(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0081294
(43) 공개일자 2024년06월07일

(51) 국제특허분류(Int. Cl.)
G16H 50/50 (2018.01) *A61B 5/021* (2006.01)
A61B 5/103 (2006.01) *A61B 5/145* (2006.01)
C12Q 1/6883 (2018.01) *G16B 20/20* (2019.01)
G16B 50/00 (2019.01) *G16H 10/20* (2018.01)
G16H 10/60 (2018.01) *G16H 15/00* (2018.01)
G16H 70/00 (2018.01)
(52) CPC특허분류
G16H 50/50 (2018.01)
A61B 5/021 (2013.01)
(21) 출원번호 10-2023-0063485
(22) 출원일자 2023년05월17일
심사청구일자 2023년05월17일
(30) 우선권주장
1020220163487 2022년11월29일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
한동대학교 산학협력단
경상북도 포항시 북구 흥해읍 한동로 558
(72) 발명자
김난희
서울특별시 서초구 신반포로 45(반포동, 반포아파트)
박소영
서울특별시 서초구 양재대로2길 90(우면동, 서초힐스)
(뒷면에 계속)
(74) 대리인
특허법인태백

전체 청구항 수 : 총 15 항

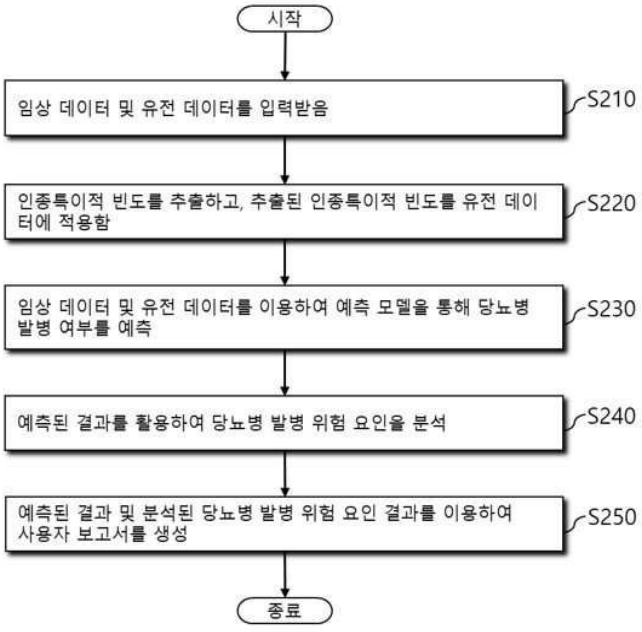
(54) 발명의 명칭 **임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법 및 그 시스템**

(57) 요 약

본 발명은 임상 데이터와 유전자형 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법 및 그 시스템에 관한 것이다. 본 발명에 따르면 임상 데이터 및 유전 데이터를 입력받는 단계; 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 상기 인종특이적 빈도를 상기 입력된 유전 데이터에 적용

(뒷면에 계속)

대 표 도 - 도2



하는 데이터 처리 단계; 및 상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 일정 기간 내의 당뇨병 발병 여부를 예측하는 단계를 포함하며, 상기 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 단계; 및 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 단계를 더 포함할 수 있다.

이와 같이 본 발명에 따르면, 임상 데이터와 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 추출한 인종특이적 빈도를 적용한 유전 데이터를 이용하여 일정 기간 내의 사용자의 당뇨병 발병 확률을 산출함으로써 당뇨병 발병 여부를 예측할 수 있다. 또한, 사용자의 환경적 요인 및 선천적 요인을 고려하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 포함한 사용자 보고서를 생성할 수 있다.

(52) CPC특허분류
A61B 5/103 (2013.01)
A61B 5/145 (2013.01)
C12Q 1/6883 (2022.01)
G16B 20/20 (2019.02)
G16B 50/00 (2019.02)
G16H 10/20 (2021.08)
G16H 10/60 (2021.08)
G16H 15/00 (2021.08)
G16H 70/00 (2021.08)

(72) 발명자
김민희
경기도 안산시 단원구 광덕동로 25(고잔동, 안산레이크타운 푸르지오)
김재영
경기도 시흥시 배곧4로 106-25(배곧동, 호반베르디움 센트로하임아파트)
안태진
서울특별시 서초구 서초중앙로29길 28, 306동 210호 (반포동, 반포미도아파트)
정윤성
경상북도 포항시 북구 장량로128번길 18-10, 101호 (장성동, 이정팰리스)
한선교
전라남도 순천시 자경3길 34, 703호 (조곡동, 금호아파트)
장은희
경상북도 포항시 북구 천마로46번길 28-9, 203호 (양덕동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711151151
과제번호	S0252-21-1001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신산업진흥원
연구사업명	AI정밀의료솔루션(닥터앤서2.0)개발(R&D)
연구과제명	AI정밀의료솔루션(닥터앤서2.0)개발
기 여 율	1/1
과제수행기관명	분당서울대학병원
연구기간	2021.04.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

임상 데이터 및 유전 데이터를 입력받는 단계;

당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 상기 인종특이적 빈도를 상기 입력된 유전 데이터에 적용하는 데이터 처리 단계; 및

상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 일정 기간 내의 당뇨병 발병 여부를 예측하는 단계를 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 2

제 1 항에 있어서,

상기 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 단계; 및

예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 단계를 더 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 3

제 1 항에 있어서,

상기 임상 데이터는,

사용자의 혈액, 소변 및 DNA 중 적어도 하나를 포함한 인체 유래물에 대한 분석 데이터 및 생활습관 문진 결과를 포함할 수 있고,

상기 유전 데이터는,

제2형 당뇨병과 관련 있는 70개의 유전 정보와 상기 유전 정보의 위험 대립 유전 정보를 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 4

제 3 항에 있어서,

상기 임상 데이터는,

사용자의 공복혈당, 트리글리세이드, 백혈구 수, 알아닌아미노전달효소, 크레아티닌, 식습관, 비만도, 나이, 성별, 허리둘레, 수축기 혈압, 이완기 혈압, 고혈압 과거력, 고지혈증 과거력, 운동습관, 당뇨병 가족력, 흡연 여부 및 음주 여부 중 적어도 하나를 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 5

제 1 항에 있어서,

상기 데이터 처리 단계는,

입력된 유전 데이터에서 주 대립 유전자, 부 대립 유전자 및 위험 대립 유전 정보를 이용하여 유전자형 인코딩 정보를 추출하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 6

제 5 항에 있어서,

상기 데이터 처리 단계는,

상기 수행된 유전자형 인코딩 정보를 이용하여 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 상기 유전 데이터에 인종특이적 빈도를 적용하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 7

제 1 항에 있어서,

상기 당뇨병 발병 여부를 예측하는 단계는,

상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 향후 5년 내의 사용자의 제2형 당뇨병 발병 확률을 0 내지 1의 값으로 출력하여 당뇨병 발병 여부를 예측하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 8

제 1 항에 있어서,

상기 당뇨병 발병 여부를 예측하는 단계에서, 상기 예측 모델은 다음의 시그모이드 함수($\sigma(x_c)$)를 이용한 수학식을 통해 하나의 데이터 c 에 대한 비용함수를 연산하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법:

$$\mathcal{L}_c = -p_c y_c \log \sigma(x_c) - (1 - y_c) \log\left(1 - \sigma(x_c)\right)$$

여기서, y는 환자의 당뇨병 유무를 이진수로 나타낸 실제 정답이고, x는 예측 모델이 입력된 임상 데이터 및 유전 데이터를 이용하여 예측한 예측 값이며, p_c 는 비용함수의 가중치 값으로 현 시점에서는 정상이나 당뇨병이 발병되는 환자를 정확히 예측했을 때 부과된다.

