



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0091408
(43) 공개일자 2022년06월30일

- (51) 국제특허분류(Int. Cl.)
G16H 50/20 (2018.01) G01N 21/65 (2006.01)
G06N 20/00 (2019.01) G16H 10/00 (2018.01)
G16H 30/00 (2018.01) G16H 30/40 (2018.01)
G16H 50/50 (2018.01) G16H 50/70 (2018.01)
- (52) CPC특허분류
G16H 50/20 (2018.01)
G01N 21/658 (2013.01)
- (21) 출원번호 10-2021-0185168
(22) 출원일자 2021년12월22일
심사청구일자 2021년12월22일
- (30) 우선권주장
1020200181367 2020년12월22일 대한민국(KR)
1020200181368 2020년12월22일 대한민국(KR)

- (71) 출원인
두에이아이(주)
경기도 성남시 수정구 창업로 54,415호,416호,707호,708호,709호(시흥동, 판교제2테크노밸리기업성장센터)
- 울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
(뒷면에 계속)
- (72) 발명자
남좌민
김송철
서울시 성동구 왕십리로 16 트리마제 104-3001
서울시 송파구 올림픽로 435 파크리오 아파트 221동 2002호
(뒷면에 계속)
- (74) 대리인
특허법인알파엠

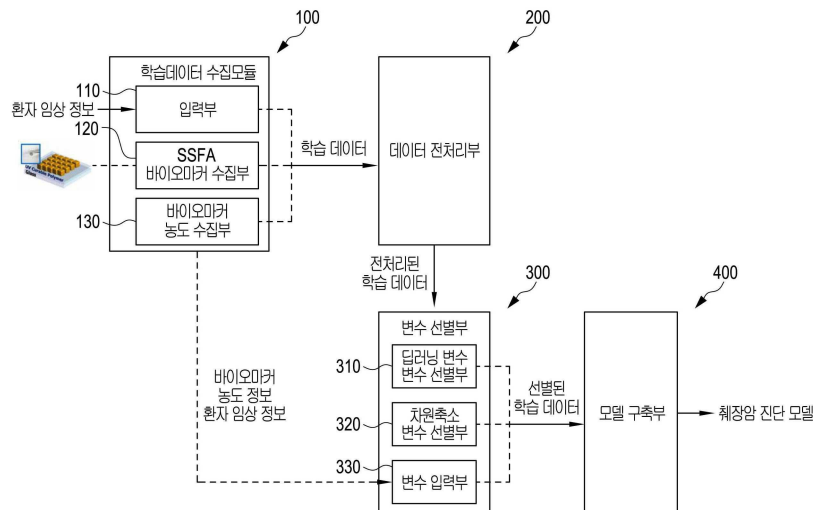
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 인공지능 라만 분석 기법을 이용한 혈장암 조기진단 방법

(57) 요약

본 발명은 인공지능 라만 분석 기법을 이용한 혈장암 조기진단 방법에 관한 것이다.

대표도



(52) CPC특허분류

G06N 20/00 (2021.08)

G16H 10/00 (2021.08)

G16H 30/00 (2021.08)

G16H 30/40 (2018.01)

G16H 50/50 (2018.01)

G16H 50/70 (2018.01)

(71) 출원인

서울대학교산학협력단

서울특별시 관악구 관악로 1 (신림동)

재단법인 아산사회복지재단

서울특별시 송파구 올림픽로43길 88 (풍납동)

(72) 발명자

이학진

경기도 광주시 오포로 701-3 오포추자서희스타힐스 103동 201호

황재호

경기도 용인시 수지구 성복2로 251 212-1303

최용준

서울시 송파구 위례중양로124, 리슈빌 퍼스트클래스 108동 704호

김지은

서울특별시 성북구 북악산로 851, 106동 901호 (정릉동, 정릉2차 e-편한세상)

이우형

서울시 송파구 올림픽로 47길9

임경목

서울시 동작구 흑석로 13마길20 303호

이연희

서울시 관악구 신림로7가길 20-14 102호

차승상

서울시 관악구 은천로 86 211동 2303호

명세서

청구범위

청구항 1

췌장암 진단 모델을 생성하는 방법으로서,

- (a) 학습 데이터 수집부(100)가 미리 설정된 제1 방법을 이용하여 환자의 혈액으로부터 라만 스펙트럼 정보를 수집하는 단계;
- (b) 상기 학습 데이터 수집부(100)가 상기 라만 스펙트럼 정보를 포함하는 학습 데이터를 생성하는 단계; 및
- (c) 모델 구축부(400)가 상기 학습 데이터를 이용하여 췌장암 진단 모델을 생성하되, 상기 췌장암 진단 모델은 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 췌장암 보유 여부를 출력하는 모델인, 단계;를 포함하는,

방법.

청구항 2

제1항에 있어서,

상기 (a)단계에서,

상기 미리 설정된 제1 방법은 큐브들 사이에 나노 갭(11)을 갖는 금 나노 큐브 어레이가 형성된 나노 큐브 기판(10)에 위치되는 환자의 혈액을, 표면증강 라만 분광분석(Surface enhanced Raman Spectroscopy, SERS) 기법을 통해 라만 스펙트럼 정보를 획득하는 방법인,

방법.

청구항 3

제2항에 있어서,

상기 (a)단계는,

상기 학습 데이터 수집부(100)가 바이오마커 농도 정보를 수집하는 단계를 더 포함하는,

방법.

청구항 4

제3항에 있어서,

상기 바이오마커 농도 정보는 CA19-9(Carbohydrate antigen 19-9), CEA (Cacinoembryonic antigen), LRG1(Leucine-rich alpha-2-glycoprotein 1), CFB (Complement Factor B), TTL (Tubulin tyrosine ligase), Thrombosondin-2 (THBS2) 중 어느 하나 이상에 대한 농도 정보를 포함하는,

방법.

청구항 5

제4항에 있어서,

상기 (b)단계 이후 상기 (c)단계 이전에,

(x) 데이터 전처리부(200)가 라만 스펙트럼 정보를이용하여 전처리를 수행하는 단계;를 더 포함하고,

상기 (x)단계는,

(x1) 상기 데이터 전처리부(200)가 상기 라만 스펙트럼 정보에 대하여 특이치(outlier)를 제거하는 단계; 및

(x2) 상기 데이터 전처리부(200)가 상기 특이치가 제거된 상기 라만 스펙트럼 정보에 대하여 기준선(baseline)을 제거하는 단계; 를 포함하는,

방법.

청구항 6

제5항에 있어서,

상기 (x)단계 이후,

(y) 변수 선별부(300)가 상기 전처리된 학습 데이터를 각각 주성분 분석(PCA, Principle components analysis) 및 딥러닝 학습 중 어느 하나 이상을 이용하여 변수를 선별하고, 상기 전처리된 학습 데이터에서 선별된 변수가 포함되도록 선별 학습 데이터를 생성하는 단계;를 더 포함하고,

상기 (c)단계에서,

상기 모델 구축부(400)는 상기 선별 학습 데이터를 학습하여, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 체장암 보유 유무를 출력하는 체장암 진단 모델을 생성하는 단계를 더 포함하는,

방법.

청구항 7

제5항에 있어서,

상기 (x)단계 이후,

(z) 변수 선별부(300)가 상기 바이오마커 농도 정보와, 환자 임상 정보 중 어느 하나 이상이 포함되도록 선별 학습데이터를 생성하는 단계;를 더 포함하고,

상기 (c)단계에서,

상기 모델 구축부(400)는 상기 선별 학습데이터를 추가 학습하여, 임의의 개체의 임상 정보, 상기 임의의 개체의 바이오마커 농도 정보와, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보 중 하나 이상을 입력하는 경우, 상기 임의의 개체의 체장암 보유 유무를 출력하는 체장암 진단 모델을 생성하는 단계를 더 포함하고,

상기 환자의 임상 정보는, 환자의 성별, 연령, 체중, 키 및 체질량 지수 중 어느 하나 이상을 포함하는,

방법.

청구항 8

제5항에 있어서,

상기 (x)단계 이후,

(o) 변수 선별부(300)가 상기 전처리된 학습 데이터를 각각 주성분 분석(PCA, Principle components analysis)으로부터 선별된 변수, 딥러닝 학습으로부터 선별된 변수, 상기 바이오 마커 농도 정보와, 환자 임상

정보 중 어느 하나 이상이 포함되도록 선별 학습데이터를 생성하는 단계;를 더 포함하고,

상기 (c)단계에서,

상기 모델 구축부(400)는 상기 선별 학습데이터를 학습하여, 임의의 개체의 임상 정보, 상기 임의의 개체의 바이오마커 농도 정보와, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼을 입력하는 경우, 상기 임의의 개체의 췌장암 보유 유무를 출력하는 췌장암 진단 모델을 생성하는 단계를 더 포함하고,

상기 환자의 임상 정보는, 환자의 성별, 연령, 체중, 키 및 체질량 지수 중 어느 하나 이상을 포함하는, 방법.

청구항 9

제1항에 있어서,

상기 (c)단계 이후,

(d) 추가 학습부(600)가 상기 췌장암 진단 모델을 추가 학습시키는 단계;를 더 포함하고,

상기 (d)단계는,

(d1) 상기 추가 학습부(600)가 상기 라만 스펙트럼 정보의 원본 신호에 잡음 생성(Jittering), 신호 크기 조정(Scaling), 신호 회전(Rotation), 구간 섞기(Permutation), 뒤틀림 추가(Magnitude warping), 선형 변환(Linear transformation), 시프팅(Shifting) 중 어느 하나 이상의 방법을 이용하여 추가 학습 데이터를 생성하는 단계;를 포함하는,

방법.

청구항 10

제9항에 있어서,

상기 (d)단계는,

(d2) 상기 추가 학습부(600)가 상기 라만 스펙트럼 정보의 기준선을 변경하여 다른 추가 학습 데이터를 생성하는 단계; 및

(d3) 상기 추가 학습부(600)가 상기 다른 추가 학습 데이터를 이용하여 상기 췌장암 진단 모델을 더 추가 학습시키는 단계;를 더 포함하는,

방법.

청구항 11

제1항, 제6항 내지 제8항 중 어느 한 항에 있어서,

상기 (c)단계는,

(c1) 상기 모델 구축부(400)가 상기 학습 데이터를 인공신경망(ANN, Artificial neural network), 서포트 벡터 머신(Support vector machine), 로지스틱 회귀(Logistic regression), 경사 부스팅(Tree based Gradient boosting), 딥러닝 학습(Deep Learning) 중 어느 하나 이상의 방법을 이용하여 학습하여, 상기 췌장암 진단 모델을 생성하는 단계;를 포함하는,

방법.

청구항 12

제1항에 따라 생성된 모델을 이용하여 췌장암을 진단하는 방법으로서,

상기 (c)단계 이후,

(e) 상기 췌장암 진단 모델을 전달받은 진단부(500)에 환자의 임상 정보와 환자의 라만 스펙트럼 정보를 입력하면, 출력부를 통해 췌장암 진행 유무가 출력되는 단계;를 더 포함하는,

방법.

