



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0150145
(43) 공개일자 2023년10월30일

(51) 국제특허분류(Int. Cl.)

G16H 50/50 (2018.01) A61B 5/00 (2021.01)
C12Q 1/6883 (2018.01) C12Q 1/689 (2018.01)
G06N 20/00 (2019.01) G16B 40/00 (2019.01)
G16H 50/20 (2018.01)

(52) CPC특허분류

G16H 50/50 (2018.01)
A61B 5/4343 (2013.01)

(21) 출원번호 10-2022-0049735

(22) 출원일자 2022년04월21일

심사청구일자 2022년04월21일

(71) 출원인

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

서울대학교산학협력단

서울특별시 관악구 관악로 1 (신림동)

(72) 발명자

김영주

서울특별시 양천구 목동서로 38, 120동 902호(목동, 목동신시가지아파트1단지)

박태성

서울특별시 서초구 효령로68길 33, 102동 1602호 (서초동, 서초동 아이파크)

(뒷면에 계속)

(74) 대리인

특허법인(유한)아이시스

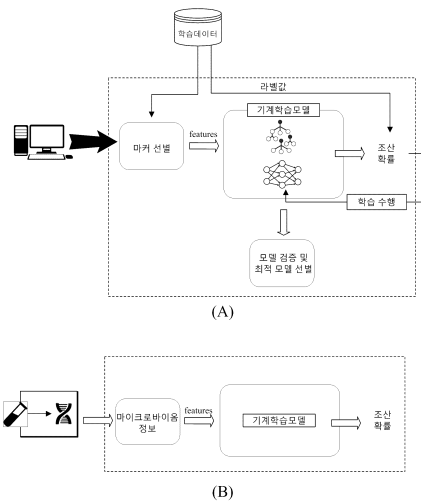
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 마이크로바이옴을 분석하는 학습 모델을 이용한 조산 예측 방법 및 분석장치

(57) 요약

마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법은 분석장치가 대상자의 샘플을 분석하여 획득한 마이크로바이옴 정보를 입력받는 단계, 상기 분석장치가 상기 마이크로바이옴 정보를 사전에 구축한 기계 학습 모델에 적용하는 단계 및 상기 분석장치가 상기 기계 학습 모델이 출력하는 값 내지 정보에 따라 상기 대상자의 조산 가능성을 예측하는 단계를 포함한다.

대표도 - 도3



(52) CPC특허분류

C12Q 1/6883 (2022.01)
C12Q 1/689 (2018.05)
G06N 20/00 (2021.08)
G16B 40/00 (2019.02)
G16H 50/20 (2018.01)

(72) 발명자

박선화

서울특별시 양천구 목동동로 257, 103-4102호(목동, 현대하이페리온)

문정섭

경기도 안양시 동안구 포도원로27번길 20, 가동 407호(호계동, 삼천리아파트)

권은진

경상북도 경주시 초당길43번길 23, 가동 1005호(동천동)

김수민

서울특별시 강서구 강서로7길 20-29, 410호(화곡동, 스위트132드림하우스)

유영아

서울특별시 광진구 독섬로 510(자양동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711145659
과제번호	2020R1A2C3011850
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	다중 오믹스 기반의 바이오빅데이터 딥러닝을 이용한 조산 예측 모델 개발
기 여 율	60/100
과제수행기관명	이화여자대학교
연구기간	2021.03.01 ~ 2022.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711157953
과제번호	2013M3A9C4078158
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	바이오. 의료기술개발(R&D)
연구과제명	전통천연물 안전성/효능 검증을 위한 실험디자인 및 통계적 품질관리기술, 검증기술

개발

기 여 율	20/100
과제수행기관명	서울대학교
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1465035531
과제번호	HI16C2037010022
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	포스트게놈 다부처 유전체사업
연구과제명	다중오믹스 기반 암 정밀 맞춤의학을 위한 통계예측 플랫폼 개발
기 여 율	20/100
과제수행기관명	서울대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명세서

청구범위

청구항 1

분석장치가 대상자의 샘플을 분석하여 획득한 마이크로바이옴 정보를 입력받는 단계;
상기 분석장치가 상기 마이크로바이옴 정보를 사전에 구축한 기계 학습 모델에 적용하는 단계; 및
상기 분석장치가 상기 기계 학습 모델이 출력하는 값 내지 정보에 따라 상기 대상자의 조산 가능성을 예측하는 단계를 포함하는 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법.

청구항 2

제1항에 있어서,
상기 기계 학습 모델은 LR(logistic regression), RF(random forest), XGBoost(extreme gradient boosting), SVM(support vector machine) 및 GUIDE(Generalized, Unbiased, Interaction Detection and Estimation) 중 적어도 하나인 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법.

청구항 3

제1항에 있어서,
상기 분석장치는 상기 대상자의 백혈구 수치 및 자궁경부 길이를 더 입력받고, 상기 백혈구 수치 및 상기 자궁경부 길이를 상기 기계 학습 모델에 더 입력하는 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법.

청구항 4

제1항에 있어서,
상기 기계 학습 모델은 GUIDE(Generalized, Unbiased, Interaction Detection and Estimation) 모델이고,
상기 GUIDE 모델은 상기 대상자의 자궁경부 길이를 기준으로 분기하는 루트 노드 및 상기 마이크로바이옴 정보에 포함된 복수의 미생물 각각의 양에 따라 분기하는 하나 이상의 자식 노드를 포함하는 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법.

청구항 5

제4항에 있어서,
상기 복수의 미생물은 *Gardnerella vaginalis*, *Lactobacillus iners*, *Peptoniphilus grossensis* 및 *Lactobacillus gasseri*를 포함하는 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법.

청구항 6

대상자의 샘플을 분석하여 획득한 마이크로바이옴 정보를 입력받는 입력장치;
상기 마이크로바이옴 정보를 분석하여 조산 가능성을 분류하는 기계 학습 모델을 저장하는 저장장치; 및
상기 마이크로바이옴 정보를 상기 기계 학습 모델에 적용하고, 상기 학습 모델이 출력하는 값 내지 정보에 따라 상기 대상자의 조산 가능성을 예측하는 연산장치를 포함하는 마이크로바이옴을 분석하여 조산을 예측하는 분석장치.

청구항 7

제6항에 있어서,
상기 기계 학습 모델은 LR(logistic regression), RF(random forest), XGBoost(extreme gradient boosting),

SVM(support vector machine) 및 GUIDE(Generalized, Unbiased, Interaction Detection and Estimation) 중 적어도 하나인 마이크로바이옴을 분석하여 조산을 예측하는 분석장치.