청구항 9

제 2 항에 있어서,

상기 위험 요인을 분석하는 단계는,

상기 임상 데이터 및 유전 데이터와 상기 예측 결과 간의 관계를 SHAP(Shapley additive explanation) 알고리즘을 이용하여 상기 예측된 결과에 영향을 미치는 복수의 주요 변수 및 각 주요 변수의 기여도를 분석하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 10

제 9 항에 있어서,

상기 위험 요인을 분석하는 단계에서,

상기 SHAP 알고리즘은 다음의 수학식을 이용하여 SHAP 값(SHAP value)을 획득하기 위한 복수의 주요 변수의 기여도의 차이를 연산하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법:

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{x_j\}) - val(S))$$

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

여기서, ϕ_j 는 j번째 주요 변수가 임상 데이터 및 유전 데이터에 대해 평균 예측 값과 입력 값을 비교 및 예측

할 때 가지는 기여도이며, p 는 전체 변수의 집합이고, S 는 상기 전체 변수의 집합에서 j번째 데이터를 제외한 부분 집합이며, x 는 적어도 하나의 입력 값에 대한 주요 변수 값의 벡터이며, $\mathrm{val}_x(S)$ 는 집합 S 에 포함되지 않은 모든 변수들을 배제한(Marginalized) 변수 값들의 기여도이다.

청구항 11

제 10 항에 있어서,

상기 위험 요인을 분석하는 단계는, 다음의 수학식을 이용하여 SHAP 값의 추정치를 연산하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법:

$$\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$$

$$\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M}\phi_j^m$$

여기서, x는 당뇨병 예측에 사용되는 임상 데이터 및 유전 데이터의 변수이고, M는 반복 횟수이며, x_{+j} 는 기여도를 구하고 싶은 변수가 포함된 입력 값이고, x_{-j} 는 j번째 변수를 제외한 입력 값이다.

청구항 12

제 9 항에 있어서,

상기 위험 요인을 분석하는 단계는,

환경적 요인 분석을 위해 상기 임상 데이터에 대해 군집화를 수행하고,

선천적 요인 분석을 위해 상기 유전 데이터에 대해 군집화를 수행하여 제2형 당뇨병 발병에 영향을 미치는 위험 그룹 정보를 생성하는 임상 데이터와 유전자형 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 13

제 12 항에 있어서,

상기 사용자 보고서를 생성하는 단계는,

산출된 제2형 당뇨병 발병 확률과 상기 분석된 결과를 이용하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 생성하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법.

청구항 14

임상 데이터 및 유전 데이터를 관리하는 데이터베이스부;

당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 상기 인종특이적 빈도를 상기 입력된 유전 데이터에 적용하는 데이터 처리부; 및

상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 일정 기간 내의 당뇨병 발병 여부를 예측하는 예측부를 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 시스템.

청구항 15

제 14 항에 있어서,

상기 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 분석부; 및 상기 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 사용자 보고서 생성부를 더 포함하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법 및 그 시스템에 관한 것으로서, 더욱 상세하게 설명하면 사용자의 임상 데이터 및 유전 데이터를 입력받고, 입력된 유전 데이터에 인종특이적 빈도를 적용하며 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 일정 기간 내의 당뇨병 발병 확률을 산출함으로써 당뇨병을 예측하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법 및 그 시스템에 관한 것이다.

배 경 기 술

[0002] 당뇨병(Diabetes mellitus)은 인슐린의 분비량이 부족하거나 정상적인 기능이 이루어지지 않는 등의 대사질환의 일종으로, 혈중 포도당의 농도가 높아지는 고혈당과 소변에서 포도당을 배출하는 증상을 대표적으로 보인다.

[0003] 당뇨병은 제1형과 제2형으로 구분되며, 제1형 당뇨병(Type 1 diabetes)은 인슐린을 전혀 생산하지 못하는 것이 원인이 되어 발생한다. 제2형 당뇨병(Type 2 diabetes)은 당뇨병 환자의 90%를 차지하며, 노화, 비만, 약물 및 스트레스와 같은 유전적 및 환경적 요인에 의해 인슐린 저항성(insulin resistance)이 감소하여 발생한다.

[0004] 당뇨병이 일단 발병하면 완치가 어렵고 다양한 합병증의 발생 가능성이 높으므로, 당뇨병 발병 전에 예방하는 것이 가장 중요하다.

[0005] 따라서, 근미래의 당뇨병 발병 여부를 예측하고, 예방 방안을 제공하는 기술이 필요한 실정이다.

발명의 내용

해결하려는 과제

[0006] 이와 같이 본 발명에 따르면, 사용자의 임상 데이터 및 유전 데이터를 입력받고, 입력된 유전 데이터에 인종특이적 빈도를 적용하며 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 일정 기간 내의 당뇨병 발병 확률을 산출함으로써 당뇨병을 예측하는 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법 및 그 시스템을 제공하기 위한 것이다.

[0007] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있으며, 본 발명의 실시예에 의해 보다 분명하게 알게 될 것이다. 또한, 본 발명의 목적 및 장점들은 특허청구범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

[0008] 이러한 기술적 과제를 이루기 위한 본 발명의 실시예에 따르면 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법에 있어서, 임상 데이터 및 유전 데이터를 입력받는 단계; 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 상기 인종특이적 빈도를 상기 입력된 유전 데이터에 적용하는 데이터 처리 단계; 및 상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 일정 기간 내의 당뇨병 발병 여부를 예측하는 단계를 포함하며, 상기 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 단계; 및 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 단계를 더 포함할 수 있다.

[0009] 상기 임상 데이터는, 사용자의 혈액, 소변 및 DNA 중 적어도 하나를 포함한 인체 유래물에 대한 분석 데이터 및 생활습관 문진 결과를 포함할 수 있고, 상기 유전 데이터는, 제2형 당뇨병과 관련 있는 70개의 유전 정보와 상기 유전 정보의 위험 대립 유전 정보를 포함할 수 있다.

[0010] 상기 임상 데이터는 사용자의 공복혈당, 트리글리세이드, 백혈구 수, 알아닌아미노전달효소, 크레아티닌, 식습관, 비만도, 나이, 성별, 허리둘레, 수축기 혈압, 이완기 혈압, 고혈압 과거력, 고지혈증 과거력, 운동습관, 당

뇨병 가족력, 흡연 여부 및 음주 여부 중 적어도 하나를 포함할 수 있다.

[0011] 상기 데이터 처리 단계는, 입력된 유전 데이터에서 주 대립 유전자, 부 대립 유전자 및 위험 대립 유전 정보를 이용하여 유전자형 인코딩 정보를 추출할 수 있다.

