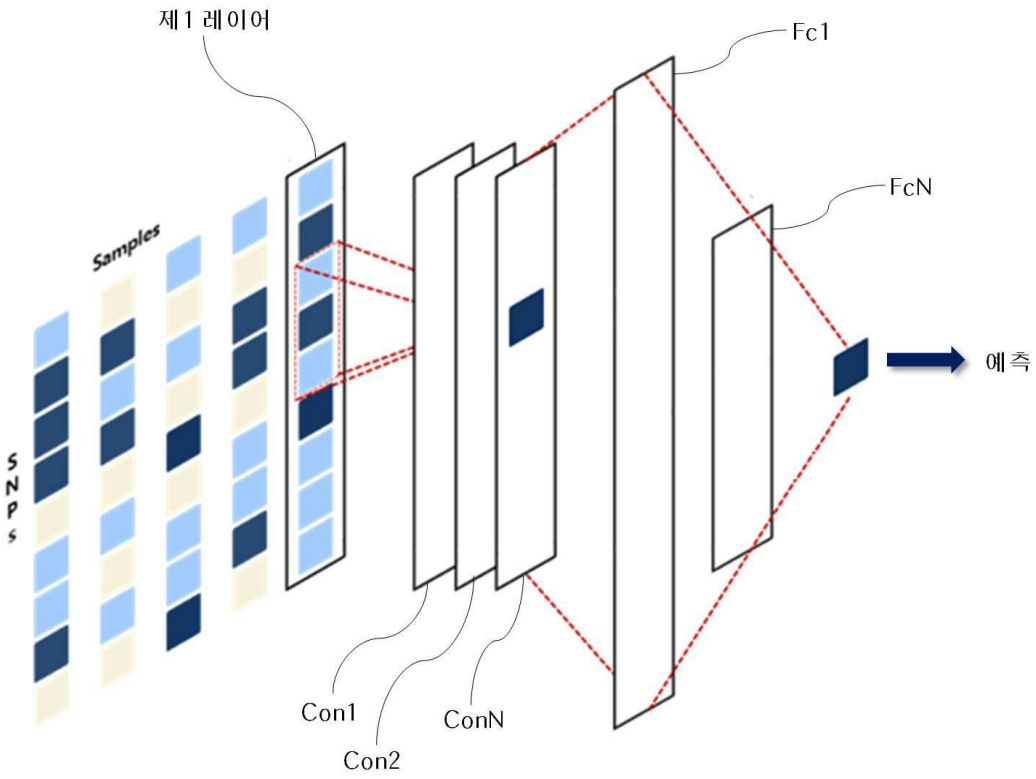

100
질환 위험도 예측 장치
프로세서
S220-3´ 단계의 수행 결과
프로세서
S220-3´ 단계의 수행 결과
프로세서
10
제1 데이터
제2 데이터

