JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 11, pp. 1-11, November 2023
https://doi.org/10.9708/jksci.2023.28.11.001

Transfer Learning based DNN-SVM Hybrid Model for Breast Cancer Classification

Gui Rae Jo*, Beomsu Baek*, Young Soon Kim**, Dong Hoon Lim***

*Master's graduate, Dept. of Information and Statistics, Gyeongsang National University, Jinju, Korea
*Master's graduate, Dept. of Information and Statistics, Gyeongsang National University, Jinju, Korea
**Professor, Dept. of Information and Statistics, Dept. of Bio & Medical Bigdata (BK4 program), Gyeongsang National University, Jinju, Korea
***Professor, Dept. of Information and Statistics, RINS, Gyeongsang National University, Jinju, Korea

[Abstract]

Breast cancer is the disease that affects women the most worldwide. Due to the development of computer technology, the efficiency of machine learning has increased, and thus plays an important role in cancer detection and diagnosis. Deep learning is a field of machine learning technology based on an artificial neural network, and its performance has been rapidly improved in recent years, and its application range is expanding.

In this paper, we propose a DNN-SVM hybrid model that combines the structure of a deep neural network (DNN) based on transfer learning and a support vector machine (SVM) for breast cancer classification. The transfer learning-based proposed model is effective for small training data, has a fast learning speed, and can improve model performance by combining all the advantages of a single model, that is, DNN and SVM. To evaluate the performance of the proposed DNN-SVM Hybrid model, the performance test results with WOBC and WDBC breast cancer data provided by the UCI machine learning repository showed that the proposed model is superior to single models such as logistic regression, DNN, and SVM, and ensemble models such as random forest in various performance measures.

▸ **Key words**: Transfer learning, Deep learning, Breast cancer, DNN-SVM Hybrid model

• First Author: Gui Rae Jo, Corresponding Author: Young Soon Kim, Dong Hoon Lim
*Gui Rae Jo (cm0602@naver.com), Dept. of Information and Statistics, Gyeongsang National University
*Beomsu Baek (qorqjatn1388@gmail.com), Dept. of Information and Statistics, Gyeongsang National University
**Young Soon Kim (youngsoonkim@gnu.ac.kr), Dept. of Information and Statistics, Dept. of Bio & Medical Bigdata (BK4 program), Gyeongsang National University
***Dong Hoon Lim (dhlim@gnu.ac.kr), Dept. of Information and Statistics, RINS, Gyeongsang National University
• Received: 2023. 08. 21, Revised: 2023. 10. 16, Accepted: 2023. 10. 24.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

유방암은 전 세계적으로 여성들 대다수에게 가장 두려워하는 질환이다. 오늘날 데이터의 증가와 컴퓨팅 기술의 향상으로 머신러닝(machine learning)의 효율성이 증대되어 암 검출 및 진단 등에 중요한 역할을 하고 있다. 딥러닝(deep learning)은 인공신경망(artificial neural network, ANN)을 기반으로 하는 머신러닝 기술의 한 분야로 최근 여러 분야에서 성능이 급속도로 개선되어 활용 범위가 확대되고 있다.

본 연구에서는 유방암 분류를 위해 전이학습(transfer learning) 기반 DNN(Deep Neural Network)과 SVM(support vector machine)의 구조를 결합한 DNN-SVM Hybrid 모형을 제안한다. 전이학습 기반 제안된 모형은 적은 학습 데이터에도 효과적이고, 학습 속도도 빠르며, 단일모형, 즉 DNN과 SVM이 가지는 장점을 모두 활용 가능토록 결합함으로써 모형 성능이 개선되었다. 제안된 DNN-SVM Hybrid 모형의 성능평가를 위해 UCI 머신러닝 저장소에서 제공하는 WOBC와 WDBC 유방암 자료를 가지고 성능실험 결과, 제안된 모형은 여러 가지 성능 척도 면에서 단일모형인 로지스틱회귀모형, DNN, SVM 그리고 앙상블 모형인 랜덤 포레스트보다 우수함을 보였다.

▸ **주제어**: 전이학습, 딥러닝, 유방암, DNN-SVM Hybrid 모형

I. Introduction

통계청에서 발표한「2021 사망원인 통계」에 의하면, 우리나라 사망원인 1위는 암으로 전체 사망자의 26%를 차지하고 있다[1]. 이는 한국인 4명 중 1명은 암으로 사망한다는 것을 의미한다. 유방암은 국제암연구소(International Agency for Research on Cancer, IARC)가 펴낸 세계 암보고서 'GLOBOCAN 2020'에 따르면, 1년간 새로 유방암 진단을 받은 환자만 226만여 명에 달하고, 이 수치는 전체 여성암의 24.5%를 차지할 정도로 전 세계 여성에서 많이 발생한 암이다[2][3]. 한편, 같은 기간 동안 약 68만 명의 여성이 유방암으로 사망하여 전 세계 여성 암 사망원인의 15.5%를 차지하였으며, 유방암은 발생 빈도와 암 사망 원인 모두 1위를 차지하였다[4].

유방암은 유방 촬영술(mammography), 유방 초음파(ultrasonography), 미세침 흡인 검사(fine needle aspiration, FNA), 자기 공명 영상(magnetic resonance imaging;MRI) 등으로 진단할 수 있다[5][6]. 유방 촬영술은 유방암 검진에 있어 가장 기본적인 검사로 종양의 크기가 작을 경우 위음성률(false negative rate)이 높고 양성 종양(benign tumor)과 악성 종양(malignant tumor)의 감별이 쉽지 않을 뿐 아니라 X-선을 이용함으로써 방사선 피폭 위험이 있다. 유방 초음파 검사는 유방조직을 통과하는 고주파의 음파를 사용하여 유방질환을 진단하는 검사로 유방 촬영술에서 이상소견을 보이거나 치밀 유방 조직을 갖는 여성에게 사용되는 검사이다. 미세침 흡인 검사는 주사침을 병소 부위에 찔러서 조직이나 세포를 흡인하여 광학현미경으로 관찰하는 진단방법으로 진단이 빠르고 정확하며 검사비가 낮은 장점이 있으나 양성 종양인지 악성 종양인지 구별하는 것은 구조적인 유사성 때문에 어려운 일이다[7].

유방 자기 공명 영상은 유방 촬영술이나 유방 초음파 검사에 비해 민감도가 높고 방사선이 아닌 자기장을 이용하기 때문에 몸에 해롭지 않고 부작용이 없으나 검사 시간과 판독 시간이 길고 높은 검사 비용으로 검진에 적용하기에는 효과적이지 않다.