청구항 13

제1항에 따라 생성된 모델을 이용하여 췌장암을 진단하는 시스템으로서,

임의의 개체의 임상 정보와, 상기 임의의 개체의 혈액으로부터 획득된 라만 스펙트럼 정보를 입력받는 입력 장치; 및

상기 췌장암 진단 모델을 전달받아, 상기 입력 장치로부터 입력된 임상 정보와 라만 스펙트럼 정보가 상기 췌장암 진단 모델에 입력되는 경우, 입력된 정보에 대한 췌장암 보유 유무를 출력하는 연산 장치;를 포함하는,

시스템.

청구항 14

제1항 내지 제10항 및 제12항 중 어느 한 항에 따른 방법을 실행하도록 컴퓨터 판독 가능한 기록 매체에 저장된, 컴퓨터 프로그램.

청구항 15

제2항에 따른 방법을 수행하기 위한 기판을 이용하여 라만 스펙트럼 정보를 수집하는 방법으로서,

(q) 상기 환자의 혈액을 기설정된 배수만큼 희석시키고, 희석된 혈액이 상기 나노 갭(11)이 형성되는 상기 나노 큐브 기판(10)에 위치되는 단계; 및

(r) 라만분광부가 상기 나노 큐브 기판(10)에 표면증강 라만 분광분석(Surface enhanced Raman Spectroscopy, SERS)하여 라만 신호를 발생시키고, 상기 SSFA 바이오마커 수집부(120)가 상기 라만 스펙트럼 정보를 수집하는 단계; 를 포함하는,

방법.

발명의 설명

기술 분야

[0001] 본 발명은 인공지능 라만 분석 기법을 이용한 췌장암 조기진단 방법에 관한 것이다.

배경 기술

[0002] 라만 산란(Raman scattering)이란 빛과 분자 진동 운동 간의 상호작용으로 발생하는 광학적 현상이다. 그러나 라만 산란은 발생 빈도가 매우 낮기 때문에 플라즈모닉 나노구조체(plasmonic nanostructure)의 표면 플라즈몬(surface plasmon)을 이용한 표면-증강 라만 산란(SERS, surface-enhanced Raman scattering)와 같은 신호의 증폭이 필수적이다.

[0003] 증폭된 신호는 분자의 다양한 진동 운동들에 대한 정보가 포함되어있고, 마치 지문과 같이 특이적인 패턴이 라만 이동(Raman shift)와 세기로써 스펙트럼 안에 나타난다. 이와 같은 특이적인 패턴은 특정 분자 검지 및 정량, 분자 검지를 통한 질병 진단과 같은 다양한 분야에 대한 응용이 제안되었다.

[0004] 다양한 분야에 대한 응용 중, SERS 기반 라만 스펙트럼을 이용하여 인체의 질병을 진단하는데 이용될 수

있으며, 특히 조기 진단이 어려운 췌장암을 진단하는데 적극 이용될 수 있다.

[0005] 가령, 조기 진단이 어려운 질병으로 손꼽히는 췌장암의 경우에는, 현재까지 증상이 나타나기 전에 췌장암을 조기에 발견할 수 있는 공인된 선별검사 방법이 없는 실정이다. 췌장암 진단을 위하여, 복부 초음파, 복부 전산화 단층촬영(CT), 자기공명영상(MRI), 내시경적 역행성 담췌관 조영술(ERCP), 내시경 초음파(EUS), 양성자방출 단층촬영(PET), 혈청종양 표지자(CA19-9) 검사에 대한 연구가 활발히 이루어지고 있다. 그러나, 종래의 췌장암 진단에 사용되는 CA19-9(Carbohydrate antigen 19-9) 지표의 경우 진단 성능이 민감도 78%, 특이도 82% 정도에 그친다는 문제가 있다. 이러한, 췌장암의 진단에도 SERS 기반 라만 스펙트럼을 이용하여 췌장암을 진단할 수 있다.

[0006] 그러나 SERS 기반 분자 검지를 통한 췌장암 진단에 있어서 구조적 불안정성 및 낮은 재현성에서 오는 신호의 편차와 이에 따른 성능 저하, 높은 난해성에 따른 정보의 제한적 이용이 기술적 한계로 지적되고 있으며, 기존 SERS를 이용한 진단 기술은 사람이 인지할 수 없는 라만 스펙트럼(spectrum) 내 많은 정보를 생략하고 선명한 특정 피크의 세기 비교로만 진단하여 정확성이 낮은 문제가 있었으며, 특정 바이오마커의 기준 스펙트럼이 존재하지 않으면 혼합물에서 분석이 불가능한 한계가 있었다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허문헌 제10-1830314호
(특허문헌 0002) 한국공개특허문헌 제10-2021-0100068호

발명의 내용

해결하려는 과제

[0008] 본 발명은 상기와 같은 문제점을 해결하기 위하여 안출된 것이다.

[0009] 구체적으로, 본 발명은 난해성 높은 정보들에서 정보 추출이 용이한 인공지능과 복잡한 정보들의 중첩으로 난해성이 높은 라만 기술을 융합하여 정보를 제한적으로 사용한 기존 라만 기반 분자 검지를 통한 췌장암 진단법의 한계를 극복하는 것을 목적으로 한다.

[0010] 또한, 많은 정보를 내포한 혈액 자체를 스펙트럼으로서 활용하여 특정한 분자 바이오마커가 없는 췌장암의 진단이 가능하도록 하며, 정상인과 췌장암 환자 사이의 스펙트럼 간 유의미한 차이를 확인하여 스펙트럴 바이오마커로서 이용하여 진단하는 것을 목적으로 한다.

[0011] 또한, 본 발명은 SERS 기판을 통해 획득된 스펙트럴 바이오마커 및 SLISA 기판을 통해 획득된 라만 스펙트럼을 통해 췌장암 조기 진단 모델을 생성하고, 이를 이용하여 환자의 췌장암 및 타종 암 여부를 정확하게 진단하는 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0012] 상기와 같은 과제를 해결하기 위한 본 발명의 일 실시예는, 췌장암 진단 모델을 생성하는 방법으로서, (a) 학습 데이터 수집부(100)가 미리 설정된 제1 방법을 이용하여 환자의 혈액으로부터 대한 라만 스펙트럼 정보를 수집하는 단계; (b) 상기 학습 데이터 수집부(100)가 상기 라만 스펙트럼 정보를 포함하는 학습 데이터를 생성하는 단계; 및 (c) 모델 구축부(400)가 상기 학습 데이터를 이용하여 췌장암 진단 모델을 생성하되, 상기 췌장암 진단 모델은 상기 학습 데이터를 학습하여 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 췌장암 보유 유무를 출력하는 모델인, 단계;를 포함하는, 방법을 제공한다.

[0013] 일 실시예에 있어서, 상기 (a)단계에서, 상기 미리 설정된 제1 방법은 큐브들 사이에 나노 갭(11)을 갖는 금 나노 큐브 어레이가 형성된 나노 큐브 기판(10)에 위치되는 환자의 혈액을, 표면증강 라만 분광분석(Surface enhanced Raman Spectroscopy, SERS) 기법을 통해 라만 스펙트럼 정보를 획득하는 방법일 수 있다.

[0014] 일 실시예에 있어서, 상기 (a)단계는, 상기 학습 데이터 수집부(100)가 바이오마커농도 정보를 수집하는 단계를 더 포함할 수 있다.

- [0015] 일 실시예에 있어서, 상기 바이오마커 농도 정보는 CA19-9(Carbohydrate antigen 19-9), CEA (Cacinoembryonic antigen), LRG1(Leucine-rich alpha-2-glycoprotein 1), CFB (Complement Factor B), TTL (Tubulin tyrosine ligase) 및Thrombosondin-2 (THBS2) 중 어느 하나 이상에 대한 농도 정보를 포함할 수 있다.
- [0016] 일 실시예에 있어서, 상기 (b)단계 이후 상기 (c)단계 이전에, (x) 데이터 전처리부(200)가 라만 스펙트럼 정보를 이용하여 전처리를 수행하는 단계; 상기 (x)단계는, (x1) 상기 데이터 전처리부(200)가 상기 라만 스펙트럼 정보에 대하여 특이치(outlier)를 제거하는 단계; 및 (x2) 상기 데이터 전처리부(200)가 상기 특이치가 제거된 상기 라만 스펙트럼 정보에 대하여 기준선 (baseline)을 제거하는 단계; 를 포함할 수 있다.
- [0017] 일 실시예에 있어서, 상기 (x)단계 이후, (y) 변수 선별부(300)가 상기 전처리된 학습 데이터를 각각 주성분 분석 (PCA, Principle components analysis) 및 딥러닝 학습 중 어느 하나 이상으로부터 변수를 선별하고, 상기 전처리된 학습데이터에서 선별된 변수가 포함하도록 선별 학습데이터를 생성하는 단계;를 더 포함하고, 상기 (c)단계에서, 상기 모델 구축부(400)는 상기 선별 학습데이터를 학습하여, 임의의 개체의 임상 정보와, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 췌장암 보유 유무를 출력하는 췌장암 진단 모델을 생성하는 단계를 더 포함할 수 있다.
- [0018] 일 실시예에 있어서, 상기 (x)단계 이후, (z) 변수 선별부(300)가 바이오마커 농도 정보와, 환자 임상 정보 중 어느 하나 이상이 포함되도록 선별 학습데이터를 생성하는 단계;를 더 포함하고, 상기 (c)단계에서, 상기 모델 구축부(400)는 상기 선별 학습데이터를 학습하여, 임의의 개체의 임상 정보와, 상기 임의의 개체의 바이오마커 농도 정보와, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 췌장암 보유 유무와 췌장암 진행 단계를 함께 출력하는 췌장암 진단 모델을 생성하는 단계를 더 포함하고, 상기 환자의 임상 정보는, 환자의 성별, 연령, 체중, 키 및 체질량 지수 중 어느 하나 이상을 포함할 수 있다.
- [0019] 일 실시예에 있어서, 상기 환자의 임상 정보는, 췌장암 진행 단계를 더 포함하고, 상기 (x)단계 이후, (o) 변수 선별부(300)가 상기 전처리된 학습 데이터를 각각 주성분 분석 (PCA, Principle components analysis)으로부터 선별된 변수, 딥러닝 학습으로부터 선별된 변수, 상기 바이오마커농도 정보와, 환자 임상 정보 중 어느 하나 이상이 포함되도록 선별 학습데이터를 생성하는 단계;를 더 포함하고, 상기 (c)단계에서, 상기 모델 구축부(400)는 상기 선별 학습데이터를 학습하여, 임의의 개체의 임상 정보와, 상기 임의의 개체의 바이오마커 농도 정보와, 상기 임의의 개체의 혈액으로부터 수집한 라만 스펙트럼 정보를 입력하는 경우, 상기 임의의 개체의 췌장암 보유 유무를 출력하는 췌장암 진단 모델을 생성하는 단계를 더 포함하고, 상기 환자의 임상 정보는, 환자의 성별, 연령, 체중, 키 및 체질량 지수 중 어느 하나 이상을 포함할 수 있다.
- [0020] 일 실시예에 있어서, 상기 (c)단계 이후, (d) 추가 학습부(600)가 상기 췌장암 진단 모델을 추가 학습시키는 단계;를 더 포함하고, 상기 (d)단계는, (d1) 상기 추가 학습부(600)가 라만 스펙트럼 정보의 원본 신호에 잡음 생성(Jittering), 신호 크기 조정(Scaling), 신호 회전(Rotation), 구간 섞기(Permutation), 뒤틀림 추가 (Magnitude warping), 선형 변환(Linear transformation), 시프팅(Shifting) 중 어느 하나 이상의 방법을 이용하여 추가 학습 데이터를 생성하는 단계; 를 포함할 수 있다.
- [0021] 일 실시예에 있어서, 상기 (d)단계는, (d2) 상기 추가 학습부(600)가 상기 라만 스펙트럼 정보의 기준선을 변경하여 다른 추가 학습 데이터를 생성하는 단계; 및 (d3) 상기 추가 학습부(600)가 상기 다른 추가 학습 데이터를 이용하여 상기 췌장암 진단 모델을 더 추가 학습 시키는 단계;를 더 포함할 수 있다.
- [0022] 일 실시예에 있어서, 상기 (c)단계는, (c1) 상기 모델 구축부(400)가 상기 학습 데이터를 인공신경망(ANN, Artificial neural network), 서포트 벡터 머신(Support vector machine), 로지스틱 회귀(Logistic regression), 경사 부스팅(Tree based Gradient boosting), 딥러닝 학습 중 어느 하나 이상의 방법을 이용하여 학습하여, 상기 췌장암 진단 모델을 생성하는 단계; 를 포함할 수 있다.
- [0023] 본 발명의 다른 실시예에서, 생성된 모델을 이용하여 췌장암을 진단하는 방법으로서, 상기 (c)단계 이후, (e) 상기 췌장암 진단 모델을 전달받은 진단부(500)에 환자의 임상 정보와 환자의 라만 스펙트럼 정보 중 하나 이상을 입력하면, 출력부를 통해 췌장암 진행 단계가 출력되는 단계;를 더 포함하는 방법을 제공한다.
- [0024] 본 발명의 다른 실시예에서, 생성된 모델을 이용하여 췌장암을 진단하는 시스템으로서, 임의의 개체의 임상 정보와, 상기 임의의 개체의 혈액으로부터 획득된 라만 스펙트럼 정보를 입력받는 입력 장치; 및 상기 췌장암 진단 모델을 전달받아, 상기 입력 장치로부터 입력된 임상 정보와 라만 스펙트럼 정보가 상기 췌장암 진단 모델에 입력되는 경우, 입력된 정보에 대한 췌장암 보유 유무를 출력하는 연산 장치;를 포함하는 시스템을 제공한다.

