(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0022314
(43) 공개일자 2021년03월03일

(51) 국제특허분류(Int. Cl.)
G16C 20/00 (2019.01) *G06N 3/02* (2019.01)
(52) CPC특허분류
G16C 20/00 (2019.02)
G06N 3/02 (2019.01)
(21) 출원번호 10-2019-0101654
(22) 출원일자 2019년08월20일
심사청구일자 2019년08월20일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
유창규
경기도 수원시 영통구 웰빙타운로 20, 8308동 202호(이의동, 호반가든하임)
허성구
경기도 용인시 기흥구 덕영대로 1732(서천동)
(74) 대리인
특허법인도담

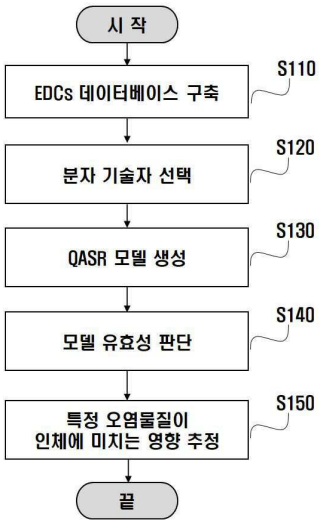
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 **내분비 교란 물질의 인체 위해성 평가를 위한 심층인공신경망 기반 계산독성학 모델 생성 방법 및 장치**

(57) 요 약

본 명세서는 내분비 교란 위해성을 가지는 화학 물질을 선정하는 단계; 상기 화학 물질을 나타내는 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하는 단계; 및 상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법을 개시한다. 일 실시 예에서 사용자는 화학물질의 분자표현자 중 부분최소자승법 기반의 VIP score와 LASSO 회귀 방법을 이용해 중요 분자표현자를 선택하여 심층 인공신경망 기반 계산독성학 모델을 개발함으로써 내분비교란물질의 인체위해성을 효율적으로 평가할 수 있다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2017R1E1A1A03070713
부처명	한국연구재단
과제관리(전문)기관명	한국연구재단
연구사업명	전략과제(산업수학)
연구과제명	환경정보 빅 데이타와 시스템 최적화 기법을 이용한 스마트 통합환경관리시스템개발
기 여 율	1/2
과제수행기관명	한국연구재단
연구기간	2017.09.01 ~ 2021.08.31

이 발명을 지원한 국가연구개발사업

과제고유번호	19QPPW-B152307-01
부처명	국토교통부
과제관리(전문)기관명	국토교통과학기술진흥원
연구사업명	지하철 미세먼지 저감 기술개발 사업
연구과제명	인공지능 기반 지하철 이용객의 (초)미세먼지 노출 저감기술 개발
기 여 율	1/2
과제수행기관명	한국철도기술연구원
연구기간	2019.04.25 ~ 2023.12.31

명 세 서

청구범위

청구항 1

위해성 평가 모델 생성 장치에 의하여 수행되는 위해성 평가모델 생성 방법에 있어서,

내분비 교란 위해성을 가지는 화학 물질을 선정하는 단계;

상기 화학 물질을 나타내는 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하는 단계; 및

상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 2

제 1 항에 있어서,

아래의 수학식을 이용하여 계산된 값이 미리 설정된 범위내에 있는 물질은 내분비 교란 위해성을 가지는 화학물질로 선정되는 위해성 평가 모델 생성 방법.

$$logRBA = log\left(\frac{IC_{50}^{C}}{IC_{50}} \times 100\right)$$

여기서, IC_{50}은 생물학적 검정에서 표적 수용체에 대한 결합의 50% 억제를 생성하는 상기 화학 물질의 농도이고,

IC_{50}^{C}은 기준 내인성 리간드의 IC_{50}이다.

청구항 3

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 생성하는 단계는,

상기 분자 표현자 집합에 속한 분자 표현자 마다 평가 점수를 부여하는 단계; 및

상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 4

제 3 항에 있어서,

상기 평가 점수는 아래의 수학식을 이용하여 계산되는 위해성 평가 모델 생성 방법.

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)(w_{jk} / \left\| w_{jk} \right\|)^2}{\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)}}$$

여기서, p는 x 변수의 개수이고, i는 선택된 잠재 변수(latent variable)의 개수이고, w는 가중치이며, $q_k^2 \cdot t_k^{'} \cdot t_k$ 는 각 컴포넌트에 의하여 해석되는 차이값이며, $(w_{jk} / \left\| w_{jk} \right\|)^2$ 는 j번째 변수의 중요도를 나타낸다.

청구항 5

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 생성하는 단계는,

상기 분자 표현자 집합의 분자 표현자에 LASSO 회귀 분석을 적용한 결과 파라미터 값이 0이 된 분자 표현자를 제외함으로써 축소된 분자 표현자 집합을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 6

제 5 항에 있어서, LASSO 회귀 분석은 아래의 수학식을 이용하여 수행되는 위해성 평가 모델 생성 방법.

$$\hat{\beta}^{LASSO} = \underset{\hat{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

여기서, $\hat{\beta}^{LASSO}$ 는 LASSO 회귀의 추정된 파라미터 벡터이고, y는 반응 변수이고, x는 예측 변수이고, n은 반응 변수의 개수이고, p는 예측 변수의 개수이다.

청구항 7

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 생성하는 단계는,

상기 분자 표현자 집합에 대하여 각각의 분자 표현자를 독립변수로, 상관 결합 친화도를 종속변수로 하여, 독립변수와 종속변수 간의 선형 상관 관계를 계산하는 단계; 및

상기 선형 상관 관계의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 8

제 7 항에 있어서,

상기 독립변수와 종속변수 간의 선형 상관 관계는 아래의 수학식을 이용하여 계산되는 위해성 평가 모델 생성 방법.

$$C_{pearson}(x,y)=\frac{\sum_{i=1}^{N}(x_i-m_x)(y_i-m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i-m_x)^2][\sum_{i=1}^{N}(y_i-m_y)^2]}}$$

여기서, x는 독립변수, y는 종속변수, $m_x=\frac{1}{N}\sum_{n=1}^{N}x_n$,

$m_y=\frac{1}{N}\sum_{n=1}^{N}y_n$ 이다.

청구항 9

제 1 항에 있어서,

상기 학습 모델은 SAE(stacked autoencoder), DBN(deep-belief network) 및 DNN(deep neural network) 중 어느 하나의 인공신경망 알고리즘을 이용하여 상기 축소된 분자 표현자 집합에 대하여 생성된 정량적 구조-활동 관계(Quantitative structure-activity relationship) 모델인 위해성 평가 모델 생성 방법.

청구항 10

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 단계는,

고정확률 p에 따라 학습 모델 네트워크를 구성하는 유닛을 무작위로 드롭하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 11

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 생성하는 단계는,

상기 화학 물질을 나타내는 분자 표현자 집합에 대하여, 부분최소자승법 기반의 VIP(variable importance in projection) 점수를 분자 표현자 마다 부여하는 단계;

상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하는 단계; 및

상기 분자 표현자 집합에 남은 분자 표현자에 LASSO 회귀 분석을 적용한 결과 파라미터의 값이 미리 설정된 값

으로 감소된 분자 표현자를 더 제외하여 축소된 분자 표현자 집합을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 12

제 1 항에 있어서,

상기 축소된 분자 표현자 집합을 생성하는 단계는,

상기 화학 물질을 나타내는 분자 표현자 집합에 대하여 각각의 분자 표현자를 독립변수로, 상관 결합 친화도를 종속변수로 하여, 독립변수와 종속변수 간의 선형 상관 관계를 계산하는 단계;

상기 선형 상관 관계의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하는 단계;

상기 분자 표현자 집합에 남은 분자 표현자에 대하여, 부분최소자승법 기반의 VIP(variable importance in projection) 점수를 분자 표현자 마다 부여하는 단계; 및

상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함하는 위해성 평가 모델 생성 방법.

청구항 13

위해성 평가 장치에 의하여 수행되는 위해성 평가 방법에 있어서,

화학 물질을 나타내는 분자 표현자 집합을 입력 받는 단계;

상기 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자를 생성하는 단계; 및

상기 축소된 분자 표현자와 제 1 항 내지 제 12 항 중 어느 한 항에 기재된 위해성 평가 모델 생성 방법에 따라 생성된 위해성 평가 모델을 이용하여 상기 화학 물질의 내분비 교란 위해성을 평가하는 단계를 포함하는 화학 물질의 내분비 교란 위해성 평가 방법.

청구항 14

화학물질의 내분비 교란 위해성 데이터가 저장된 메모리; 및

상기 데이터베이스로부터 내분비 교란 위해성을 가지는 화학 물질을 선정하고,

상기 화학 물질을 나타내는 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하며,

상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 프로세서를 포함하는 위해성 평가 모델 생성 장치.

청구항 15

화학 물질을 나타내는 분자 표현자 집합이 저장된 메모리; 및

상기 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하고,

상기 축소된 분자 표현자 집합과 제 1 항 내지 제 12 항 중 어느 한 항에 기재된 위해성 평가 모델 생성 방법에 따라 생성된 위해성 평가 모델을 이용하여 상기 화학 물질의 내분비 교란 위해성을 평가하는 프로세서를 포함하는 화학 물질의 내분비 교란 위해성 평가 장치.

청구항 16

제 1 항 내지 제 12 항 중 어느 한 항에 기재된 위해성 평가 모델 생성 방법을 수행하기 위한 컴퓨터 프로그램.

청구항 17

제 1 항 내지 제 12 항 중 어느 한 항에 기재된 위해성 평가 모델 생성 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터 판독 가능 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 내분비 교란 물질의 인체 위해성 평가를 위한 계산독성학 모델을 생성하기 위한 방법, 장치, 컴퓨터 프로그램 및 컴퓨터 판독 가능 기록매체에 관한 것이다.

배 경 기 술

[0002] 도 1a에 도시된 바와 같이 사람이 만들어낸 EDC(endocrine-disrupting chemicals)는 제약, 개인 관리 제품, 플라스틱 파편, 산업 화학 물질, 살충제, 연소 부산물 및 계면 활성제와 같은 형태로 하수로 배출된다. 이 EDC는 물고기에 의해 섭취되거나 흡수되어 먹이사슬에 축적된다. 따라서 EDC는 인간의 건강, 특히 내분비 계에 악영향을 미칠 수 있다.

[0003] EDC(endocrine-disrupting chemicals)는 인간 내분비계에 악영향을 미치는 외인성 화합물이다. 내분비 계는 뇌하수체, 송과선, 갑상선, 부갑상선 및 부신과 같은 호르몬 방출 기관을 일컫는 일반적인 용어이다. 내분비 계의 활동은 호르몬 합성, 호르몬 분비선으로부터의 호르몬 분비, 단백질을 수송하기 위한 호르몬 결합과 표적 조직으로의 호르몬 수송을 포함한다. 이러한 과정을 통해 내분비 시스템은 다양한 유기 기능을 조절할 수 있다. 살아있는 유기체의 항상성, 생식, 대사, 발달 및 행동을 유지하는 것은 내분비 계의 중요한 역할이다.

[0004] 그러나 EDC는 호르몬의 메커니즘을 모방함으로써 호르몬의 역할을 방해한다. EDC는 시상 하부-뇌하수체-갑상선 축선 기능, 호르몬 합성, 호르몬 수송 단백질, 호르몬 수용체 및 호르몬 대사에 영향을 줄 수 있다고 보고된 바 있다.

[0005] 합성 및 천연 화학 화합물의 매우 큰 집합체는 EDC 인 것으로 보고 되었다. 그럼에도 불구하고 EDC의 활동을 보여주는 화학 물질의 정의와 범위를 명확히 하는 것은 매우 중요하다. 따라서, 다양한 EDC의 생물학적 효과를 평가하기 위해 실험적이지 않은 모델을 개발하는 것이 시급하다.

발명의 내용

해결하려는 과제

[0007] 본 명세서에서는 심층인공신경망을 이용하여 화학물질이 인체 내분비계에 미치는 영향을 평가할 수 있는 있는 효율적인 방안을 제시하고자 한다.

과제의 해결 수단

[0009] 상기한 과제를 해결하기 위한 일 실시예에 따른 위해성 평가 모델 생성 장치에 의하여 수행되는 위해성 평가모델 생성 방법은 내분비 교란 위해성을 가지는 화학 물질을 선정하는 단계; 상기 화학 물질을 나타내는 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하는 단계; 및 상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 단계를 포함한다.

