pISSN 2005-8063
eISSN 2586-5854
2022. 12. 31.
Vol.14 No.4
pp. 35-43

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2022.14.4.035

Performance comparison on vocal cords disordered voice discrimination via machine learning methods*

Cheolwoo Jo[1,**] · Soo-Geun Wang[2] · Ickhwan Kwon[3]

[1]*School of Electrical, Electronics and Control Engineering, Changwon National University, Changwon, Korea*
[2]*Department of Otolaryngology, Pusan National University Hospital, Busan, Korea*
[3]*Department of Applied IT and Engineering, Pusan National University Hospital, Miryang, Korea*

Abstract

This paper studies how to improve the identification rate of laryngeal disability speech data by convolutional neural network (CNN) and machine learning ensemble learning methods. In general, the number of laryngeal dysfunction speech data is small, so even if identifiers are constructed by statistical methods, the phenomenon caused by overfitting depending on the training method can lead to a decrease the identification rate when exposed to external data. In this work, we try to combine results derived from CNN models and machine learning models with various accuracy in a multi-voting manner to ensure improved classification efficiency compared to the original trained models. The Pusan National University Hospital (PNUH) dataset was used to train and validate algorithms. The dataset contains normal voice and voice data of benign and malignant tumors. In the experiment, an attempt was made to distinguish between normal and benign tumors and malignant tumors. As a result of the experiment, the random forest method was found to be the best ensemble method and showed an identification rate of 85%.

Keywords: diagnosis, glottic cancer, vocal cords disorder, machine learning, convolutional neural network

1. 서론

본 논문에서는 후두 장애음성 식별기의 성능을 앙상블 방법에 의해 개선한 사례에 대해 보고하고자 한다. 후두 장애음성 식별기(identifier)는 후두 부위의 질환에 의해 음성의 변이가 발생한 경우 그 음성을 분석하여 질병의 유무를 식별하여 조기 진단하는 것을 목적으로 한다. 후두 장애음성의 구분과 식별에 관하여는 다양한 연구가 머신러닝의 다양한 방법을 수단으로 하여 수행되어 왔다(Al-Nasheri et al., 2017; Jo et al., 2001; Saldanha et al., 2014). 최근에는 인공신경망의 다양한 도구가 보급되고 공통 데이터 획득이 용이해짐에 따라 CNN(convolutional neural network)을 적용한 식별 사례들이 자주 보고되고 있다(Fang et al., 2019; Kim et al., 2020; Lee, 2021; Roy et al., 2019; Wu et al., 2018). 이러한 연구들은 주로 획득 가능한 장애음성 데이터를

* This research was supported by the Changwon National University in 2021-2022.
** cwjo@changwon.ac.kr, Corresponding author
Received 1 August 2022; Revised 2 November 2022; Accepted 4 November 2022

분류하기 위해 적절한 CNN 네트워크를 구성하고 분류의 정확도를 주어진 데이터 내에서 확인하는 과정으로 수행하고 있다. Hegde et al.(2019)은 CNN을 포함하여 다양한 기계학습 방법에 의해 진행된 장애음성 식별 및 진단 관련 연구를 조사하고 그 결과를 발표하였다. 이 문헌에 의하면 후두질환 분석에 사용된 기계학습 방법은 HMM(hidden markov model), MLP(multi-layered platron), SVM(support vector machine), ANN(artificial neural network) 등 다양하나, SVM과 ANN(CNN의 전 단계)이 주종을 이루고 있다. 문헌의 표 1에 의하면 지금까지의 연구들 중 cancer를 대상으로 한 경우로 분류된 사례는 5개 정도였으며, 모두 이 중 한 곳에서 공개 DB인 SVD(Saarbruecken DB)를 사용한 것으로 표기되었으나 문헌을 확인한 결과 cancer가 아닌 전 단계의 질병의 경우를 대상으로 한 것으로 확인되었고, 다른 곳은 모두 미공개 자체 수집 데이터를 사용하였다(Aicha, 2018).

CNN 등 통계적 방법에 의한 장애음성 식별에서는 중요한 부분인 관련 데이터의 확보가 중요한데, 사례는 자체 수집한 경우를 제외하면 대부분이 Massachusetts Eye and Ear Infirmary(1994)나 Saarbruecken Voice Database(2020)와 같은 공개된 데이터를 이용한 연구이다. 이러한 연구들은 관련 수행 성능을 비교해 볼 수 있다는 장점이 있는 반면 그 연구가 해당 데이터에 한정된 경우만으로 실험이 제한될 수밖에 없는 단점이 있기도 하다. 또한 공개 DB에는 악성 종양의 경우가 현저히 적다. 따라서 본 연구에서는 연구자들이 직접 수집한 PNUH(Pusan National University Hospital) 데이터 셋을 이용하여 연구를 수행한다.