최근 인공지능의 한 분야인 머신러닝(machine learning)에 대한 효율성이 개선되면서 의료분야에서 질병 예측과 여러 질환의 진단에 활용되고 있다. 여기서 머신러닝이란 데이터를 기반으로 컴퓨터가 스스로 학습하여 규칙을 생성하도록 하는 기술을 말하는데 대표적인 분류 방법으로는 로지스틱회귀모형(logistic regression model), 인공 신경망(artificial neural network, ANN)과 서포트 벡터 머신(support vector machine, SVM)이 있다[8][9][10][11]. 로지스틱회귀모형은 종속변수가 범주형인 경우 사용하는 회귀분석의 확장형이고, 인공 신경망은 사람 두뇌의 신경세포를 모방하여 만든 알고리즘으로 주요 장점으로는 높은 예측 정확도와 확장성이 뛰어난 반면에 과대적합(over-fitting) 문제, 로컬 최적화(local optimization) 그리고 모형 해석의 어려움 등의 단점을 갖

고 있다. SVM은 서포트 벡터(support vector)와 초평면(hyperplane)을 이용하여 분류하는 방법으로 전통적인 기법들이 경험적 위험(empirical risk)을 최소화하는 것에 기초한 반면, 구조적 위험(structural risk)을 최소화하는 것에 기초하므로 과대적합 문제를 어느 정도 피할 수 있고, 또한, 적은 학습 데이터만으로 신속하게 분류학습을 수행할 수 있고 무엇보다도 글로벌 최적화(global optimization)를 얻을 수 있는 장점을 갖고 있다[12]. 최근에는 머신러닝보다 더 복잡한 데이터를 처리할 수 있는 딥러닝(deep learning)까지 등장하여 의료분야의 발전을 한층 더 가속화하고 있다. 가장 대표적인 딥러닝 모형에는 DNN(deep neural network), CNN(conventional neural network), RNN(recurrent neural network) 그리고 GAN(generative adversarial network) 등이 있다[13][14]. 여기서 DNN은 ANN 기법의 여러 문제점 해결을 통해 은닉층 수 확대로부터 학습의 결과를 향상시키는 모형이고, CNN 모형은 영상인식, 비디오 분석 등과 같은 영상처리 문제에 특화된 모형이고, RNN 모형은 순차 데이터 분석을 위한 모형이고 GAN 모형은 생성자(generator)와 판별자(discriminator)가 적대적 학습을 통해 실제 데이터와 비슷한 데이터를 생성하는 모형이다. 최근 의료 영상분석에서 많이 사용하는 딥러닝은 CNN 모형이며, 흉부와 유방 방사선 영상으로 시작하여 CT, MRI 등 다양한 의료영상데이터들을 대상으로 분류 모델이 개발되고 실제 의료 현장에서 사용되고 있다[15][16].

본 논문에서는 다루고자 하는 유방암 데이터는 영상 형태가 아닌 Tabular 데이터이다. 여기서 Tabular 데이터란 테이블(table) 형태의 행(row)과 열(column)로 표현되는 데이터를 말하며, 주로 정형 데이터(structured data)라고 한다. 하나의 행은 하나의 데이터 인스턴스를 나타내며, 각 열은 피처(feature)를 나타낸다. CNN에서 다루는 영상은 대표적인 비정형 데이터(unstructured data)이며, 이는 테이블 형식으로 표현할 수 없는 데이터를 말한다. CNN 모형은 영상의 공간정보를 유지한 채 학습이 이루어진다. 하지만 공간적인 관계가 없는 Tabular 데이터에 CNN 모형을 그대로 적용할 수 없다. 따라서 본 논문에서 유방암과 같은 Tabular 데이터 처리는 CNN 모형보다 DNN 모형을 사용하는 것이 더 적절하다[17][18][19].

DNN 모형은 기본적으로 다수의 은닉층을 포함하여 복잡한 비선형 관계(non-linear relationship)를 모델링할 수 있지만, 항상 좋은 성능을 얻을 수 있는 것은 아니다. DNN 모형은 모든 유형의 문제에 대한 만능 해결사가 아니며, 어떤 문제에 대해서는 전통적인 머신러닝, 여기서는 SVM과의 결합을 통해 모델을 개선시킬 수 있다. 영상분석을 위해 CNN 모형을 단일모형으로 많이 사용되지만, 성능 향상을 위해 전통적인 머신러닝, 예를 들면, SVM을 결합하는 Hybrid 형태의 모형 즉, CNN-SVM 모형들이 많이 사용되고 있다[20][21]. 여기서 CNN-SVM 모형에서는 특징 추출을 위해서 CNN 모형을 사용하고, 영상분류를 위해서 CNN의 FC 층(fully connected layer) 대신 SVM 모형을 사용한다. 본 논문에서는 Tabular 형태의 유방암 데이터에서 CNN 대신 DNN을 사용하여 DNN과 SVM 모형의 결합 형태인 DNN-SVM Hybrid 모형을 제안한다. 지금까지 DNN-SVM Hybrid 모형은 화자인식(speaker recognition), 음성인식(speech recognition), 네트워크 트래픽 분류(network traffic classification) 등에 사용되어 왔다[22][23][24]. 기존 DNN-SVM Hybrid 모형에서는 단지 DNN 모형의 마지막 층에서 다중 클래스 분류를 위해 소프트맥스 함수(softmax function) 대신 SVM를 사용하였다. 본 논문에서는 유방암 데이터에 대해 DNN 전체모형을 새로 학습시키지 않고 DNN 기반 전이학습 즉, 딥 전이학습(deep transfer learning)을 이용하여 일부 하위층(lower layer)은 전이하여 재사용하고, 상위층(upper layer)은 SVM 모형(optimal SVM model)을 이용하여 암의 악성 혹은 양성 여부를 분류함으로써 유방암과 같은 학습 데이터가 부족한 상황에서도 학습 성능을 높일 수 있는 장점이 있다. 여기서 전이학습이란 새로운 문제에 대해 사전 학습된 모형(pre-trained model)을 재사용하는 것을 말한다[25][26][27]. 제안된 DNN-SVM Hybrid 모형의 성능은 선택된 SVM 모형의 여러 파라미터에 의해 영향을 많이 받기 때문에 적절한 파라미터 선택은 매우 중요하다. 본 논문에서는 그리드 탐색법(grid search)을 사용하여 파라미터 최적화(parameter optimization)를 통해 얻어진 최적의 SVM 모형을 사용한다[28][29].

본 논문에서는 제안된 DNN-SVM Hybrid 모형의 성능을 평가하기 위해 UCI 머신러닝 저장소(machine learing repository)[30]에서 제공하는 두 개의 유방암 데이터인 WOBC (Wisconsin Original Breast Cancer) 와 WDBC (Wisconsin Diagnostic Breast Cancer)를 가지고 여러 가지 성능 척도 즉, 정확도(accuracy), 정밀도(precision), 재현율 (recall), 그리고 F1-스코어(F1-score)를 가지고 비교하고자 한다.

본 논문의 구성은 다음과 같다. 제2절에서는 DNN과 SVM에 대한 이론적 배경을 살펴보고 제3절에서는 DNN-SVM Hybrid 모형을 제안한다. 제4절에서는 성능실험을 위한 환경과 실험결과를 살펴보고 제5절에서 결론을 맺고자 한다.