[0012] 상기 데이터 처리 단계는, 상기 수행된 유전자형 인코딩 정보를 이용하여 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 상기 유전 데이터에 인종특이적 빈도를 적용할 수 있다.

[0013] 상기 당뇨병 발병 여부를 예측하는 단계는, 상기 인종특이적 빈도가 적용된 유전 데이터 및 상기 임상 데이터를 이용하여 기 학습된 예측 모델을 통해 향후 5년 내의 사용자의 제2형 당뇨병 발병 확률을 0 내지 1의 값으로 출력하여 당뇨병 발병 여부를 예측할 수 있다.

[0014] 상기 당뇨병 발병 여부를 예측하는 단계에서, 상기 예측 모델은 다음의 시그모이드 함수($\sigma(x_c)$)를 이용한 수학식을 통해 하나의 데이터 c 에 대한 비용함수를 연산할 수 있다:

[0015] $$\mathcal{L}_c = -p_c y_c \log \sigma(x_c) - (1 - y_c) \log\left(1 - \sigma(x_c)\right)$$

[0016] 여기서, y는 환자의 당뇨병 유무를 이진수로 나타낸 실제 정답이고, x는 예측 모델이 입력된 임상 데이터 및 유전 데이터를 이용하여 예측한 예측 값이며, p_c 는 비용함수의 가중치 값으로 현 시점에서는 정상이나 당뇨병이 발병되는 환자를 정확히 예측했을 때 부과된다.

[0017] 상기 위험 요인을 분석하는 단계는, 상기 임상 데이터 및 유전 데이터와 상기 예측 결과 간의 관계를 SHAP(Shapley additive explanation) 알고리즘을 이용하여 상기 예측된 결과에 영향을 미치는 복수의 주요 변수 및 각 주요 변수의 기여도를 분석할 수 있다.

[0018] 상기 위험 요인을 분석하는 단계에서, 상기 SHAP 알고리즘은 다음의 수학식을 이용하여 SHAP 값(SHAP value)을 획득하기 위한 복수의 주요 변수의 기여도의 차이를 연산할 수 있다:

[0019] $$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{x_j\}) - val(S))$$

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

[0020] 여기서, ϕ_j 는 j번째 주요 변수가 임상 데이터 및 유전 데이터에 대해 평균 예측 값과 입력 값을 비교 및 예측할 때 가지는 기여도이며, p 는 전체 변수의 집합이고, S 는 상기 전체 변수의 집합에서 j번째 데이터를 제외한 부분 집합이며, x 는 적어도 하나의 입력 값에 대한 주요 변수 값의 벡터이며, $val_x(S)$ 는 집합 S 에 포함되지 않은 모든 변수들을 배제한(Marginalized) 변수 값들의 기여도이다.

[0021] 상기 위험 요인을 분석하는 단계는 다음의 수학식을 이용하여 SHAP 값의 추정치를 연산할 수 있다:

[0022] $$\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$$

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$$

[0023] 여기서, x는 당뇨병 예측에 사용되는 임상 데이터 및 유전 데이터의 변수이고, M은 반복 횟수이며, x_{+j} 는 기여도를 구하고 싶은 변수가 포함된 입력 값이고, x_{-j} 는 j번째 변수를 제외한 입력 값이다.

[0024] 상기 위험 요인을 분석하는 단계는 환경적 요인 분석을 위해 상기 임상 데이터에 대해 군집화를 수행하고, 선천적 요인 분석을 위해 상기 유전 데이터에 대해 군집화를 수행하여 제2형 당뇨병 발병에 영향을 미치는 위험 그룹 정보를 생성할 수 있다.

[0025] 상기 사용자 보고서를 생성하는 단계는, 산출된 제2형 당뇨병 발병 확률과 상기 분석된 결과를 이용하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 생성할 수 있다.

[0026] 본 발명의 다른 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 시스템에 있어서 임상 데이터 및 유전 데이터를 관리하는 데이터베이스부; 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 상기 인종특이적 빈도를 상기 입력된 유전 데이터에 적용하는 데이터 처리부; 및 상기 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 일정 기간 내의 당뇨병 발병 여부를 예측하는 예측부를 포함하고, 상기 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 분석부; 및 상기 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 사용자 보고서 생성부를 더 포함할 수 있다.

발명의 효과

[0027] 이와 같이 본 발명에 따르면, 임상 데이터와 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 추출한 인종특이적 빈도를 적용한 유전 데이터를 이용하여 일정 기간 내의 사용자의 당뇨병 발병 확률을 산출함으로써 당뇨병 발병 여부를 예측할 수 있다. 또한, 사용자의 환경적 요인 및 선천적 요인을 고려하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 포함한 사용자 보고서를 생성할 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 시스템의 구성도이다.

도 2는 본 발명의 다른 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법의 순서도이다.

도 3은 도 2의 S210 단계의 임상 데이터 및 유전 데이터의 일 예를 도시하는 예시도이다.

도 4는 도 2의 S220 단계의 한국인의 인종특이적 빈도를 추출하는 과정을 흐름도이다.

도 5는 도 4의 유전자형 인코딩 정보로부터 인종특이적 빈도를 추출하는 일 예를 도시하는 예시도이다.

도 6은 도 4의 S230 단계에서 예측부의 구현 예를 도시하는 도면이다.

도 7은 본 발명의 일 실시예에 따른 분석부의 구현 예를 도시하는 도면이다.

도 8은 도 2의 S240 단계의 제2형 당뇨병 발병 확률 위험 요인을 분석하는 일 예를 도시하는 도면이다.

도 9는 본 발명의 다른 실시예에 따른 예측 모델과 종래의 예측 모델을 비교하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다.

[0030] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0031] 또한 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서, 이는 사용자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0032] 후술하는 본 발명의 실시예에서는 당뇨병 예측 시스템(100)은 사용자의 향후 5년 내의 사용자의 제2형 당뇨병 발병 여부를 예측하는 것으로 구체적인 예를 들어 설명한다. 그러나, 본 발명이 반드시 이로 제한되는 것은 아니다.

[0033] 도 1은 본 발명의 일 실시예에 따른 본 발명의 일 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 시스템(100)의 구성도이다.

[0034] 도 1을 참조하면, 본 발명의 일 실시예에 따른 당뇨병 예측 시스템(100)은 데이터베이스부(110), 데이터 처리부(120), 예측부(130), 분석부(140) 및 사용자 보고서 생성부(150)를 포함하여 구성될 수 있다.

[0035] 데이터베이스부(110)는 임상 데이터 및 유전 데이터를 관리할 수 있다.

[0036] 구체적으로, 데이터베이스(110)는 사용자의 임상 데이터 및 유전 데이터를 입력받을 수 있다.

[0037] 또한, 데이터베이스부(110)는 GWAS(Genome-wide Association studies)와 당뇨관련 문헌들을 통해 1,184개의 유전 정보(Single nucleotide polymorphism, SNP) 및 한국인 맞춤형 유전체칩(Korean Chip, K-Chip)을 사용하여 3,065,958개의 유전 정보를 저장하고, 저장한 두 종류의 유전 정보를 교집합 하여 제2형 당뇨병과 관련있는 70개의 유전 정보와 각 유전 정보의 위험 대립 유전 정보(risk-allele)를 매칭(matching)할 수 있다.