[0010] 아래의 수학식을 이용하여 계산된 값이 미리 설정된 범위내에 있는 물질은 내분비 교란 위해성을 가지는 화학물질로 선정될 수 있다.

$$logRBA = log\left(\frac{IC_{50}^{C}}{IC_{50}} \times 100\right)$$

[0011]

[0012] 여기서, IC_{50}은 생물학적 검정에서 표적 수용체에 대한 결합의 50% 억제를 생성하는 상기 화학 물질의 농도이고, IC^{C}_{50}은 기준 내인성 리간드의 IC_{50}이다.

[0013] 상기 축소된 분자 표현자 집합을 생성하는 단계는, 상기 분자 표현자 집합에 속한 분자 표현자 마다 평가 점수를 부여하는 단계; 및 상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함할 수 있다.

[0014] 상기 평가 점수는 아래의 수학식을 이용하여 계산될 수 있다.

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)(w_{jk} / \|w_{jk}\|)^2}{\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)}}$$

[0015]

[0016] 여기서, p는 x 변수의 개수이고, i는 선택된 잠재 변수(latent variable)의 개수이고, w는 가중치이며, $q_k^2 \cdot t_k^{'} \cdot t_k$는 각 컴포넌트에 의하여 해석되는 차이값이며, $(w_{jk} / \|w_{jk}\|)^2$는 j번째 변수의 중요도를 나타낸다.

[0017] 상기 축소된 분자 표현자 집합을 생성하는 단계는, 상기 분자 표현자 집합의 분자 표현자에 LASSO 회귀 분석을 적용한 결과 파라미터 값이 0이 된 분자 표현자를 제외함으로써 축소된 분자 표현자 집합을 생성하는 단계를 포함할 수 있다.

[0018] 상기 LASSO 회귀 분석은 아래의 수학식을 이용하여 수행될 수 있다.

$$\hat{\beta}^{LASSO} = \underset{\hat{\beta}}{\operatorname{argmin}}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right)$$

[0019]

[0020] 여기서, $\hat{\beta}^{LASSO}$는 LASSO 회귀의 추정된 파라미터 벡터이고, y는 반응 변수이고, x는 예측 변수이고, n은 반응 변수의 개수이고, p는 예측 변수의 개수이다.

[0021] 상기 축소된 분자 표현자 집합을 생성하는 단계는, 상기 분자 표현자 집합에 대하여 각각의 분자 표현자를 독립변수로, 상관 결합 친화도를 종속변수로 하여, 독립변수와 종속변수 간의 선형 상관 관계를 계산하는 단계; 및 상기 선형 상관 관계의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함할 수 있다.

[0022] 상기 독립변수와 종속변수 간의 선형 상관 관계는 아래의 수학식을 이용하여 계산될 수 있다.

[0023]
$$C_{pearson}(x,y)=\frac{\sum_{i=1}^{N}(x_i-m_x)(y_i-m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i-m_x)^2][\sum_{i=1}^{N}(y_i-m_y)^2]}}$$

[0024] 여기서, x는 독립변수, y는 종속변수, $m_x=\frac{1}{N}\sum_{n=1}^{N}x_n$,

$m_y=\frac{1}{N}\sum_{n=1}^{N}y_n$

이다.

[0025] 상기 학습 모델은 SAE(stacked autoencoder), DBN(deep-belief network) 및 DNN(deep neural network) 중 어느 하나의 인공신경망 알고리즘을 이용하여 상기 축소된 분자 표현자 집합에 대하여 생성된 정량적 구조-활동 관계(Quantitative structure-activity relationship) 모델일 수 있다.

[0026] 상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 단계는, 고정확률 p에 따라 학습 모델 네트워크를 구성하는 유닛을 무작위로 드롭하는 단계를 포함할 수 있다.

[0027] 상기 축소된 분자 표현자 집합을 생성하는 단계는, 상기 화학 물질을 나타내는 분자 표현자 집합에 대하여, 부분최소자승법 기반의 VIP(variable importance in projection) 점수를 분자 표현자 마다 부여하는 단계; 상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하는 단계; 및 상기 분자 표현자 집합에 남은 분자 표현자에 LASSO 회귀 분석을 적용한 결과 파라미터의 값이 미리 설정된 값으로 감소된 분자 표현자를 더 제외하여 축소된 분자 표현자 집합을 생성하는 단계를 포함할 수 있다.

[0028] 상기 축소된 분자 표현자 집합을 생성하는 단계는, 상기 화학 물질을 나타내는 분자 표현자 집합에 대하여 각각의 분자 표현자를 독립변수로, 상관 결합 친화도를 종속변수로 하여, 독립변수와 종속변수 간의 선형 상관 관계를 계산하는 단계; 상기 선형 상관 관계의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하는 단계; 상기 분자 표현자 집합에 남은 분자 표현자에 대하여, 부분최소자승법 기반의 VIP(variable importance in projection) 점수를 분자 표현자 마다 부여하는 단계; 및 상기 평가 점수의 값이 미리 설정된 기준값보다 작은 분자 표현자를 상기 분자 표현자 집합에서 제거하여 축소된 분자 표현자 집합을 생성하는 단계를 포함할 수 있다.

[0029] 또한, 상기한 과제를 해결하기 위한 일 실시 예에 따른 위해성 평가 장치에 의하여 수행되는 위해성 평가 방법은 화학 물질을 나타내는 분자 표현자 집합을 입력 받는 단계; 상기 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자를 생성하는 단계; 및 상기 축소된 분자 표현자와 제 1 항 내지 제 12 항 중 어느 한 항에 기재된 위해성 평가 모델 생성 방법에 따라 생성된 위해성 평가 모델을 이용하여 상기 화학 물질의 내분비 교란 위해성을 평가하는 단계를 포함한다.

[0030] 또한, 상기한 과제를 해결하기 위한 일 실시 예에 따른 위해성 평가 모델 생성 장치는 화학물질의 내분비 교란 위해성 데이터가 저장된 메모리; 및 상기 데이터베이스로부터 내분비 교란 위해성을 가지는 화학 물질을 선정하고, 상기 화학 물질을 나타내는 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하며, 상기 축소된 분자 표현자 집합을 기반으로 인공신경망 알고리즘을 이용하여 화학 물질의 위해성을 분석하는 학습모델을 생성하는 프로세서를 포함한다.

[0031] 또한, 상기한 과제를 해결하기 위한 일 실시 예에 따른 내분비 교란 위해성 평가 장치는 화학 물질을 나타내는 분자 표현자 집합이 저장된 메모리; 및 상기 분자 표현자 집합에서 적어도 하나의 분자 표현자를 제거하여 축소된 분자 표현자 집합을 생성하고, 상기 축소된 분자 표현자 집합과 제 1 항 내지 제 12 항 중 어느 한 항에 기

재된 위해성 평가 모델 생성 방법에 따라 생성된 위해성 평가 모델을 이용하여 상기 화학 물질의 내분비 교란 위해성을 평가하는 프로세서를 포함한다.

[0032] 또한, 본 명세서는 상기한 과제를 해결하기 위한 위와 같이 기재된 위해성 평가 모델 생성 방법을 수행하기 위한 컴퓨터 프로그램과, 컴퓨터 프로그램이 기록된 컴퓨터 판독 가능 기록매체를 개시한다.

발명의 효과

[0034] 본 명세서에 기재된 기술적 사항에 따라, 사용자는 화학물질의 분자표현자 중 부분최소자승법 기반의 VIP score와 LASSO 회귀 방법을 이용해 중요 분자표현자를 선택하여 심층 인공신경망 기반 계산독성학 모델을 개발함으로써 내분비교란물질의 인체위해성을 효율적으로 평가할 수 있게 된다.

도면의 간단한 설명

[0036] 도 1a는 EDC의 순환 시스템을 도시하는 도면이다.

도 1b는 EDC가 인체 건강에 미치는 잠재적 영향을 평가하는 정량적 구조-활동 관계(Quantitative structure-activity relationship, QSAR) 모델을 활용하는 방안을 개념적으로 도시하는 개념도이다.

도 2는 일 실시예에 따른 모델링 장치가 DL-QSAR 모델을 생성하는 방법을 설명하는 도면이다.

도 3은 EDC 데이터베이스를 구축하는 방법을 설명하는 도면이다.

도 4는 VIP 점수와 LASSO 회귀를 이용하여 분자 표현자를 선택하는 일 실시예를 설명하는 도면이다.

도 5는 분류 (classification) 모델을 위한 분자 표현자를 선택하는 일 실시예를 나타내는 도면이다.

도 6은 예측(prediction) 모델을 위한 분자 표현자를 선택하는 일 실시예를 나타내는 도면이다.

도 7은 SAE(stacked autoencoder), DBN(deep-belief network) 및 DNN(deep neural network) 알고리즘의 개념을 도시하는 도면이다.

도 8a~8d는 분자 표현자의 VIP 점수를 나타내는 그래프이다.

도 9는 ER 및 SHBG의 분류 및 예측 성능에 대해 선택된 중요한 표현자의 일부 예시를 도시하는 도면이다.

도 10은 ER과 SHBG에 대한 트레이닝 및 테스트 세트에 대한 측정 및 예측값의 대응 QQ 플롯을 도시하는 그래프이다.

도 11a 및 도 11b는 DL 방법을 포함한 다른 ML 기술과의 성능 비교결과를 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0037] 이하의 내용은 단지 발명의 원리를 예시한다. 그러므로 당업자는 비록 본 명세서에 명확히 설명되거나 도시 되지 않았지만 발명의 원리를 구현하고 발명의 개념과 범위에 포함된 다양한 장치를 발명할 수 있는 것이다. 또한, 본 명세서에 열거된 모든 조건부 용어 및 실시 예들은 원칙적으로, 발명의 개념이 이해되도록 하기 위한 목적으로만 명백히 의도되고, 이와 같이 특별히 열거된 실시 예들 및 상태들에 제한적이지 않는 것으로 이해되어야 한다.

[0038] 상술한 목적, 특징 및 장점은 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다.

[0039] 명세서 및 청구범위에서 "제 1", "제 2", "제 3" 및 "제 4" 등의 용어는, 만약 있는 경우, 유사한 구성요소 사이의 구분을 위해 사용되며, 반드시 그렇지는 않지만 특정 순차 또는 발생 순서를 기술하기 위해 사용된다. 그와 같이 사용되는 용어는 여기에 기술된 본 발명의 실시예가, 예컨대, 여기에 도시 또는 설명된 것이 아닌 다른 시퀀스로 동작할 수 있도록 적절한 환경하에서 호환 가능한 것이 이해될 것이다. 마찬가지로, 여기서 방법이 일련의 단계를 포함하는 것으로 기술되는 경우, 여기에 제시된 그러한 단계의 순서는 반드시 그러한 단계가 실행

될 수 있는 순서인 것은 아니며, 임의의 기술된 단계는 생략될 수 있고/있거나 여기에 기술되지 않은 임의의 다른 단계가 그 방법에 부가 가능할 것이다.

[0040] 또한 명세서 및 청구범위의 "왼쪽", "오른쪽", "앞", "뒤", "상부", "바닥", "위에", "아래에" 등의 용어는, 설명을 위해 사용되는 것이며, 반드시 불변의 상대적 위치를 기술하기 위한 것은 아니다. 그와 같이 사용되는 용어는 여기에 기술된 본 발명의 실시예가, 예컨대, 여기에 도시 또는 설명된 것이 아닌 다른 방향으로 동작할 수 있도록 적절한 환경하에서 호환 가능한 것이 이해될 것이다. 여기서 사용된 용어 "연결된"은 전기적 또는 비 전기적 방식으로 직접 또는 간접적으로 접속되는 것으로 정의된다. 여기서 서로 "인접하는" 것으로 기술된 대상은, 그 문구가 사용되는 문맥에 대해 적절하게, 서로 물리적으로 접촉하거나, 서로 근접하거나, 서로 동일한 일반적 범위 또는 영역에 있는 것일 수 있다. 여기서 "일실시예에서"라는 문구의 존재는 반드시 그런 것은 아니지만 동일한 실시예를 의미한다.

[0041] 또한 명세서 및 청구범위에서 '연결된다', '연결하는', '체결된다', '체결하는', '결합된다', '결합하는' 등과 이런 표현의 다양한 변형들의 지칭은 다른 구성요소와 직접적으로 연결되거나 다른 구성요소를 통해 간접적으로 연결되는 것을 포함하는 의미로 사용된다.

[0042] 또한, 본 명세서에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.