[0025] 본 발명의 다른 실시예에서, 상기의 방법을 실행하도록 컴퓨터 판독 가능한 기록 매체에 저장된, 컴퓨터 프로그램을 제공한다.

[0026] 본 발명의 다른 실시예에서, 상기의 방법을 수행하기 위한 기관을 이용하여 라만 스펙트럼 정보를 수집하는 방법으로서, (q) 상기 환자의 혈액을 기설정된 배수만큼 희석시키고, 희석된 혈액이 상기 나노 캡(11)이 형성되는 상기 나노 큐브 기관(10)에 위치되는 단계; 및 (r) 라만분광부가 상기 나노 큐브 기관(10)에 표면증강 라만 분광분석(Surface enhanced Raman Spectroscopy, SERS)하여 라만 신호를 발생시키고, 상기 SSFA 바이오마커 수집부(120)가 상기 라만 스펙트럼 정보를 수집하는 단계; 를 포함하는, 방법을 제공한다.

발명의 효과

[0027] 본 발명에 따라, 다음과 같은 효과가 달성된다.

[0028] 본 발명에서는, 바이오마커의 정량적 정보뿐만 아니라 혈액 내에서 확인되는 전체적인 라만 스펙트럼의 패턴(pattern)을 통해 암과 관련된 정보를 스펙트럼 내에 포함시켜 스펙트럼 자체를 바이오마커로서 사용이 가능하고, 기존의 액체 생검에서 놓칠 수 있는 물질 정보들까지 사용할 수 있다는 장점이 있으며, 높은 진단 정확도, 단순하고 빠른 검사 방법, 저렴한 검사비용을 제공할 수 있다.

[0029] 또한, 본 발명은 SERS 기관을 통해 획득된 스펙트럼 바이오마커를 통해 체장암 조기 진단 모델을 생성하고, 이를 이용하여 환자의 체장암 및 타종 암 여부를 정확하게 진단할 수 있다.

도면의 간단한 설명

[0030] 도 1은 본 발명에 따른 체장암 진단 모델 생성 방법을 설명하기 위한 모식도이다.

도 2는 SERS 소재에 따른 스펙트럼 비교 결과와 체장암과 정상인의 라만 스펙트럼 차이를 도시한다.

도 3은 SSFA 기술을 이용하여, 라만스펙트럼을 획득하는 방법을 설명하기 위한 도면이다.

도 4는 동일한 SERS기관의 동일 샘플에서 획득된 50개의 라만 스펙트럼을 도시한 도면이다.

도 5는 데이터 전처리부가 학습 데이터를 전처리하는 방법을 설명하기 위한 도면이다.

도 6은 데이터 전처리 결과, 라만 스펙트럼의 원본 신호와 기준선이 제거된 신호를 도시한 그래프이다.

도 7은 본 발명에 따라 생성된 체장암 진단 모델의 정확도를 설명하기 위한 도면이다.

도 8은 본 발명에 따른 진단부에서 체장암 진행 단계를 출력하는 것을 설명하기 위한 도면이다.

도 9 내지 도 12는 추가 학습부가 체장암 진단 모델을 추가 학습하는 것을 설명하기 위한 도면이다.

도 13은 나노 큐브 기관을 설명하기 위한 도면이다.

도 14는 나노 큐브 기관에서 라만 스펙트럼 획득하는 과정을 설명하기 위한 도면이다.

도 15는 혈장이 있을 때와 없을 때의 라만 스펙트럼을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0031] 몇몇 경우, 본 발명의 개념이 모호해지는 것을 피하기 위하여 공지의 구조 및 장치는 생략되거나, 각 구조 및 장치의 핵심기능을 중심으로 한 블록도 형식으로 도시될 수 있다.

[0032] 또한, 본 발명의 실시예들을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 그리고 후술되는 용어들은 본 발명의 실시예에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0033] 이하, 본 발명에서는 체장암을 진단하는 것을 기준으로 설명하고 있으나, 이에 한정되지 않고, 1차 적용 분야는 간체담도 외과, 진단검사의학 분야가 될 것이며 다른 종류의 암을 비롯한 질병을 진단하는 기술로 확장이 가능하다.

[0034] 한편, 본 발명에서 개시된 각각의 설명 및 실시형태는 각각의 다른 설명 및 실시 형태에도 적용될 수 있다. 즉, 본 발명에서 개시된 다양한 요소들의 모든 조합이 본 발명의 범주에 속한다. 또한, 하 기 기술된 구체적인 서술

에 의하여 본 발명의 범주가 제한된다고 볼 수 없다.

- [0036] 본 발명은 암 환자의 혈액 내에 존재하는 화학 성분들과 구성요소들에 관련된 라만 스펙트럼(Raman spectrum)을 인공지능으로 분석하여 췌장암 환자의 조기진단을 쉽고 빠르게 확인할 수 있도록 할 수 있다. 본 발명은 표면증강 라만 분광분석(surface enhanced Ramanspectroscopy; SERS) 기술과 인공지능 기술을 융합한 신규 액체 생검 기술을 적용하여 췌장암 환자의 조기진단을 수행할 수 있다.
- [0037] 기존의 진단 검사를 받기 위해서는 대형장비와 전문인력을 보유한 대형 병원에 방문해야 했으나, 본 발명의 다양한 실시예에 따르면 보건소나 의원 같은 1차 의료기관에도 보급 가능하다는 이점이 있다.
- [0038] 본 발명에 따른 방법에서, 환자의 혈장을 채취하여 SERS기관에 담아 라만 분광기에 넣으면 인공지능이 분석한 진단 결과가 출력될 수 있다.
- [0040] 도 1 내지 도 15를 참조하여, 본 발명에 따른 방법을 설명한다.
- [0041] 본 발명에 따른 방법은, 학습 데이터 수집부(100), 데이터 전처리부(200), 라벨링부(300), 모델 구축부(400), 진단부(500) 및 추가 학습부(600)를 포함한다.
- [0043] **1. 췌장암 진단 모델 생성**
- [0044] **(1)학습 데이터 수집부(100)**
- [0045] 학습 데이터 수집부(100)는 본 발명의 췌장암 진단 모델을 생성하기 위한 데이터를 수집하고, 이를 이용하여 학습 데이터를 생성한다.
- [0046] 학습 데이터 수집부(100)는 입력부(110), SSFA 바이오마커 수집부(120) 및 바이오마커 농도 수집부(130)를 포함한다.
- [0047] 입력부(110)에 환자 임상 정보가 입력된다.
- [0048] 환자 임상 정보는 환자의 성별, 연령, 체중, 키, 체질량지수 및 췌장암 진행 단계 중 어느 하나 이상을 포함할 수 있다. 하지만 이에 제한되지 않고, 환자 임상과 관련된 모든 정보, 예를 들어 혈압 등이 여기에 포함될 수 있다.
- [0050] 학습 데이터 수집부(100)가 환자 임상 정보와, 미리 설정된 제1 방법을 이용하여 환자의 혈액으로부터 대한 라만 스펙트럼 정보를수집한다. 또한, 후술하는 바와 같이 바이오마커 농도 정보를 더 수집한다. 본 발명에서는 환자 임상 정보, 라만 스펙트럼 정보, 바이오마커 농도 정보 중 어느 하나 이상을 이용하여 췌장암 진단 모델을 생성할 수 있다.
- [0052] SSFA 바이오마커 수집부(120)가 환자의 혈액(혈장)을 나노 큐브 기관(10)에서 표면증강 라만 분광분석(Surface enhanced Raman Spectroscopy, SERS)하여 라만 스펙트럼 정보를 수집한다. 학습 데이터 수집부(100)는 상기의 과정을 반복 수행하여 다수의 학습 데이터를 생성할 수 있다.
- [0053] 도 3(a)에서는 나노 큐브 기관(10)으로부터 라만 스펙트럼 정보를 수집하는 것이 도시되며, 도 4는 동일한 나노 큐브 기관(10)의 동일 샘플에서 획득된 라만 스펙트럼을 도시한 도면이다.
- [0054] SSFA 바이오마커 수집부(120)는 동일한 기관에 놓인 환자의 혈액에 라만 분광분석법을 다수 회 적용하여 동일한 대상에 대한 복수의 라만 스펙트럼을 수집할 수 있다. 도 4에서는 SSFA 바이오마커 수집부(120)는 동일한 기관에 놓인 환자의 혈액에 라만 분광분석법을 다수 회 적용하여 50개의 라만 스펙트럼을 수집한 것을 도시한다.
- [0055] 나노 큐브 기관(10)은 나노 갭(11)이 형성된다. 나노 큐브 기관(10)에 대한 보다 구체적인 설명에 대해서는 후술하기로 한다.
- [0056] 이 때, 라만 분광분석법으로 얻어진 스펙트럼은 일정한 파장 범위 내에서 각각의 파수(wave number) 별로 라만

산란의 세기 값을 갖는 일차원 신호 벡터(vector)로 정의될 수 있다.