II. Preliminaries

1. DNN

Figure 2.1은 입력층(input layer)-은닉층(hidden layer)-출력층(output layer) 구조에서 은닉층이 여러 개인 DNN 구조를 보여주고 있다.

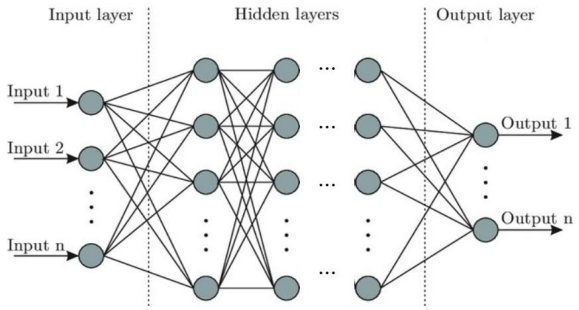

Fig. 2.1 Structure of DNN model

Figure 2.1의 DNN 구조는 층 $\ell = 0, 1, ..., L$에 대해 $\ell = 0$은 입력층, $\ell = L$는 출력층, 그리고 $\ell = 1, ...L-1$은 은닉층을 나타낸다. DNN은 주어진 데이터를 이용하여 학습 과정을 통해 찾을 모형의 파라미터(parameter)는 다음과 같다.

$$\boldsymbol{W} = \{W^{(0)}, W^{(1)}, ..., W^{(L)}\}$$
$$\boldsymbol{b} = \{b^{(1)}, b^{(2)}, ..., b^{(L)}\}$$

여기서 $W^{(\ell-1)}$는 층 $\ell-1$과 층 ℓ사이 가중치 행렬(weight matrix), $b^{(\ell)}$는 층 ℓ에 연결된 바이어스 벡터(bias vector)를 나타낸다. 즉, $W^{(\ell-1)}$과 $b^{(\ell)}$은 다음과 같다.

$$W^{(\ell-1)} = \begin{pmatrix} w_{1,1}^{(\ell-1)} & w_{1,2}^{(\ell-1)} & \cdots & w_{1,n}^{(\ell-1)} \\ w_{2,1}^{(\ell-1)} & w_{2,2}^{(\ell-1)} & \cdots & w_{2,n}^{(\ell-1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1}^{(\ell-1)} & w_{m,2}^{(\ell-1)} & \cdots & w_{m,n}^{(\ell-1)} \end{pmatrix}, \quad b^{(\ell)} = \begin{pmatrix} b_1^{(\ell)} \\ b_2^{(\ell)} \\ \vdots \\ b_m^{(\ell)} \end{pmatrix}$$

이때 $W_{i,j}^{(\ell-1)}$는 층 $\ell-1$에서 j번째 유닛과 층 ℓ에서 i번째 유닛을 연결하는 가중치이고 $b_i^{(\ell)}$는 층 ℓ에서 i번째 유닛의 바이어스(bias)이다.

DNN을 통해 학습할 N개의 훈련 데이터는 다음과 같다.

$$\{(x^1, y^1), (x^2, y^2), ..., (x^N, y^N)\}$$

우리는 DNN의 학습은 확률적 경사 하강법(stochastic gradient descent)을 이용하며, 이를 위해 사용되는 비용함수(cost function)는 식 (2.1)과 같이 나타낼 수 있다.

$$J(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{2}\|h_{W,b}(x^k) - y^k\|^2\right) + \frac{\lambda}{2}\sum_{\ell=1}^{L}\sum_{j=1}^{m}\sum_{i=1}^{n}(W_{i,j}^{(\ell)})^2 \quad (2.1)$$

여기서 첫 번째 항은 평균 제곱 오차(mean square error)를 나타내는 항이고, 두 번째 항은 정규화 패널티(regularization penalty)를 나타내고 λ은 정규화 파라미터로서 두 항의 상대적인 중요도를 조절한다. 그리고 비선형 활성화 함수 $h_{\boldsymbol{W},\boldsymbol{b}}(x)$는 다음과 같이 정의한다.

$$h_{\boldsymbol{W},\boldsymbol{b}}(x) = f(\boldsymbol{W}^{\boldsymbol{T}}x + \boldsymbol{b})$$

확률적 경사 하강법은 식 (2.1)에 주어진 비용함수를 최소화하는 파라미터 $\boldsymbol{W}, \boldsymbol{b}$을 다음과 같이 반복적인 방법에 의해 구한다.

$$W_{i,j}^{(\ell)} = W_{i,j}^{(\ell)} - \alpha\frac{\partial}{\partial W_{i,j}^{(\ell)}}J(\boldsymbol{W}, \boldsymbol{b})$$

$$b_i^{(\ell)} = b_i^{(\ell)} - \alpha\frac{\partial}{\partial b_i^{(\ell)}}J(\boldsymbol{W}, \boldsymbol{b})$$

여기서 α는 학습률을 나타낸다.

2. SVM

SVM은 1992년 Vladimir Vapnik[8]에 의해 개발된 통계적 학습이론으로, 뛰어난 일반화 성능을 보여주며 패턴 인식 분야에서 널리 사용되고 있다. SVM의 분류모형을 위한 학습 데이터의 집합 D가 다음과 같이 주어졌다고 가정한다.

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\},$$
$$y_i \in \{-1, +1\}$$

여기서 p차원 벡터 $\boldsymbol{x_i}$는 클래스 y_i에 대한 학습용 튜플(tuple)이다.

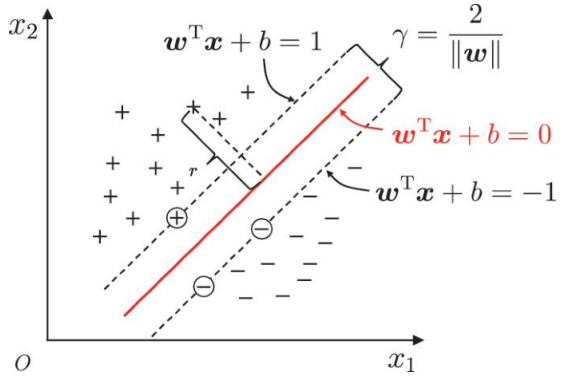

Fig. 2.2 SVM optimal separating hyperplane

SVM은 학습 데이터가 Figure 2.2과 같이 $y_i = +1$, $y_i = -1$로 표시되는 두 클래스를 분류하기 위한 최적 분리 초평면(optimal separating hyperplane)은 마진(margin)을 최대로 하는 서포트 벡터를 찾는 것이다. 이를 위해 임의의 선형 분리 초평면을 식 (2.2)와 같이 나타낼

수 있다.

$$d(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}+b = 0, \quad (2.2)$$

여기서 벡터 $\boldsymbol{w}$는 초평면에 수직인 정규벡터(normal vector)이다. 그리고 식 (2.2)와 평행인 초평면은 식 (2.3)과 식 (2.4)에 의해 표현할 수 있다.