[0043] 또한 본 명세서에서 사용된 용어들은 실시예를 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 '포함한다(comprise)' 및/또는 '포함하는(comprising)'은 언급된 구성요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0044] 또한, 발명을 설명함에 있어서 발명과 관련된 공지 기술에 대한 구체적인 설명이 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에 그 상세한 설명을 생략하기로 한다. 이하에는 첨부한 도면을 참조하여 본 발명의 바람직한 실시 예에 대해 상세하게 설명한다.

[0046] 8 만 개가 넘는 내분비 장애 화학 물질 (endocrine-disrupting chemicals, EDC)이 신흥 오염 물질 (emerging contaminants, EC)로 고려되고 있다. 이는 인체 건강에 미치는 영향으로 인해 이에 대한 위험성이 상당히 우려되고 있다.

[0047] 본 명세서에서는 도 1b에 도시된 바와 같이 인체 건강(c)에 미치는 EDC(a)의 잠재적 영향을 평가하기 위해 비-실험적인 화학 물질의 독성 영향 예측 방법으로 정량적 구조-활동 관계(Quantitative structure-activity relationship, QSAR) 모델(b)을 활용하는 방안을 제시한다.

[0048] 이하에서는 인간 내분비 시스템 중에서 성 호르몬 결합 글로불린 (sex-hormone binding globulin; SHBG) 및 에스트로겐 수용체(estrogen receptor, ER)를 예시로 EDCs의 질적 및 양적 효과를 예측하기 위해 딥러닝 기반 QSAR (DL-QSAR) 모델을 생성하고, 모델의 유효성을 확인함으로써, EDCs의 위험성을 판단하는 위해성 평가모델 생성 방법 및 이를 이용한 위해성 평가 방법을 설명한다. 그러나, 이하의 설명에 따른 기술적 사항은 SHBG나 ER이 아닌 다른 인간 내분비 시스템에 대한 DL-QSAR 모델을 생성하기 위하여도 적용될 수 있다.

[0050] 도 2는 일 실시예에 따른 모델링 장치가 DL-QSAR 모델을 생성하는 방법을 설명하는 도면이다. 모델링 장치는 프로세서와 메모리를 구비하고 있으며, 그 외 사용자 인터페이스, 통신부 등을 포함할 수 있다. 먼저, 모델링 장치는 EDCs 데이터베이스를 구축한다(S110). 다음으로, 모델링 장치는 데이터베이스에서의 변수 선택을 통해 분자 표현자(Molecular descriptors)를 선택한다(S120). 다음으로, 모델링 장치는 선택된 분자 표현자에 대한 QASR 모델을 생성한다(S130). 다음으로, 모델링 장치는 생성된 QASR 모델의 유효성을 판단하며(S140), 유효성을 갖춘 경우 생성된 QASR 모델을 활용하여 화학 물질이 인체이 미치는 영향을 추정한다(S150). 이하 도 2의 각 단계를 보다 구체적으로 설명한다.

[0052] **EDC 데이터베이스 구축(S110)**

[0053] 도 3은 EDC 데이터베이스를 구축하는 방법을 설명하는 도면이다. 모델링 장치는 반응 변수와 예측 변수를 구분

하여 데이터베이스화를 수행한다(S111).

[0054] 입력 데이터가 반응 변수인 경우, 모델링 장치는 정성적 반응(활성 또는 비활성 결합)을 평가하고 EDC에 대한 상대 결합 친화도 값의 로그(logarithm of relative binding affinity; logRBA)를 보고하는 실험 데이터 세트를 사용자 입력으로부터 또는 각종 문헌으로부터 수집하여 데이터베이스화 할 수 있다(S112). logRBA는 아래의 수식에 따라 계산될 수 있다.

[0055] [수학식1]

[0056]
$$logRBA = log\left(\frac{IC_{50}^{C}}{IC_{50}} \times 100\right)$$

[0057] 여기서 IC_{50}은 생물학적 검정에서 표적 수용체에 대한 결합의 50% 억제를 생성하는 시험 화학 물질의 농도를 나타낸다. IC^{C}_{50}은 기준 내인성 리간드의 IC_{50}이며, 일예에서 이는 SHBG에 대하여는 테스토스테론이고 ER에 대하여는 17β-estradiol이다. logRBA 값은 표 1에 요약되어 있다. SHBG에서 활성을 갖는 80 가지 화학 물질의 logRBA 값은 -4.437 내지 2.796의 범위를 가지고, ER에서 활성을 갖는 89 가지 활성 화학 물질의 logRBA 값은 -4.09 내지 2.6의 범위를 가지는 것으로 조사되었다.

[0058] [표 1]

ID	Name	CAS	SHBG	ER
1	Testosterone	53-16-7	2	
2	17β-Estradiol	50-28-2	1.33	2
3	6α-OH-Estradiol	1229-24-9	-0.514	-0.15
4	17-Desoxyestradiol	53-63-4	2	1.14
5	3-OH-Estra-1,3,5(10)-trien-16-one	3601-97-6	1.193	-0.29
6	Estriol	50-27-1	-1.286	0.99
7	17α-Estradiol	57-91-0	0.796	0.49
8	Ethinylestradiol	57-63-6	-0.722	2.28
9	Mestranol	72-33-3		0.35
10	Moxestrol	34816-55-2	-1.548	1.14
11	ICI164-384	98007-99-9		1.16
12	Fulvestrant	129453-61-8		1.57
13	Dihydrotestosterone	521-18-6	2.796	-2.89
14	Progesterone	57-83-0	-0.467	
15	Estrone	58-22-0	1	0.86
16	3α-Androstanediol	1852-53-5	1.897	-2.67
17	3β-Androstanediol	571-20-0	2.796	-0.92
18	Corticosterone	50-22-6	-0.467	
19	Norethynodrel	68-23-5	0.176	-0.67
20	Flavone	525-82-6		
21	Quercetin	117-39-5	-2.623	
22	Kaempferol	520-18-3		-1.61
23	Apigenin	520-36-5	-2.186	-1.55
24	Myricetin	529-44-2		-2.75
25	Fisetin	528-48-3		-2.35
26	3,6,4'-Trihydroxyflavone	253195-19-6		-0.35
27	Chrysin	480-40-0	-2.415	
28	Baicalein	491-67-8		-3.05

[0059]

[0060] 생성된 RBA 값은 모델 성능 개선을 위해 정규화되어 전처리될 수 있다(S113).

[0062] 입력 데이터가 예측 변수인 경우 모델링 장치는 분자 표현자를 데이터베이스화한다(S114). 분자 표현자는 분자의 중요한 구조적 특징을 인코딩하는 데 사용되는 분자의 간단한 수학적 표현이다. DRAGON 소프트웨어를 사용하여 테스트된 EDC의 분자 표현자를 생성할 수 있다. 일 예에서, 분자의 최적화된 구조적 특성을 기반으로 29 개

의 서로 다른 블록에서 4,885 개의 분자 표현자가 화학 물질의 구조적 다양성을 설명하기 위해 계산될 수 있다. 이를 위해, 정보 지수, 2D 자기 상관(autocorrelation)과, 구성적(constitutional), 위상적(topological), 고리(ring) 및 기하학적(geometrical) 표현자가 고려될 수 있다. 그 후, 전처리를 위해, 모델링 장치는 계산된 표현자를 정규화하고, DRAGON에 내장된 변수 제외 절차를 사용하여 표준 편차가 0 인 상수 표현자를 제외시켜 전처리할 수 있다(S115). 나아가, 전체 데이터셋은 트레이닝 세트와 테스트 세트로 7:3의 비율로 무작위 분할될 수 있다(S116).

[0064] **분자표현자 선택(S120)**

[0065] QSAR 모델링의 목적은 EC의 활동을 예측하는 것이며, 특히 인체 건강에 미치는 영향에 대한 잠재적 영향을 평가하기 위한 것이다. 이를 위해 분자 표현자와 목표 활동 사이의 상관 관계는 신중하게 고려되어야 한다. 분자 표현자의 대규모 데이터 세트에서 QSAR 모델을 개발할 수도 있지만, 분자 표현자의 차원이 높으면 과적합 문제(overfitting problem)와 다중공선성(multicollinearity)이 발생할 수 있다. 따라서, 다중공선성을 감소시키고 과적합을 방지하기 위한 분자 표현자의 선택은 QSAR 모델링의 예측 성능 및 해석성을 향상시키는 핵심 단계이다. 분자 표현자에서 중복성(redundancy) 및 다중공선성을 감소시키기 위해, 변수들은 부분 최소 자승법(partial least square, PLS) 기반의 변수중요도 척도(variable importance in projection, VIP) 및 LASSO(least absolute shrinkage and selection operator) 회귀 기술에 기반하여 선택될 수 있다.

[0066] VIP는 각 구성 요소에 가중치를 부여하여 각 표현자의 중요도를 축적한다. VIP 점수가 1보다 크면, 해당 변수는 PLS 모델에서 중요한 변수로 선택될 수 있다. VIP는 아래의 수학식을 이용하여 계산될 수 있다.

[0067] [수학식 2]

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)(w_{jk} / \|w_{jk}\|)^2}{\sum_{k=1}^{l}(q_k^2 \cdot t_k^{'} \cdot t_k)}}$$

[0068]

[0069] 여기서, p는 x 변수의 개수이고, i는 선택된 잠재 변수(latent variable)의 개수이고, w는 가중치이며, $q_k^2 \cdot t_k^{'} \cdot t_k$ 는 각 컴포넌트에 의하여 해석되는 차이값이며, $(w_{jk} / \|w_{jk}\|)^2$ 는 j번째 변수의 중요도를 나타낸다.

[0071] VIP 점수를 사용하여 응답과 관련이 없는 표현자를 줄인 후, 중요하지 않은 분자 표현자의 파라미터를 0으로 감소시키기(shrinkage) 위해 LASSO 회귀법을 적용할 수 있다. LASSO 회귀는 L_1-norm을 기반으로 한 벌칙화된(penalized) 방법의 클래스에서 가장 일반적인 정규화 기술 중 하나이며, 아래의 수학식을 이용하여 계산될 수 있다.

[0072] [수학식 3]

$$\hat{\beta}^{LASSO} = \underset{\hat{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

[0073]

[0074] 여기서, $\hat{\beta}^{LASSO}$ 는 LASSO 회귀의 추정된 파라미터 벡터이고, y는 반응 변수이고, x는 예측 변수이고, n은 반응 변수의 개수이고, p는 예측 변수의 개수이다. λ 파라미터는 감소(shrinkage) 수준을 설정하기 위해 조정

될 수 있다. λ값이 높을수록 더 많은 계수가 0으로 줄어든다. 따라서, λ는 일부 계수를 정확히 0으로 축소함으로써 변수를 선택할 수 있게 한다.

[0076] 도 4는 VIP 점수와 LASSO 회귀를 이용하여 분자 표현자를 선택하는 일 실시예를 설명하는 도면이다. 도 4의 예와 같이, 모델링 장치는 VIP 점수가 1.21보다 높은지를 판단하고(S121a), LASSO 회귀의 파라미터 벡터가 0이 아닌지를 판단함으로써(S122a), 분자 표현자를 선택할 수 있다(S123a). 모델링 장치는 VIP 점수가 1.21보다 높지 않거나, LASSO 회귀의 파라미터 벡터가 0인 변수는 분자 표현자의 선택에서 배제할 수 있다(S124a).

[0078] 도 5는 분류 (classification) 모델을 위한 분자 표현자를 선택하는 일 실시예를 나타내며 표준편차(standard deviation; STD)와 VIP를 이용한다. 분류 모델은 EDCs가 내분비 교란에 영향을 미치는지 또는 미치지 않는지 판단하는 모델이다. 도 5의 예와 같이, 분류 모델을 위한 분자 표현자를 선택하는 경우 모델링 장치는 표준편차가 0보다 큰지를 판단하고(S121b), VIP 점수가 1보다 높은지를 판단함으로써(S122b), 분자 표현자를 선택할 수 있다(S123b). 모델링 장치는 표준편차가 0이거나, VIP 점수가 1보다 크지 않은 변수는 분자 표현자의 선택에서 배제할 수 있다(S124b).

[0080] 도 6은 예측(prediction) 모델을 위한 분자 표현자를 선택하는 일 실시예를 나타내며 피어슨 상관 계수(Pearson's Correlation Coefficient)와 VIP를 이용하여 분자 표현자를 선택한다. 예측 모델은 EDCs가 내분비 교란에 얼마나 영향을 미치는지 추정하는 모델이다.

[0081] 피어슨 상관 계수 $C_{pearson}$은 독립 변수와 종속 변수 간의 선형 상관 관계를 나타내며 아래의 수식을 이용하여 계산될 수 있다. $C_{pearson}$의 값이 0.6보다 크면, 양 변수간에 높은 상관관계가 있는 것으로 해석할 수 있다.