[0057] 도 2를 참조하면, 혈장이 투입되는 기관 디자인(소재/형태)에 따른 스펙트럼 비교 결과(도 2(a), (b))와 채장압과 정상인의 라만 스펙트럼 차이(도 2(c))를 도시한다. 도 2(a)에서, 여러 기관 소재(금 필름, 나노 큐브 등)에 따른 스펙트럼 변화를 비교해본 결과 나노 큐브-혈장 혼합물 사용과 같은 라만 증폭 기술이 필요함을 확인하였다, 이러한 신호 증폭이 없는 유리나 금필름의 경우 라만 신호가 발생하지 않은 것을 확인할 수 있었다.

[0058] 다만, 도 2(b)에서, 나노큐브-혈장 혼합물에서 랜덤 스팟의 스펙트럼 각각 측정하였을 때, 스팟 별로 큰 변화가 발생하며, 재현성 낮은 것이 도시된다. 이에 따라, 조절성이 좋은 패턴을 통한 기관이 필요한 것을 알 수 있다.

[0060] 또한, 바이오마커 농도 수집부(130)는 채장압 진단을 위한 바이오마커 농도 정보를 확인할 수 있다. 이는, 후술하는 변수 선별부(300)로 전송되고, 변수 선별부(300)에서 채장압을 진단하기 위한 하나의 변수로 사용될 수 있다.

[0061] 바이오마커 농도 수집부(130)에서 바이오마커 농도 정보를 수집하는 방법은 특정한 방법에 제한되는 것은 아니다. 가령, 바이오마커 농도 정보 수집부(130)는 기존에 공지된 채장압 진단에 이용되는 바이오마커 농도 정보를 수집할 수 있다. 또한, 바이오마커 농도 수집부(130)는 환자의 혈액으로부터 ELISA(enzyme-linked immunoassay)의 방법을 이용하여 수집된 정보일 수 있으나, 특정한 방법에 제한되지 않는다.

[0062] 바이오마커 농도 정보는 CA19-9(Carbohydrate antigen 19-9), CEA (Cacinoembryonic antigen), LRG1(Leucine-rich alpha-2-glycoprotein 1), CFB (Complement Factor B), TTL (Tubulin tyrosine ligase), Thrombospondin-2 (THBS2) 중 어느 하나 이상에 대한 농도 정보를 포함할 수 있다.

[0064] (2) 데이터 전처리부(200)

[0065] 데이터 전처리부(200)는 학습 데이터 수집부(100)로부터 학습 데이터를 전송받고, 학습 데이터를 전처리한다.

[0067] 데이터 전처리부(200)는 라만 스펙트럼의 특이치(outlier)를 제거한다.

[0068] 데이터 전처리부(200)는 라만 스펙트럼 정보의 특이치(outlier)를 제거한다.

[0069] 특이치(outlier)는 이동창 (moving window) 마다 정상 범위를 계산하고, 정상 범위를 벗어나는 값으로, 특이치는 정상범위보다 낮거나 높은 세기 값이다.

[0070] 이 때, 도 5를 참조하면, 이동창은 100개 길이일 수 있고, 이동창 사이의 간격은 50개이나, 이에 한정되는 것은 아니다.

[0071] 데이터 전처리부(200)는 정상 범위 계산을 위해 이동창 별로 신호 세기의 평균 값과 표준 편차 값을 연산한다.

[0072] 이 때, 정상 범위는 평균값 - a * 표준 편차부터 평균값 + a * 표준편차로 가정하였으며, 여기서는 a는 5의 값으로 설정되었으며 1부터 10 사이의 값으로 변경하여 정상 범위를 수정할 수 있다.

[0073] 데이터 전처리부(200)는 라만 스펙트럼 별로 특이치들을 제외한 이동창 범위 내 평균값을 수정값으로 확인한다.

[0074] 데이터 전처리부(200)는 특이치에 해당하는 세기 값들은 이 수정값으로 대체할 수 있다.

[0075] 그러나, 본 발명에 따른 진단 모델은 이에 한정 되지 않고, 다수의 스펙트럼 각각의 특이치를 제거하지 않고, 특이치를 무시하고 스펙트럼에 대한 분석을 수행할 수 있는 인공지능 모델을 생성할 수 있다

[0077] 이후, 데이터 전처리부(200)는 각각의 스펙트럼에 대해 기준선(baseline)을 제거한다.

[0078] 기준선(baseline)은 불이익 가중치 기반 최소 자승법에 따라서 추정될 수 있다. 최소 자승법은 라만 스펙트럼의 원본 스펙트럼이 기준선(baseline), 잡음(noise), 여과된 스펙트럼(filtered spectrum)으로 구성된다고 가정한다.

- [0079] 이 때, 기준선(baseline)은 환자의 혈액에 대한 스펙트럼을 추출할 시, 혈액과 관계없이 지속적으로 출력되는 선을 의미한다.
- [0080] 이 때, 잡음(noise)은 SERS 기관에서 스펙트럼을 추출 시, 기관에 의해 발생하는 각종 노이즈를 의미한다.
- [0081] 이 때, 여과된 스펙트럼(filtered spectrum)은 원본 스펙트럼에서 기준선(baseline)과 잡음(noise)이 제거된 스펙트럼을 의미한다.
- [0082] 데이터 전처리부(200)는 반복적으로 신호 내 잡음 정도를 추정하고 이에 따라서 가중치를 조정한다.
- [0083] 데이터 전처리부(200)는 원본신호에 대해서는 기준선보다 낮은 경우에 한하여 가중치를 부여한다. 이러한 가중치를 반영한 기준선과 원본 신호 사이의 잔차(residual)의 제곱이 최소가 되도록 반복해서 기준선을 추정한다.
- [0084] 데이터 전처리부(200)는 최종적으로 잔차가 최소일 때의 기준선을 원본 신호에서 제거한다.
- [0085] 도 6을 참조하면, 원본 신호와 기준선이 제거된 신호를 모두 도시한다. 원본 신호는 상측의 신호로 도시되고, 기준선이 제거된 신호는 하측의 신호로 도시된다.
- [0086] 다양한 실시예에서, 원본 스펙트럼에 포함된 기준선 및 잡음을 추출하고, 추출된 기준선과 잡음을 학습 데이터로 하여 기 학습된 인공 지능 모델을 이용해 원본 스펙트럼으로부터 기준선과 잡음을 추출 및 제거할 수 있다.
- [0088] **(3) 변수 선별부(300)**
- [0089] 변수 선별부(300)는 전술한 데이터 전처리부(200)에서 전처리된 데이터를 전송받는다.
- [0090] 변수 선별부(300)는 전처리된 데이터에서 변수를 선별하여, 선별 학습 데이터를 생성한다. 변수 선별부(300)는 선별 학습 데이터를 모델 구축부(400)로 전송한다.
- [0092] 변수 선별부(300)는 주성분 분석 (PCA, Principle components analysis) 및 딥러닝 학습 중 어느 하나 이상으로 변수를 선별할 수 있다.
- [0093] 이 때, 변수 선별부(300)의 차원 축소 변수 선별부(320)에서 해당 과정이 이루어질 수 있다.
- [0094] 또한, 변수 선별부(300)는 학습데이터 수집부(100)로부터 바이오마커 농도 정보를 전달받고, 전달된 바이오마커 농도 정보를 변수로 이용할 수 있다. 또한, 변수 선별부(300)는 학습데이터 수집부(100)로부터 환자 임상 정보를 전달받아 이를 변수로 선별할 수 있다.
- [0095] 이 때, 변수 선별부(300)의 변수 입력부(330)가 학습데이터 수집부(100)로부터 바이오마커 농도와 환자 임상 정보를 입력받을 수 있다.
- [0096] 본 발명에서는 주성분 분석으로부터 선별된 변수, 딥러닝 학습으로부터 선별된 변수, 상기 바이오 마커 농도 정보와, 환자 임상 정보중 어느 하나 이상이 포함되도록 선별 학습데이터를 생성할 수 있다. 후술하는 모델 구축부(400)는 선별 학습데이터를 이용하여 복수의 체장암 진단 모델을 생성할 수도 있다.
- [0098] 변수 선별부(300)는 학습 데이터를 주성분 분석 (PCA, Principle components analysis)하여 변수를 선별할 수 있다.
- [0099] 주성분 분석은 하기의 [식 1] 내지 [식 5]의 연산으로 구현되는 것을 특징으로 할 수 있다.
- [0100] 데이터 집합의 전체 평균이 0이라고 가정하면, (아닐 경우 평균을 데이터 집합에서 뺀다) 데이터 집합 x 의 주성분 w_1 은 다음과 같이 정의된다.
- [0101] 여기서 x 는 혈청을 라만분광기에서 분석하여 획득한 라만 스펙트럼 데이터에 해당한다.
- [0102] [식 1]

$$w_1 = \arg \max_{\|w\|=1} E \left\{ (w^T x)^2 \right\}$$

[0103]

[0104] ($\arg \max$ 는 함수 $f(x)$ 의 값을 최대로 만드는 x 를 가리킨다.)

[0105] $k-1$ 개의 주성분이 이미 주어져 있을 때 k 번째 주성분은 앞의 $k-1$ 개 주성분을 뺀으로써 찾을 수 있다:

[0106] [식 2]

$$\hat{\mathbf{x}}_k = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x}$$

[0107]

[0108] 그리고 이 값을 데이터 집합에서 뺀 다음 주성분을 새로 찾는다.

[0109] [식 3]

$$\mathbf{w}_k = \arg \max_{\|\mathbf{w}\|=1} E \left\{ (\mathbf{w}^T \hat{\mathbf{x}}_k)^2 \right\}.$$

[0110]

[0111] 따라서 주성분 분석은 데이터 행렬 $\mathbf{X}$ 의 특잇값 분해를 찾은 다음, $\mathbf{X}$ 를 L 개의 특이 벡터, $\mathbf{W}_L$ 로 정의된 부분공간으로 사상시켜 부분 데이터 집합 $\mathbf{Y}$ 를 찾는 것과 같다.