[0038] 이때, 임상 데이터는 사용자의 혈액, 소변 및 DNA 중 적어도 하나를 포함한 인체 유래물에 대한 분석 데이터와 생활습관 문진 결과를 포함한 데이터로 구성될 수 있다.

[0039] 예를 들어, 임상 데이터는 사용자의 공복혈당(Fasting blood sugar, FBS), 트리글리세이드(Triglycerides, Tg), 백혈구 수(white blood cell counts, WBC), 알라닌아미노전달효소(alanine aminotransferase, ALT), 크레아티닌(Creatinine, Cr), 식습관, 비만도(BMI), 나이(Age), 성별(Sex), 허리둘레(Waist), 수축기 혈압(Systolic blood pressure, SBP), 확장기 혈압(Diastolic blood pressure, DBP), 고혈압 과거력, 고지혈증 과거력, 운동(Exercise) 습관, 당뇨병 가족력(Diabetes mellitus Family history, DM_FH), 흡연(smoke) 여부 및 음주(Drink) 여부 중 적어도 하나를 포함한 데이터일 수 있다.

[0040] 또한, 유전 데이터는, 제2형 당뇨병과 관련 있는 70개의 유전 정보와 유전 정보의 위험 대립 유전 정보를 포함할 수 있다.

[0041] 데이터 처리부(120)는 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 인종특이적 빈도를 입력된 유전 데이터에 적용할 수 있다.

[0042] 구체적으로, 데이터 처리부(120)는 주 대립 유전자(Major allele), 부 대립 유전자(Minor allele) 및 위험 대립 유전 정보를 이용하여 유전자형 인코딩(Genotype encoding) 정보를 추출하고, 추출된 유전자형 인코딩 정보를 기 설정된 알고리즘을 통해 한국인의 인종특이적 빈도(race-specific frequency)를 산출하여 추출할 수 있다. 데이터 처리부(120)는 추출된 한국인의 인종특이적 빈도를 유전 데이터에 적용할 수 있다.

[0043] 또한, 데이터 처리부(120)는 데이터베이스부(110)에 입력된 70개의 제2형 당뇨병 관련 유전 정보의 주 대립 유전자(Major allele)를 기준으로 주 대립 유전자 개수에 따라 표기된 유전자형 데이터(Genotype data)(예를 들어 주 대립 유전자가 0개면 유전자형 데이터는 0개로 표기되고, 주 대립 유전자가 1개면 유전자형 데이터가 1개로 표기되는 등)를 구성할 수 있다.

[0044] 예측부(130)는 임상 데이터 및 한국인의 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 당뇨병 발병 여부를 예측할 수 있다.

[0045] 구체적으로, 예측부(130)는 임상 데이터와 한국인의 인종특이적 빈도를 적용한 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 향후5년 내의 제2형 당뇨병 발병 확률을 0 내지 1의 값으로 산출하여 향후 5년 내의 사용자의 제2형 당뇨병 발병 여부를 예측할 수 있다. 여기서, 예측 모델은 인공지능 모델(예를 들어, 딥러닝(Deep learning) 모델, 합성곱 신경망(Convolution neural network, CNN) 모델 및 기계학습(Machine learning) 모델 등) 중 하나일 수 있다.

[0046] 분석부(140)는 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석할 수 있다.

[0047] 구체적으로, 분석부(140)는 임상 데이터 및 한국인의 인종특이적 빈도가 적용된 유전 데이터와 예측부(130)의 예측 결과 간의 관계를 SHAP(Shapely additive explanation) 알고리즘을 이용하여 복수의 주요 변수 및 각 주요 변수의 기여도를 분석할 수 있다.

[0048] 사용자 보고서 생성부(150)는 예측된 결과 및 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성할 수 있다.

[0049] 부연하자면, 사용자 보고서 생성부(150)는 산출된 제2형 당뇨병 발병 확률과 분석부(140)에서 분석된 결과를 이용하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 포함한 사용자 보고서를 생성할 수 있다.

[0050] 이하에서는 도 2 내지 도 9를 이용하여 본 발명의 실시예에 따른 당뇨병 예측 시스템(100)에 대하여 상세하게 설명한다.

[0051] 도 2는 본 발명의 다른 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법의 순서도이다.

[0052] 도 2를 참조하면, 본 발명의 다른 실시예에 따른 임상 데이터와 인종특이적 빈도를 적용한 유전 데이터를 활용한 당뇨병 예측 방법은, 데이터베이스부(110)는 사용자로부터 임상 데이터 및 유전 데이터를 입력받는 단계(S210), 데이터 처리부(120)는 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 인종특이적 빈도를 유전 데이터에 적용하는 단계(S220), 예측부(130)는 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 당뇨병 발병 여부를 예측하는 단계(S230), 분석부(140)는 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석하는 단계(S240) 및 사용자 보고서 생성부(150)는 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성하는 단계(S250)를 포함할 수 있다.

[0053] 먼저, 임상 데이터 및 유전 데이터를 입력받는 단계(S210)에서, 데이터베이스부(110)는 사용자로부터 임상 데이터 및 유전 데이터를 입력받을 수 있다.

[0054] 도 3은 도 2의 S210 단계의 임상 데이터 및 유전 데이터의 일 예를 도시하는 예시도이다.

[0055] 도 3의 (a)를 참조하면, 임상 데이터 및 유전 데이터를 입력받는 단계(S210)에서, 데이터베이스부(110)는 사용자의 건강 및 생활습관 관련 및 검진을 통해 획득한 복수의 데이터 중 제2형 당뇨병과 관련된 18개의 임상 변수를 포함한 임상 데이터를 입력받을 수 있다.

[0056] 도 3의 (b)를 참조하면, 임상 데이터 및 유전 데이터를 입력받는 단계(S210)에서, 데이터베이스부(110)는 제2형 당뇨병과 관련된 70개의 유전 정보와 각 유전 정보의 위험 대립 유전 정보를 포함한 유전 데이터를 입력받을 수 있다.

[0057] 다음으로, 추출된 유전자형 빈도를 유전 데이터에 적용하는 단계(S220)에서, 데이터 처리부(120)는 당뇨 위험을 예측하고자 하는 환자가 속한 인종에서의 인종특이적 빈도를 추출하고, 추출된 인종특이적 빈도를 유전 데이터에 적용할 수 있다.

[0058] 도 4는 도 2의 S220 단계의 한국인의 인종특이적 빈도를 추출하는 과정을 흐름도이고, 도 5는 도 4의 유전자형 인코딩 정보로부터 인종특이적 빈도를 추출하는 일 예를 도시하는 예시도이다.

[0059] 도 4에 도시된 바와 같이, 데이터 처리부(120)는 데이터베이스부(110)에 입력된 유전 데이터에서 주 대립 유전자, 부 대립 유전자 및 위험 대립 유전 정보를 이용하여 유전자형 인코딩 정보를 추출할 수 있다.

[0060] 구체적으로, 데이터 처리부(120)는 입력된 유전 데이터 중 제2형 당뇨병과 관련된 유전 정보에 대해 위험 대립 유전 정보를 기준으로 주 대립 유전자 및 부 대립 유전자를 재구성함으로써 유전자형 인코딩 정보를 추출할 수 있다.