후두 질환 음성 데이터의 특성상 관련 질환의 환자 확보가 대량으로 이루어지기 어렵기 때문에 본질적으로 그 수가 제한될 수밖에 없다. 따라서 식별기의 성능은 훈련에 사용된 데이터의 크기에 좌우될 수밖에 없다. 식별기 구성과 관련한 또 한 가지 문제는, 훈련에 사용된 데이터에 한정하여 높은 식별률을 보인다고 하더라도 훈련에 참여하지 않은 데이터에 대하여는 동등한 식별률을 보장할 수 없다는 문제점도 있다. 두 번째 문제는 근본적으로 장시간에 걸친 대량의 데이터 수집과 함께 여러 의료기관에서의 협력 등 식별기 구현과는 별도의 조치가 요구되어, 기존의 데이터를 이용한 식별률을 높이는 데 집중할 수밖에 없게 된다. 따라서 획득 가능한 범위에서 주어진 크기의 데이터 셋을 이용하여 어떻게 하면 최선의 식별 결과를 얻을 수 있는가가 연구의 방향이 된다.

이번 연구에서는 기존의 CNN 및 다양한 기계학습 방법을 이용한 식별기 구성 방식의 식별률을 높일 수 있는 방안으로, 각 식별기 훈련 과정에서 구해진 다양한 정확도의 CNN 모델 및 기계학습 방법을 통합하여 향상된 결과를 얻을 수 있는 방안을 도출하고자 한다. 또한 본 연구의 중점 사항으로 다양한 질병에 대한 분류가 아닌 정상 및 양성 질환과 악성 종양 간의 구분에 초점을 맞추어 수행하였다. 이는 의료 현장에서의 실질적인 적용 사례를 고려할 때 양성 질환 간을 구분하기보다는 악성질환의 조기 선별이 훨씬 더 필요하다는 판단에 따른 것이다. 그러므로 본 연구에서의 식별기 구성은 정상 및 양성 종양과 악성종양 간의 구분을 하는 데 중점을 두고 수행하였다.

2. 후두장애 음성 데이터

2.1. 음성 데이터 개요

실험에 사용한 음성은 부산대학교 병원 이비인후과에서 수집한 PNUH 데이터 셋으로 /a/ 음성 데이터를 포함한다. 정상 및 양성종양 221사례와 악성종양 146사례를 포함하고 있으며, 정상 및 6가지 부류의 양성 종양 그리고 악성종양의 사례가 포함되어 있다. 이 데이터 셋의 특징은 다른 데이터 셋과는 다르게 상대적으로 악성 질환의 데이터 비중이 높다. 본 연구에서도 통계적 방법을 적용하기에 부족한 수를 보완해줄 가능성을 염두에 두고 외부 데이터를 포함하여 적은 데이터 수를 보완하고자 하였으나, 검토 결과 악성종양의 경우 본 연구에 포함된 질병의 종류와 일치하는 경우가 거의 없어서 포함할 수 없었고, 이번 연구에서는 자체 데이터만으로 실험을 진행하였으므로 데이터를 얻기 어려운 실정이다. 표 1은 사용한 음성 데이터셋의 구성이다. 이 데이터 셋은 자체 데이터를 이용한 Kim et al(2020)의 사례(양성: 180, 악성: 45)에 비해 상대적으로 더 많은 데이터와 사례를 포함하고 있다. 데이터 셋에 포함된 질병의 종류는 대학병원에 내원하는 환자의 빈도를 고려하여 사례가 많은 경우로 한정하였다. 실험에 포함된 각 질병에 관해 간단히 요약하면 다음과 같다. Cyst는 낭종으로 일종의 혹이 성대 주변에 발생한 경우이고, palsy는 성대의 마비, edema는 부종, polyp은 용종, nodule은 성대결절을 말한다. 이들은 모두 성대 주변에 발생하여 음성의 비정상적인 변화를 초래하는 질병들이다.

표 1. 데이터셋의 구성
Table 1. Structure of voice dataset

Label: group		Number
Cancer		146
Non-cancer	Normal	49
	Cyst	16
	Palsy	45
	Edema	46
	Polyp	42
	Nodule	20
	Others	3
	Sub-total	221
Total		367

2.2. 음성 발화자

음성데이터는 1998년부터 2020년까지 부산대학교 병원 이비인후과에 내원하여 전문의로부터 질병명을 진단받은 사람을 대상으로 수집하였다. Cancer 데이터는 40세 이상의 남성으로부터 수집하였고 정상 데이터와 non-cancer 데이터도 40세 이상의 남성으로부터 수집하였다. 남성의 데이터만을 이용한 이유는 후두암의 경우 중년의 남성에게서 빈발하고 여성의 경우가 거의 없기 때문이다. 또한 여성의 경우 음성 특성이 달라 별도의 고려가 필요하므로 이번 실험에서는 제외하였다.