청구항 8

제6항에 있어서,

상기 입력장치는 상기 대상자의 백혈구 수치 및 자궁경부 길이를 더 입력받고,

상기 연산장치는 상기 백혈구 수치 및 상기 자궁경부 길이를 상기 기계 학습 모델에 더 입력하는 마이크로바이옴을 분석하여 조산을 예측하는 분석장치.

청구항 9

제6항에 있어서,

상기 기계 학습 모델은 GUIDE(Generalized, Unbiased, Interaction Detection and Estimation) 모델이고,

상기 GUIDE 모델은 상기 대상자의 자궁경부 길이를 기준으로 분기하는 루트 노트 및 상기 마이크로바이옴 정보에 포함된 복수의 미생물 각각의 양에 따라 분기하는 하나 이상의 자식 노트를 포함하는 마이크로바이옴을 분석하여 조산을 예측하는 분석장치.

청구항 10

제9항에 있어서,

상기 복수의 미생물은 *Gardnerella vaginalis*, *Lactobacillus iners*, *Peptoniphilus grossensis* 및 *Lactobacillus gasseri*를 포함하는 마이크로바이옴을 분석하여 조산을 예측하는 분석장치.

발명의 설명

기술 분야

[0001] 이하 설명하는 기술은 산모의 시료에서 검출한 마이크로바이옴(microbiome)을 기준으로 조산을 예측하는 기법에 관한 것이다.

배경 기술

[0002] 조산(preterm birth)은 임신 20주에서 37주 사이에 분만하는 것을 의미한다. 조산으로 태어난 신생아는 정상적으로 출생한 신생아에 비하여 사망율, 합병증 또는 장애 발생 가능성이 높다. 조산아의 생존율은 임신 주수(성숙도)에 영향을 받는 것으로 알려졌다. 따라서, 조산은 조산 징후를 사전에 파악하고 적절한 치료를 통하여 분만 주수를 지연시키는 처치가 중요하다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 미국등록특허 US 11,062,808호

발명의 내용

해결하려는 과제

[0004] 이하 설명하는 기술은 비침습적으로 획득한 마이크로바이옴을 이용하여 산모의 조산 가능성을 예측하는 기법을 제공하고자 한다. 이하 설명하는 기술은 학습 모델을 이용하여 자동으로 산모의 조산 가능성을 예측하는 기법을 제공하고자 한다.

과제의 해결 수단

[0005] 마이크로바이옴을 분석하는 기계 학습 모델을 이용한 조산 예측 방법은 분석장치가 대상자의 샘플을 분석하여 획득한 마이크로바이옴 정보를 입력받는 단계, 상기 분석장치가 상기 마이크로바이옴 정보를 사전에 구축한 기계 학습 모델에 적용하는 단계 및 상기 분석장치가 상기 기계 학습 모델이 출력하는 값 내지 정보에 따라 상기 대상자의 조산 가능성을 예측하는 단계를 포함한다.

[0006] 마이크로바이옴을 분석하여 조산을 예측하는 분석장치는 대상자의 샘플을 분석하여 획득한 마이크로바이옴 정보를 입력받는 입력장치, 상기 마이크로바이옴 정보를 분석하여 조산 가능성을 분류하는 기계 학습 모델을 저장하는 저장장치 및 상기 마이크로바이옴 정보를 상기 기계 학습 모델에 적용하고, 상기 학습 모델이 출력하는 값 내지 정보에 따라 상기 대상자의 조산 가능성을 예측하는 연산장치를 포함한다.

발명의 효과

[0007] 이하 설명하는 기술은 산모가 병원에서 비침습적인 방법으로 간단하게 채취한 질분비물을 이용하여 높은 정확도로 조산 가능성을 예측한다.

도면의 간단한 설명

[0008] 도 1은 마이크로바이옴을 이용한 조산 예측 시스템에 대한 예이다.

도 2는 정상군과 조산군의 미생물 다양성의 차이를 분석한 결과이다.

도 3은 마이크로바이옴을 분석하여 조산을 예측하는 기계 학습 모델 구축 과정과 조산 예측 과정의 예이다.

도 4는 조산을 예측하는 GUIDE 모델을 구축한 예이다.

도 5는 마이크로바이옴을 이용하여 조산을 예측하는 분석장치에 대한 예이다.

발명을 실시하기 위한 구체적인 내용

[0009] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0010] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0011] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설명된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[0012] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.

[0013] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0014] 이하 설명에서 사용되는 용어에 대하여 설명한다.

[0015] 샘플은 산모에서 수집한 시료를 의미한다. 샘플은 분변, 혈액, 질분비물 등 산모에서 수집가능한 다양한 시료

중 어느 하나일 수 있다. 이하 연구자는 질분비물에서 수집한 마이크로바이옴을 이용하여 조산을 예측하는 연구를 하였다. 따라서, 이하 시료는 산모의 질분비물이라고 가정한다. 다만, 이하 설명하는 기술이 특정 사이트(site)에서 수집한 시료에 한정되는 것은 아니다.