[0082] [수학식 4]

$$C_{pearson}(x,y)=\frac{\sum_{i=1}^{N}(x_i-m_x)(y_i-m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i-m_x)^2][\sum_{i=1}^{N}(y_i-m_y)^2]}}$$

[0083]

[0084] [수학식 5]

$$m_x=\frac{1}{N}\sum_{n=1}^{N}x_n$$

[0085]

[0086] [수학식 6]

$$m_y=\frac{1}{N}\sum_{n=1}^{N}y_n$$

[0087]

[0089] 분자 표현자가 독립 변수로, 정성분석결과(활성화/비활성화) 또는 RBA 값이 종속 변수로 사용될 수 있다. 도 6의 예와 같이, 예측 모델을 위한 분자 표현자를 선택하는 경우 모델링 장치는 $C_{pearson}$이 0.6보다 큰지를 판단하고(S121c), VIP 점수가 1보다 높은지를 판단함으로써(S122c), 분자표현자를 선택할 수 있다(S123c). 모델링 장치는 $C_{pearson}$이 0.6보다 크지 않거나, VIP 점수가 1보다 크지 않은 변수는 분자 표현자의 선택에서 배제할 수 있다

(S124b).

[0091] **DL-QSAR 모델을 위한 기반 알고리즘(SAE, DBN, DNN)**

[0092] 분류 및 예측을 위한 QSAR 모델은 예측 변수로서 선택된 분자 표현자를 갖는 SAE(stacked autoencoder), DBN(deep-belief network) 및 DNN(deep neural network) 알고리즘을 사용하여 모델링 될 수 있으며, 각 모델의 개념 구조가 도 7의 (a), (b), (c)에 도시되어 있다. 각 모델의 더 나은 예측 성능을 얻기 위해 각 구조의 하이퍼-파라미터에 정규화 기술이 적용될 수 있다. 이하 도 7을 참조하여 각 알고리즘을 간단히 설명한다.

[0094] SAE (stacked autoencoder)는 자동 연결 신경망 (또는 자동 인코더; AE)을 기반으로하는 딥 러닝 기술이다. AE는 비-지도 학습에 기반한 일종의 피드 포워드 신경망이다. 이 방법은 일반적으로 인공 신경망(artificial neural network; ANN) 모델에서 X 변수의 차원을 줄이기 위해 사용된다. SAE는 입력 및 출력과 신경망과 유사한 추가적인 은닉층으로 구성된다. SAE의 주요 특징은 각각의 은닉 스택-AE층이 리콜된 입력에 의해 학습된다는 것이다. 따라서 입력 및 출력층에는 동일한 갯수의 유닛이 있게 된다.

[0095] 은닉층은 입력 변수와 출력 변수 사이의 오류를 최소화함으로써 출력층에서 입력 변수를 재구성하려고 시도한다. 경사 하강을 갖는 역 전파 알고리즘은 SAE를 트레이닝시키는 데 사용된다. 인코더라고 불리는 훈련된 은닉층은 중복없이 입력 변수의 차원을 줄일 수 있다. 오토 인코더의 심층 네트워크는 입력 변수의 기능을 점차적으로 압축하고 신경 네트워크의 심층 아키텍처를 개발할 수 있는 여러 개의 인코더로 구성되어 있다.

[0097] DBN(deep-belief network)은 빠르고(fast) 그리디(greedy)한 알고리즘을 사용하여 심층 신경망 아키텍처를 구축하기 위해 도출되었다. DBN 네트워크는 생산적 무방향 그래픽 모델(generative undirected graphical model)인 제한 볼츠만 머신(restricted Boltzmann machine; RBM)의 계층으로 구성된다. X층은 입력층이고 h층은 은닉층이다. RBM을 트레이닝시키는 목적은 KLD(Kullback-Leibler divergence)를 사용하여 X와 h층의 확률 분포를 비교하여 결정되는 X와 h층 사이의 값 가중치를 계산하는 것이다. KLD는 일반적으로 두 확률 분포 간의 유사성을 측정하는 데 사용된다. RBM의 목적 함수는 두 KLD 간의 차이를 최소화하는 것이다.

[0098] DBN 트레이닝에는 지속적인 RBM 트레이닝 과정이 포함되며, 은닉층 사이에 위치한 가중치는 RBM 트레이닝 과정에 의해 계산된다. DBN이 완료된 후, 마지막 은닉층과 출력층 사이에 위치한 가중치를 계산하기 위해 ANN이 DBN에 적용될 수 있다.

[0100] DNN(deep neural network)은 과적합없이 대규모 데이터를 처리할 수 있는 장점이 있다. ANN의 성능을 향상시키기 위해 DNN에는 하나 보다 많은 은닉층과, 모든 층에 더 많은 유닛이 포함된 심층적 계층 구조 모델이 있을 수 있다. DNN 구조가 깊을수록 모델이 과도하게 적합되기 쉽다. 그 결과, 과적합 문제를 피하기 위해 많은 기술(일반적으로 "정규화"라고 함)이 개발되었다. 예를 들어, 드롭 아웃 및 배치 정규화 기술이 구현될 수 있다. 드롭 아웃의 주요 개념은 고정 확률 p에 따라 네트워크를 트레이닝하는 동안 네트워크에서 유닛을 무작위로 드롭하는 것이다.

[0101] [수학식 7]

[0102]
$$\begin{cases} a_i^{(l+1)} = f(z_i^{(l+1)}) \\ z_i^{(l+1)} = \sum_{i=1}^{n} (w_i^{(l)} \cdot r_i^{(l)} \cdot a_i^{(l)}) + b \\ r_i^{(l)} \sim Bernoulli(p) \end{cases}$$

[0103] 여기서 $r_i^{(l)}$는 확률 p가 1 인 독립적인 베르누이 랜덤 변수(independent Bernoulli random variables)이다. 배

치 정규화는 공변량 이동에 대한 활성화를 정규화함으로써 딥 네트워크의 교육을 가속화하기 위해 개발되었다. 그러나 이 기법은 과적합을 줄임으로써 DNN 네트워크의 일반화에 유리한 영향을 미치는 것으로 보고되었다.

[0104] 과적합 문제 외에도 DNN 개발의 또 다른 한계는 기울기 값의 소실 문제이다. 네트워크는 연쇄 규칙과 같이 활성화 함수로부터 파생된 값의 곱에 기반한 역 전파에 의해 학습된다. 그러나, 종래의 신경망에서, 활성화 함수는 전체 변수 범위에 걸쳐 1보다 작은 기울기를 갖는 시그모이드 함수로부터 도출된다. 따라서 연쇄 규칙에 따라 결정되듯이 DNN의 기울기는 0에 수렴한다. 기울기 값 소실 문제를 해결하기 위해, ReLU(rectified linear unit)이 본 실시예에서의 활성화 함수로 구현되었다. 이는 다음과 같이 표현될 수 있다.

[0105] [수학식 8]

[0106] $f(x) = \max(0, x)$

[0107] 여기서 x는 뉴런에 대한 입력이다. ReLU의 미분 값은 0 (if $x \leq 0$) 또는 1 (if $x>1$)이다.

[0109] **DL-QSAR 모델 생성(S130)**

[0110] 일 실시 예에서 SAE, DBN 및 DNN 알고리즘은 최적의 구조를 찾기 위해 Keras 라이브러리를 사용하여 Python에서 구현되었으며, MATLAB R2016b는 LR, MLR 및 SVM 모델을 개발하는 데 사용되었다. SVM 모델의 경우, 아래의 수학식 9와 같은 선형 커널 함수가 활용되었다.

[0111] [수학식 9]

[0112] $$y(x) = sign[\sum_{k=1}^{N} \alpha_k \times y_k \times \Psi(x, x_k) + b] = w_T \times \varphi(x_k) + b$$

[0113] 여기서 α_k와 w 는 각각 가중값과 벡터이며, b는 바이어스 항이다. $\Psi(x, x_k)$와 $\varphi(x_k)$는 입력 데이터를 더 높은 차원 공간에 매핑하는 커널 함수이다.

[0115] 분류 및 예측을 위한 QSAR 모델은 SAE, DBN 및 DNN 알고리즘과 예측 변수로 선택된 분자 표현자를 사용하여 개발되었다. 각 개발 모델에 대해 더 나은 예측 성능을 얻기 위해 Relu 함수가 입력 및 각각의 은닉층의 활성화 함수로써 활용되었고, 정규화 기법은 각 구조의 하이퍼-파라미터들과 결합되었다.

[0117] 각 QSAR 모델에 대한 최적의 구조를 개발하기 위해, 스택 자동 인코더 기반 QSAR (SAE-QSAR) 모델의 구조는 분류를 위해 16 개의 노드로 구성된 입력층과 8개, 4개, 8개의 노드로 구성된 각각의 은닉층으로 총 4개의 층(즉, 하나의 입력층과 3 개의 은닉층)으로 구성될 수 있다.

[0118] 예측을 위한 SAE-QSAR 모델의 구조는 입력층에 대해 32 개의 노드와 각각의 은닉층에 대해 16개, 4개 및 16개의 노드로 구성된 총 4개의 계층으로 구성될 수 있다.

[0120] DBN-QSAR(deep-belief network-based QSAR) 모델의 최적 구조를 결정하기 위해, 은닉층, 은닉 노드, 에포크 수, 배치 크기 및 학습 속도가 정의될 수 있다. 분류를 위한 DBN-QSAR 모델은 8, 4, 4 및 8 노드로 구성된 4 개의 층으로 구성될 수 있다. 일 실시 예에서, 학습 속도는 0.1로 설정되고, 에포크 (epoch)의 수는 150이며 노드의 10%가 드롭 될 수 있다.

[0121] 예측을 위한 DBN-QSAR 모델은 하나의 계층으로 구성되며 32 개의 노드를 가질 수 있다. 학습률은 0.01로, 에포크의 수는 100으로, 노드의 10%가 드롭될 수 있다.

[0123] 분류를 위한 DNN-QSAR(deep-neural-network-based QSAR) 모델은 0.002 학습 속도로 300 에포크(epoch)로 트레

이닝될 수 있다. 최적의 구조는 16 개의 노드가 있는 하나의 입력층과 8과 4 개의 노드가 있는 두 개의 은닉층으로 구성될 수 있다. 정규화를 위해, 10 %의 노드를 드롭하고 0.002의 L2 정규화기를 각 층에 사용할 수 있다.

[0124] 예측을 위한 DNN-QSAR 모델은 0.002 학습 속도로 100 에포크에서 학습되며, 구조는 32 노드를 가진 하나의 입력층과, 각각 16, 8 및 4 노드를 가진 3 개의 은닉층으로 조정될 수 있다. 정규화를 위해, 입력층의 10 % 노드를 드롭하고, 0.002의 L2 정규화기 및 배치 정규화 방법을 각 층에 사용할 수 있다.

[0126] **성능 평가**

[0127] 각 QSAR 모델에 의한 분류 예측 모델을 로지스틱 회귀 (LR), MLR 및 SVM을 포함한 기존 ML 방법과 비교하여 검증 지수를 사용하여 제안된 모델의 성능을 평가할 수 있다.

[0128] 각 분자 표현자의 분류 모델에 대한 최상의 방법은 아래의 수학식 10 내지 13을 사용하여 계산된 정확도(Q), 정밀도(Pr), 리콜(Rc) 및 밸런스(Bl)로 구성되는 분류 성능에 대한 통계적 파라미터를 비교함으로써 선택될 수 있다.

[0129] [수학식 10]

[0130] $$Q = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

[0131] [수학식 11]

[0132] $$Pr = \frac{TP}{TP + FP} \times 100\%$$

[0133] [수학식 12]

[0134] $$Rc = \frac{TP}{TP + FN} \times 100\%$$

[0135] [수학식 13]

[0136] $$Bl = \frac{TN}{TN + FP} \times 100\%$$

[0138] 여기서 TP(True Positive)와 TN(True Negative)은 활성화(active)와 비활성화(inactive)로 분류된 화학 물질의 개수이다. FN(False Negative)은 비활성 화학 물질로 분류된 활성 화학 물질의 개수이고, FP(False Positive)는 활성 화학 물질로 분류된 비활성 화학 물질의 개수이다. 이 네 가지 통계 매개 변수의 값이 높으면 모델의 분류 성능이 우수함을 나타낸다.