2.3. 수집 및 전처리

음성 녹음 및 분석의 장비는 Kay Computer Speech Lab(CSL) (Model 4300B and 4150B; KayPENTAX, Montvale, NJ, USA)을 사용하였다. 음성 신호는 16 bit, 44.1 kHz로 표본화하였고 Shure-Prolog 마이크로폰을 이용하여 화자가 10–15 cm 떨어진 곳에 위치하도록 하여 수집하였다. 발성 시 화자에게 편안한 상태에서 /a/ 음성을 4초 이상 발성하도록 요청하였다. 수집된 음성은 이비인후과 전문의의 검토를 거쳐 유효성을 확인하였다. 분석은 매 20 ms 구간을 대상으로 40차의 MFCC(mel-frequency cepstral coefficient) 분석을 행하여 구간당 40개의 MFCC 계수를 구하였다. 분석에는 Tensorflow(2021)와 음성 분석용 Python라이브러리 Librosa(2021) 패키지를 사용하였다.

3. 식별기의 구성

3.1. 기본 CNN 식별기

본 연구에서는 CNN을 기본 식별기로 사용하였고 사용한 CNN 모델의 아키텍처는 그림 1에 요약되어 있다. 기존의 연구들을 조사한 문헌(Hegde et al., 2019)에서 후두 질환을 포함한 다양한 경우에 대한 식별기로 SVM, HMM, GMM(Gaussian maixture model), CNN 등 여러 가지 방법이 시도되었으나, 사용한 데이터의 종류나 그 중에서 선별한 질환의 종류 등이 모두 달라 절대적인 식별률로 식별기 성능의 우열을 가리기는 어렵다. 최근 후두 질환을 음성을 통해 판별한 유사 연구로는 Kim(2021)이 자체 데이터 셋을 이용하여 여러 가지 방법을 시도하고, 그중 1D-CNN을 이용하여 85%의 민감도를 갖는 분류기를 구성한 결과를 보고한 바가 있다. 본 연구에서는 최근 각광을 받고 있는 CNN 모델에 의한 식별을 시도하였다. 실험에 사용한 2D-CNN은 그림 1에 보인 바와 같이 4층 구조로 이루어져 있다. 이 모델은 MFCC 분석된 자료로부터 받은 스펙트로그램 이미지를 40×40 크기로 순차적으로 입력으로 받고 3개의 컨볼루션 레이어, 2개의 완전히 연결된 레이어, 1개의 출력 레이어를 포함한다. 제1 컨볼루션 층은 크기 3×3의 필터 16개, 그 다음에 배치 정규화 층을 갖는다. 제2 및 제3 컨볼루션 층은 각각 크기 3×3의 필터 32개 및 64개를 갖는다. 마지막 층은 평탄화를 거쳐 128개의 뉴런과 64개의 뉴런으로 완전히 연결된 두 개의 층을 제공한다. 최종 출력층은 시그모이드 활성화 기능을 가진 이진 분류를 위한 단일 뉴런이다. CNN 훈련 시 loss 척도는 categorical_crossentropy를, 최적화는 adm 옵션을 사용하였다. CNN의 훈련에는 실험에 의해 최대 100회의 epoch를 설정하였으며 주어진 데이터에 대해 충분히 수렴하는 결과를 얻을 수 있었다. 그림 2는 개별 CNN 훈련과정 및 검증 정확도와 검증 손실에 대한 추이를 나타낸 그래프이다. 그림 2에 의해 개별 CNN이 데이터에 적합하게 충분히 훈련되었음을 확인할 수 있다.