- [0016] 마이크로바이옴은 인체에 서식, 공생하는 개체 수준의 세균, 바이러스 그리고 곰팡이 등 모든 미생물 군집과 이들 미생물 군집이 가지는 유전정보로 정의된다.
- [0017] 인간 마이크로바이옴은 미생물의 16s rRNA와 함께 전체 게놈(Whole genome) 등의 정보를 포함한다. 16s rRNA의 양은 사람으로부터 나온 여러 가지 샘플에 대해 메타게놈(metagenome) 분석을 하여 알 수 있다. 메타게놈 분석은 여러 가지 메타게놈 분석과 관련된 생명정보학적 도구를 이용해 수행할 수 있다. 분석의 목적은 일반적으로 16s rRNA의 양을 측정하여 사용자의 인체 유래 미생물의 양과 분포를 확인하는 것이다. 인간 마이크로바이옴은 16s rRNA분석을 위한 원시 정보부터 획득되는 OTU(Operational Taxonomic Unit)나 ASV(Amplicon Sequence Variant) 정보를 포함한다. 나아가, 전장 유전체 분석(whole-genome sequencing)은 샘플의 염기서열 전체를 해독하는 방식으로, 분석 결과는 인간 유전 정보와 미생물의 유전 정보를 모두 포함한다. 따라서, 연구자는 미생물의 유전 정보만 추출(인간 유전 정보 필터링)하여 미생물 군집의 유전 정보를 확인할 수 있다. 따라서, 마이크로바이옴 정보는 산출하는 방법론은 다양할 수 있다. 이하 설명하는 기술은 특정한 시료 분석 방법에 제한되지 않는다.
- [0018] 이하 설명하는 기술은 분석 대상의 시료에서 추출한 마이크로바이옴을 분석하여 산모의 조산 가능성을 예측하는 기술이다. 이하, 분석장치가 마이크로바이옴을 분석한다고 설명한다. 분석장치는 마이크로바이옴 정보 추출(서열 분석), 마이크로바이옴 기반 조산 예측 등을 수행할 수 있다. 분석장치는 데이터 처리가 가능한 컴퓨터 장치를 의미한다. 예컨대, 분석장치는 PC, 네트워크상의 서버, 스마트 기기, 전용 프로그램이 임베딩된 칩셋 등으로 구현될 수 있다.
- [0019] 이하 설명하는 기술은 기계 학습(machine learning)을 이용하여 마이크로바이옴을 분석한다. 기계 학습은 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 알고리즘을 개발하는 분야를 의미한다. 학습 모델은 결정 트리, 랜덤 포레스트(random forest), KNN(K-nearest neighbor), 나이브 베이즈(Naive Bayes), SVM(support vector machine), 인공신경망(artificial neural network) 등이 있다.
- [0021] 도 1은 마이크로바이옴을 이용한 조산 예측 시스템(100)에 대한 예이다.
- [0022] 도 1에서 분석장치(130, 140, 150)가 산모의 조산 여부를 예측한다. 도 1에서 분석장치는 서버(130) 및 컴퓨터 단말(140, 150) 형태로 도시하였다. 나아가, 분석장치(130, 140, 150)는 모바일 기기 등 다양한 형태로 구현될 수 있다.
- [0023] 시료 분석장치(110)는 산모의 샘플을 분석하여 마이크로바이옴을 분석할 수 있다. 시료 분석장치(110)는 샘플의 미생물 서열을 산출할 수 있다. 나아가 시료 분석 장치(110)는 미생물들의 종류 및 미생물의 양이라는 정보를 산출할 수도 있다. 시료 분석장치(110)는 생성한 데이터를 별도의 DB(120)에 저장할 수도 있다.
- [0024] 서버(130)는 시료 분석장치(110) 또는 DB(120)로부터 유전자 서열 또는 미생물 정보를 수신한다. 서버(130)는 유전자 서열을 분석하여 미생물 정보를 산출할 수도 있다. 서버(130)는 샘플의 미생물 정보를 사전에 구축한 학습 모델에 입력하여 조산 여부를 예측할 수 있다. 사용자(10)는 사용자 단말을 통해 서버(130)에 접속하여 분석 결과를 확인할 수 있다.
- [0025] 컴퓨터 단말(140)은 시료 분석장치(110) 또는 DB(120)로부터 유전자 서열 또는 미생물 정보를 수신한다. 컴퓨터 단말(140)은 유전자 서열을 분석하여 미생물 정보를 산출할 수도 있다. 컴퓨터 단말(140)은 샘플의 미생물 정보를 사전에 구축한 학습 모델에 입력하여 조산 여부를 예측할 수 있다. 컴퓨터 단말(140)은 사용자(20)에게 분석 결과를 제공할 수 있다.
- [0026] 컴퓨터 단말(150)은 시료 분석장치(110)가 생성한 유전자 서열 또는 미생물 정보가 저장된 매체(예컨대, USB, SD 카드 등)를 통해 유전자 서열 또는 미생물 정보를 입력받는다. 컴퓨터 단말(150)은 유전자 서열을 분석하여 미생물 정보를 산출할 수도 있다. 컴퓨터 단말(150)은 샘플의 미생물 정보를 사전에 구축한 학습 모델에 입력하여 조산 여부를 예측할 수 있다. 컴퓨터 단말(150)은 사용자(30)에게 분석 결과를 제공할 수 있다.

[0028] 이하 연구자가 학습모델을 구축하는 과정을 설명하고자 한다.

[0029] 연구자는 2018년~2020년에 연세대학교 세브란스병원과 이화여자대학교 목동 병원에서 참여자들을 모집하여 데이터를 수집하였다. 참여자들은 임신 주수가 15주에서 36주 사이의 임산부로서 정상 및 조산 징후가 있는 산모들로 구성되었다. 연구자는 참여자들에 대한 임상적 치료 전에 질분비물(cervicovaginal fluid, CVF)을 수집하였다. 또한, 연구자는 참여자들에 대한 질분비물 수집시에 백혈구 수치(WBC) 확인을 위한 혈액 검사와 자궁경부길이(cervical length, CL)를 측정하였다. 참여자들에 대한 정보는 아래 표 1과 같다. 조산 그룹(Preterm birth)은 37주 미만에 분만한 피험자이고, 정상 그룹(Term birth)은 37주 이후 분만한 피험자로 구분하였다. 아래 표 1에서 약어로 표현된 GAB(gestational age at birth)는 출산 시점을 나타내고, APGAR(Appearance, Pulse, Grimace, Activity, and Respiration) 점수는 신생아 건강상태를 나타내는 점수 지표를 말한다.

표 1

Characteristics	Preterm birth (n = 56)	Term birth (n = 99)	P-value
Maternal age (year)	32.32 (±4.35)	33.02 (±3.40)	0.237
Pre-pregnancy BMI (kg/m ²)	21.60 (±2.89)	21.36 (±2.77)	0.657
Education level			>0.999
High school graduation or below	6 (13.3)	11 (15.3)	
University graduates	39 (86.7)	61 (84.7)	
History of PTB			<0.015*
No	79 (88.8)	94 (97.9)	
Yes	10 (11.2)	2 (2.1)	
WBC (1×10 ³ /μl) ^a	9.40 (8.8-11.5)	8.85 (7.5-10.2)	<0.001*
GAB (wks)	34.0 (30.4-35.6)	39.3 (38.2-39.9)	<0.001*
Birth Weight (g)	1975.4 (±780.5)	3234.9 (±316.8)	<0.001*
APGAR score at 1 min	6.99 (5-9)	9.31 (9-10)	<0.001*
APGAR score at 5 min	8.29 (7-10)	9.74 (10-10)	<0.001*

[0031]

[0033] 연구자는 NucleoSpin Tissue Kit을 사용하여 마이크로바이옴 분석을 위한 미생물 DNA를 추출하였다. 16S rRNA 라이브러리는 16S rRNA 유전자의 가변영역 V3 및 V4를 타겟으로 하는 16S 메타게놈 시퀀싱 라이브러리 프로토콜에 따라 구축되었다. 연구자는 PCR을 통해 대상 서열을 증폭하고 정규화한 후 시퀀싱을 수행하였다.