$$d(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}+b = 1 \quad (2.3)$$

$$d(\boldsymbol{x}) = \boldsymbol{w}^{T}\boldsymbol{x}+b = -1 \quad (2.4)$$

학습 데이터가 선형적으로 분리 가능하면 두 개의 초평면 사이의 거리는 $\frac{2}{\| \boldsymbol{w} \|}$이므로 다음의 조건을 만족하는 $\|\boldsymbol{w}\|$을 최소화하는 초평면을 찾는다.

$$d(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}+b \geq 1 \quad \forall y_i = 1 \quad (2.5)$$

$$d(\boldsymbol{x}) = \boldsymbol{w}^{T}\boldsymbol{x}+b \leq -1 \quad \forall y_i = -1 \quad (2.6)$$

식 (2.5)과 식 (2.6)의 두 제약식을 식 (2.7)과 같이 하나의 제약식으로 표현할 수 있다.

$$y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}+b) \geq 1 \quad (2.7)$$

따라서 SVM은 식 (2.7)에 주어진 제약조건 하에서 $\frac{1}{2}\| \boldsymbol{w} \|^2$을 최소화하는 유일한 초평면을 찾는다. 즉, 수식으로 표현하면 다음과 같다.

$$\text{minimize} \quad \frac{1}{2}\| \boldsymbol{w} \|^2$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^{T}\boldsymbol{x}+b) \geq 1$$

학습데이터가 선형분리 불가능한 경우는 슬랙변수(slack variable) $\xi_i(\geq 0)$을 도입하여 다음과 같이 최적화 문제를 형식화할 수 있다.

$$\text{minimize} \quad Q(\boldsymbol{w},b,\boldsymbol{\xi}) = \frac{1}{2}\| \boldsymbol{w} \|^2 + C\sum_{i=1}^{n}\xi_i \quad (2.8)$$

$$\text{subject to} \quad \begin{cases} y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}-b) \geq 1-\xi_i, & i=1,\ldots,n \\ \xi_i \geq 0, & i=1,\ldots,n \end{cases}$$

여기서 C는 마진폭과 분류 오류 사이 타협점(trade-off)을 결정하는 모수이다.

SVM은 커널 함수(kernel function) $K(\boldsymbol{x_i},\boldsymbol{x_j})$을 도입하여 원래 공간에서 데이터를 고차원 공간으로 변환하여 식 (2.8)에 주어진 최적화 문제를 식 (2.9)와 같이 쌍대문제(dual problem)로 나타낼 수 있다.

$$\text{maximize} \quad Q(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i},\boldsymbol{x_j}) \quad (2.9)$$

$$\text{subject to} \quad \begin{cases} y_i\alpha_i = 0, & i=1,\ldots,n \\ C \geq \alpha_i \geq 0, & i=1,\ldots,n \end{cases}$$

여기서 α_i는 라그랑지 배수(Lagrange multiplier)이다. 결국, 입력공간에서 식 (2.10)의 비선형 결정함수를 이용하여 최적의 선형함수를 결정한다.

$$f(x) = sign(\sum_{ij}^{N}\alpha_i y_i K(\boldsymbol{x_i},\boldsymbol{x_j}) + b), \quad (2.10)$$

여기서 $sign(x)$는 x가 양수이면 1, 0이면 0, 음수이면 -1을 갖는 함수이다.

Table 2.1은 자주 사용되는 커널함수 $K(\boldsymbol{x_i},\boldsymbol{x_j})$을 나타내고 있다.

Table 2.1 Representative kernel functions

Kernel functions	Function expressions
linear	$K(\boldsymbol{x_i},\boldsymbol{x_j}) = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_j}$
polynomial	$K(\boldsymbol{x_i},\boldsymbol{x_j}) = (\gamma\boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_j}+r)^d$
radial basis function (RBF)	$K(\boldsymbol{x_i},\boldsymbol{x_j}) = \exp(-\gamma\|\boldsymbol{x_i}-\boldsymbol{x_j}\|^2)$
sigmoid	$K(\boldsymbol{x_i},\boldsymbol{x_j}) = \tanh(\gamma\boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_j}+r)$

여기서 γ, r 그리고 d는 커널의 형태를 결정하는 모수들이다.

III. Transfer Learning based DNN-SVM Hybrid Model

1. Deep Transfer Learning

딥 전이학습은 DNN 모형에서 이미 학습된 모형을 이용하여 새로운 데이터에 대해 적용하는 것으로, DNN 모형에서 하위층은 입력 데이터에 대한 공통적인 특징을 추출하고, 상위층으로 갈수록 점차 복잡한 정보를 추출하므로 새로운 데이터가 사전 학습시킨 데이터셋과 유사한 특성을 가질수록 전이학습의 효과는 크다. 예를 들면, 0 ~ 4의 숫자로 이루어진 MNIST 데이터셋에서 학습된 DNN 모형을 5 ~ 9의 숫자를 분류하는데 적용할 수 있는 것은 전이학습의 좋은 예이다[31][32][33][34].

Figure 3.1은 딥 전이학습 모형을 나타내고 있다.

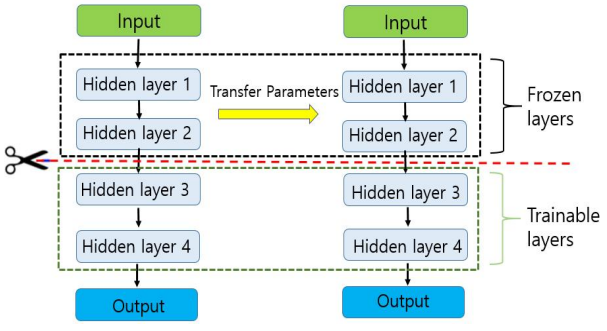

Fig. 3.1 Deep transfer learning model

Figure 3.1에서 입력층에 가까운 은닉층의 일부 즉, 모형의 하위층은 동결(freezing)하고, 나머지 은닉층 즉, 모형의 상위층만 학습이 이루어진다. Figure 3.1의 딥 전이학습에서 4가지 경우의 동결 수준, 즉, DNN[:1]에서부터 DNN[:4]까지를 생각할 수 있다. DNN[:1]은 첫 번째 층을 동결하고 나머지 3개의 층에서 학습이 이루어지고, DNN[:2]은 Figure 3.1에서 보는 것처럼 두 번째 층까지 동결하고 나머지 2개의 층에서 학습이 이루어지고, DNN[:3]은 세 번째 층까지 동결하고 나머지 1개의 층에서 학습이 이루어지고, DNN[:4]은 네 번째 층까지 모두 동결하는 경우이다. 여기서 일부 동결층(frozen layers)을 제외하고 학습이 가능한 상위층에서는 역전파 알고리즘을 통해 학습이 이루어진다. 이처럼 DNN의 하위층은 주어진 데이터에 대한 일반적인 학습 정보를 포함하고 있기 때문에 재사용하고 상위층은 주어진 데이터에 대한 특화된 패턴을 학습하기 때문에 학습이 이루어진다. 학습할 새로운 데이터의 종류가 사전 학습에 사용된 데이터와 특성이 매우 다른 경우, 모든 층을 재사용해서는 안되고, 일부 하위층 만을 재사용해야한다[35].