[0112] [식 4]

$$\begin{aligned} \mathbf{X} &= \mathbf{W} \Sigma \mathbf{V}^T, \\ \mathbf{Y} &= \mathbf{W}_L^T \mathbf{X} = \Sigma_L \mathbf{V}_L^T \end{aligned}$$

[0113]

[0114] $\mathbf{X}$ 의 특잇값 벡터 행렬 $\mathbf{W}$ 는 공분산 $\mathbf{C}=\mathbf{X}\mathbf{X}^T$ 의 고유 벡터 행렬 $\mathbf{W}$ 와 동일하다.

[0115] [식 5]

$$\mathbf{X}\mathbf{X}^T = \mathbf{W}\Sigma^2\mathbf{W}^T$$

[0116]

[0117] 가장 큰 고유값을 갖는 고유 벡터는 데이터 집합에서 가장 강한 상관성을 갖는 차원에 대응된다.

[0118] 변수 선별부(300)는 상기의 과정을 통해 고유 벡터를 변수로 선별할 수 있다.

[0120] 변수 선별부(300)는 학습 데이터를 딥러닝 학습하여 변수를 선별할 수 있다. 이 때, 변수 선별부(300)의 딥러닝 변수 선별부(310)에서 해당 과정이 이루어질 수 있다.

[0121] 딥러닝 학습은 입력과 출력 사이의 수많은 은닉층 (hidden layer)으로 이뤄진 인공 신경망으로, 각각의 은닉층이 일종의 자질추출을 담당하기 때문에 전처리 학습 데이터를 이용하여 딥러닝 분류 모델을 바로 학습시킬 수 있다. 일차원 합성곱 신경망(1-D CNN)은 일차원 데이터의 국소적(local) 특성을 반영하는데 특화되어 있다.

[0122] 또한, 재귀 신경망 (RNN)의 일종인 장단기메모리(LSTM) 네트워크는 음성 및 문자열과 같은 연속적인 데이터를 분석하는 데 탁월하다. 아울러 이 두 가지 모델을 결합한 딥러닝 모델은 심전도 신호, 동작감지 신호등의 분류 혹은 회귀 문제에서 좋은 성능을 보인 바 있어 이를 활용할 수 있다.

[0123] 본 발명에서 사용되는 딥러닝 학습은 DNN(Deep Neural Network), CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), GAN(Generative Adversarial Network), RL(Reinforcement Learning)을 포함할 수 있으나, 이에 제한되는 것은 아니고, 다양한 딥러닝 학습이 적용될 수 있다.

[0125] (4) 모델 구축부(400)

[0126] 모델 구축부(400)는 학습 데이터를 이용하여 체장암 진단 모델을 생성한다.

[0127] 이 때, 모델 구축부(400)는 선별 학습 데이터를 이용하여 체장암 진단 모델을 생성할 수 있다.

[0128] 모델 구축부(400)는 인공신경망(ANN, Artificial neural network), 서포트 벡터 머신(Support vector machine), 로지스틱 회귀(Logistic regression), 경사 부스팅(Tree based Gradient boosting), 딥러닝 학습 중

어느 하나 이상의 방법을 이용하여 학습하여, 체장암 진단 모델을 생성할 수 있다.

[0129] 이 때, 모델 구축부(400)는 학습 데이터를 학습하는 방법은 전술한 바에 제한되는 것은 아니고, 다양한 방식의 머신러닝과 딥러닝 학습이 적용될 수 있음은 물론이다.

[0130] 가령, 본 발명에서 사용되는 딥러닝 학습은 DNN(Deep Neural Network), CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), GAN(Generative Adversarial Network), RL(Reinforcement Learning)을 포함할 수 있으나, 이에 제한되는 것은 아니고, 다양한 딥러닝 학습이 적용될 수 있다.

[0132] 이하, 모델 구축부(400)가 학습 데이터를 인공신경망 모델을 이용하여 체장암 진단 모델을 생성하는 것으로 설명하나, 본 발명에서 체장암 진단 모델을 생성하는 방법은 특정한 방법에 제한되는 것은 아니다.

[0133] 인공 신경망은 생물학적 뉴런의 작동과정을 흉내낸 것으로 데이터를 입력으로 받아, 가중치와 입력을 곱한 후 활성화 함수(Activation) f 의 결과를 다음 뉴런으로 보내는 프레임워크를 갖는다.

[0134] 활성화함수 (Activation Function)는 가중치와 입력(Node)을 곱하는 함수로 은닉층 구성요소를 얻을 경우 활성화 함수로 표현한다.

[0135] 출력함수 (Output Function)는 가중치와 입력을 곱하는 함수로 출력층 결과를 얻을 경우 출력함수로 표현한다.

[0136] 손실함수 (Loss Function)는 가중치의 학습을 위해 출력함수의 결과와 예측(y)값 간의 오차를 측정하는 함수를 의미한다.

[0138] 본 발명의 인공 신경망 기반 조기진단 시스템은, 선별 학습 데이터를 받아들이는 층인 입력층(Input Layer), 가중치를 곱하고 활성화함수의 결과를 얻은 데이터로 구성된 은닉층(Hidden Layer), 최종 은닉층/입력층에 가중치를 곱하고, 출력함수의 결과를 생성하는 출력층(Output Layer)로 구성된다.

[0139] 본 발명의 일 실시 예에 따른 조기진단 시스템에서, 인공 신경망 모델은 하기의 [식 6]의 연산으로 구현되는 것을 특징으로 할 수 있다.

[0140] [식 6]

$$(\text{Input})x \rightarrow f^{(1)}(w^{(0)T}x + b^{(0)}) \rightarrow h^{(1)} \rightarrow f^{(2)}(w^{(1)T}h^{(1)} + b^{(1)}) \rightarrow h^{(2)} \rightarrow g(w^{(2)T}h^{(2)} + b^{(2)}) \rightarrow$$

[0141] $\alpha(\text{Output})$

$x = h^{(0)} \in \mathbb{R}^p$: 입력 벡터.

$w_k^{(l)}$: 가중치 벡터로 l 레이어의 모든 노드에 곱해져, $l + 1$ 레이어 k 번째 노드로 연결됨.

• $l \in \{1, \dots, L\}, k \in \{1, \dots, m_l\}$

$b_k^{(l)}$: $l + 1$ 레이어 k 번째 노드에 대한 bias.

$f^{(l)}(\cdot)$: l 은닉층에 대한 활성화함수 (Activation Function).

$h_k^{(l)} = f^{(l)}\left(\sum_{j=1}^{m_{l-1}} w_{kj}^{(l-1)} h_j^{(l-1)} + b_k^{(l-1)}\right) = f^{(l)}\left(w_k^{(l-1)T} h^{(l-1)} + b_k^{(l-1)}\right)$: l 레이어 k 번째 은닉 노드.

$g(\cdot)$: 출력함수 (Output Function).

$o_k = g\left(w_k^{(L)T} h^{(L)} + b_k^{(L)}\right)$: k 번째 출력 노드.

[0142]

[0144] 또한, 모델 구축부(400)는 선별 학습 데이터를 이용하여 딥러닝 학습을 수행하여 체장암 진단 모델을 생성할 수 있다.

[0145] 또한, 모델 구축부(400)는 심층 신경망을 둘 이상 학습시킨 뒤 이를 통합하여 최종적인 진단 결과를 출력하도록

앙상블 학습(Ensemble learning)이 가능하다.

- [0146] 구체적인 실시예로 다중 경로 일차원 합성곱 신경망, 이차원 합성곱 신경 망, 합성곱 신경망과 순환 신경망을 결합한 모델 등을 개별적으로 학습시킨 후 출 력되는 진단 결과를 수집할 수 있다. 수집된 결과 값들은 최종 병 합 모델의 입력 데이터로서 설정되고 이전 입력 스펙트럼에 대응하는 진단 결과를 목적 변수로 설 정할 수 있다.
- [0147] 앙상블 학습은 각 심층 신경망의 출력 별로 가중치를 학습시켜 최종 통합 모델이 높은 성능으로 통합된 진단 결 과를 만들어 내도록 한다. 학습된 통합 모델은 스펙트럼 데이터를 입력 받아 포함된 심층신경망들 각각의 진단 결과 값을 가중치에 따라 수집한 뒤 최종적인 진단결과를 출력한다.
- [0148] 다양한 실시예에서, 둘 이상의 심층 신경망을 각각 이용하여 둘 이상의 결과를 도출할 수 있고, 도출된 결과 를 취합하여 과반수 이상이 가리키는 결과를 최종 결과로 결정할 수 있다(hard voting).
- [0149] 다양한 실시예에서, 둘 이상의 심층 신경망을 각각 이용하여 둘 이 상의 결과(확률 값)를 도출할 수 있고, 도출 된 둘 이상의 결과의 평균을 산출하여 산출된 평균을 최종 결과로 결정할 수 있다(soft voting),
- [0150] 즉, 본 발명은 다중 경로 일차원 합성곱 신경망, 이차원 합성곱 신경망, 합성곱 신경망과 순환 신경망을 결합한 모델 중 어느 하나의 신경망을 이용하여 결과를 도출할 수 있으나, 경우에 따라 둘 이상의 신경망 모델을 조합 하여 결과를 도출할 수 있다. 예를 들어, 제어부는 각각의 신경망의 성능을 판단하여, 각각의 신경망의 성능이 기준 이하일 경우, 복수의 신경망 중 둘 이상의 신경망을 이용하는 앙상블 모델을 이용하여 결과를 도출할 수 있고, 각각의 신경망의 성능이 기준 이상일 경우, 복수의 신경망 중 어느 하나의 신경망을 이용하여 결과를 도 출할 수 있다.
- [0152] **2. 췌장암 진단 모델 생성 및 실험예**
- [0153] 진단부(500)는 췌장암 진단 모델을 전달받아, 진단부(500)에 환자의 임상 정보와 환자의 라만 스펙트럼 정보 중 하나 이상을 입력하면, 출력부를 통해 췌장암 유무가출력될 수 있다.
- [0154] 또한, 췌장암 진단 모델은 환자 임상 정보 중 췌장암 진행 단계를 함께 학습하여, 췌장암 진행 단계를 출력할 수도 있다.
- [0156] 도 8을 참조하여, 본 발명에 따라 생성된 췌장암 진단 모델의 정확도를 설명한다.
- [0157] 도 8은 총 80 샘플의 라만 스펙트럼 측정(췌관선암 40명, 담석증40명)한 결과를 도시하며, 1기 9명, 2기 초반 10명, 2기 후반 21명이고, 아산병원에서 전달 받은 수술 전 CA19-9와 CEA 수치의 췌장암 진단 성능과 비교하였 다.
- [0158] 본 발명에 따라 나노 큐브 기관(10)로 측정한 라만스펙트럼을 PCA로 차원감소 후 ANN (artificial neural network)을 이용하여 학습된 모델에 따르면, AUROC가 0.969, Accuracy가 0.900, Sensitivity가 0.825, Specificity가 0.975로 도시된다.
- [0159] 또한, 본 발명에 따라 나노 큐브 기관(10)로 측정한 라만스펙트럼을 PCA로 차원감소 후 획득된 변수와, ELISA로 측정한 CA19-9, CEA정보를 변수로 추가하여 췌장암 진단하는 ANN (artificial neural network)을 이용하여 학 습된 모델에 따르면, AUROC가 0.957, Accuracy가 0.913, Sensitivity가 0.85, Specificity가 0.975로 도시된다.
- [0160] 또한, 본 발명에 따라 나노 큐브 기관(10)로 측정한 라만스펙트럼을 DNN(deep neural network)를 이용하여 학습 된 모델에 따르면, AUROC가 0.944, Accuracy가 0.938, Sensitivity가 0.900, Specificity가 0.975로 도시된다.
- [0161] 이는, 종래의 ELISA로 측정한 CEA에 따른 AUROC가 0.600, Accuracy가 0.600, Sensitivity가 0.250, Specificity가 0.950와,
- [0162] 또한, 종래의 ELISA로 측정한 CA19-19의 AUROC가 0.762, Accuracy가 0. 762, Sensitivity가 0.575, Specificity가 0.950 보다 AUROC가, Accuracy, Sensitivity, Specificity 측면에서 모두 높은 값을 보이는 것 으로 도시되며, 이에 따라 본 발명에 따른 진단모델의 췌장암 진단의 정확도가 높음을 알 수 있다.