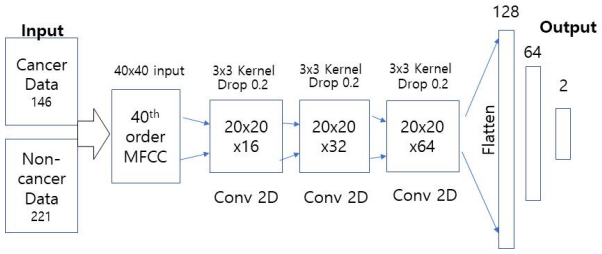
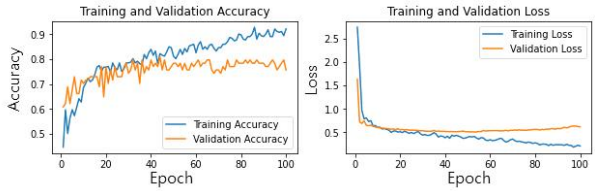
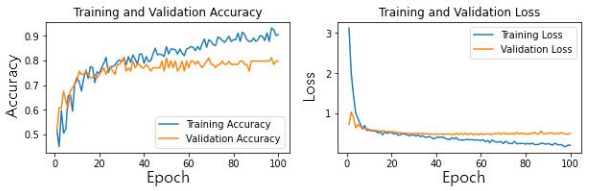


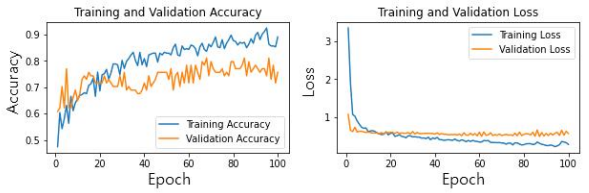
그림 1. CNN 식별기의 구조
Figure 1. Structure of CNN identifier



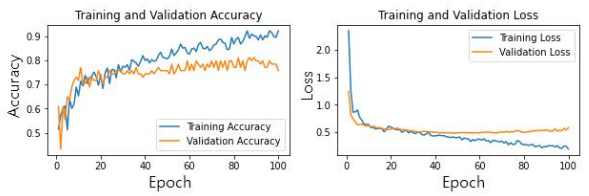
(a) CNN1 훈련 시 정확도 추이 (b) CNN1훈련 시 검증 손실 추이



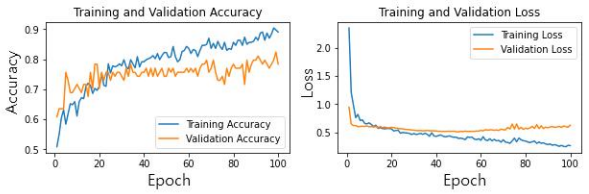
(c) CNN2 훈련 시 정확도 추이 (d) CNN2훈련 시 검증 손실추이



(e) CNN3 훈련 시 정확도 추이 (f) CNN3 훈련 시 검증 손실 추이



(g) CNN4 훈련 시 정확도 추이 (h) CNN4 훈련 시 검증 손실 추이



(i) CNN5훈련 시 정확도 추이 (j) CNN5훈련 시 검증 손실 추이

그림 2. 각 CNN훈련 검증 정확도와 훈련 검증 손실 추이 그래프
Figure 2. Training validation accuracy and training validation loss graph of each CNN model

이러한 CNN을 통한 훈련의 경우 성대 장애음성과 같이 훈련

에 사용된 데이터의 양이 적을 경우는 훈련 시기마다 다른 성능의 모델을 얻게 되며 주어진 전체 데이터 셋의 랜덤한 부분 집합에 의해 훈련이 진행되기 때문에 적은 수의 데이터에 과적합되도록 훈련된 모델이 생성되게 된다. 동일한 난수 발생기에 의해 동일하게 훈련될 경우는 앙상블 효과를 기대하기 어려울 수 있기 때문에 훈련 시기마다 다른 조합에 의한 다른 식별률을 갖는 모델을 생성하도록 하였다. 이 경우 훈련 모델에 따라 다른 식별 결과 또는 식별률이 나올 수 있다. 일반적으로는 주어진 데이터에 대해 여러 번의 훈련을 실시한 뒤 가장 최선의 결과를 보여주는 모델을 택하나 장애음성 데이터 셋과 같이 소규모 데이터를 적용하는 경우에는 경우에 따라서 최종 훈련된 모델과 다른 모델 사이에 식별되는 결과가 달라질 수도 있다는 문제점을 발견하였다. 이는 훈련 시 발생하는 과적합에 의한 영향으로 판단된다. 이와 같은 단점을 극복하기 위해 여러 가지 정확도를 갖도록 훈련된 모델을 생성하고 이 모델들로부터 도출된 결과를 융합하여 향상된 분류 결과를 얻고자 시도하였다.

본 논문에서 사용한 방법은 다양한 정확도로 훈련된 모델을 동시에 통과시켜 개별 모델로부터 식별 결과를 얻고 이들을 통해 다수 판정을 받은 결과를 최종 결과로 정하는 방법이다.