[0140] 각 QSAR 모델에 의한 예측 성능을 평가하기 위해, 결정 계수 (R^2), 근 평균 제곱 오차 (RMSE) 및 예측 제곱 상관 계수 (Q2)를 포함하여 세 가지 통계 파라미터가 계산될 수 있다. R2, RMSE 및 Q2 값은 아래의 수학식 14 내지 16를 사용하여 얻을 수 있다.

[0142] [수학식 14]

$$R^2 = 1 - \frac{\left[\sum_{i-1}^{n_{TR}} (\hat{y}_i - y_i)^2\right] / n_{TR}}{\left[\sum_{i-1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}}$$

[0143]

[0144] 여기서, n_{TR}은 트레이닝 세트의 수이고, $\bar{y}_{TR}$는 트레이닝 세트에서의 변수의 평균값이며, $\left[\sum_{i-1}^{n_{TR}} (\hat{y}_i - y_i)^2\right]$는 잔차 제곱합(residual sum of squares)이고 $\left[\sum_{i-1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right]$는 총 제곱합(total sum of square)이다.

[0146] [수학식 15]

$$\text{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} (y_{predicted} - y_{observed})^2\right)}$$

[0147]

[0148] [수학식 16]

$$Q^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_{TR})^2}$$

[0149]

[0150] 여기서 n은 총 데이터 개수이고 $y_{predicted}$는 모델의 예측 값이고 $y_{observed}$는 모델에 사용된 관측값이다.

[0152] **실시예 - 핵심 분자 서술자 선정**

[0153] 처음에는 DRAGON 내장 변수 제외 프로시저를 적용하여 4,885 개의 분자 표현자를 약 2,100개로 줄였다. 그러나, 데이터 세트에서 수집된 EDC에 비해, 이러한 분자 표현자의 매트릭스 크기는 각 분석에 대한 QSAR 모델을 개발하기에 너무 컸다. 큰 입력과 작은 출력 데이터를 가진 데이터 세트는 다중 공선성을 불러 일으킬 수 있기 때문에, 여분의 데이터 세트로부터의 응답 변수를 효율적으로 설명할 수 있는 정보 표현자를 추출해야 한다. VIP 점수 및 LASSO 회귀의 두 단계로 분자 표현자 선택이 선행되었다.

[0154] 먼저, SHBG와 ER의 활성/비활성 및 정량적 응답인 반응 변수에 대한 중요한 표현자를 정렬하기 위해 각 표현자의 VIP 점수를 계산하였다. 도 8a는 SHBG의 분류 분석을 위한 분자 표현자의 VIP 점수를 나타내고, 도 8b는 ER의 분류 분석을 위한 분자 표현자의 VIP 점수를 나타내고, 도 8c는 SHBG의 예측 분석을 위한 분자 표현자의 VIP 점수를 나타내고, 도 8d는 ER의 예측 분석을 위한 분자 표현자의 VIP 점수를 나타낸다.

[0155] 결과에서 VIP 점수가 1.21 이상인 표현자는 각 예측 변수와 관련이 높기 때문에 선택되었다. 분류 성능에서 SHBG 및 ER에 대해 각각 949 및 595 표현자가 선택되었다. 또한, SHBG 및 ER의 정량적 반응을 예측하기 위해 536 및 515 표현자가 선택되었다. SHBG 및 ER의 분류 성능에서 가장 중요한 분자 표현자는 VIP 점수가 5.01 인 토폴로지 인덱스인 ZM2Kup이고, VIP 점수가 18.47 인 CATS 2D의 CATS2D_05_AA다. 또한 SHBG 및 ER의 성능 예측

에서 가장 중요한 분자 표현자는 R3s +, GETAWAY 표현자 및 Mor30v로 VIP 점수는 각각 11.60, 18.29 및 18.29였다.

[0156] 둘째, 중복성을 더욱 줄이기 위해, ER과 SHBG의 분류 및 예측 성능에 대한 예측 변수 및 반응 변수의 LASSO 회귀 ($\hat{\beta}^{LASSO}$)를 별도로 수행하였다. 중요한 분자 표현자를 선택하기 위해 관련 없는 표현자를 0으로 줄였다. 트레이닝 세트는 테스트 세트와 함께 7:3의 비율로 설정되어 각 분석에 대해 LASSO 회귀의 최적 튜닝 파라미터를 검색하여 주요 표현자를 선택하도록 하였다. 그 후, 최적의 구속 값 λ를 검색하기 위해 K값으로 5를 선택한 K-폴드 교차 검증 방법을 수행하였다. 분류 분석에 대한 λ값은 ER 및 SHBG에 대해 각각 0.0111 및 0.0142였다. 그러나 SHBG와 ER의 예측 분석의 λ는 각각 0.0022와 0.0071이었다.

[0157] 마지막으로, ER 및 SHBG에 대한 예측 성능을 위한 핵심 분자 표현자로서 52 및 58의 분자 표현자가 각각 선택되었고, ER 및 SHBG에 대한 분류 성능을 위해 33 및 36의 분자 표현자가 각각 선택되었다. ER 및 SHBG의 분류 및 예측 성능에 대해 선택된 중요한 표현자의 일부 예시를 도 9에 기재하였다.

[0158] 위의 4 가지 분석을 위해 선택된 주요 그룹은 도 9에 제시된 바와 같이 토폴로지 인덱스, 2D 매트릭스 기반 표현자(2D-matrix-based descriptors), 2D 자기 상관(2D autocorrelations), 부담 고유값(burden eigenvalues), 에지 인접 인덱스(edge adjacency indices), RDF 표현자(RDF descriptors), 3D-MoRSE 표현자(3D-MoRSE descriptors), WHIM 표현자(WHIM descriptors), GETAWAY 표현자(GETAWAY descriptors) 및 약물-유사 인덱스(drug-like indices)를 포함한다. 분류 성능을 위해 2D 매트릭스 기반 표현자, 3D-MoRSE 및 GETAWAY 표현자는 최종 분자 표현자로의 기여도가 높았다. 마찬가지로, 예측 분석의 경우, 2D 자기 상관 및 3D-MoRSE 표현자는 분자 표현자의 기여도가 높았다. 2D 자기 상관은 고려된 경로 길이에서 모든 경로의 말단 원자의 원자량의 산술 평균에 의한 분자 그래프의 계산을 설명하였다. 원자량의 기하 평균으로부터 얻은 3D-MoRSE 서술자들은 다른 각도 산란 함수에 의해 관찰되었다. GETAWAY 표현자는 기하학적 정보로부터 나타나는 중심 원자 좌표 및 원자량으로부터 레버리지 값을 나타냈다.

[0160] **실시예 - DL-QSAR 모델의 분류 성능**

[0161] 앞서 설명된 DL-QSAR 모델을 검증하기 위해, 표 2에 도시 된 바와 같이 테스트 세트로부터의 분류 결과를 나타내는 오차행렬(confusion matrix)을 생성하였다.

[0162] [표 2]

Descriptors	Assay	Actual / Predicted	SAE		DBN		DNN	
			positive	Negative	positive	Negative	positive	Negative
Selected	ER	Positive	23	1	25	2	25	0
		Negative	3	6	1	5	1	7
	SHBG	Positive	24	2	24	2	26	1
		Negative	2	9	2	9	0	10
Non-selected	ER	Positive	22	6	26	7	18	3
		Negative	4	1	0	0	8	4
	SHBG	Positive	22	8	24	10	18	3
		Negative	2	2	0	0	6	7

[0163]

[0165] 표에 따르면, 대각 매트릭스(diagonal matrix)는 정정된 분류 번호를 제공하고 비대각 행렬(off-diagonal matrix)은 수정되지 않은 분류된 숫자를 나타낸다.

[0166] 모델 성능에 대한 변수 선택의 영향을 식별하기 위해, 선택되지 않은 분자 표현자로부터 제안된 DL-QSAR 모델을 사용한 분류 결과를 계산하였다. SAE-QSAR 모델의 분류 결과는 ER 응답에 대하여 평균 69.70% 정확도

(accuracy), 78.57% 정밀도(precision), 84.62% 리콜(recall) 및 14.29% 밸런스(balance)를 나타냈으며, SHBG 응답에 대하여 70.59% 정확도, 73.33% 정밀도, 91.67% 리콜 및 20.00 % 밸런스를 나타냈다. DBN-QSAR 모델은 ER 및 SHBG 응답의 분류에 대해 각각 78.79% 및 70.58% 정확도, 78.79% 및 70.58% 정밀도, 100.00% 리콜 및 0.00% 밸런스를 가졌다. ER 및 SHBG 응답에 대한 DNN-QSAR 모델의 정확도, 정밀도, 리콜 및 밸런스 비율은 각각 66.67% 및 73.53%, 85.71%, 69.23% 및 75.00%, 57.14% 및 70.0%이다.

[0167] 대조적으로, 선택된 서술자에 대한 SAE-QSAR 모델에 대한 분류 결과는 ER에 대해 평균 87.88% 정확도 및 95.83% 정밀도, SHBG 분류에 대해 89.19% 정확도 및 92.31% 정밀도를 나타냈다. 또한 SAE-QSAR 모델은 ER에 대해 88.46% 리콜율, 85.71%의 밸런스를 나타냈고, SHBG의 92.31% 리콜율 및 81.82% 밸런스를 나타냈다. DBN-QSAR 모델은 표 3에 나타낸 바와 같이 각각 ER 및 SHBG 반응의 분류에 대해 90.91% 및 89.19%의 높은 정확도를 가졌다. 그러나 DNN-QSAR의 정확성, 정밀도, 리콜 및 밸런스는 ER 반응의 경우 96.97%, 100%, 96.15% 및 100%이고, SHBG 반응의 경우 94.74%, 96.3%, 100% 및 90.91% 였다.

[0168] [표 3]

Assay	DL-QSAR	Accuracy Q (%)	Precision Pr (%)	Recall Rc (%)	Balance Bl (%)
ER	DNN	96.97	100.00	96.15	100
	DBN	90.91	92.59	96.15	71.42
	SAE	87.88	95.83	88.46	85.71
SHBG	DNN	94.74	96.15	100	90.91
	DBN	89.19	92.31	92.31	81.82
	SAE	89.19	92.31	92.31	81.82

[0169]

[0170] 선택된 주요 분자 표현자를 갖는 DL-QSAR 모델의 분류 성능은 비 선택된 분자 표현자로부터 제안된 모델의 분류 성능보다 우수하였다. 리콜 및 밸런스의 값은 중요 표현자를 먼저 추출하지 않으면 큰 입력 데이터 세트가 다중 공선성 및 과적합 문제를 일으킨다는 것을 보여주었다. 따라서 가비지 데이터(garbage data)에서 중요한 표현자를 선택하는 것은 모델 성능을 향상시키는 것에 결정적 의미를 가지는 것을 볼 수 있다.

[0172] **실시예 - DL-QSAR 모델의 예측 성능**

[0173] 제안된 예측 모델의 견고성을 분석하기 위해 서로 다른 배치 크기, 은닉층 및 에포크가 검사되었다. 도 10은 ER과 SHBG에 대한 트레이닝 및 테스트 세트에 대한 측정 및 예측값의 대응 QQ 플롯을 도시하며, ER에 대하여 (a)는 DNN, (b)는 DBN, (c)는 SAE의 성능 비교를 나타내고, SHBG에 대하여 (d)는 DNN, (e)는 DBN, (f)는 SAE의 성능 비교를 나타낸다.

[0174] 트레이닝 및 테스트 세트에 대한 측정 값과 예측값 사이의 시각적 상관 관계는 ER에는 만족스럽지 않았다. 그러나 SHBG는 더 나은 시각적 상관 관계를 보여주었다. 수치 평가에서 R^2 및 RMSE의 값은 각각 ER 및 SHBG에 대한 트레이닝 세트에 대해 0.69, 0.53, 0.85, 0.39이며, 표 4에 제시되어있다. 결과는 ER에 대하여 낮은 예측 상관 관계와 SHBG에 대하여 좋은 상관 관계를 나타낸다. SAE-QSAR 모델의 예측 능력을 추가로 평가하기 위해 Q^2가 계산되었다. Q^2의 값 0.68 및 0.63는 SAE를 사용한 제안된 EDC가 만족스럽다는 것을 보여준다.