[0164] **3. 채장암 진단 모델 추가 학습**

[0165] 또한, 본 발명에 따른 채장암 진단 모델은 추가 학습부(600)에 의해 추가 학습될 수 있다.

[0166] 도 9 내지 도 12를 참조하여, 추가 학습부(600)가 추가 학습 데이터를 이용하여 채장암 진단 모델을 추가 학습시키는 것을 설명한다.

[0167] 추가 학습부(600)가 라만 스펙트럼 각각의 원본 신호에 잡음 생성(Jittering), 신호 크기 조정(Scaling), 신호 회전(Rotation), 구간 섞기(Permutation), 뒤틀림 추가(Magnitude warping), 선형 변환(Linear transformation), 시프팅(Shifting) 중 적어도 하나의 방법을 적용하여 추가 학습 데이터를 생성할 수 있다.

[0168] "잡음 생성(Jittering)"이란, 원본 신호에 가우시안 분포(Gaussian distribution)를 기반으로 한 잡음 신호를 추가하는 것을 의미한다.

[0169] "신호 크기 조정(Scaling)"이란, 원본 신호에 무작위한 실수 값을 곱해주는 것을 의미한다.

[0170] "신호 회전(Rotation)"이란, 라만 스펙트럼 내 무작위로 선택된 한 점을 기준으로 기설정된 각도만큼 회전시킨다. 가령, 기설정된 각도는 -10~10도일 수 있다. 원본 데이터에 회전변환행렬
$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$
 을 곱한 결과를 사용하고, 이때 θ 는 회전할 라디안 값을 의미한다.

[0171] "구간 섞기(Permutation)"란 원본 신호를 무작위한 개수의 조각으로 분할한 뒤 섞어서 다시 조합하는 것을 의미한다.

[0172] "뒤틀림 추가(Magnitude warping)"란, 무작위한 선분을 신호에 더해주는 것을 의미한다.

[0173] "선형 변환(Linear transformation)"이란, 원본 수식은 $y=a*x+b$ 형태로 변환하나 범위 설정을 위해 $y= a*n+b$ 형태의 선분을 원본 스펙트럼에 더하는 것을 의미한다 (이때 n은 wave number).

[0174] "시프팅(Shifting)"이란, 스펙트럼을 -2칸에서 2칸 사이로 랜덤하게 x축 이동시키는 것을 의미하고, 이때 데이터 범위를 벗어나는 데이터는 삭제되며, 부족한 부분은 가장 가까운 데이터를 이용하여 채워진다.

[0175] 이와 같이, 원본 신호를 다양한 방법으로 변화시켜 복수의 가상 라만 스펙트럼을 생성하고, 이를 추가 학습 데이터로 포함 하여 채장암 진단 모델을 지속적으로 학습시킴으로써, 채장암 진단 모델은 잡음이 제거되지 않은 원본의 신호를 입력받더라도 입력된 원본 신호에서 잡음을 무시하고 분석을 수행할 수 있고, 이를 통해, 잡음을 제어하는 과정 없이도 보다 정확한 진단 결과를 도출할 수 있다.

[0177] 도 12를 참조하여, 추가 학습부(600)가 라만 스펙트럼 정보의 기준선을 변경하여 적용하여 다른 추가 학습 데이터를 생성할 수 있다.

[0178] 추가 학습부(600)는 라만 스펙트럼에서 추측된 기준선을 활용하여, 가상의 라만 스펙트럼을 생성하고 이를 이용하여 추가 학습한다.

[0179] 추가 학습부(600)는 라만 스펙트럼의 원본 스펙트럼에서 전술한 기준선 추정방법으로 얻어진 기준선과, 이러한 기준선을 제거한 여과 신호를 수집한다.

[0180] 이후, 추가 학습부(600)는 기준선을 무작위로 뒤섞어 다른 라만 스펙트럼에서 얻어진 기준선을 여과 신호에 더해준다.

[0181] 도 12(a)에서 원본 스펙트럼이 도시되며, 기준선이 도시된다. 도 12(b)는 도 12(a)의 여과 신호이다. 도 12(a)와 도 12(b)의 상단의 좌우 스펙트럼들의 기준선을 교체할 경우 두 개 스펙트럼이 새롭게 생성되고, 새롭게 생성된 하단의 도 12(c), 도 12(d)에서 기준선이 교체된 스펙트럼이 도시된다.

[0182] 이에 따라, 추가 학습부(600)는 새로운 가상의 라만 스펙트럼을 인공적으로 만들 수 있다.

[0183] 이와 같이 여과 신호에 기준선을 더해줌으로써 생성된 가상의 라만 스펙트럼을 인공지능 모델의 학습 데이터로 하여 지속적으로 학습시킴으로써, 추후 채장암 진단 모델은 기준선이 제거되지 않은 원본의 스펙트럼을 입력받

더라도, 입력된 원본의 스펙트럼에서 기준선을 무시하고 분석을 수행할 수 있고, 이를 통해 기준선을 제거하는 과정 없이도 보다 정확한 진단 결과를 도출할 수 있다

[0185] **4. 나노 큐브 기관(10)의 구조/ 검증 실험/ 임상 샘플 측정**

[0186] 본 발명의 제1 실시예에 따라, SSFA 바이오마커 수집부(120)에서 학습 데이터를 수집하기 위해 사용되는 나노 큐브 기관(10)에 대해 보다 구체적으로 설명한다.

[0187] 혈액 검체의 경우 다양한 핵산(nucleotide), 단백질, 저분자(small molecule)등 수많은 생체분자들이 섞여 있다. 검체의 스펙트럼 자체를 바이오마커로 사용하기 위해서는 검체 내의 기계학습 검지의 특이도(specificity)와 민감도(sensitivity)와 관계되어 있는 분자들을 라만 활성 부위에 흡착시키는 것이 중요하다.

[0188] 혈액 검체 안에 존재하는 분자들의 라만 스펙트럼(spectrum)을 얻기 위해서는 신호대잡음비(signal-to-noise ratio, S/N ratio)를 높이는 과정과 재현성을 위해 강하고 일정한 세기의 전기장이 필수적이며, 이를 위해 혈액 검체에 최적화된 새로운 금속 라만 기관이 요구된다.

[0190] 본 발명에서는 임상샘플을 측정하기 전 나노 큐브 기관(10)의 광학적 특성을 로다민 분자(rhodamine 6G)를 통해 확인하였다.

[0191] 일 실시예에서는 0.1 mM 로다민 용액 0.5 μ L를 스팟팅(spotting)한 뒤, 건조하고 포인트 매핑(point mapping)을 통해 라만 신호가 발생하는 위치와 그 세기 차이에 대해서 확인하였다. SERS 기관으로서 효용성이 확인된 나노 큐브 기관(10)은 후술하는 바와 같이 임상샘플을 진행한다.

[0192] 일 실시예에서는 혈장(blood plasma)을 10배 묽히고, 나노 큐브 기관(10)에 5 μ L씩 샘플링한 뒤 30분간 생분자(biomolecules)들의 흡착을 기다린 뒤 진공 상태에서 건조시켰다. 이후 시료 내 분자들이 흡착된 임상샘플의 라만 스펙트럼을 측정하였다.

[0194] **나노 큐브 기관(10)의 구조와 나노 큐브 기관(10) 검증실험**

[0195] 먼저, 체장암 진단의 목적을 위하여 나노 큐브 기관(10)이 혈장 스펙트럼 획득에 사용될 수 있는지 확인하였다. 다만, 혈장 내의 분자들은 라만세기가 크지 않으므로, 라만 관측을 용이하게 하기 위하여 신호가 큰 로다민(rhodamine 6g) 분자를 먼저 사용하여 증강된 라만신호가 얻어졌음을 확인한 후, 임상샘플의 스펙트럼을 획득하였다.

[0196] 본 발명에서는 나노 겹(11)이 형성되는 나노 큐브 기관(10)을 통해 라만 스펙트럼을 획득할 수 있다.

[0197] 도 13에서는 나노 큐브 기관(10)을 도시한다.

[0198] 도 13(a)에서는 기설정된 간격의 나노 겹(11)이 형성된 나노 큐브 기관(10)을 도시한다.

[0199] 나노 큐브 기관(10)은 금 나노큐브 어레이 구조체를 형성시켜 나노 겹(11)을 만들었다. 나노 겹(11)에서 라만 증강이 발생한다고 알려져 있기 때문에 이와 같은 구조체를 SERS 기관으로 선정하였다.

[0200] 나노입자에는 표면안정제 외에 다양한 화학물질들이 흡착되어 있다. 이러한 화학물질들은 분석물의 흡착 방해, 불필요한 신호 중첩 등의 이유로 핫스팟을 막고 라만 신호를 방해한다. 이러한 물질들은 산소 플라즈마를 통해서 제거하였고, 라만 활성 분자를 통해 SERS 기관으로서 분광학적 효용성 확인을 하였다.

[0201] 도 13(b)는 나노 큐브 기관(10)의 주사 전자현미경(SEM) 이미지를 도시한다.

[0202] 도 13(c)는 나노 큐브 기관(10)의 효용성 확인을 위해 0.1 mM 로다민 분자 매핑 결과 1509.46 cm^{-1} 위치에서 라만 세기를 초록색으로 표시하여 광학 현미경 이미지에 투사한 이미지를 도시한다.

[0203] 핫스팟 노출 전후의 로다민(rhodamine 6g) 분자의 라만 세기를 비교하여 화학물질 제거 여부와 체장암 진단용 SERS 기관으로서의 효용성을 확인하였다.