[0034] 연구자는 참여자의 시료에서 마이크로바이옴 서열을 분석하였고, 미생물 군집에 대하여 샤논 지수(Shannon index)를 이용한 알파 다양성(α-diversity)과 Bray-Curtis 거리를 이용한 베타 다양성(β-diversity)을 분석하였다. 도 2는 정상군과 조산군의 미생물 다양성의 차이를 분석한 결과이다. 도 2(A)는 샤논 지수에 따른 알파 다양성 결과이다. 도 2(A)를 살펴보면 정상군(Term)과 조산군(preterm) 사이에 미생물 종류의 다양성에 분명한 차이가 있음을 알 수 있다. 도 2(B)는 베타 다양성 결과인데 정상군과 조산군의 차이가 뚜렷하지 않았다.

[0036] 연구자는 마이크로바이옴에서 마커를 선별하여 이용하였다.

[0037] 연구자는 전체 수집된 데이터를 학습 데이터(n = 101)와 검증 데이터(n = 49)로 구분하였다. 연구자는 학습 데이터에 대하여 7개의 통계적 기법들을 적용하여 그룹간 유의미한 차이를 나타내는 마커(들)를 선별하였다. 연구자가 사용한 통계적 기법들은 ZIG(zero-inflated Gaussian mixture model), ZIBSeq(zero-inflated beta regression), ANCOM(analysis of composition of microbiomes), CLR_P(centered log-ratio transformation and permutation logistic regression model), Wilcoxon(Wilcoxon rank-sum test), DESeq2 및 edgeR이다. 연구자는 두 개 이상의 통계적 기법들에서 마커 빈도가 25% 이상, 비례중향(mean proportion)이 0.001% 이상이고 p-값이 0.05 미만인 경우 마커로 선별하였다. 이 과정에서 연구자는 10개의 마커들을 선별하였다. 나아가, 연구자는 조산과 관련된다고 알려진 마커들을 선택하였다. 이를 통해 연구자는 7개의 마커들을 추가하였다. 이를 통해 연구자는 초기 17개의 마커들을 선별하였다. 아래 표 2는 선별한 마커들이다. 처음 10개(1~10)가 통계적 기법으로 선별한 마커들이고, 나머지 7개(11~17)가 추가적으로 선별한 마커들이다.

표 2

[0039]

No.	Species
1	<i>Atopobium vaginae</i>
2	<i>Gardnerella vaginalis</i>
3	<i>Lactobacillus crispatus/gallinarum</i>
4	<i>Lactobacillus fornicalis</i>
5	<i>Lactobacillus gasseri</i>
6	<i>Lactobacillus iners</i>
7	<i>Lactobacillus jensenii</i>
8	<i>Peptoniphilus grossensis</i>
9	<i>Prevotella timonensis</i>
10	<i>Ureaplasma parvum</i>
11	<i>Bifidobacterium breve</i>
12	<i>Dialister propionicifaciens</i>
13	<i>Lactobacillus paracasei</i>
14	<i>Mobiluncus curtisii</i>
15	<i>Prevotella disiens</i>
16	<i>Staphylococcus aureus</i>
17	<i>Streptococcus anginosus</i>

[0041] 연구자는 17개의 마커들 대상으로 가능한 조합에 대한 회귀분석을 통해 유효한 마커 세트를 선별하였다. 또한, 연구자는 후술할 기계 학습모델을 이용하여 5차 교차 검증(five-fold cross-validation)을 하면서 일정 개수의 마커로 구성된 마커 세트들 중 AUC(Area under the ROC Curve)가 가장 높은 마커 세트를 선별하였다.

[0042] 나아가, 연구자는 전체 마커들을 대상으로도 동일한 과정을 통해 최적의 성능을 갖는 마커 세트를 선별하는 과정을 수행하였다.

[0043] 연구 결과 조산과 연관된 20개의 후보 미생물을 식별하였고, 정상 분만에 연관된 3개의 미생물을 식별하였다. *Lactobacillus* spp. 및 *Gardnerella vaginalis*는 정상 그룹에서 주로 발견되었다. *Bacterioides* spp., *Ureaplasma* spp., *Staphylococcus* spp., *Escherichia fergusonii*, *Prevotella* spp., 및 *Sphingomonas* spp. ctc는 조산 그룹에서 증가하는 추세를 보였다.

[0045] 도 3은 마이크로바이옴을 분석하여 조산을 예측하는 기계 학습 모델 구축 과정과 조산 예측 과정의 예이다.

[0046] 도 3(A)는 기계 학습 모델을 구축하는 과정의 예이다. 기계 학습 모델의 학습은 분석장치 또는 별도의 컴퓨터 장치에서 수행된다. 이하 기계 학습 모델을 구축하는 장치를 학습 장치라고 명명한다.

[0047] 학습 장치는 학습을 위한 학습 데이터를 이용한다. 전술한 바와 같이 학습 데이터는 만삭분만 산모와 조산한 산

모에서 추출한 마이크로바이옴 정보를 포함한다.