[0175] [표 4]

Descriptors	Assay	DL-QSAR	Training		Test	
			RMSE	R^2	RMSE	Q^2
Selected	ER	DNN	0.28	0.91	0.47	0.84
		DBN	0.38	0.84	0.57	0.75
		SAE	0.53	0.69	0.61	0.68
	SHBG	DNN	0.38	0.80	0.48	0.86
		DBN	0.44	0.72	0.80	0.63
		SAE	0.39	0.85	0.61	0.63
Non-selected	ER	DNN	0.86	0.76	1.25	0.47
		DBN	1.33	0.45	1.30	0.33
		SAE	1.27	0.41	1.92	0.04
	SHBG	DNN	0.66	0.78	1.21	0.29
		DBN	1.00	0.53	1.03	0.39
		SAE	0.53	0.86	1.29	0.20

[0176]

[0178] DBN-QSAR 모델의 네트워크는 예측 모델의 정확성을 향상시키기 위해 최적화되었다. 인체 건강에 대한 EDC 효과를 예측하기위한 DBN-QSAR 및 DNN-QSAR 모델은 도 10과 같이 다른 모델보다 우수한 성능을 나타낸다. DBN-QSAR 모델을 사용하여 ER 응답을 예측하기위한 R^2 및 RMSE 값은 0.84 및 0.38이다. 이 모델은 표 3에 요약된 바와 같이 Q^2가 높은 값인 0.75를 가진 테스트 세트에서 우수한 예측성을 보여준다. 그러나 SHBG 응답을 예측하기 위해, DBN-QSAR 모델은 Q^2 값이 0.63 인 과적합 문제가 있었다. EDC 예측을 위한 DNN-QSAR 모델은 트레이닝 세트에서 ER (0.91) 및 SHBG (0.80)에 대해 높은 상관 계수를 가졌다. DNN-QSAR 모델은 인간 건강에 대한 EDC 효과에 대한 더 나은 예측 가능성을 보여주었다.

[0180] 주요 분자 표현자를 선택하지 않은 DL-QSAR 모델의 예측 성능 또한 분류 및 과적합 문제를 일으키게 된다. 표 3에 제시된 DL-QSAR 모델의 Q^2 값으로 표시된 바와 같이, 모델의 예측력은 R^2에 비해 매우 열악하다. 모델의 구조와 매개 변수가 일관되기 때문에, 이 결과는 불필요한 표현자로 인하여 제안된 모델의 예측 성능이 저하되는 것으로 해석될 수 있다. DL-QSAR 모델의 분류 및 예측 결과로부터, 중요한 표현자를 선택하는 것이 과적합 문제를 피하고 모델의 예측 성능을 향상시키는 데 중요하다고 해석할 수 있다.

[0182] **제안된 모델의 비교 분석**

[0183] 제안된 딥 러닝 방법의 성능을 평가하기 위해, DL-QSAR 모델을 LR, MLR 및 SVM과 비교하였다. 도 11a는 DL 방법을 포함한 다른 ML 기술과의 분류 성능 비교결과를 도시하고, 도 11b는 DL 방법을 포함한 다른 ML 기술과의 예측 성능 비교결과를 도시한다.

[0184] QSAR 모델의 분류 성능과 관련하여, 도 11a는 DL-QSAR 모델이 두 가지 생물 분석에서 기존 머신 러닝 기반 QSAR (기존 ML-QSAR) 모델보다 우수했음을 나타낸다. 결과는 DNN-QSAR 모델이 ER에 대해 Q, Pr, Rc 및 B1 값이 각각 96.97 %, 100 %, 96.15 % 및 100 %이고, SHBG에 대해 각각 94.74 %, 96.3%, 100 % 및 90.91 % 인 최고의 분류 성능을 보였다. 이러한 발견은 DNN-QSAR 모델이 분석테스트에 대한 EDC의 대부분의 정성적 반응을 포착할 수 있

음을 의미한다.

[0185] 도 11a와 같이 DBN-QSAR 모델은 90% 정확하고 SAE-QSAR 모델은 거의 90 % 정확하다. 두 모델 모두 LR 모델의 정확도보다 크다. 따라서 DL은 EDC의 정성적 반응을 충분히 포착할 수 있다고 보여진다.

[0187] 마찬가지로 예측 분석에서 DL-QSAR 모델은 도 11b에 표시된 것처럼 MLR 및 SVM을 포함한 기존 ML-QSAR 모델보다 우수했다. DNN-QSAR 모델은 DBN, SAE, MLR 및 SVM 기반 QSAR 모델들에 비해 뛰어난 예측 성능을 보여준다. DNN 모델의 R^2 및 Q^2 값은 ER의 경우 0.91 및 0.84, SHBG의 경우 0.80 및 0.86이다. 비록, 트레이닝 세트에서 MLR이 높은 R^2 값을 보였지만, 테스트 세트에서 Q^2 값이 가장 낮은 최저의 예측 성능을 보였다.

[0188] DBN-QSAR 모델은 적합도에서 우수함을 보여주었다. 그러나 DNN-QSAR 모델은 최상의 예측 성능을 보여주었다. 세 가지 DL 방법 모두 SVM 및 MLR에 비해 해석성이 좋으며 과적합을 방지했다. 따라서 DL-QSAR 모델은 새로운 화학 물질의 정량적 반응을 예측할 수 있는 능력이 있다고 해석된다. 그러나 전체적으로 DNN-QSAR 모델은 분류 및 예측 성능 모두에서 열등한 값을 생성했다.

[0189] DNN 및 DBN 기반 QSAR 모델은 예측 가능성이 매우 효율적인 방법으로 내분비 교란 화학 물질에 대한 ER 및 SHBG의 정성적 및 정량적 반응을 예측하는 데 적합하다. 이는 DNN 및 DBN 기반 QSAR 모델의 로컬 매핑 접근 방식이 기존 통계적 접근 방식보다 더 안정적이고 높은 분류 결과를 제공할 수 있음을 보여준다. DL-QSAR 모델은 복잡한 분자 정보를 추출할 수 있고, 국소적 최소점에 갇히거나 과적합되는 경향이 없다. 복잡한 정보는 통계 기법의 성능에 부정적인 영향을 미쳐 낮은 Q^2 값을 나타내게 하기 때문에, 제안된 DL-QSAR 모델, 특히 DNN 및 DBN은 인체 건강에 미치는 EDC의 부작용을 예측할 수 있다.

[0190] 이러한 점에서 분류 모델과 예측 모델을 위한 중요 분자 표현자를 선택하기 위해 VIP 기법을 사용하고, 분류 모델을 구축하도록 DBN을 사용하며, 예측 모델로 DNN을 피어슨 상관 계수의 적용 이후 사용하는 방안등을 고려할 수 있다.

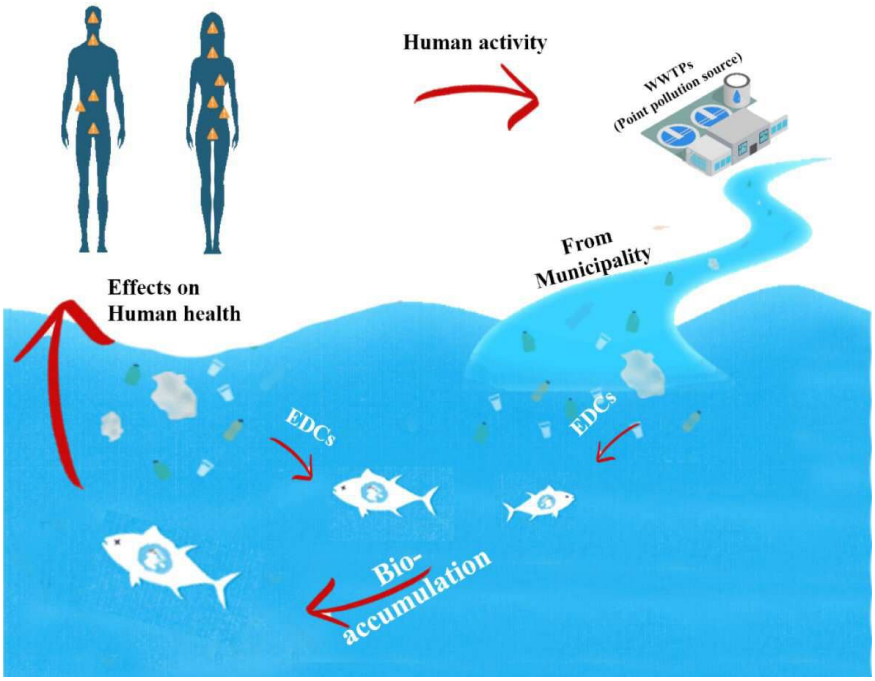
[0192] 앞서 설명된 일 실시 예에 따른 위해성 평가 모델 생성 방법 및 위해성 평가 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시 예에 따라 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광 기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0193] 앞선 실시예에 대한 설명에서 참조된 도면 각각은 설명의 편의를 위해 도시된 일 실시 예에 불과하며, 각 화면에 표시된 정보들의 항목, 내용과 이미지들은 다양한 형태로 변형되어 표시될 수 있다.

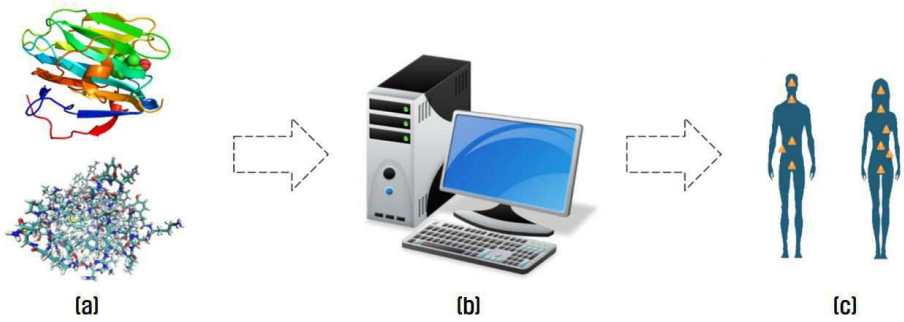
[0194] 본 발명은 도면에 도시된 일 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

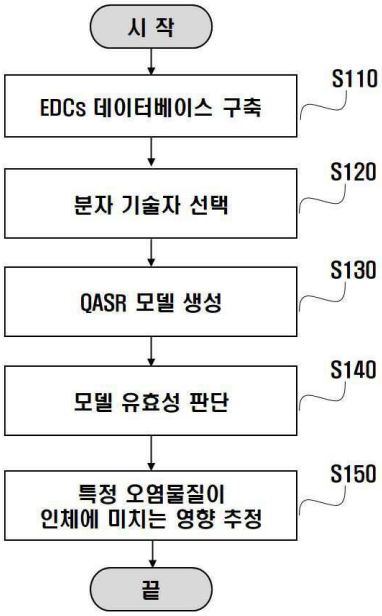
도면1a


Human activity
WWTPs
(Point pollution source)
From
Municipality
Effects on
Human health
EDCs
EDCs
Bio-
accumulation

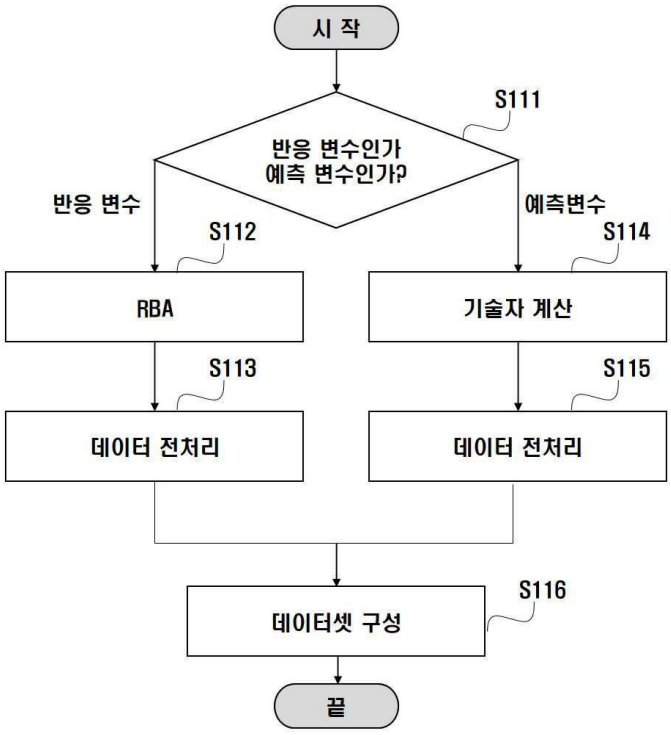
도면1b


(a)
(b)
(c)