[0204] 나노 큐브 기관(10)에는 핫스팟이 형성되어있는 영역(노란색)과 그렇지 않은 영역(검정색)이 함께 존재하였고, 핫스팟이 형성된 영역에서만 라만 신호가 나오는 것을 확인하기 위해 SERS 기관을 포인트 매핑(point mapping)

을 하였다.

[0205] 매핑이 진행된 각 위치별 스펙트럼 내 1509.4cm^{-1} 부근의 면적의 크기를 초록색으로 나타내었고, 광학 현미경 이미지에 투사하여 도 13 (c)와 같은 이미지를 만들었다. 도 13(c)에서 나타나있는 바와 같이 핫스팟이 형성된 영역의 모양으로 초록색이 생겼고 이를 통해서 핫스팟이 형성된 영역에서만 로다민 분자의 라만 신호가 발생하였음을 알 수 있었다.

[0207] 도 13(d)는 핫스팟(hotspot) 노출 전을 하측에 위치되는 그래프인 빨간 색으로, 노출 후를 상측에 위치되는 파란색으로 도시되며, 그래프의 평균 세기(실선)과 표준편차(음영)를 도시한다.

[0208] 핫스팟의 노출 전후에 따른 로다민 분자의 라만 세기 변화를 확인하기 위해 라만 매핑을 통해 획득한 라만 스펙트럼을 도 13 (d)에 나타내었다. 로다민 분자의 대표적인 피크들(1311.74 , 1363.91 , 1509.46 , 1651.76cm^{-1})에서 핫스팟 노출 후 최소 13배에서 최대 28배까지 증가하였고, 평균 상대표준편차(relative standard deviation)은 38.3 %였다.

[0209] 기관 내에 나노구조체가 존재하지 않고 폴리머만 존재하는 마이크로 크랙(micro crack)에 의한 영향을 고려한다면 라만증강이 일어나는 곳에서의 신호는 더욱 균일하게 나올 것이고, 본 발명에 따른 나노 큐브 기관(10)이 체장암 진단용 SERS 기관으로서 기능이 가능하다는 것을 확인하였다.

[0211] 나노 큐브 기관(10)을 이용한 임상샘플 측정

[0212] 나노 큐브 기관(10)을 통해 혈장 임상샘플의 스펙트럼을 획득할 수 있는지 알아보기 위한 연구를 진행하였다.

[0213] 혈장을 바로 사용할 경우 기관에 습윤(wetting)되기 전 응고가 되어 기관에서 라만 신호가 발생하지 않아, 혈장을 희석시켜 사용한다.

[0214] 이 때, 혈장을 10배 묽혀서 사용하였으나, 해당 수치에 제한되는 것은 아니다.

[0215] 임상샘플을 기관에 샘플링한 뒤 기설정된 시간 동안 건조시켜, 생분자 흡착이 발생하도록 기다린다.

[0216] 이 때, 임상샘플을 기관에 샘플링한 뒤 30분간 건조시키나, 해당 수치에 제한되는 것은 아니나, 혈장은 상온 보관의 시간이 길어질수록 대사반응이 진행되어 hypoxanthine와 같은 분자의 농도가 증가한다고 알려져있고, 해당 분자는 스펙트럼 정보로서 관측이 가능하고 혈장 정보의 왜곡을 가져올 수 있다. 이와 같은 이유로 샘플링의 시간은 한시간을 넘기지 않을 수 있다.

[0218] 혈장 내 분자들이 흡착이 잘 되는지 확인하기 위해 혈장을 샘플링한 기관과 그렇지 않은 기관의 라만 스펙트럼을 측정한다.

[0219] 혈장이 샘플링되지 않은 경우 도 15의 검정색 실선(하측 실선)과 같은 스펙트럼을 얻었고, 혈장이 샘플링된 기관에서는 도 15의 붉은색 실선(상측 실선) 및 음영과 같은 스펙트럼을 얻었다.

[0220] 두 스펙트럼 간에는 다양한 위치에서 차이를 보였다. 혈장 내 분자를 특정할 수는 없지만, 대략 $800-900\text{cm}^{-1}$, $1400-1500\text{cm}^{-1}$ 은 DNA, lipids, proteins, 1000cm^{-1} 은 phenylalanine, $1100-1200\text{cm}^{-1}$, $1550-1650\text{cm}^{-1}$ 은 proteins, $1300-1400\text{cm}^{-1}$ 은 lipid 및 protein, 1500cm^{-1} 은 DNA에 해당되는 정보로 알려져 있다. 이에 해당하지 않는 영역에서 발생한 신호는 대사체, 잔류 세포, 소낭과 같은 분자에서 발생했을 것으로 추측된다.

[0221] 따라서, 나노 큐브 기관(10)을 이용하면 특정 조건 하에서 혈장 스펙트럼이 측정됨을 보였고 전술한 인공지능모델을 생성하는 학습 데이터로 사용되고, 인공지능 모델에 입력되는 입력 데이터로 사용되어 체장암 진단에 사용될 수 있다.

[0222] 본 발명에서는, 생체 신호로부터 사람 눈에는 보이지 않는 정보를 추출할 수 있다는 점을 고려하여, 세계 최초로 표면증강 라만 분광법과 인공지능을 결합한 진단 기술을 개발하여 기존 액체 생검의 문제였던 낮은 정확도를 보완할 수 있다

[0223] 빛의 산란정도를 이용해 혈액 내 고유한 화학성분 정보를 한번에 획득할 수 있고, 인공지능을 이용하여 복잡한 라만신호로부터 사람 눈에 보이지 않는 정보를 자동으로 추출할 수 있다

[0224] 이상, 본 명세서에는 본 발명을 당업자가 용이하게 이해하고 재현할 수 있도록 도면에 도시한 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 당업자라면 본 발명의 실시예로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 보호범위는 특허청구범위에 의해서 정해져야 할 것이다.