전이학습은 데이터 수집이 어려운 의료분야에 유용하게 활용될 수 있다. 의료 데이터는 민감한 개인정보로서 개인정보보호 규제를 받는다. 따라서 특정 질환 유병자의 데이터가 필요한데, 데이터 확보에 한계가 있는 경우, 유사한 질환에 대한 데이터의 양이 많거나 혹은 유사한 데이터로 학습한 모형이 존재하는 경우 전이학습이 이용될 수 있다[36][37].

2. Proposed DNN-SVM Hybrid Model

Figure 3.2는 전이학습 기반 DNN 모형과 SVM 모형의 결합 모형을 나타내고 있다.

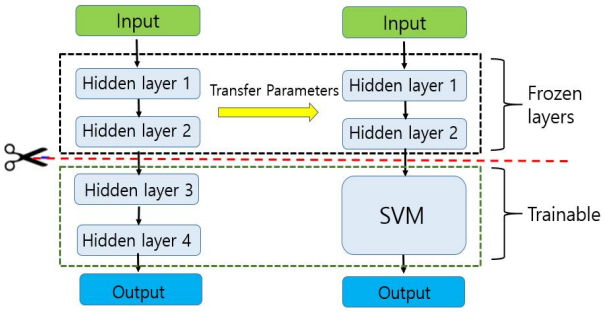

Fig. 3.2 Structure of DNN-SVM Hybrid model

Figure 3.2에서 보면, DNN에서 은닉층의 4개 중 하위층 2개를 동결하여 재사용하고, 나머지 은닉층 2개는 SVM으로 대체하여 학습한다. DNN-SVM Hybrid 모형을 단계별로 설명하면 다음과 같다.

STEP 1: 학습 데이터셋으로 DNN 모형을 학습한다. 여기서 학습 데이터셋은 WOBC, WDBC 데이터를 포함하여 UCI 저장소에 있는 4 종류의 유방암 데이터를 말한다[38].

STEP 2: 사전학습된 DNN 모형에서 새로운 학습 데이터셋을 순전파하여 특정 은닉층의 출력을 계산한다.

STEP 3: 은닉층의 출력을 SVM 모형의 입력으로 사용해 SVM 모형을 학습한다.

STEP 4: 테스트 데이터셋과 학습된 DNN, SVM을 통해 STEP 2, STEP 3과 같은 방식으로 결합 모형의 예측을 진행하고 평가한다.

따라서 DNN-SVM Hybrid 모형은 Figure 3.2과 같이 사전학습된 DNN 모형을 통한 특징 추출(feature extraction)하고, SVM 모형을 이용하여 최종 분류가 이루어진다.

본 논문에서 유방암 자료인 WOBC와 WDBC 자료를 가지고 양성 종양과 악성 종양을 분류하기 위해 사용한 DNN 모형은 총 4개의 은닉층으로 구성되고 각각의 은닉층 노드는 16개, 16개, 16개, 4개로 구성된다. 본 논문에서는 Figure 3.2과 같이 두 번째 은닉층의 출력을 SVM의 입력으로 사용하는 DNN-SVM Hybrid 모형 즉, DNN[:2]+SVM을 포함하여, 첫 번째, 세 번째, 네 번째 은닉층의 출력을 SVM의 입력으로 사용하는 DNN-SVM Hybrid 모형 즉, DNN[:1]+SVM, DNN[:3]+SVM, DNN[:4]+SVM을 고려하였다.

IV. Performance experiment and results

1. Performance experiment

1.1 Breast Cancer data

WOBC 데이터는 699명을 대상으로 조사되었고 Table 4.1과 같이 독립변수로는 FNA의 세포 특성을 나타내는 9개의 변수와 종속변수로는 이를 양성(benign)인지 악성(malignant)을 나타내는 클래스 변수로 구성되어 있다. 여기서 세포 특성값은 1부터 10까지 정수값을 가지고 그 값이 1에 가까울수록 양성일 확률이 높고 10에 가까울수록 악성일 확률이 높다는 것을 나타낸다.

Table 4.1. WOBC data description

Attribute	Domain
lump thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
marginal attachment	1-10
Single epithelial cell size	1-10
naked nucleus	1-10
soft chromatin	1-10
normal nucleolus	1-10
Mitoses	1-10
Class	2 = Benign, 4 = Malignant

WDBC 데이터는 569개 표본 수로 이루어져 있고 Table 4.2과 같이 독립변수는 10가지 세포의 특징들에 대한 각각의 평균, 표준편차, 최대값을 나타내는 총 30개의 변수로 구성되어 있다. 종속변수는 유방암의 양성과 악성을 나타내는 클래스 변수이다.

Table 4.2 WDBC data description

Attribute	Descriptions
Radius	Mean of the distance from the centroid to a point on the circumference
Texture	standard deviation of gray-scale values
Perimeter	nucleus circumference
Area	cell nucleus area
Smoothness	Local variations in radius length
Compactness	cell nucleus compactness, $\text{perimeter}^2/\text{area} - 1.0$
Concavity	Severity of contour concavity
Concave points	Number of concavities in the contour
Symmetry	cell nucleus shape
Fractal dimension	fractal dimension, "coastline approximation" - 1
Diagnosis	cancer diagnosis M=Malignant, B=Benign

1.2 K-Fold Cross-Validation

본 연구에서는 모형의 성능평가를 위해 전체 데이터를 train, validation, test 데이터로 3:1:1 비율로 분할하여 K-폴드 교차 검증(K-fold cross-validation)을 수행하였다[39]. Table 4.3은 K=5인 5-폴드 교차검증을 위한 데이터 구조를 보여주고 있다.

Table 4.3 5-Fold Cross-Validation

train	train	train	validation	test	$\rightarrow E_1$
train	train	train	test	validation	$\rightarrow E_2$
validation	train	test	train	train	$\rightarrow E_3$
train	test	validation	train	train	$\rightarrow E_4$
test	validation	train	train	train	$\rightarrow E_5$

최종적으로 5개의 폴드된 데이터를 train, validation, test를 위한 데이터로 번갈아 변경하면서 5가지의 경우의 수를 다 활용한 뒤, 이 5개의 예측 평가를 구했으면 이를 평균해서 K-폴드 평가 결과로 반영한다. 즉,

$$\text{Model Evaluation} = \frac{1}{5}\sum_{i=1}^{K=5} E_i$$

1.3 Parameters for DNN-SVM model

Table 4.4는 WOBC 데이터에서 DNN의 전이 파라미터(transfer parameters)를 나타내고 있다[40].