3.2. 앙상블 식별기

식별기의 성능 향상을 위한 앙상블 방법은 융합이 발생하는 방법과 단계에 따라 특성 융합, 데이터 단계 융합, 식별기 단계 융합의 세 가지로 구분할 수 있다(Bezdek et al., 2005). 이 중 특성 융합과 데이터 단계의 융합은 특성 파라미터 또는 상이한 데이터를 통한 융합이 필요할 때 적용할 수 있으므로, 본 연구에서 적합한 방법으로는 CNN 및 여러 가지 기계학습 식별기의 융합 및 기존의 앙상블 알고리즘 모델에 의한 접근을 시도하였다.

앙상블 방법에 의한 CNN의 성능 개선 시도는 최근 다양한 영역에서 다양한 방식으로 시도되고 있다. 영상 식별 영역에서는 Jung et al.(2020), Ko et al.(2019), Szmurlo & Osowski(2021) 등의 사례가 있고, hard voting 방식에 의한 앙상블 수행 사례는 Morvant et al.(2014), Liu et al.(2020), Lv et al.(2018), Su et al.(2019) 등이 있고 soft voting 방식에 의한 앙상블 사례로 Jeon et al.(2021)의 사례가 있다. 이들 사례를 통해 사용 데이터에 적절하도록 개별 훈련된 CNN 모델의 결과를 융합할 경우 수행 능력이 개선됨을 알 수 있다.

또한 기존의 단일 방법에 의한 한계를 극복하기 위해 학습방법을 개선하거나 여러 가지 학습법을 결합한 앙상블 학습법을 사용한다. 그림 3은 CNN에 의한 앙상블 학습 구조를 나타낸 것이다. 개별 CNN은 랜덤 훈련에 의해 서로 다른 식별률을 갖도록 훈련되고 각각으로부터 도출된 결과는 voting에 의해 다수 득표한 출력으로 최종 결정된다.

또한 기존의 기계학습 방법과 이들의 조합에 의한 앙상블 방법도 시도하였다. 기본 기계학습법으로는 SVM, KNN(K-nearest neighbor), DT(decision tree) 방법을 적용하였다. 또한 기존에 기계학습 분야에서 널리 사용되고 있는 앙상블 방법 중 bagging, random forest, adaboost, gradient boosting 방법과 아울러 기본 기계학습의 결과를 voting에 의해 결합한 방법을 시도하였다. 그림 4는 실험에 사용한 기계학습 방법들과 앙상블 학습 구조를 나타낸 것이다. 앙상블에 의한 성능 향상 시도는 여러 가지 다른 유형의 식별기를 결합하는 방법과 개선된 학습 알고리즘을 이용하는 경우가 있다. 여기서는 SVM, KNN, decision tree의 3가지 학습법에 대해 개별 학습과 앙상블 학습을 진행하였다.

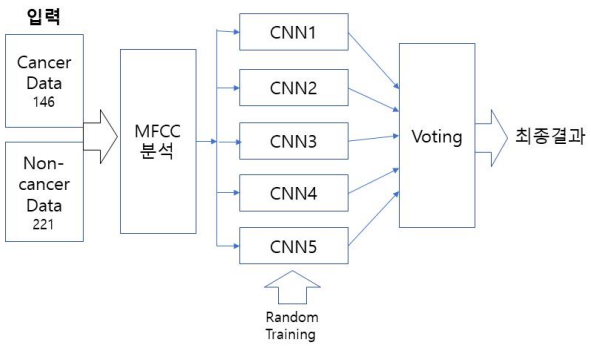


그림 3. 다수 투표에 의한 CNN 앙상블 학습 구조
Figure 3. Ensemble structure of majorityi-voting for CNN

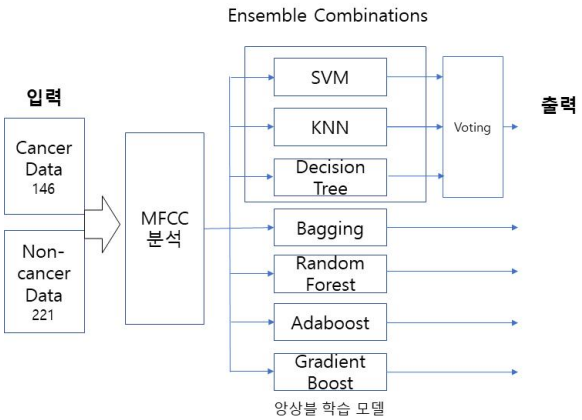


그림 4. 다양한 앙상블 기계학습 모델과 투표를 이용한 식별 실험 셋업
Figure 4. Experimental setup for voting and various ensemble machine learning models

본 연구에서 적용한 식별기 간의 앙상블은 다음과 같은 과정을 통해 구성하였다. 먼저 MFCC 분석한 데이터 셋을 이용하여 랜덤 시드에 의해 서로 다른 성능을 갖도록 훈련된 5개의 CNN 분류기 모델을 구성한다. 5개의 분류기 모델을 통해 입력 데이터 셋을 분류하고 그 결과를 앙상블 모델의 입력으로 사용한다. 또한 다양한 기계학습 모델로부터 결과를 얻고 이들을 결합한 식별기를 구성한 뒤 CNN 앙상블의 결과와 비교하여 최적의 결과를 내는 식별기를 도출하고자 한다.