[0061] 도 5에 도시된 바와 같이, 데이터 처리부(120)는 추출된 유전자형 인코딩 정보를 기 설정된 알고리즘을 이용한 프로그램을 통해 한국인의 인종특이적 빈도를 산출함으로써 한국인의 인종특이적 빈도를 추출할 수 있다.

[0062] 또한, 데이터 처리부(120)는 입력된 유전 데이터에 추출된 한국인의 인종특이적 빈도를 적용할 수 있다.

[0063] 다음으로, 당뇨병 발병 여부를 예측하는 단계(S230)에서, 예측부(130)는 임상 데이터와 한국인의 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 당뇨병 발병 여부를 예측할 수 있다.

[0064] 구체적으로, 예측부(130)는 임상 데이터 및 인종특이적 빈도가 적용된 유전 데이터를 이용하여 기 학습된 예측 모델을 통해 사용자의 향후 5년 내의 제2형 당뇨병 발병 확률을 0 내지 1의 값으로 출력하여 당뇨병 발병 여부를 예측할 수 있다.

[0065] 도 6은 도 4의 S230 단계에서 예측부(130)의 구현 예를 도시하는 도면이다.

[0066] 도 6을 참조하면, 예측부(130)는 임상 데이터 및 유전 데이터를 이용하여 완전 연결 계층(Fully-connected layer)으로 구성되고, 이진 교차 엔트로피 손실(Binary cross entropy)을 비용함수로 사용하여 학습된 예측 모델을 통해 향후 5년 내의 사용자의 제2형 당뇨병 발병 확률을 산출할 수 있다.

[0067] 여기서, 예측부(130)는 하기의 시그모이드 함수(Sigmoid function, $\sigma(x_c)$)를 이용한 [수학식 1]을 통해 하나의 데이터 c 에 대한 비용함수($\mathcal{L}_c$)를 사용하여 기 학습된 예측 모델이 오진할 경우 더 많은 가중치(p_c)를 부여함으로써 역전파 과정을 수행할 수 있다.

수학식 1

[0068] $$\mathcal{L}_c = -p_c y_c \log \sigma(x_c) - (1 - y_c) \log\left(1 - \sigma(x_c)\right)$$

[0069] 여기서, y는 환자의 당뇨병 유무를 이진수로 나타낸 실제 정답이고, x는 예측 모델이 입력된 임상 데이터 및 유전 데이터를 이용하여 예측한 예측 값이며, p_c 는 비용함수의 가중치 값으로 현 시점에서는 정상이나 당뇨병이 발병되는 환자를 정확히 예측했을 때 부과된다.

[0070] 또한, y가 1인 경우, $-p_c y_c log\sigma(x_c)$ 는 예측부(130)의 예측 값이 1에 근접할수록 0에 근접한 오차 값을 가지고, 예측부(130)는 오차 값이 커짐으로 인하여 예측 모델이 오진할 경우 예측 모델의 역전파 과정에 $-p_c y_c log\sigma(x_c)$ 을 전달할 수 있다.

[0071] 또한, y가 0인 경우, 예측부(130)는 예측 값이 0에 근접할수록 $-(1-y_c)log(1-\sigma(x_c))$ 을 연산했을 시 0에 가까운 손실 값을 나타낼 수 있다.

[0072] 즉, 예측부(130)는 비용함수의 연산 값이 0에 근접하도록 예측 모델을 학습시킬 수 있다. 이때, 가중치의 비율은 당뇨병이 발병하지 않는 정상 환자의 임상 데이터 및 유전 데이터와 당뇨병이 발병되는 환자의 임상 데이터 및 유전 데이터의 비율을 계산하여 부여된다.

[0073] 일 예로, 예측부(130)는 18개의 임상 변수와 70개의 유전 정보를 사용하여 학습된 예측 모델에 가중치(p)를 8.0으로 부여할 수 있다. 이때, 예측 모델은 선형 레이어 3개가 쌓아진 형태로 각 레이어에서 선형 연산을 수행한 후 비선형 활성화 함수를 적용하여 다음 레이어로 값을 전달할 수 있다.

[0074] 그 다음, 당뇨병 발병 위험 요인을 분석하는 단계(S240)에서, 분석부(140)는 예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석할 수 있다.

[0075] 구체적으로, 분석부(140)는 임상 데이터 및 유전 데이터와 예측부(130)에서 예측된 결과 간의 관계를 SHAP(Shapley additive explanation) 알고리즘을 이용하여 예측된 결과에 영향을 미치는 복수의 주요 변수 및 각 주요 변수의 기여도를 분석할 수 있다.

[0076] 여기서, 분석부(140)는 하기의 [수학식 2]를 이용하여 SHAP 값(SHAP value)을 획득하기 위한 복수의 주요 변수의 기여도의 차이를 연산할 수 있다.

수학식 2

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S))$$

$$val_x(S) = \int \hat{f}(x_1,\ldots,x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

[0077]

[0078] 여기서, ϕ_j 는 j번째 주요 변수가 임상 데이터 및 유전 데이터에 대해 평균 예측 값과 입력 값을 비교 및 예측할 때 가지는 기여도이며, p 는 전체 변수의 집합이고, S 는 전체 변수의 집합에서 j번째 데이터를 제외한 부분집합이며, x 는 적어도 하나의 입력 값에 대한 주요 변수 값의 벡터이며, $\mathrm{val}_x(S)$ 는 집합 S 에 포함되지 않은 모든 변수들을 배제한(Marginalized) 변수 값들의 기여도이다.

[0079] 일 예로, j가 당뇨병 예측에 사용되는 임상 변수 중 성별인 경우, S는 성별을 제외한 나머지 변수들의 모든 조합의 집합이며, val(S)는 성별을 제외한 변수들의 부분 집합의 기여도일 수 있다. 분석부(140)는 전체 변수 기여도에서 성별을 제외한 변수들의 기여도의 차를 산출하여 성별이 가지고 있는 기여도를 연산할 수 있다. 이를 모든 변수에 각각 적용하여 임상 데이터 및 유전 데이터의 변수들의 각 기여도를 연산할 수 있다.

[0080] 이를 통해, 분석부(140)는 모형을 구성하고 있는 임상 변수에 대하여 개인의 당뇨 발병의 위험에 대한 기여도가 높은 기 설정된 개수의 변수를 계산하여 제시할 수 있다.

[0081] 또한, 분석부(140)는 [수학식 2]를 이용하여 주요 변수의 예측 값과 평균 예측 값의 차이 분포를 나타낼 수 있다.

[0082] 또한, 분석부(140)는 주요 변수 수가 증가함에 따라 SHAP 값을 연산하는데 사용되는 주요 변수의 조합이 기하급수적으로 증가할 수 있다.

[0083] 여기서, 분석부(140)는 하기의 [수학식 3]을 이용하여 SHAP 값의 추정치를 연산할 수 있다.