Table 4.4 DNN Transfer parameters for WOBC data

parameter	type
optimizer	SGD
activation function	ReLU
learning rate	0.001
batch size	32
weight initialization	Xavier
dropout rate	0.3

Table 4.5는 WOBC 데이터에서 SVM에서 사용된 커널과 최적의 파라미터를 나타내고 있다.

Table 4.5 SVM kernel and parameters for WOBC data

Model	Kernel	C	γ
DNN[:1]+SVM	RBF	0.1	1
DNN[:2]+SVM	RBF	0.1	1
DNN[:3]+SVM	RBF	0.1	0.01
DNN[:4]+SVM	RBF	0.1	0.01

Figure 4.1은 WOBC 데이터에서 에포크 500인 경우 모형의 정확도와 손실함수 그래프를 보여주고 있다.

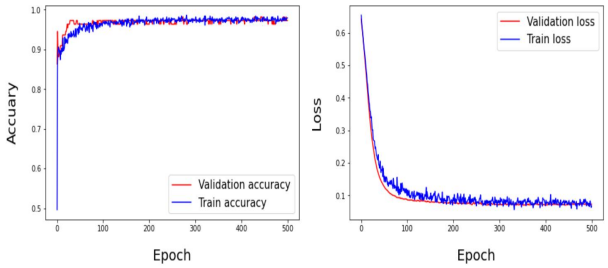

Fig. 4.1 Model accuracy and loss curve in WOBC data

Figure 4.1로부터 WOBC 데이터에서 DNN 모형은 에포크가 증가함에 따라 정확도와 손실함수 측면에서 학습이 잘 이루어진 것을 알 수 있다.

Table 4.6은 WDBC 데이터에서 DNN의 전이 파라미터를 나타내고 있다.

Table 4.6 DNN transfer parameters for WDBC data

parameter	type
optimizer	SGD
activation function	ReLU
learning rate	0.01
batch size	64
weight initialization	Xavier
dropout rate	0

Table 4.7은 WDBC 데이터에서 SVM에서 사용된 커널과 최적의 파라미터를 나타내고 있다.

Table 4.7 SVM kernel and parameters for WDBC data

Model	Kernel	C	γ
DNN[:1]+SVM	RBF	10	0.01
DNN[:2]+SVM	RBF	1	0.01
DNN[:3]+SVM	RBF	1	0.01
DNN[:4]+SVM	RBF	0.1	0.1

Figure 4.2는 WDBC 데이터에서 에포크 500인 경우 모형의 정확도와 손실함수 그래프를 보여주고 있다.

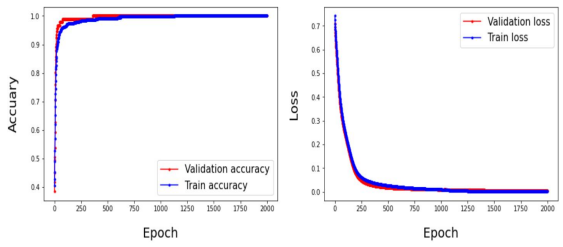

Fig. 4.2 Model accuracy and loss curve in WDBC data

Figure 4.2로부터 WDBC 데이터에서 DNN 모형은 에포크가 증가함에 따라 정확도와 손실함수 측면에서 학습이 잘 이루어진 것을 알 수 있다.

1.4 Model Evaluation Metrics

분류모형의 성능 척도는 정확도, 정밀도, 재현율, F1-score 등이 있다[41]. Table 4.8의 Confusion matrix로부터 평가지표가 얻어진다. Confusion matrix는 실제 데이터에서 실제 클래스(actual class)와 모형이 예측한 예측 클래스(predicted class)가 일치하는지를 갯수로 나타낸 표이다.

Table 4.8 Confusion matrix for Classification

		Predicted Class	
		Positive(1)	Negative(0)
Actual Class	Positive(1)	TP	FP
	Negative(0)	FN	TN

정확도는 전체 데이터 중에서 옳게 예측된 것의 비율을 의미한다. 즉, 수식으로 표현하면 다음과 같다.

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})}$$

정밀도는 Positive(양성)로 예측한 것 중 실제 Positive(양성)인 비율을 의미한다. 즉, 수식으로 표현하면 다음과 같다.

$$\text{precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

재현율은 실제 Positive인 것 중에 Positive로 예측한 것의 비율을 의미한다. 즉, 수식으로 표현하면 다음과 같다.

$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

F1-스코어는 정밀도와 재현율을 결합한 지표로서 둘 중 어느 한쪽에 치우치지 않았을 때 높은 값을 가진다. 즉, 수식으로 표현하면 다음과 같다.

$$\text{F1}-\text{score} = \frac{2\times \text{precision}\times \text{recall}}{(\text{precision}+\text{recall})}$$

2. Experiment results

여러 가지 DNN-SVM Hybrid 모형과 단일모형인 로지스틱회귀모형, DNN, SVM 그리고 랜덤 포레스트(Random Forest, RF)과 성능비교하고자 한다. 여기서 랜덤 포레스트는 의사결정나무 (decision tree)에 배깅(bagging)이라는 앙상블 학습(ensemble learning)을 적용한 모형이다[42].

Table 4.9는 WOBC 데이터에서 여러 모형들을 적용하여 얻은 성능 수치이다.

Table 4.9 Comparison of evaluation metrics with WOBC data

	Accuracy	Precision	Recall	F1-score
LOGISTIC	0.9692	0.9549	0.9591	0.9566
RF	0.9722	0.9518	0.9719	0.9611
SVM	0.9707	0.9536	0.9625	0.9577
DNN[:1]+SVM	0.9766	0.9790	0.9563	0.9667
DNN[:2]+SVM	0.9751	0.9923	0.9400	0.9651
DNN[:3]+SVM	0.9780	0.9836	0.9552	0.9691
DNN[:4]+SVM	0.9795	0.9961	0.9495	0.9718
DNN	0.9751	0.9836	0.9483	0.9654

Table 4.9에서 각 평가지표에서 최대값은 굵은 글자로 표시하였다. DNN-SVM Hybrid 모형은 모든 성능 척도에서 단일모형인 로지스틱 회귀모형, SVM과 DNN보다 좋은 수치를 보였고, 랜덤 포레스트와는 재현율을 제외한 나머지 3개의 척도에서 좋은 수치를 보였다. 특히, DNN[:4]+SVM 모형이 가장 좋은 성능 수치를 보였다.

Table 4.10은 WDBC 데이터에서 여러 모형들을 적용하여 얻은 성능 수치이다.