- [0048] 학습 장치는 전술한 바와 같이 학습 데이터를 활용하여 마이크로바이옴 중 조산 예측에 유용한 마커 세트를 선별할 수 있다.
- [0049] 학습 장치는 학습 데이터를 이용하여 기계 학습 모델을 학습시킨다. 기계 학습 모델이 복수인 경우, 학습 장치는 복수의 학습 모델별로 학습 과정을 수행한다. 학습 장치는 하나의 시료에서 추출한 마커 세트를 기계 학습 모델에 입력하고, 기계 학습 모델은 입력 데이터의 특징에 기반하여 조산 확률을 출력한다. 학습 장치는 해당 데이터의 라벨값과 학습 모델이 출력한 값을 비교하면서 파라미터를 갱신하거나 결정 트리를 구현해 간다.
- [0050] 기계 학습 모델들의 학습이 완료되면, 학습 장치는 학습 데이터 또는 검증 데이터를 이용하여 각 기계 학습 모델의 성능을 검증할 수 있다. 경우에 따라 학습 장치는 성능이 가장 높은 기계 학습 모델을 선별하여 최종 기계 학습 모델을 선별할 수 있다. 또는 학습 장치는 복수의 기계 학습 모델을 앙상블 모델로 이용할 수도 있다.
- [0052] 연구자는 복수의 기계 학습 모델들을 구축하여 각 모델의 정확도를 검증하였다. 복수의 학습 모델들은 LR(logistic regression), RF(random forest), XGBoost(extreme gradient boosting), SVM(support vector machine) 및 GUIDE(Generalized, Unbiased, Interaction Detection and Estimation)이다.
- [0053] 연구자는 18개의 서로 다른 마커 세트를 이용하여 조산 예측 모델을 구축하였다. 연구자는 기계 학습 모델별로 마커 세트를 달리하면서 가장 성능이 높은 마커 세트를 선정하였다. 예컨대, SVM은 17개의 선별된 마커 중 6개의 마커를 이용하여 학습하였고, 이때 AUC가 0.70이었다. RF는 17개의 사전 선택된 마커 전체를 사용하는 경우 AUC가 0.69였다. RF는 *Lactobacillus* spp. 및 *Ureaplasma Parvum*가 가장 영향을 많이 주는 중요한 마커로 확인되었다.
- [0054] 나아가, 연구자는 WBC(백혈구 수치)를 변수로 더 이용하는 경우 대부분의 학습 모델에서 성능이 향상되는 것을 확인하였다. 예컨대, 17개의 마커들을 사용하는 RF는 AUC가 0.75로 증가하였다. RF 및 SVM도 사전 선택된 10개의 마커를 사용하는 경우 AUC가 각각 0.72 및 0.73으로 증가하는 경향을 보였다.
- [0055] 또한, 연구자는 WBC 및 자궁 경부 길이를 더 이용하여 모델을 구축하는 경우, 해당 변수를 이용하지 않는 경우보다 AUC가 증가하는 것을 확인하였다.
- [0056] GUIDE 모델은 10개의 사전 선택된 마커를 사용하는 경우, 전체 마커에서 7개를 선별한 경우 및 전체 마커에서 LASSO로 17개 마커들을 선별한 경우 AUC가 0.77로 비교적 높은 성능을 보였다.
- [0058] 도 3(B)는 학습된 학습 모델을 이용하여 조산을 예측하는 과정의 예이다.
- [0059] 분석 장치는 분석 대상의 시료에서 추출한 서열 데이터 정보를 입력받는다. 이 경우 분석 장치가 서열 데이터를 분석하여 시료에 포함된 마이크로바이옴 정보를 추출할 수 있다. 마이크로바이옴 정보는 시료에 포함된 미생물 종류 및 각 미생물의 양을 포함한다.
- [0060] 분석 장치는 마이크로바이옴 정보에서 특정 마커 세트를 선별할 수 있다. 분석장치는 마커 세트를 학습 모델에 입력하여 학습 모델이 분류하는 결과를 기준으로 조산 여부를 예측할 수 있다.
- [0062] GUIDE는 결정트리 기반의 방식으로, 변수 선별 과정에 분기점을 많이 가지는 변수가 선택될 확률이 높아지는 편향성을 줄이기 위하여 개발되었다. GUIDE는 카이제곱 검정에서 중요 변수를 선택하는 방식으로 변수 선택의 편향성을 줄인다.
- [0063] GUIDE 모델을 중심으로 설명한다. GUIDE는 카이제곱 검정을 통해 노드를 선택한다. 이 과정을 통해 GUIDE는 모든 예측 변수를 고려하여, 완전 탐색(Exhaustive search) 알고리즘으로 모든 경우의 수를 고려하여 노드를 선택하고 최적의 분기점을 찾아 의사결정나무를 만들게 된다.
- [0064] (1) 단변량 노드 선정 과정은 다음과 같다.
- [0065] a. 선택된 예측 변수의 평균값 $\bar{x}$ 과 표준편차 s 를 구한다.
- [0066] b. 카이제곱 검정을 위해 예측 변수 X 를 아래와 같이 3개 혹은 4개의 카테고리로 나눈다. J 는 학습 데이터에

클래스 개수이고, J_t 는 노드 t 에서의 클래스 개수이다. $N(t)$ 는 노드 t 에서의 샘플 크기를 말한다.

$$\text{If } N(t) \geq 20J_t, X \text{ interval ranges : } \bar{x}, \bar{x} \pm \frac{s\sqrt{3}}{2} \\ \text{If } N(t) < 20J_t, X \text{ interval ranges : } \bar{x} \pm \frac{s\sqrt{3}}{3}$$

c. 카이제곱 검정 통계량이 가장 높은 예측 변수를 선택한다.

(2) 노드 분기점 선정 과정은 다음과 같다.

a. 노드의 분기점은 지니 불순도(Gini impurity)가 최소화되는 지점을 계산하여 정한다. J 개의 클래스가 있다 가정했을 때, t 노드에 해당하는 j 클래스 샘플의 수 $N_j(t)$ 와 해당 노드의 전체 샘플 수 $N(t)$ 는 다음과 같이 표현할 수 있다.

$$N(t) = \sum_j N_j(t)$$

b. 클래스 J 의 사전 확률을 $\pi(j)$ 라 한다면, 노드 t 에서 j 클래스가 나올 예측 확률 $p(j, t)$ 은 다음과 같이 계산된다.

$$\pi(j) = \frac{N_j}{N}, \quad p(j, t) = \frac{\pi(j)N_j(t)}{N}$$

c. 노드 t 에 대한 확률 $p(t)$ 와 t 가 주어졌을 때 j 클래스의 확률 $p(j | t)$ 을 정의한다면, 지니 계수 $g(t)$ 를 계산할 수 있다.

$$p(t) = \sum_j p(j, t), \quad p(j | t) = \frac{p(j, t)}{p(t)} \\ g(t) = \sum_{i \neq j} p(i | t)p(j | t)$$

d. 위 계산된 지니 계수를 최소화 하는 분기점을 찾아 정하여 의사결정나무를 만들게 된다.

각각의 의사결정나무는 CART(Classification And Regression Trees)에서 제시한 방법과 동일하게 예측오분류 감소율이 작은 최적의 모형이 되도록 가지치기를 수행할 수 있다. 나아가, 과적합을 방지하기 위해 K -차 교차 검증을 사용하고, 평균값을 사용하여 최적의 모델을 선정할 수 있다. 연구자는 5차 교차 검증을 사용하였다.