여러 종류의 분류기로부터 얻은 결과를 융합하는 여러가지 방법은 Ruta & Gabrys(2000)의 문헌에서 소개하고 있다. 본 논문에서 사용한 앙상블 방법은 voting 방법으로 hard voting을 이용한다.

Hard voting 방법에 의한 앙상블 과정은 다음과 같다. Hard

voting 기법으로서 다수 투표에 의한 클래스 선택 방법은 다음과 같이 정의할 수 있다(Ruta & Gabrys, 2000). N개의 차원을 갖는 결정 벡터 d를 $d=[d_1,d_2,...,d_n]^T$와 같이 정의한다. 여기서 $d_i \in \{c_1,c_{2,...,}c_m,r\}$이며 c_i는 i번째 클래스의 레이블이고 r은 어떤 클래스에도 속하지 않을 경우의 선택값이다. J번째 클래스를 선택하는 함수 B는 다음과 같이 정의할 수 있다.

$$B_j(c_i)=\begin{cases}1 & \text{if } d_j=c_j\\ 0 & \text{if } d_j\neq c_j\end{cases} \quad (1)$$

여기서 일반적인 voting 함수 $E(d)$는 다음과 같이 정의된다.

$$E(d)=\begin{cases}c_i & \text{if } \forall_{t\in\{1,...,m\}}\sum_{j=1}^{n}B_j(c_i)\leq\sum_{j=1}^{n}B_j(c_i)\geq\alpha\cdot m+k(d)\\ r & otherwise\end{cases} \quad (2)$$

여기서 α는 다수의 비율을 지정하는 파라미터이고 $k(d)$는 voting의 제약함수이다. 만약 $k(d)$=0이고 α=1이라면 전체 분류기를 모두 택하는 경우가 된다. 만약 α=0.5라면 과반수의 결정에 따라 전체 분류가 결정되는 경우가 된다. 본 실험에서는 $k(d)$=0으로 두고 α값을 0.5 이상으로 두어 3개(α=0.5), 4개(α=0.7), 5개(α=1)값이 일치하는 경우에 대해 식별 결과를 구하였다.

실험에 포함한 식별기 조합에서 개별 모델의 출력을 hard voting의 다수 득표 전략을 적용하여 3개 이상이 일치한 경우 결과를 택하는 것으로 하였다.

3.3. 실험 환경

CNN과 각 머신러닝 방법에 의한 식별 실험을 위해 Tensorflow (2021)(ver.2.9.2)와 Scikit learn(2022)(ver.1.0.2)의 함수들을 사용하였고 다음과 같이 데이터 및 식별기를 설정하였다.

CNN 파라미터:
Input MFCC spectrogram: 40×1
Input window size: 40×40
Activation function: ReLU
Kernel size: 2
Optimizer Adam
Epoch 100
Loss function categorical cross-entropy
Dropout 0.2
Pooling window Maxpooling(2,2)
N0. of Layers 4
No. of Batch Size 64
No. of Epochs 100

기계학습 파라미터:
SVM: kernel rbf
Decision Tree : max depth 4
Adm: n_estimator=5, Random state = 42
Random Forest : n_estimator=20
Bagging: estimator SVM
cross_validation: k_fold=4

훈련을 위한 데이터는 훈련:검증:시험을 6:2:2의 비율로 적용하였다. 소량의 데이터인 단점을 보완하기 위하여 기계학습 알고리즘의 경우 4-fold cross validation을 적용하였다.

4. 실험 결과

개별 분류기에 의한 훈련 결과는 다음과 같다. 그림 5는 각 CNN 식별기에 의한 시험 음성의 시험 데이터에 따른 혼동행렬을 표시한다. CNN 모델별로 non-cancer와 cancer로 구분하였다. 가로축은 입력, 세로축은 모델의 정확도와 손실을 나타낸다. 각 모델은 CNN을 훈련할 때 동일한 구조에 대해 서로 다른 정확도를 갖도록 훈련한 것을 구분하기 위해 (A, B, C, D, E)와 같이 이름을 붙였다. 표 2는 각 분류기 및 앙상블 방법에 의한 최종 분류 결과에 대한 정확도값을 비교한 것이다. 모델에 따라 정확도가 달라지며 앙상블 방법인 voting에 의해 구해진 결과의 파라미터값이 향상된 것을 알 수 있다.