Table 4.10 Comparison of evaluation metrics with WDBC data

	Accuracy	Precision	Recall	F1-score
LOGISTIC	0.9666	0.9587	0.9529	0.9551
RF	0.9596	0.9578	0.9343	0.9450
SVM	0.9718	0.9528	0.9720	0.9619
DNN[:1]+SVM	0.9789	0.9671	0.9771	0.9716
DNN[:2]+SVM	0.9789	0.9671	0.9768	0.9716
DNN[:3]+SVM	0.9825	0.9671	0.9858	0.9761
DNN[:4]+SVM	0.9842	0.9623	0.9952	0.9784
DNN	0.9807	0.9576	0.9904	0.9735

Table 4.10에서 DNN-SVM Hybrid 모형은 단일모형인 로지스틱 회귀모형, SVM과 DNN 그리고 앙상블 모형인 랜덤 포레스트보다 모든 척도에서 좋은 수치를 보였다. 특히, DNN[:4]+SVM 모형이 가장 좋은 성능 수치를 보였다.

V. Conclusions

현재 유방암은 한국 여성에게 가장 많이 발생하는 암이다. 유방암을 조기 검진을 위해 X-선 검사로 하는 유방촬영술과 병행하여 초음파 검사가 가장 널리 사용되지만 이러한 방법으로 유방암을 정확하게 진단하기는 충분하지 않다. 최근 인공지능 특히, 딥러닝 기술을 이용해 유방암으로 오인되는 위양성(false positive) 진단을 획기적으로 줄이는 기술이 개발되고 있다.

본 논문에서는 유방암 분류를 위해 전이학습(transfer learning) 기반 DNN과 SVM의 구조를 결합한 DNN-SVM Hybrid 모형을 제안하였다. 딥러닝에게 학습시킬 암 환자 관련 의료 빅데이터의 중요성이 날로 높아지고 있다. 데이터가 부족하면 구현한 딥러닝 모형 성능도 떨어질 수밖에 없다. 전이학습은 의료분야의 학습 데이터가 부족한 경우 소규모 데이터만으로도 효율적인 학습 모형 생성을 가능하게한다.

DNN-SVM Hybrid 모형은 단일모형, 즉 DNN과 SVM의 단점을 극복하고, 이들 장점을 모두 활용할 수 있도록 결합을 통해 모형의 성능을 개선하였다.

UCI 머신러닝 저장소에서 제공하는 WOBC와 WDBC 유방암 자료를 가지고 성능 실험 결과, 제안된 DNN-SVM Hybrid 모형은 단일모형인 로지스틱회귀분석, DNN, SVM 그리고 앙상블 모형인 랜덤 포레스트보다 여러 성능 척도 면에서 우수한 것으로 나타났다.

REFERENCES

[1] Statistics Korea, “Causes of Death Statistics in 2021”, Korea National Statistical Office, September 2022.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”, CA: A Cancer Journal for Clinicians, Vol. 71, No. 3, pp. 209-249, February 2021. DOI:10.3322/caac.21660.

[3] I. Zemni, M. Kacem, W. Dhouib, C. Bennasrallah, R. Hadhri, H. Abroug, M. B. Fredj, M. Mokni, I. Bouanene, and A. S. Belguith, “Breast cancer incidence and predictions (Monastir, Tunisia: 2002-2030): A registry based study”, PLoS One. Vol. 17, No. 5, pp. 1-12, May 2022, DOI: 10.1371/journal.pone.0268035.

[4] Korea Central Cancer Registry, “Annual report of cancer statistics in Korea in 2020”, Ministry of Health and Welfare, May 2023.

[5] M. Sewak, P. Vaidya, C. C. Chan, and Z. H. Duan, “SVM approach to breast cancer classification”, International Multisymposium on Computer and Computational Sciences, 2007, pp. 32-37, DOI: 10.1109/IMSCCS.2007.46.

[6] M. Fiuzy, J. Haddadnia, N. Mollania, M. Hashemian, and K. Hassanpour, “Introduction of a New Diagnostic Method for Breast Cancer Based on Fine Needle Aspiration (FNA) Test Data and

Combining Intelligent Systems", Iranian journal of cancer prevention, Vol. 5, No. 4, pp. 169-177, Autumn 2012.

[7] K. Das, S. Conjeti, J. Chatterjee, and D. Sheet, "Detection of Breast Cancer From Whole Slide Histopathological Images Using Deep Multiple Instance CNN", IEEE Access, Vol. 8, No.20163105, November 2020. pp.213502-213511, DOI: 10.1109/ACCESS.2020.3040106.

[8] V. N. Vapnik, "The Nature of Statistical Learning Theory", John Wiley & Sons, New York, 1996.

[9] G. P. Zhang, "Neural Networks for Classification: A Survey", IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews, Vol. 30, No. 4, pp. 451-462, November 2000, DOI: 10.1109/5326.897072.

[10] S. Gupta, D. Kumar, and A. Sharma, "Date Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering, Vol. 2, No. 2, pp. 188-195, April 2011.

[11] Z. Khandezamin, M. Naderan, and M. J. Rashti. "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," Journal of Biomedical Informatics, Vol. 111, No. 103591, pp. 1-16, November 2020, DOI:10.1016/j.jbi.2020.103591.

[12] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," Pattern Recognition Letters, Vol. 141, pp. 61-67, January 2021, DOI:10.1016/j.patrec.2020.07.042.

[13] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," SN Computer Science, Vol. 2, No. 420, pp. 1-20, May 2021, DOI:10.1007/s42979-021-00815-1.

[14] L. Alzubaidi1, J. Zhang, A. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, Vol. 8, No. 53, pp. 1-20, March 2021, DOI:10.1186/s40537-021-00444-8.

[15] D. R. Sarvamangala1, and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," Evolutionary Intelligence, Vol. 15, pp. 1-22, January 2022, DOI:10.1007/s12065-020-00540-3.

[16] Z. Zhu, S. H. Wang, and Y. D. Zhang, "A Survey of Convolutional Neural Network in Breast Cancer," Computer Modeling in Engineering & Sciences, Vol. 136, No. 3, pp. 2127-2172, March 2023, DOI:10.32604/cmes.2023.025484.

[17] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, and R. L. Stevens, "Converting tabular data into images for deep learning with convolutional neural networks," Scientific Report, Vol. 11, No. 11325, pp. 1-11, May 2021, DOI:10.1038/s41598-021-90923-y.

[18] V. Borisov, T. Leemann, K. SeßlerR, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, pp. 1-21, December 2022, DOI:10.1109/TNNLS.2022.3229161.

[19] V. Borisov, K. Broelemann, E. Kasneci, and G. Kasneci, "DeepTLF: robust deep neural networks for heterogeneous tabular data," International Journal of Data Science and Analytics, Vol. 16, pp. 85-100, August 2022, DOI:10.1007/s41060-022-00350-z.

[20] J. Aurelia, and Z. Rustam, "A Hybrid Convolutional Neural Network-Support Vector Machine for X-ray Computed Tomography Images on Cancer," Open Access Macedonian Journal of Medical Sciences, Vol. 9, No. B, pp.1283-1289, 2021, DOI:10.3889/oamjms.2021.6955.