도 4는 조산을 예측하는 GUIDE 모델을 구축한 예이다. 도 3은 연구자가 학습 데이터를 이용하여 GUIDE 모델을 구축한 예이다. 도 3은 사전에 선별한 10개의 마커들을 이용하여 모델을 구축한 예이다. GUIDE 모델은 경부 길이, V1 ~ V4에 해당하는 미생물의 양을 기준으로 조산 여부를 예측하는 결정 트리에 해당한다. V1 = *Gardnerella vaginalis*, V2 = *Lactobacillus iners*, V3 = *Peptoniphilus grossensis*, V4 = *Lactobacillus gasseri*이다. 결정 트리에서 각 노드의 조건이 참일 경우 왼쪽으로 이동하고, 거짓일 경우 오른쪽으로 이동한다. 리프 노드(Leaf Node)에 해당하는 숫자는 해당 노드의 인원 중 조산한 산모의 비율을 의미한다.

도 5는 마이크로바이옴을 이용하여 조산을 예측하는 분석장치(200)에 대한 예이다. 분석장치(200)는 도 1의 분석 장치(130, 140 또는 150)에 해당하는 장치이다.

분석장치(200)는 전술한 학습 모델을 이용하여 조산을 예측할 수 있다. 분석장치(200)는 물리적으로 다양한 형태로 구현될 수 있다. 예컨대, 분석장치(200)는 PC, 스마트 기기, 네트워크상의 서버, 데이터 처리 전용 칩셋 등의 형태를 가질 수 있다.

분석장치(200)는 저장장치(210), 메모리(220), 연산장치(230), 인터페이스 장치(240), 통신장치(250) 및 출력장치(260)를 포함할 수 있다.

- [0085] 저장장치(210)는 전술한 학습 모델들 중 적어도 하나를 저장할 수 있다.
- [0086] 저장장치(210)는 데이터 처리에 필요한 프로그램 내지 소스 코드 등을 저장할 수 있다.
- [0087] 저장장치(210)는 분석 대상자의 미생물 서열 정보 또는 마이크로바이옴 정보를 저장할 수 있다.
- [0088] 메모리(220)는 분석장치(200)가 마이크로바이옴을 분석하고, 조산을 예측하는 과정에서 생성되는 데이터 및 정보 등을 저장할 수 있다.
- [0089] 인터페이스 장치(240)는 외부로부터 일정한 명령 및 데이터를 입력받는 장치이다. 인터페이스 장치(240)는 물리적으로 연결된 입력 장치 또는 외부 저장장치로부터 산모의 서열 데이터 또는 마이크로바이옴 정보를 입력받을 수 있다. 서열 데이터는 시료에서 분석한 미생물의 서열 데이터를 의미한다. 마이크로바이옴 정보는 미생물 서열을 이용하여 산출한 미생물 종류 및 미생물의 양을 포함한다.
- [0090] 인터페이스 장치(240)는 추가적으로 분석 대상자의 백혈구 수치 및 자궁경부길이를 입력받을 수 있다.
- [0091] 인터페이스 장치(240)는 데이터 분석을 위한 학습모델을 입력받을 수 있다. 인터페이스 장치(240)는 학습모델 훈련을 위한 학습데이터, 정보 및 파라미터값을 입력받을 수도 있다.
- [0092] 통신장치(250)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미한다. 통신장치(250)는 외부 객체로부터 산모의 서열 데이터 또는 마이크로바이옴 정보를 수신할 수 있다.
- [0093] 통신장치(250)는 분석 대상자의 백혈구 수치 및 자궁경부길이도 수신할 수 있다.
- [0094] 통신장치(250)는 모델 학습을 위한 데이터도 수신할 수 있다.
- [0095] 통신장치(250)는 샘플에 대한 분석 결과를 외부 객체로 송신할 수 있다. 분석 결과는 분석 대상자의 조산 가능성을 포함한다.
- [0096] 통신장치(250) 내지 인터페이스 장치(240)는 외부로부터 일정한 데이터 내지 명령을 전달받는 장치이다. 데이터를 입력받는 기능에 한정하면 통신장치(250) 내지 인터페이스 장치(240)를 입력장치라고 명명할 수 있다.
- [0097] 출력장치(260)는 일정한 정보를 출력하는 장치이다. 출력장치(260)는 데이터 처리 과정에 필요한 인터페이스, 분석 결과 등을 출력할 수 있다.
- [0098] 연산 장치(230)는 시료에 대하여 획득한 데이터가 서열 데이터인 경우 서열 데이터를 분석하여 미생물 종류 및 양을 확인할 수 있다.
- [0099] 연산 장치(230)는 학습모델에 필요한 입력데이터를 입력하여 조산 가능성을 예측할 수 있다. 입력 데이터는 마이크로바이옴 정보를 포함한다. 또한, 입력데이터는 백혈구 수치 및/또는 자궁경부 길이를 더 포함할 수도 있다. 학습 모델은 학습 과정에서 필요한 입력 데이터에 따라 사전에 학습되어야 한다.
- [0100] 연산 장치(230)는 학습 모델이 출력하는 값 내지 학습 모델이 결정하는 값에 따라 분석 대상자의 조산 가능성을 예측할 수 있다.
- [0101] 연산 장치(230)는 GUIDE 모델을 이용하면 사전에 구축된 결정 트리 구조에 따라 필요한 정보를 비교하면서 최종적으로 결정된 리프노드의 정보에 따라 조산 가능성을 예측할 수 있다.
- [0102] 연산 장치(230)는 복수의 학습 모델이 각각 출력하는 값을 기준으로 최종적인 조산 가능성을 예측할 수도 있다. 이 경우 복수의 학습 모델은 앙상블 모델을 구성하게 된다.
- [0103] 연산 장치(230)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.
- [0105] 또한, 상술한 바와 같은 마이크로바이옴 기반 조산 예측 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.
- [0106] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM

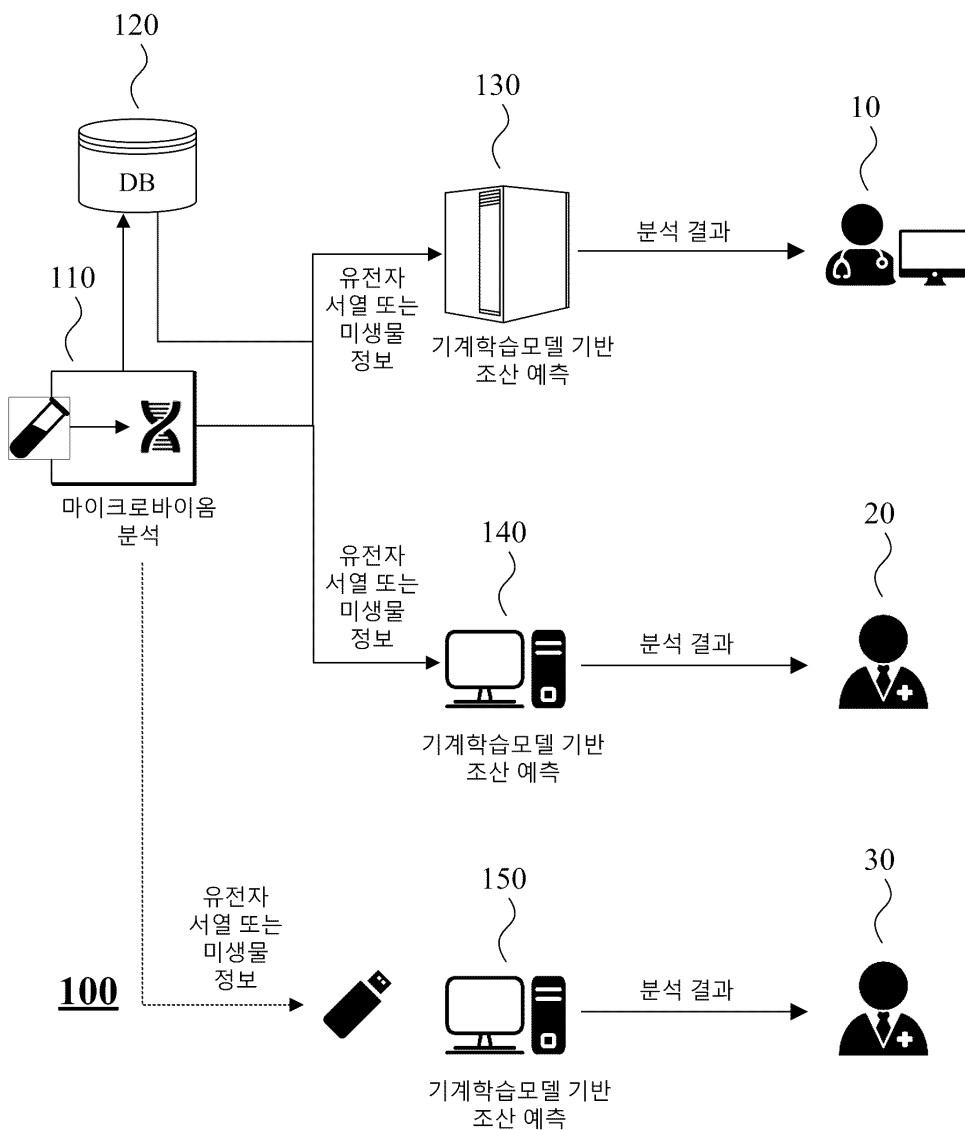
(read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[0107] 일시적 판독 가능 매체는 스테틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

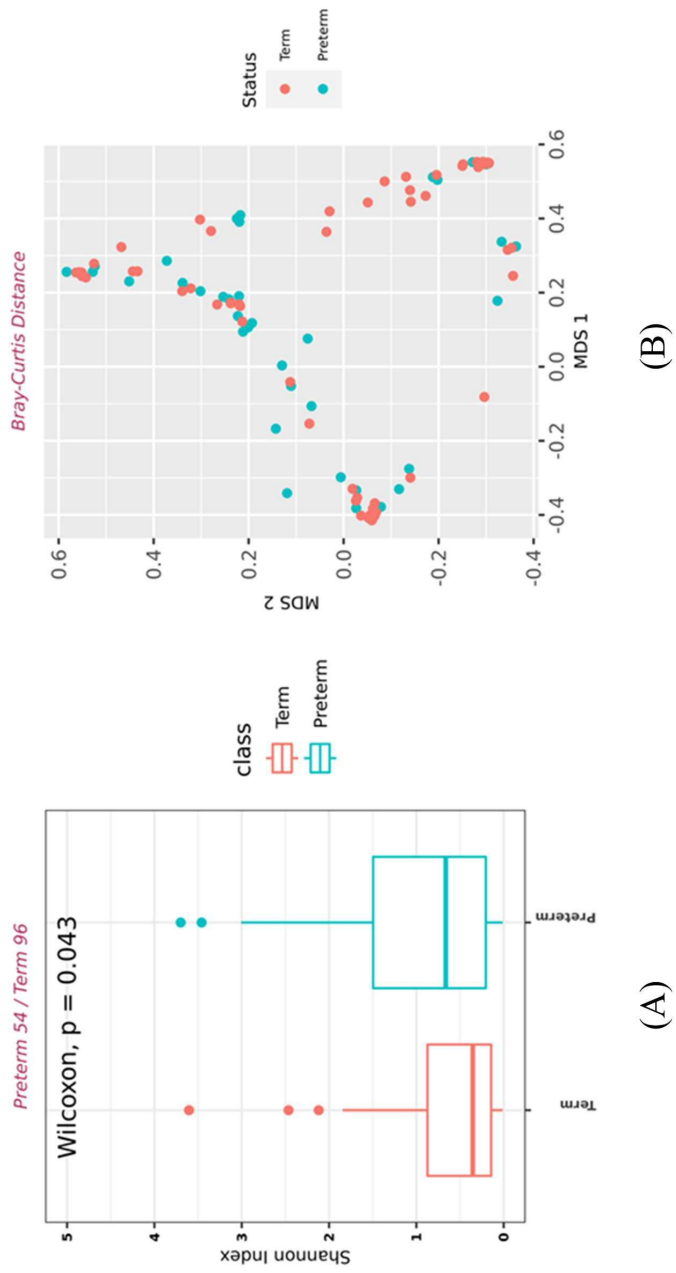
[0108] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