부호의 설명

[0225] 10: 나노 큐브 기관

11: 나노 갭

100: 학습 데이터 수집부

110: 입력부

120: SSFA 바이오마커 수집부

130: 바이오마커 농도 수집부

200: 데이터 전처리부

300: 변수 선별부

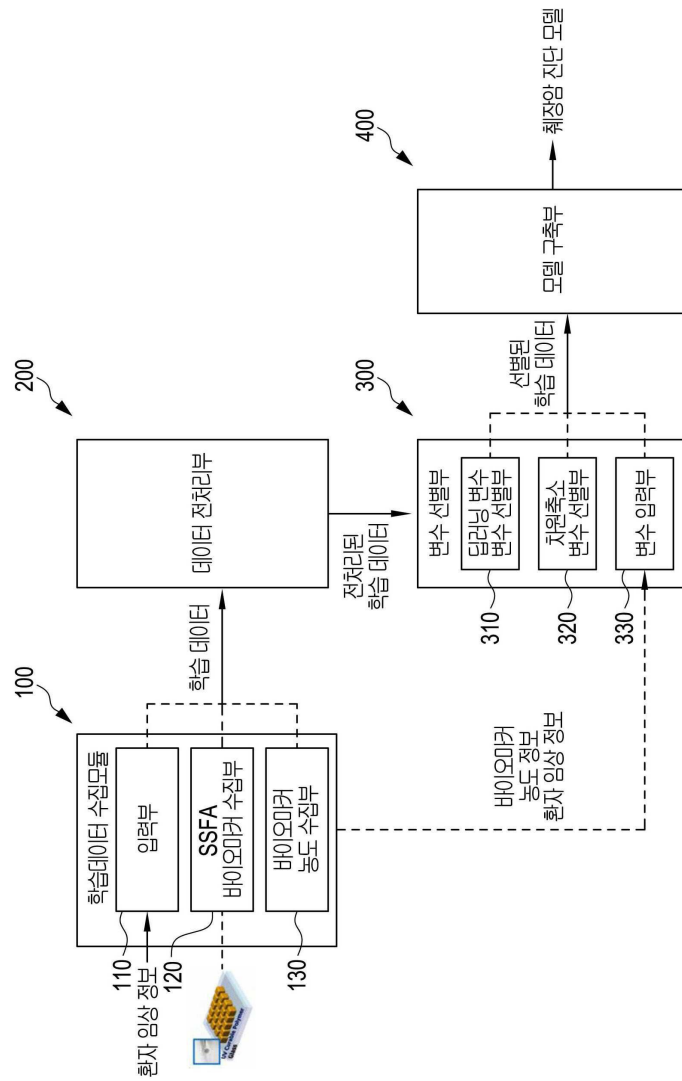
400: 모델 구축부

500: 진단부

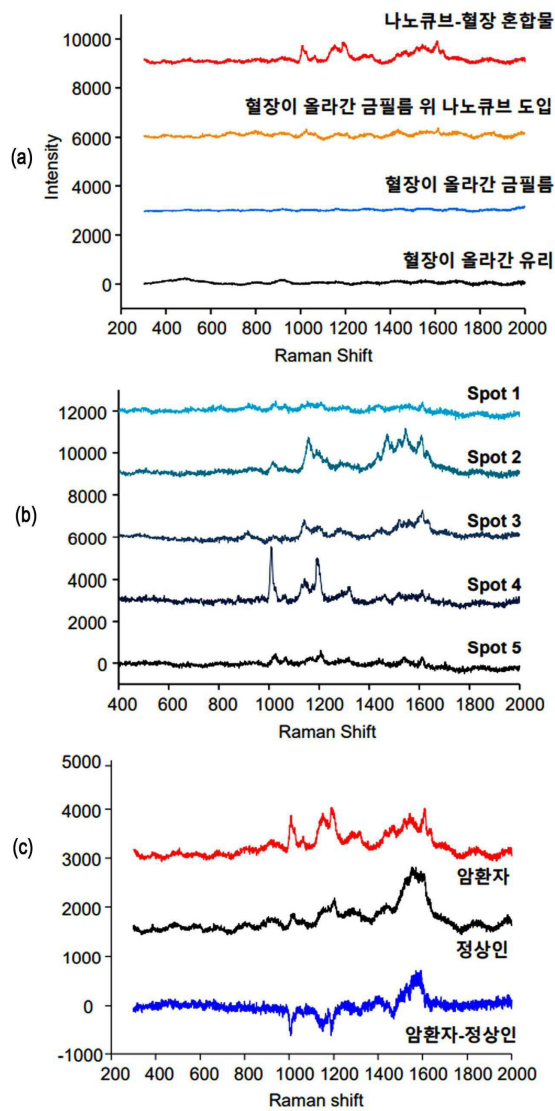
600: 추가 학습부

도면

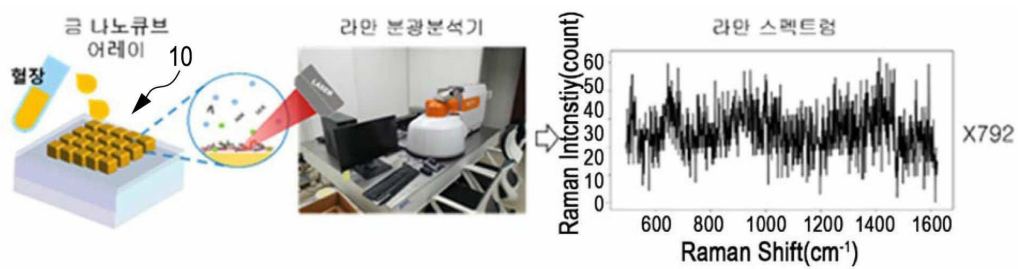
도면1



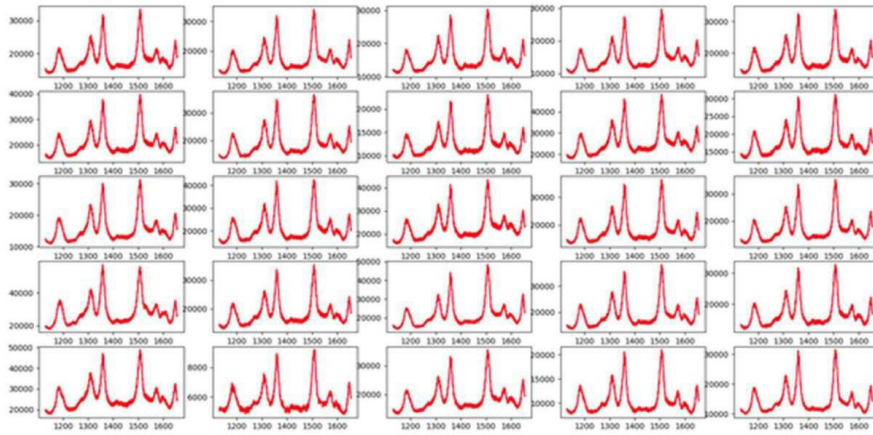
도면2



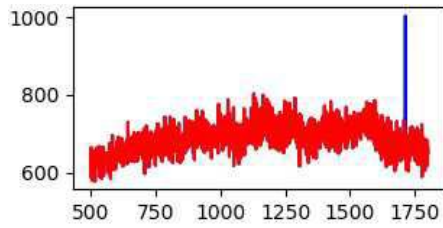
도면3



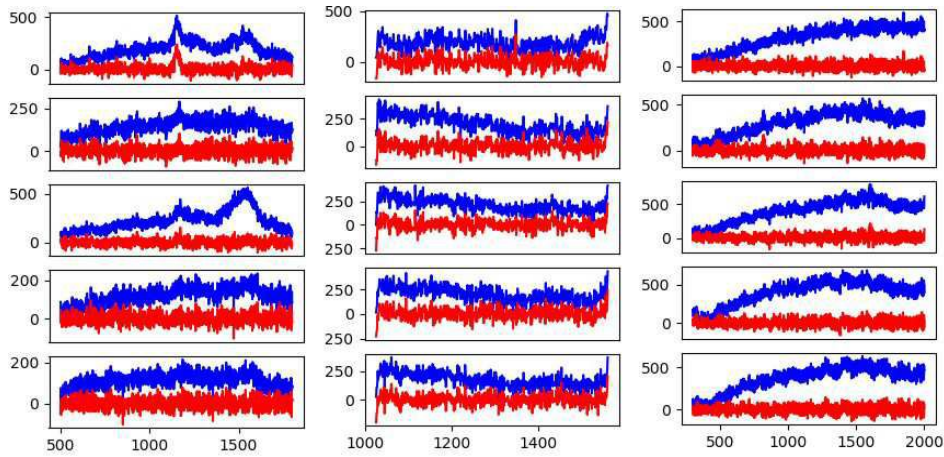
도면4



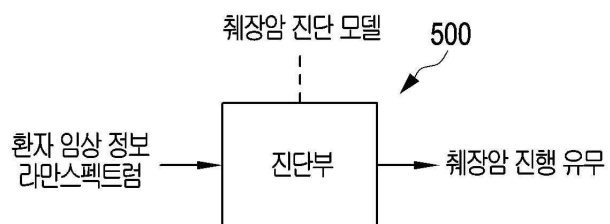
도면5



도면6



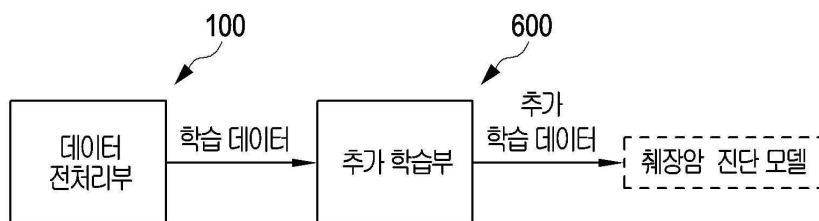
도면7



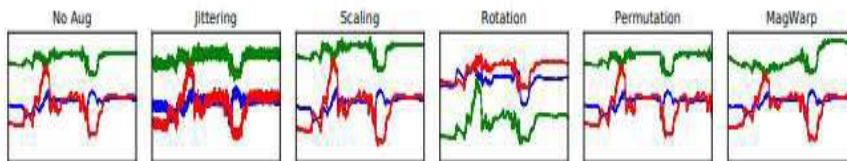
도면8

구분	AUROC	Accuracy	Sensitivity	Specificity
CEA (>5 ng/mL)	0.600	0.600	0.250	0.950
CA19-9 (>37 U/ml)	0.762	0.762	0.575	0.950
PCA-ANN	0.969	0.900	0.825	0.975
PCA+ CEA +CA19-9- ANN	0.957	0.913	0.85	0.975
DNN	0.944	0.938	0.900	0.975

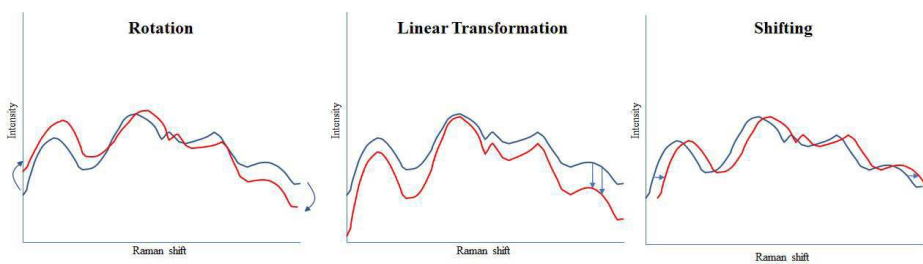
도면9



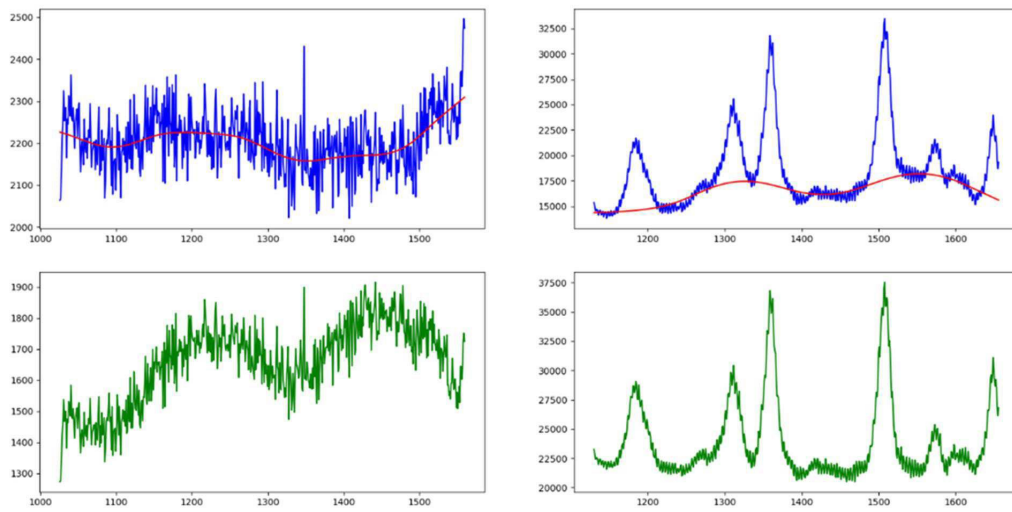
도면10



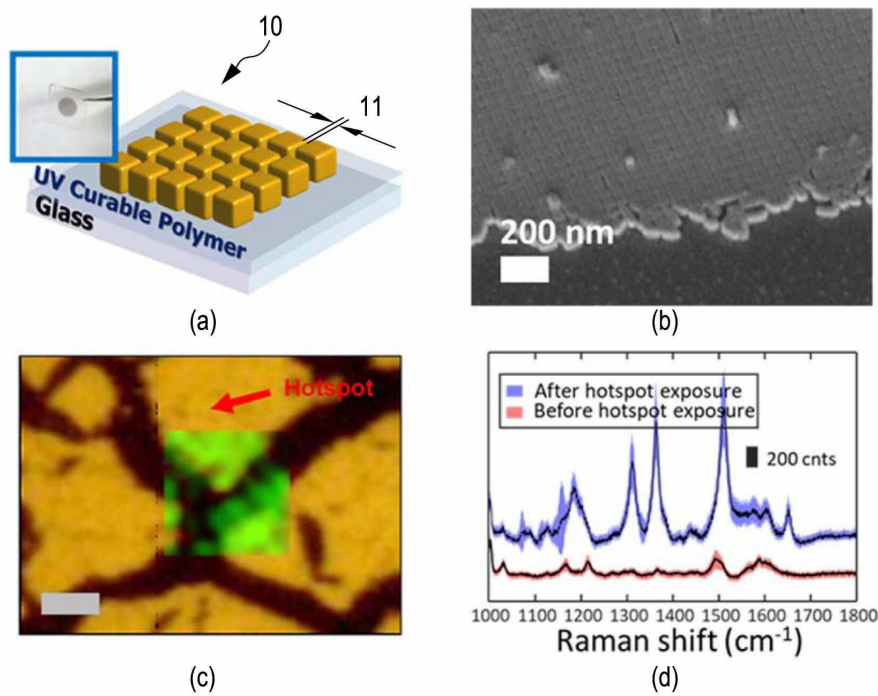
도면11



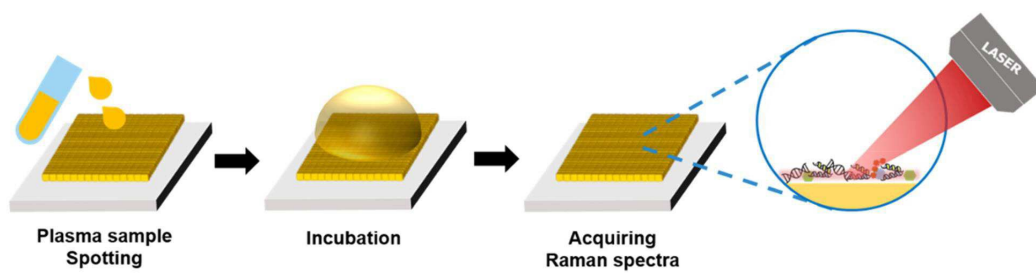
도면12



도면13



도면14



도면15