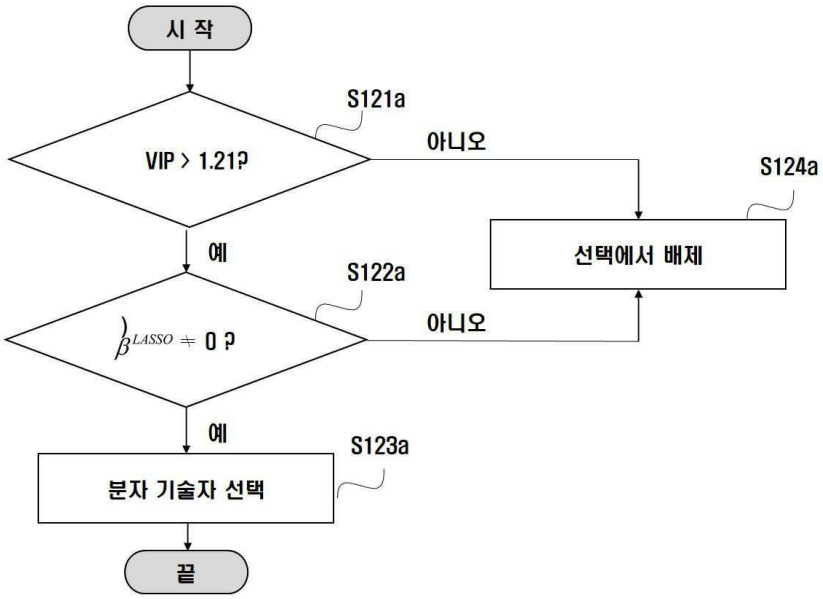
도면2


시 작
EDCs 데이터베이스 구축
S110
분자 기술자 선택
S120
QASR 모델 생성
S130
모델 유효성 판단
S140
특정 오염물질이
인체에 미치는 영향 추정
S150
끝

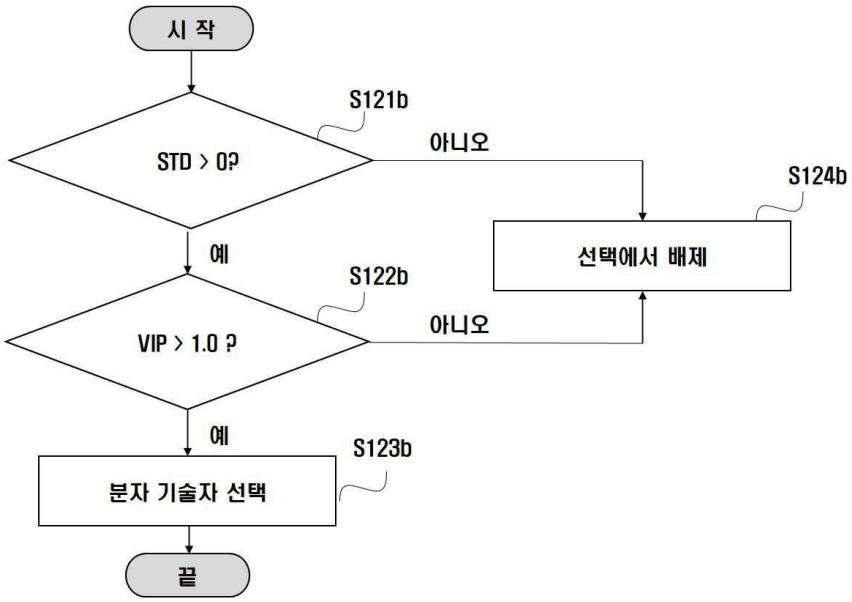
도면3


시 작
S111
반응 변수인가
예측 변수인가?
반응 변수
예측변수
S112
S114
RBA
기술자 계산
S113
S115
데이터 전처리
데이터 전처리
S116
데이터셋 구성
끝

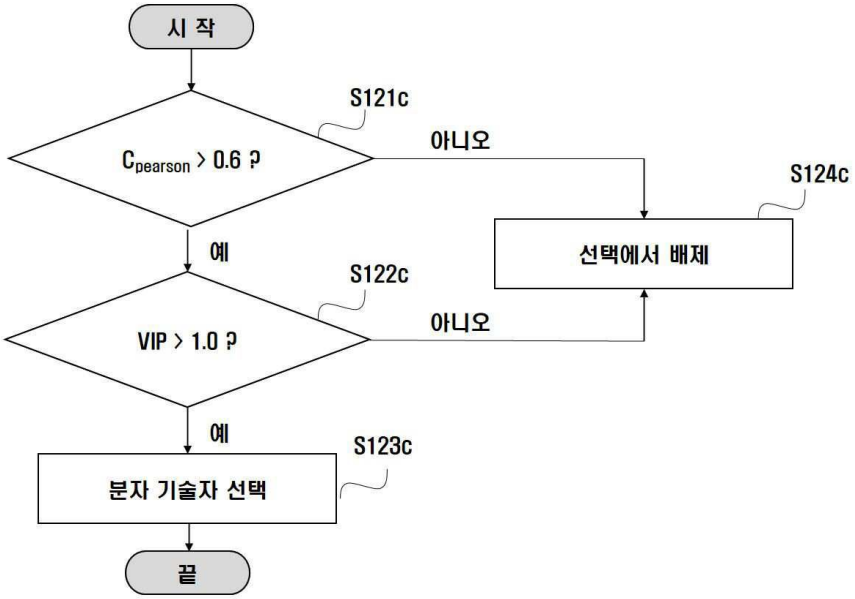
도면4


시 작
S121a
VIP > 1.21?
아니오
S124a
선택에서 배제
예
S122a
$\beta^{LASSO} \neq 0$?
아니오
예
S123a
분자 기술자 선택
끝

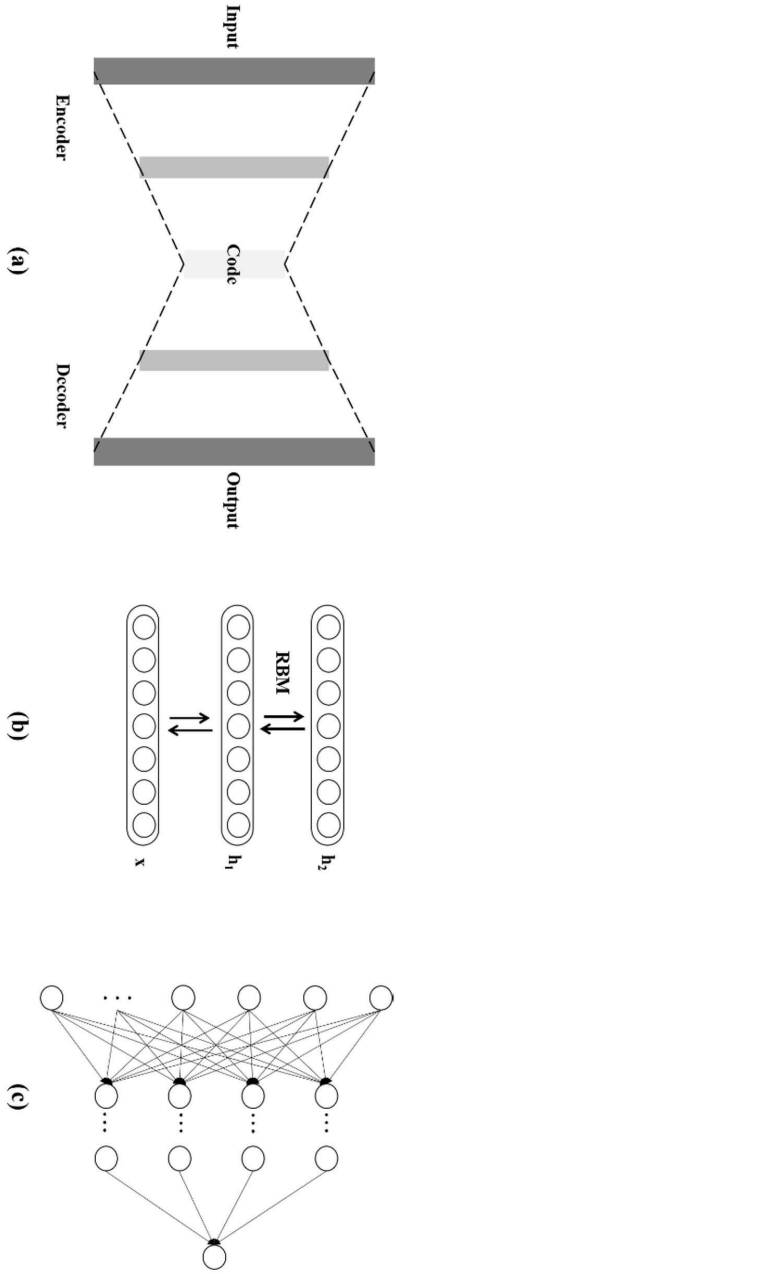
도면5


시 작
S121b
STD > 0?
아니오
S124b
선택에서 배제
예
S122b
VIP > 1.0 ?
아니오
예
S123b
분자 기술자 선택
끝

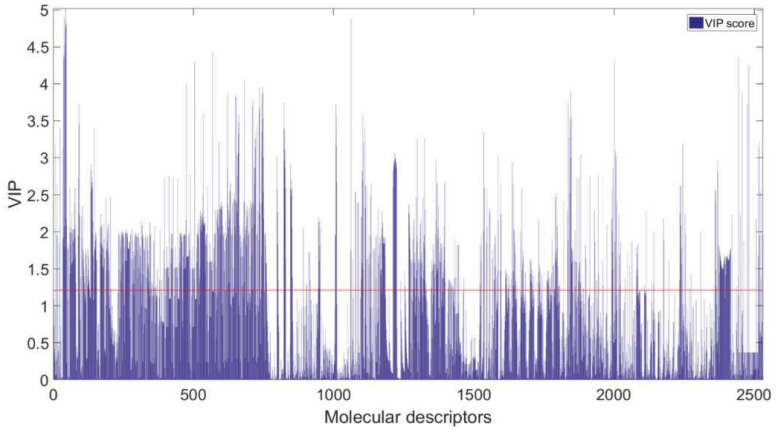
도면6


시 작
S121c
$C_{pearson}$ > 0.6 ?
아니오
S124c
선택에서 배제
예
S122c
VIP > 1.0 ?
아니오
예
S123c
분자 기술자 선택
끝

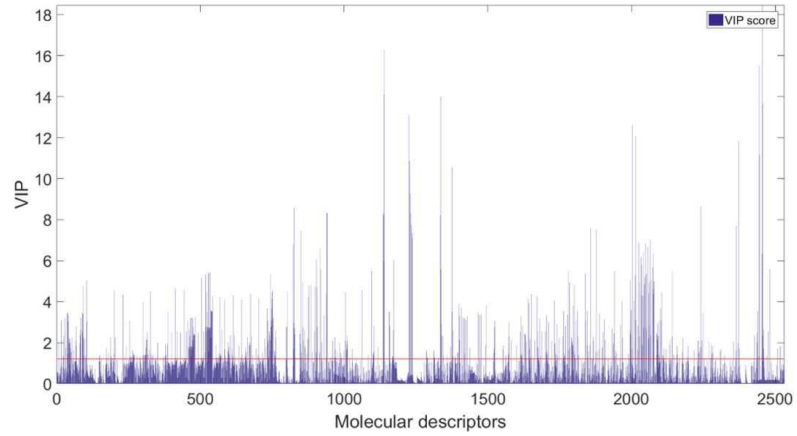
도면7


(a)
Input
Encoder
Code
Decoder
Output
(b)
RBM
x
h1
h2
(c)

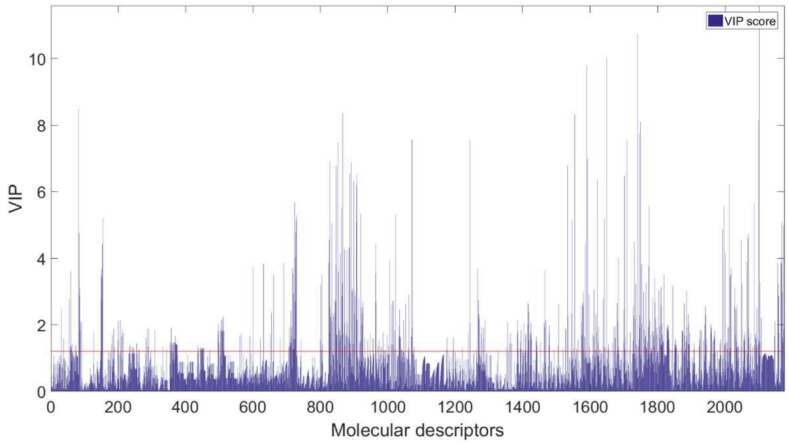
도면8a


VIP score
VIP
Molecular descriptors
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
500
1000
1500
2000
2500