각 분류기의 비교 지표로는 정확도(accuracy), 특이도(specificity), 정밀도(precision), 민감도(sensitivity)를 사용한다. 정확도(accuracy)는 전체 데이터 셋 중 올바르게 분류한 데이터 수의 비율을 말한다. 특이도(specificity)는 음성 데이터 중에서 제대로 음성으로 분류된 비율을 말한다. 정밀도(precision)는 분류기가 양성이라고 분류한 대상 중 실제 양성인 경우를 말한다. 민감도(sensitivity)는 재현율(recall)이라고도 하며 양성 환자 중 분류기가 올바르게 양성으로 분류한 비율을 말한다. 이러한 파라미터들은 발성 음성을 이용하여 후두 장애 진단을 할 경우 중요한 파라미터로서 분류기의 유용성을 검증하는 데 사용할 수 있다. 각 앙상블 모델별로 구한 4가지 파라미터를 비교하였다.

$$Accuracy=\frac{TP+TN}{FP+FN+TP+TN} \quad (3)$$

$$Specificity=\frac{TN}{TN+FP} \quad (4)$$

$$Precision=\frac{TP}{FP+TP} \quad (5)$$

$$Sensitivity(Recall)=\frac{TP}{TP+FN} \quad (6)$$

이 네 가지 중 정확도(accuracy)와 함께 의학적 진단에 중요한 것은 민감도(sensitivity)와 특이도(specificity)이다. 민감도는 질병이 있는 사람을 얼마나 잘 찾아내는가와 관련한 값이고 특이도는 정상을 얼마나 잘 찾아내는가에 대한 값이다. 일반적으로 민감도가 높아지면 특이도가 낮아지는 경향을 보인다. 좋은 진단기는 민감도와 특이도가 동시에 높은 경우이다.

표 2는 동일한 데이터에 대해 전체 식별기의 결과로부터 구한 4가지 파라미터들을 보인 것이다. 표의 수치는 각 파라미터의 값 1을 기준으로 한 비율이다. 크기가 1에 가까우면 성능이 좋은 것이고 0에 가까우면 좋지 않은 것이다. 그림 6은 그 값들의 비교를 위해 동일한 수치를 모델별로 그래프로 나타낸 것이다.

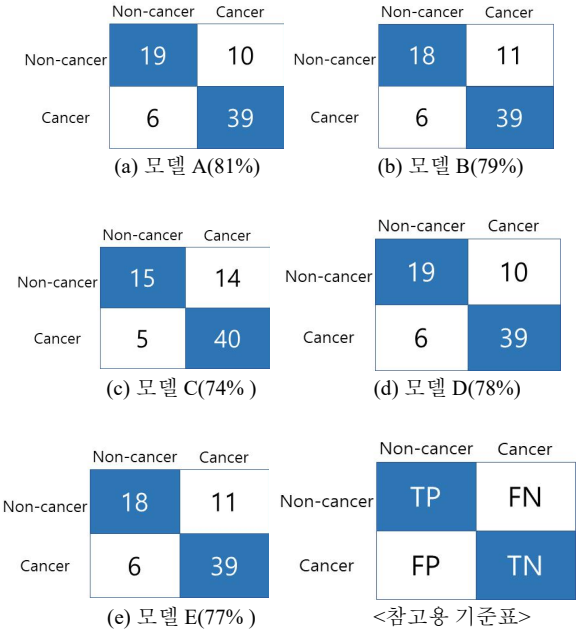


그림 5. 각 CNN 식별기 모델의 혼동 행렬
Figure 5. Confusion matrix of each classifier model

비교를 위해 정확도, 민감도, 특이도 면에서 높은 순위를 따져 비교해 보기 위해 표 3과 같이 파라미터 값을 중심으로 정리하였다. 표 2와 표 3으로부터 모든 파라미터에서 상위에 포함된 random forest 방법을 가장 뛰어난 앙상블 식별기로 볼 수 있다. CNN 식별기를 결합한 CNN-vote는 모두 상위에 위치하지만 민감도 면에서 16위를 차지하여 적절한 식별기가 될 수 없다고 판단된다.