수학식 3

$$\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$$

$$\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M} \phi_j^m$$

[0084]

[0085] 여기서, x는 당뇨병 예측에 사용되는 임상 데이터 및 유전 데이터의 변수이고, M는 반복 횟수이며, x_{+j} 는 기여도를 구하고 싶은 변수가 포함된 입력 값이고, x_{-j} 는 j번째 변수를 제외한 입력 값이다.

[0086] 즉, 분석부(140)는 x_{+j} 및 x_{-j} 의 예측 값의 차를 산출하고, 복수의 산출된 예측 값의 차를 합산한 뒤 반복 횟수로 나눔으로써 SHAP 값의 추정치를 연산할 수 있다.

[0087] 이때, 분석부(140)는 변수 x의 입력 값, j번째 변수 및 반복 횟수(M)을 이용하여 반복 수행시마다 입력되는 입

력 값 z를 산출한다. 이때, z는 랜덤한 값을 가질 수 있다.

[0088] 일 예로, 분석부(140)는 당뇨병을 예측하는 데 사용된 임상 데이터 및 유전 데이터의 변수 x의 입력 값과 z를 조합하여 새로운 입력 값을 산출하고, z는 j번째 변수로 대체될 수 있다.

[0089] 그 다음, 분석부(140)는 [수학식 3]을 이용하여 SHAP 값의 추정치를 연산할 수 있다. 이때, 분석부(140)는 무작위로 추출된 복수의 샘플을 사용하여 각 변수의 중요도를 추정하고, 이를 기초로 예측 모델의 예측 값을 산출할 수 있다.

[0090] 이를 통해, 변수의 개수 증가에 따라 기하급수적으로 증가하는 조합의 수에 대해 임의의 샘플을 추출하여 적은 수의 샘플로 SHAP 값의 추정치를 연산함으로써 계산 시간과 비용을 감소하며 실제 값과 근사한 SHAP 값의 추정치를 연산할 수 있다.

[0091] 도 7은 본 발명의 일 실시예에 따른 분석부의 구현 예를 도시하는 도면이다.

[0092] 도 7을 참조하면, 분석부(140)는 임상 데이터 및 유전 데이터에서 복수의 변수를 추가 또는 제거하여 임의의 데이터 세트를 만들고, 이를 예측부(130)의 딥러닝 모델을 통해 예측 결과를 산출함으로써 특정 변수가 추가 또는 제거되었을 경우의 변화량을 SHAP 값으로 표현할 수 있다. 분석부(140)는 SHAP 값을 통해 임상 데이터 및 유전 데이터와 예측된 결과 간의 변수 영향도를 방향과 크기로 나타냄으로써 복수의 주요 변수를 분석할 수 있고, 예측된 결과에 대한 각 주요 변수의 기여도를 확인할 수 있다.

[0093] 또한, 분석부(140)는 환경적 요인 분석을 위해 상기 임상 데이터에 대해 군집화를 수행하고, 선천적 요인 분석을 위해 상기 유전 데이터에 대해 군집화를 수행하여 제2형 당뇨병 발병에 영향을 미치는 위험 그룹 정보(Endotype)를 생성할 수 있다.

[0094] 도 8은 도 2의 S240 단계의 제2형 당뇨병 발병 확률 위험 요인을 분석하는 일 예를 도시하는 도면이다.

[0095] 도 8을 참조하면, 분석부(140)는 임상 데이터 및 유전 데이터 각각에 대해 비만도가 높고 이완기 혈압이 높은 그룹(보라색, C1), 크레아티닌 수치가 높은 흡연자 그룹(초록색, C2), 크레아티닌 수치가 낮고 당뇨병 가족력이 있는 그룹(빨간색, C3), 음주를 하며 수축기 혈압이 높은 그룹(노란색, C4) 및 크레아티닌 수치가 낮고 음주를 하는 그룹(파란색, C5)으로 군집화를 수행하여 제2형 당뇨병 발병에 영향을 미치는 위험 그룹 정보(Endotype)를 생성할 수 있다.

[0096] 다음으로, 사용자 보고서를 생성하는 단계(S250)에서, 사용자 보고서 생성부(150)는 예측된 결과 및 상기 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성할 수 있다.

[0097] 구체적으로, 사용자 보고서 생성부(150)는 사용자의 유전적인 위험 요인 및 임상적인 위험 요인을 제공하고, 사용자가 속하는 위험 인자 군집에 대한 정보를 생성하여 제공할 수 있다.

[0098] 이를 이용하여, 사용자 보고서 생성부(150)는 예측된 결과 및 분석된 결과를 이용하여 개인 맞춤형의 제2형 당뇨병 진단 및 예방 방안을 생성할 수 있다.

[0099] 도 9는 본 발명의 다른 실시예에 따른 예측 모델과 종래의 예측 모델을 비교하는 도면이다.

[0100] 도 9에 도시된 바와 같이, 본 발명의 다른 실시예에 따른 예측 모델(주황색으로 표시)과 종래 모델(파란색으로 표시)에 대해 100회의 반복 시행을 수행하였고, 평균 성능 및 T-검증(T-test)을 이용하여 예측 모델의 신뢰성을 확인할 수 있다. 여기서, 종래 모델은 한국인의 인종특이적 빈도 추출 및 적용이 되지 않은 모델을 의미한다.

[0101] 도 9에 표시된 결과에서 종래 모델의 평균 성능은 Validation AUC가 86.33% 및 Test AUC가 85.84%이고, 본 발명의 다른 실시예에 따른 예측 모델의 평균 성능은 Validation AUC가 88.32% 및 Test AUC가 87.53%로 종래 모델보다 약 2%의 성능 향상을 보인다.

[0102] 따라서, 본 발명의 일 실시예에 따르면, 한국인의 인종특이적 빈도가 적용된 임상 데이터 및 유전 데이터를 이용하여 향후 5년 내의 사용자의 당뇨병 발병 여부를 예측할 수 있다.

[0103] 본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나 이는 예시적인 것에 불과하며, 당해 기술이 속하는 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