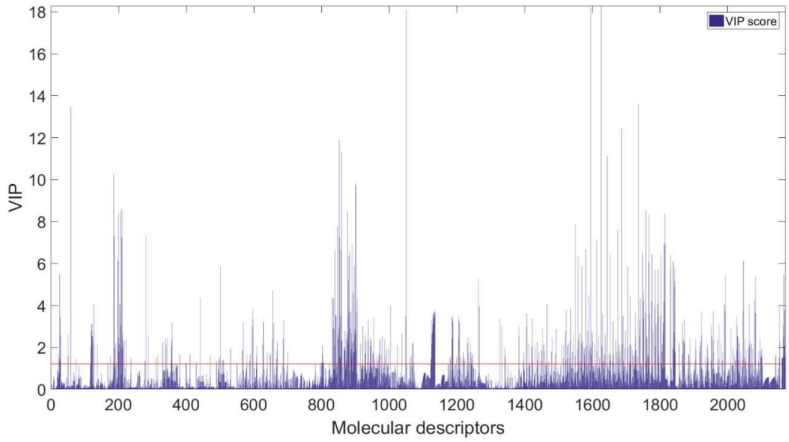
도면8b



도면8c



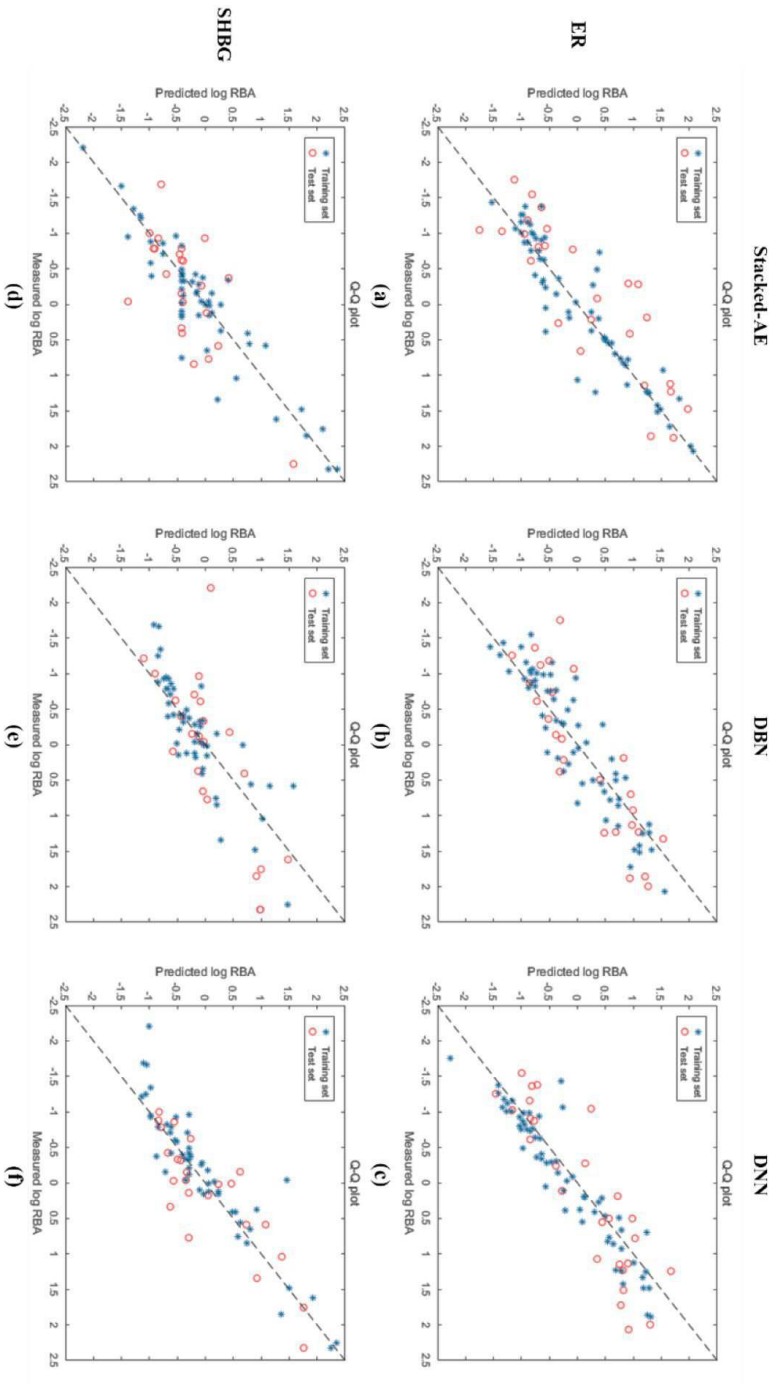
도면8d



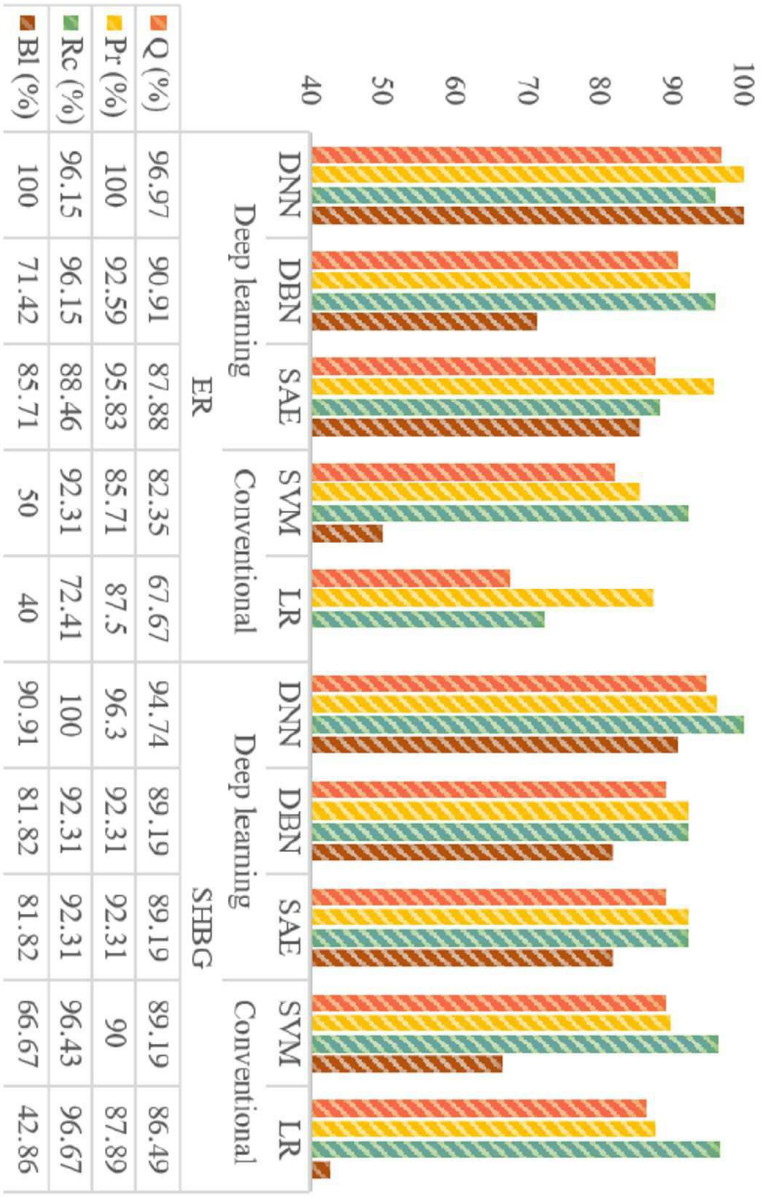
도면9

	Name	Block	Classification		Prediction	
			ER	SHBG	ER	SHBG
106	PJI2	Topological indices			○	
130	MAXDN	Topological indices				○
131	MAXDP	Topological indices				○
174	MPC05	Walk and path counts		○		
224	X3sol	Connectivity indices		○		
272	CIC2	Information indices		○		
349	VE3_L	2D-matrix-based descriptors		○		
356	ChiA_X	2D-matrix-based descriptors			○	
361	SpPosLog_X	2D-matrix-based descriptors		○		
364	SpDiam_X	2D-matrix-based descriptors				○
366	SpMAD_X	2D-matrix-based descriptors				○
443	J_D/Dt	2D-matrix-based descriptors			○	
446	SpPosA_D/Dt	2D-matrix-based descriptors			○	
515	EE_Dz(m)	2D-matrix-based descriptors		○		
538	SpPosA_Dz(v)	2D-matrix-based descriptors		○		
681	VR3_B(m)	2D-matrix-based descriptors		○		
750	SpPos_B(p)	2D-matrix-based descriptors		○		
755	SpDiam_B(p)	2D-matrix-based descriptors		○		
828	VE3_B(s)	2D-matrix-based descriptors		○		
903	ATSC8e	2D autocorrelations				○
922	ATSC3s	2D autocorrelations				○
934	MATS7m	2D autocorrelations				○
935	MATS8m	2D autocorrelations			○	
949	MATS6e	2D autocorrelations	○			○

도면10

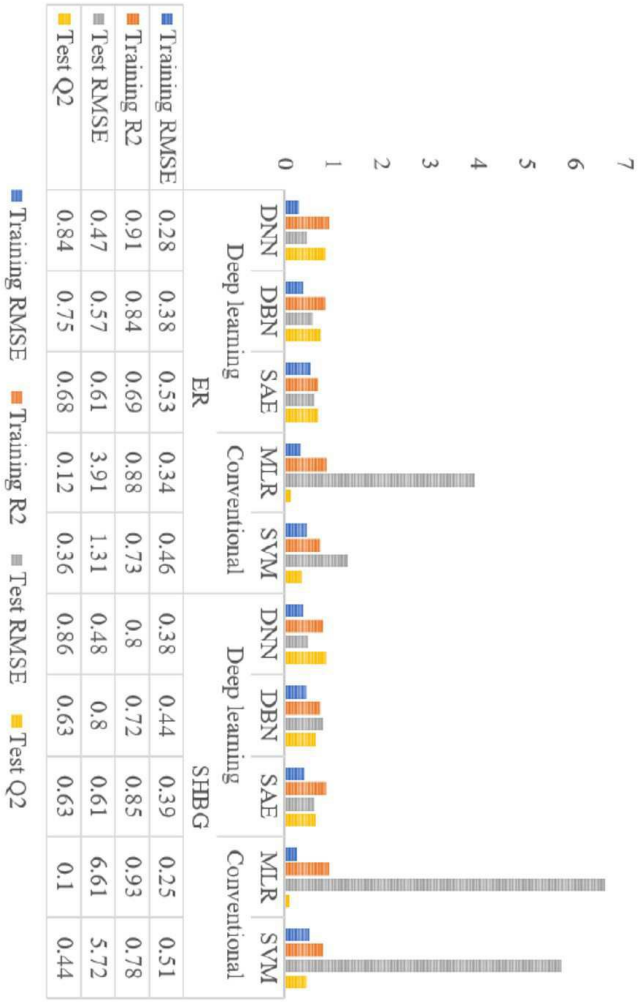

Stacked-AE
DBN
DNN
ER
SHBG
(a)
(b)
(c)
(d)
(e)
(f)
Q-Q plot
Predicted log RBA
Measured log RBA
Training set
Test set

도면11a



	ER						SHBG					
	Deep learning			Conventional			Deep learning			Conventional		
	DNN	DBN	SAE	SVM	LR		DNN	DBN	SAE	SVM	LR	
Q (%)	96.97	90.91	87.88	82.35	67.67		94.74	89.19	89.19	89.19	86.49	
Pr (%)	100	92.59	95.83	85.71	87.5		96.3	92.31	92.31	90	87.89	
Rc (%)	96.15	96.15	88.46	92.31	72.41		100	92.31	92.31	96.43	96.67	
Bl (%)	100	71.42	85.71	50	40		90.91	81.82	81.82	66.67	42.86	

도면11b



	ER					SHBG				
	Deep learning			Conventional		Deep learning			Conventional	
	DNN	DBN	SAE	MLR	SVM	DNN	DBN	SAE	MLR	SVM
Training RMSE	0.28	0.38	0.53	0.34	0.46	0.38	0.44	0.39	0.25	0.51
Training R2	0.91	0.84	0.69	0.88	0.73	0.8	0.72	0.85	0.93	0.78
Test RMSE	0.47	0.57	0.61	3.91	1.31	0.48	0.8	0.61	6.61	5.72
Test Q2	0.84	0.75	0.68	0.12	0.36	0.86	0.63	0.63	0.1	0.44