도면8

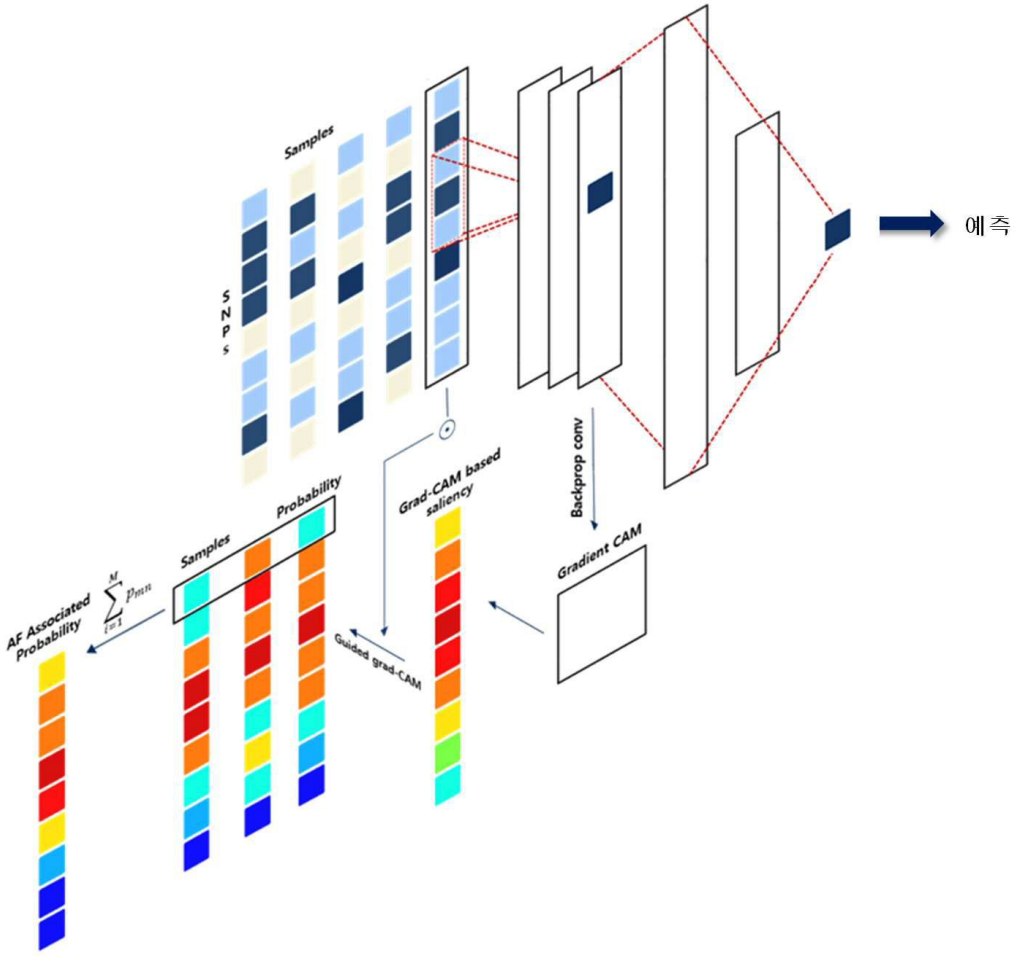


도면9

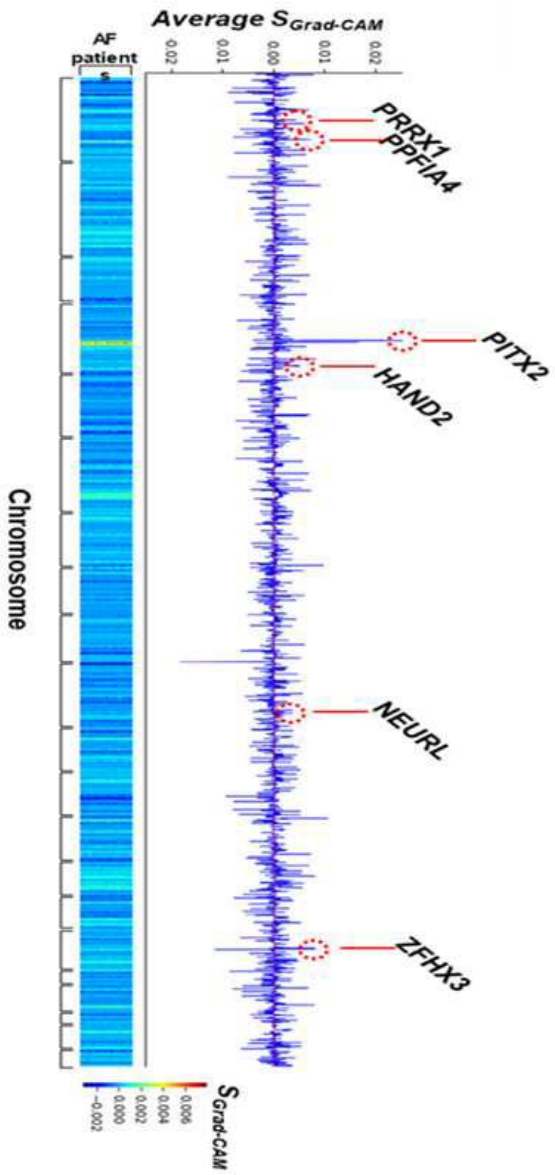
도면10



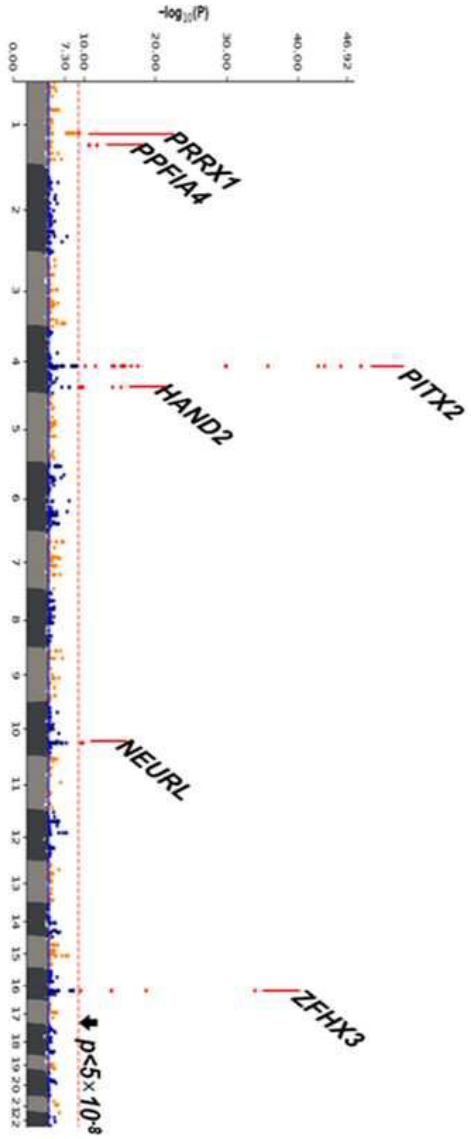
도면11


Samples
SNPs
예측
Backprop conv
Grad-CAM based saliency
Gradient CAM
Guided grad-CAM
Probability
Samples
AF Associated Probability

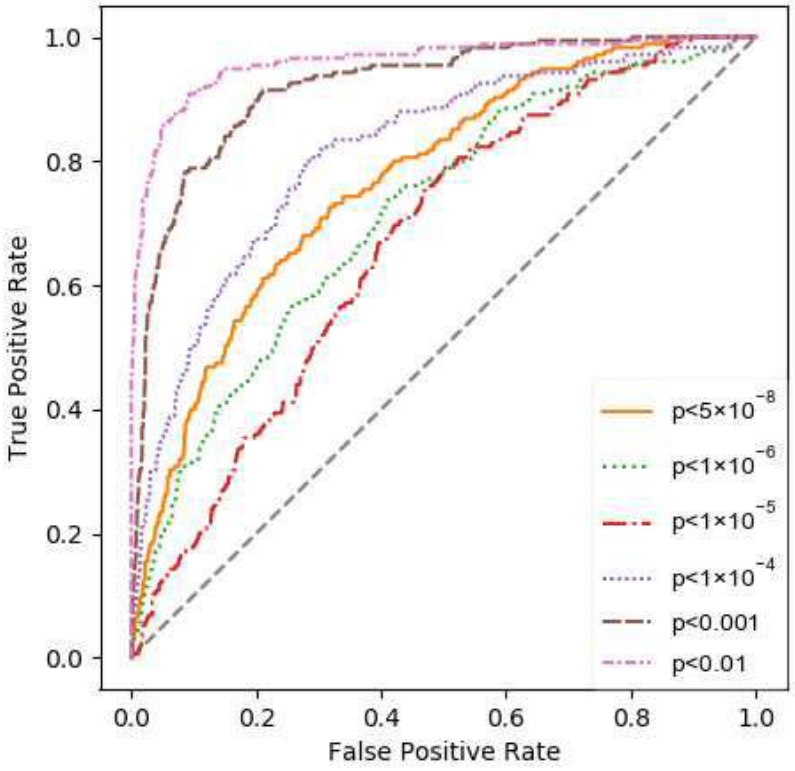
도면12


Average $S_{Grad-CAM}$
AF patient s
0.02
0.01
0.00
0.01
0.02
PRRX1
PPFIA4
PITX2
HAND2
NEURL
ZFHX3
Chromosome
−0.002
0.000
0.002
0.004
0.006
$S_{Grad-CAM}$

도면13


-log10(P)
0.00
7.30
10.00
20.00
30.00
40.00
46.92
PRRX1
PPFIA4
PITX2
HAND2
NEURL
ZFHX3
p<5×10-8

도면14


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
p<5×10^-8
p<1×10^-6
p<1×10^-5
p<1×10^-4
p<0.001
p<0.01