부호의 설명

[0104] 100: 당뇨병 예측 시스템

110: 데이터베이스부

120: 데이터 처리부

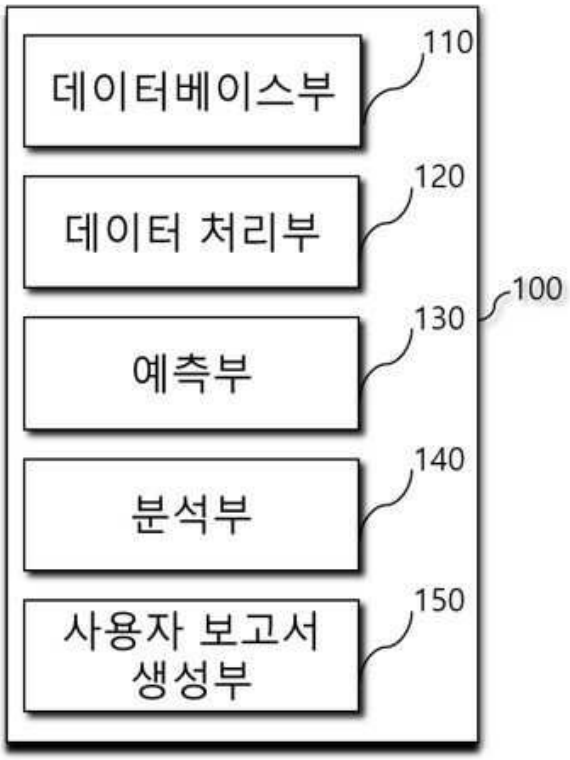
130: 예측부

140: 분석부

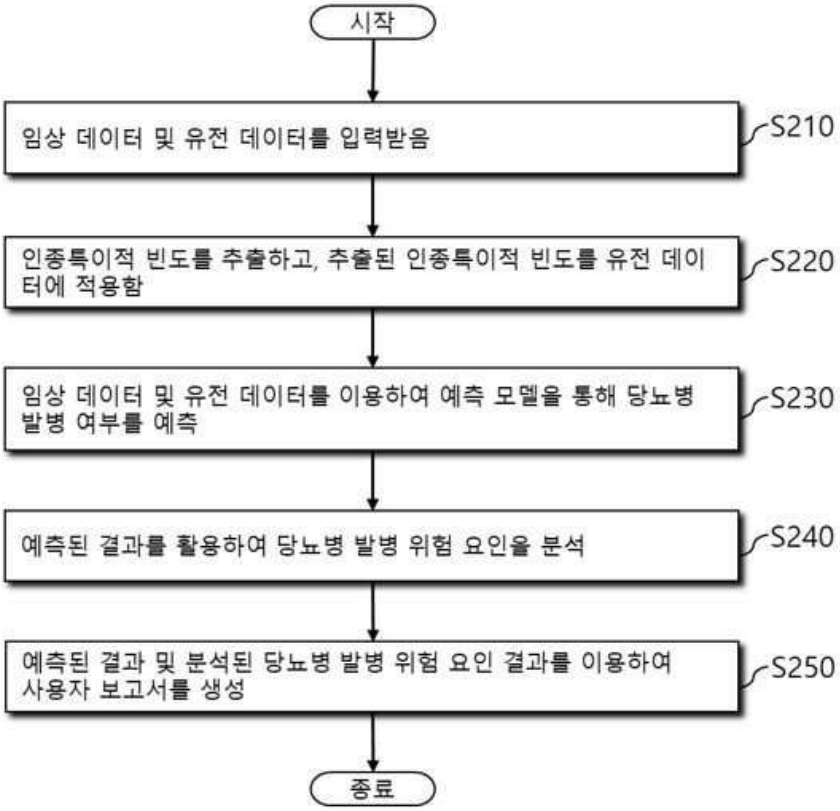
150: 사용자 보고서 생성부

도면

도면1



도면2


시작
임상 데이터 및 유전 데이터를 입력받음
S210
인종특이적 빈도를 추출하고, 추출된 인종특이적 빈도를 유전 데이터에 적용함
S220
임상 데이터 및 유전 데이터를 이용하여 예측 모델을 통해 당뇨병 발병 여부를 예측
S230
예측된 결과를 활용하여 당뇨병 발병 위험 요인을 분석
S240
예측된 결과 및 분석된 당뇨병 발병 위험 요인 결과를 이용하여 사용자 보고서를 생성
S250
종료

도면3

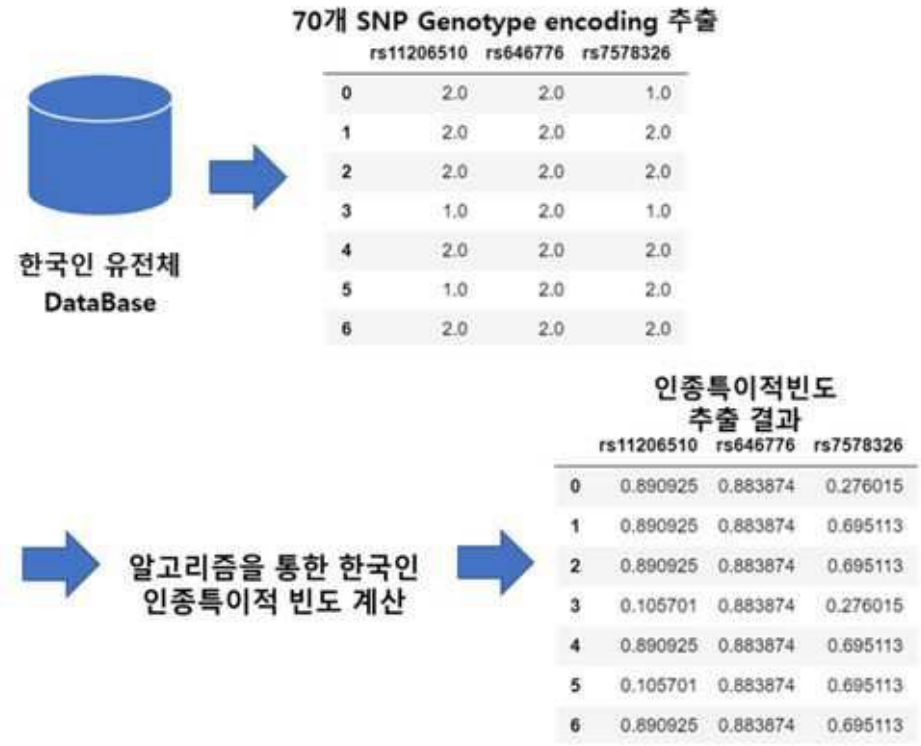
(a)

임상 변수
FBS
Hba1c
Tg
Bmi
SBP(sys)
DBP(dia)
Age
Sex
고혈압과거력
고지혈증과거력
Exercise
Smoke
Drink
WBC
ALT
Cr
DM_FH
Waist

(b)

rs_ID		
rs11206510	rs1470579	rs16826069
rs646776	rs16902871	rs2296172
rs7578326	rs2307111	rs17106184
rs2943640	rs10946398	rs72906810
rs2943641	rs7754840	rs12740374
rs734312	rs9368222	rs12041243
rs1173771	rs3869115	rs515135
rs7756992	rs75080135	rs7558502
rs3130283	rs9470794	rs2255703
rs4711750	rs4273712	rs4402960
rs583769	rs1799884	rs2028299
rs730497	rs4607517	rs10852123
rs11191548	rs35452727	rs6416749
rs10886471	rs13292136	rs781831
rs2237892	rs635634	rs75418188
rs6265	rs75872811	rs58542926
rs56200889	rs1236816	rs429358
rs1127787	rs7111341	rs10408179
rs10773000	rs2237895	rs1800437
rs9316500	rs1552224	rs10423928
rs7178762	rs2277339	rs475002
rs1378942	rs7172432	rs6102059
rs2925979	rs762551	
rs1051168	rs7178572	