도면

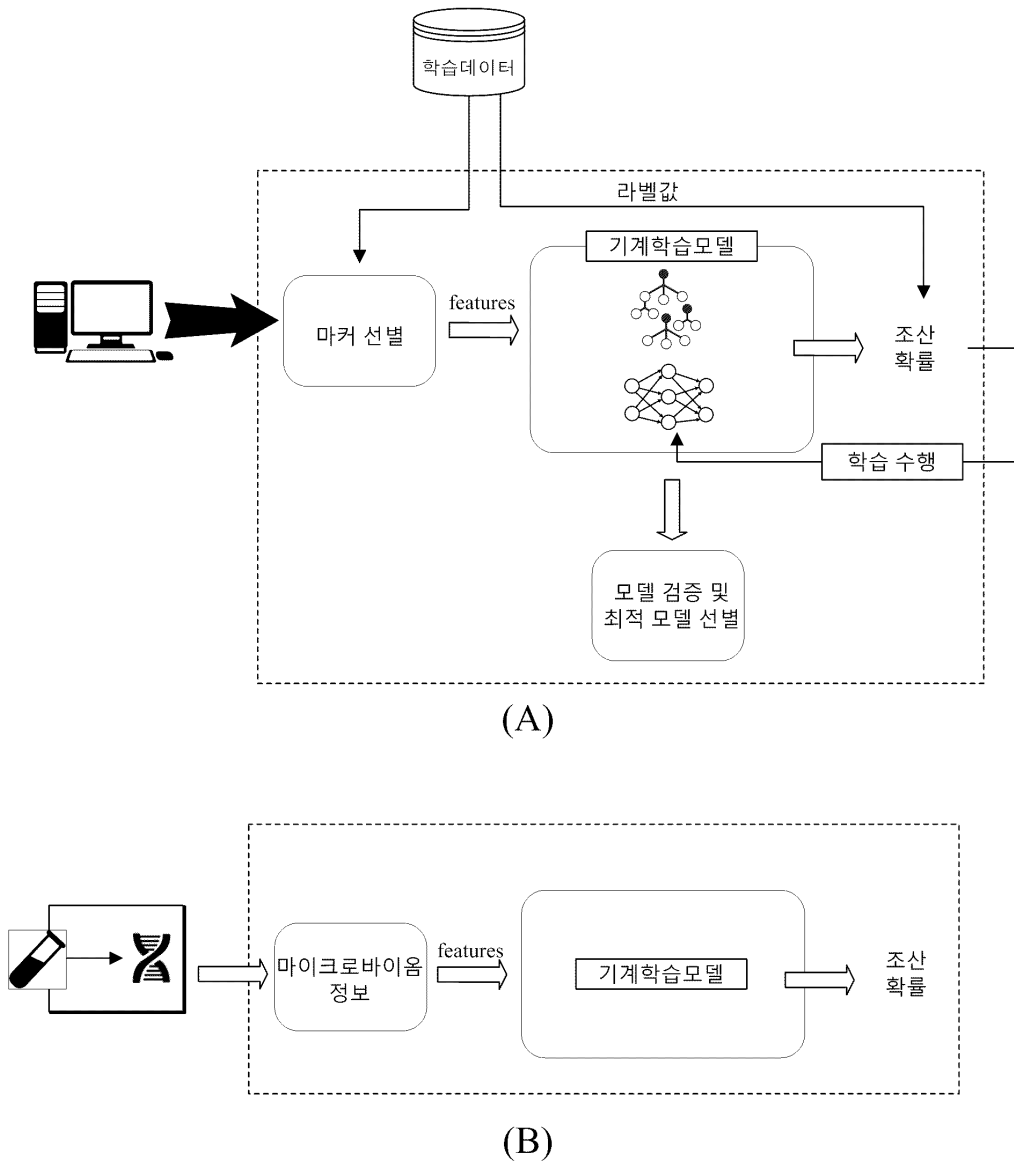
도면1



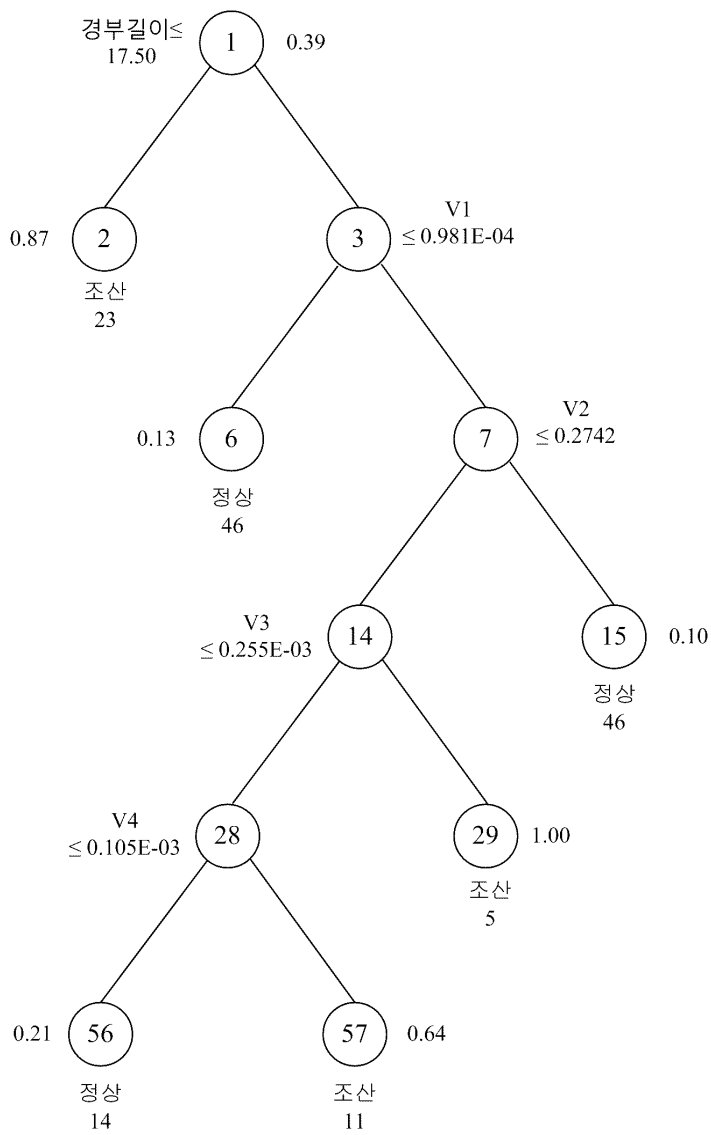
도면2



도면3



도면4



도면5