[21] D. Keerthana, V. Venugopal,, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," Biomedical Engineering Advances, Vol. 5, pp. 1-8, June 2023, DOI:10.1016/j.bea.2022.100069.

[22] S. X. Zhang, C. Liu, K. Yao, and Y. Gong, "Deep neural support vector machines for speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, April 2015, DOI: 10.1109/ICASSP.2015.7178777.

[23] V. Karthikeyan, S. S. Priyadharsini, K. Balamurugan, and M. Ramasamy, "Speaker identification using hybrid neural network support vector machine classifier," International Journal of Speech Technology, Vol. 25, No. 4, December 2022, pp 1041–1053, DOI:10.1007/s10772-021-09902-3.

[24] S. Ramraj, G. Usha, "Hybrid feature learning framework for the classification of encrypted network traffic," Connection Science, Vol. 35, No.1, April 2023, pp. 1-20, DOI: 10.1080/09540091.2023.2197172.

[25] M. Iman, K. Rasheed, and H. R. Arabnia "A Review of Deep Transfer Learning and Recent Advancements," Technologies, Vol. 11, No. 40, No. 1, January 2022, pp. 1-18, DOI: 10.48550/arXiv.2201.09679.

[26] S. Mondal, A. Chattopadhyay, A. Mukhopadhyay, and A. Ray, "Transfer learning of deep neural networks for predicting thermoacoustic instabilities in combustion systems," Energy and AI, Vol. 5, pp. 1-12, September 2021, DOI:10.1016/j.egyai.2021.100085.

[27] A. Ebbehoj, M. Ø. Thunbo, O. E. Andersen, M. V. Glindtvad, A. Hulman, "Transfer learning for non-image data in clinical research: A scoping review," PLOS Digit Health, Vol. 1, No. 2, No.1, February 2022, pp. 1-18, DOI: 10.1371/journal.pdig.0000014

[28] S. Kalyani, and K.S. Swarup, "Static security evaluation in power systems using multi-class SVM with different parameter selection methods," International Journal of Machine Learning and Computing, Vol. 1, No. 2, pp.193-198, June 2011, DOI:10.7763/

IJCTE.2013.V5.731.

[29] J. S. Lim, J. Y. Sohn, J. T. Sohn and D. H. Lim "Breast Cancer Classification Using Optimal Support Vector Machine," Journal of The Korea Society of Health Informatics and Statistics, Vol. 38, No. 1, 2013, pp. 108-121.

[30] UCI Machine Learning Repository. University of California, Center for Machine Learning and Intelligent Systems. http://archive.ics.uci.edu/ml/datasets.html [access on 1987].

[31] Y. Sawada and K. Kozuka, "Whole Layers Transfer Learning of Deep Neural Networks for a Small Scale Dataset", International Journal of Machine Learning and Computing, Vol. 6, No. 1, February 2016 DOI: 10.18178/ijmlc.2016.6.1.566.

[32] M. Swamynathan, "Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python", January 2019, Apress, DOI:10.1007/978-1-4842-2866-1.

[33] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent Backdoor Attacks on Deep Neural Networks", Proceedings of the 26th ACM Conference on Computer and Communications Security, pp. 2041–2055, November 2019, DOI: 10.1145/3319535.3354209.

[34] S. S. Tirumala, "Exploring Neural Network Layers for Knowledge Discovery", Procedia Computer Science, Vol. 193, pp. 173-182, November 2021, DOI: 10.1016/j.procs.2021.10.01710.1145/3319535.3354209.

[35] Y. Dar, L. Luzi, and R. G. Baraniuk, "Frozen Overparameterization:A Double Descent Perspective on Transfer Learning of Deep Neural Networks", Procedia Computer Science, Vol. 193, pp. 173-182, November 2021, DOI: 10.48550/arXiv.2211.11074.

[36] Z. Zhu, S. H. Wang, and Y. D. Zhang, "A Survey of Convolutional Neural Network in Breast Cancer," Computer Modeling in Engineering & Sciences, Vol. 136, No. 3, pp. 2127-2172, March 2023, DOI:10.32604/cmes.2023.025484.

[37] Y. H. Kwon, "Corroboration of Skin Diseases: Measuring the Severityof Vitiligo Using Transfer Learning", Journal of KIISE, Vol. 50, No. 1, pp. 72-79, January 2023, DOI: 10.5626/JOK.2023.50.1.72.

[38] A. Saifudin, T. Desyani, T. Desyani, and Y. Yulianti, "Bagging Techniques to Reduce Misclassification of Breast Cancer Prediction Base on Gradient Boosted Trees (GBT) Algorithm", Journal of Physics: Conference Series, Vol. 1477, No. 2, pp. 1-6, March 2020, DOI: 10.1088/1742-6596/1477/2/022029.

[39] Y. H. Kwon, "Machine Learning Developments in ROOT", Journal of Physics: Conference Series, 898, pp. 1-8, January 2017, DOI: 10.1088/1742-6596/898/7/072046.

[40] D. Soydaner, "A Comparison of Optimization Algorithms for Deep Learning," International Journal of Pattern Recognition and Artificial Intelligence, 34. pp. 1-26, July 2020, DOI: 10.48550/arXiv.2007.14166.

[41] S. Orozco-Arias, J. S. Pina, R. Tabares-Soto, L. F.Castillo-Ossa, R. Guyot, and G. Isaza, "Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements", Processes, Vol. 8, No. 6, pp. 1-19, May 2020, DOI:10.3390/pr8060638.

[42] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning", Sustainability, Vol. 14, No. 21, pp. 1-26, October 2022, DOI: 10.3390/su142113998.

Authors

Gui Rae Jo received the B.S. and M.S. degrees in Information Statistics from Gyeongsang National University, South Korea, respectively. He is currently working at Korea Kyongnam Taiyo Yuden Co.,Ltd.

He is interested in information statistics and big data analysis.

Beomsu Baek received the B.S degree in statistics from Gyeongsang National University, South Korea, in 2023. He currently pursuing M.S degree in statistics from Gyeongsang National University.

He is interested in machine learning, data mining, and big data analytics.

Young Soon Kim received the B.S., and M.S., and Ph.D. degree in statistics from Gyeongsang National University, South Korea, in 1994, 1998 and 2005, respectively. From 2009 to 2020, she was as senior researcher at GyeongNam Institute, South Korea, and she was a postdoctoral researcher and part-time Assistant Professor at Kennesaw State University from 2017 to 2019. Dr. Kim is currently an Assistant Professor with Department of Information and Statistics and Department of Bio & Medical Bigdata (BK4 program), Gyeongsang National University, South Korea. Her research interests include bioinformatics, machine learning, data mining, and big data analytics.

Dong Hoon Lim received the B.S. and M.S. degrees in Computer Science and Statistics, and Ph.D. degree in Statistics from Busan National University, Korea, respectively. Dr. Lim is currently a Professor in the Department of Information and Statistics, Gyeongsang National University. He is interested in AI, image processing, information statistics, and big data analysis.