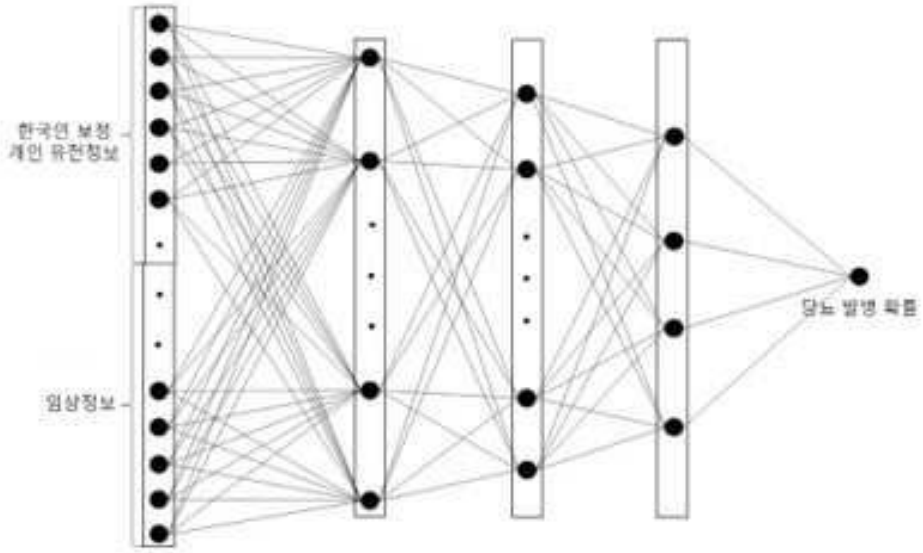
도면4



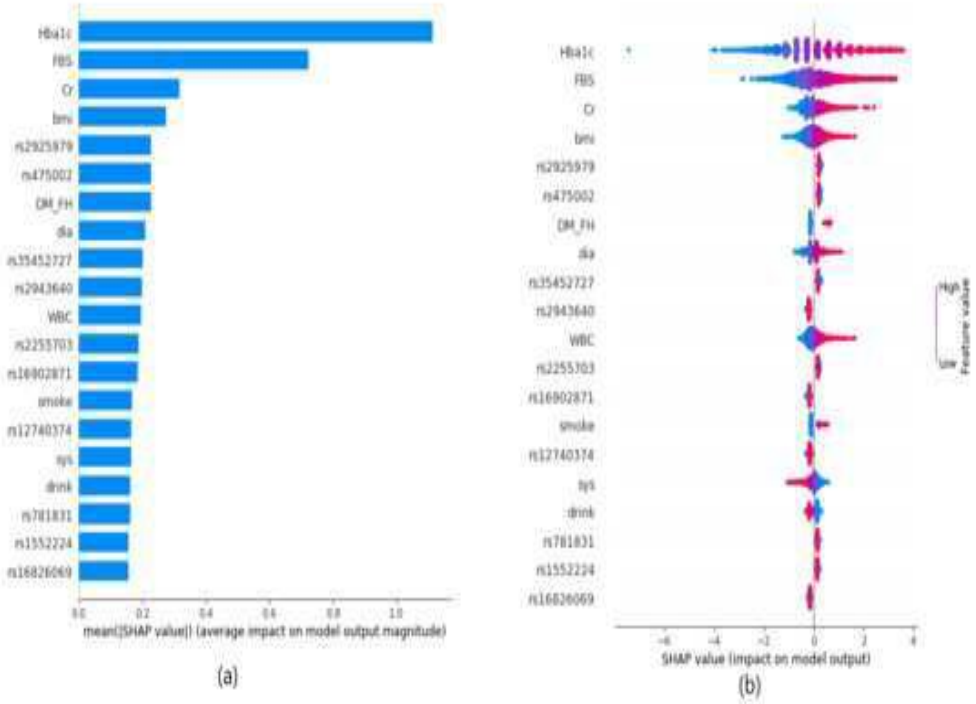
도면5

```
Input : RAW rs sample data
total_sample=Number of the RAW rs sample data
For rsID in RAW rs sample data:
        #Calculates frequency for genotype in rsID
        Frequency for genotype 0=Total number of genotype 0 in rsID/total_sample
        Frequency for genotype 1=Total number of genotype 1 in rsID/total_sample
        Frequency for genotype 2=Total number of genotype 2 in rsID/total_sample
        #Subtitutes each genotype value to frequency in rsID
        Genotype 0 in rsID=Frequency for genotype 0
        Genotype 1 in rsID=Frequency for genotype 1
        Genotype 2 in rsID=Frequency for genotype 2
End For
```

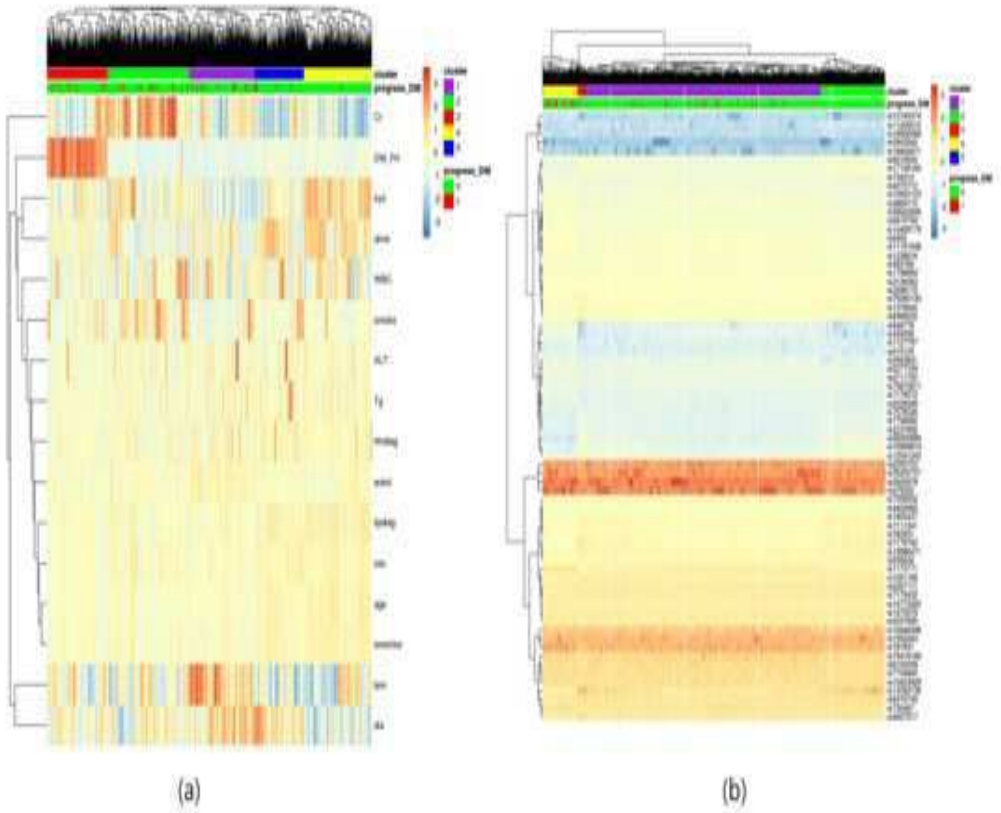
도면6


한국인 보정
개인 유전정보
임상정보
당뇨 발병 확률

도면7


HbA1c
FBS
Cr
bmi
rs2925979
rs475002
DM_FH
dia
rs35452727
rs2943640
WBC
rs2255703
rs16902871
smoke
rs12740374
tg
drink
rs781831
rs1552224
rs16826069
mean(|SHAP value|) (average impact on model output magnitude)
(a)
SHAP value (impact on model output)
(b)

도면8



도면9