CNN의 결합 모델인 CNN-vote는 포함되는 CNN의 정확도가 높은 경우가 있을 경우라도 앙상블값이 오히려 낮게 나오는 경우도 있었다. 여러 번 수행 결과 CNN 모델은 정확도는 현재와 같은 방법에 의해 voting에 의한 향상을 기대하기 어렵다는 결론을 내리게 되었다. 다만 정확도는 향상되지 않더라도 정밀도(precision), 민감도(sensitivity)는 가장 높은 정확도를 갖는 CNN 모델에 비해 향상되었으므로 모델 개선에 활용할 여지가 있다. CNN1 모델의 경우 특이도(specificity)를 제외하면 비교적 좋은 성능을 보여주었으나 특이도가 너무 낮게 나오는 것이 문제가 된다. CNN의 투표 모델인 CNN-vote는 민감도를 제외하고 상위의 성능을 보여주고 있어서 앙상블 방법에 의한 성능 개선을 기대할 수 있게 해 준다.

앙상블 알고리즘을 이용한 식별기는 잘 알려진 대로 random forest 방법이 전체적으로 우수한 성능을 보여주고 있다. 또 다른 앙상블 알고리즘 적용 모델인 bagging은 두 번째로 우수한 결과를 보여준다. 다른 기계학습 방법의 경우 CNN과 비슷한 범위의 파라미터값을 보여주고 있다. 앙상블 방법 중 SVM, decision tree, KNN을 voting에 의해 결합한 모델이 상위권으로 나쁘지 않은 결과를 얻었다. 이는 본 연구에서 기대했던 결과로 개별 식별기를 조합하였을 경우 특정 응용에서 성능 개선의 여지가 있음을 확인할 수 있었다. 또 다른 성능의 척도로 각 식별기에 대해 구한 ROC(receiver operation characteristic) 곡선을 그림 7에 보였다. ROC 곡선은 질병 진단과 같은 2분법적 판단에 대한 타당성을 검증하는 데 자주 사용되는 지표이다. 가로축을 false positive rate, 세로축을 true positive rate로 나타낸 그래프로 이진 분류기에서 곡선 부가 좌측 상단을 지향할수록 좋은 시스템으로 이야기한다. 이 그림에서 앙상블 방법은 점선으로, 단일 기계학습 방법은 실선으로 나타내었다. 시험용 데이터의 수가 많지 않아 부드러운 곡선은 얻지 못했지만 앙상블 방법들에 대한 대략적인 경향은 확인할 수 있었다. 곡선의 형태를 보면 전체적으로 좋은 식별기라고 할 수는 없겠으나 그중에서도 가장

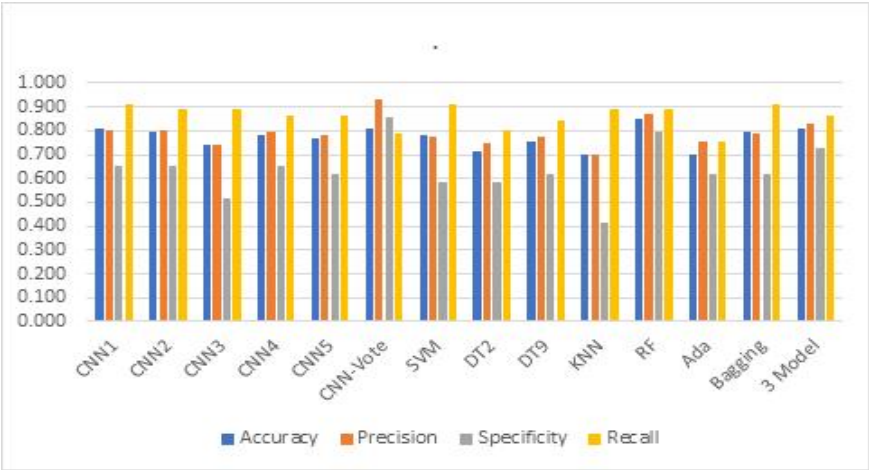


그림 6. 식별기별 정확도, 정밀도, 특이도, 민감도
Figure 6. Accuracy, precision, specificity, sensitivity of identifiers

성능이 좋은 것은 표 2에서와 같이 random forest 모델이다. 최상단의 점선이 그 모델이다. CNN-vote 모델은 직선 위에 점으로 표기되어 있는데 상단에 위치하고 있어서 voting 방법에 의한 성능 개선의 가능성을 보여주고 있다. 다른 식별기들의 경우 SVM과 CNN이 비교적 모서리 쪽에 가까운 궤적을 보여주고 있다. 반면 다른 앙상블 식별기는 일반적인 기계학습 방식의 경우와 비슷한 성능을 보이는 것을 알 수 있다.

표 2. 각 식별기별 파라미터값
Table 2. Parameter values for each identifier

모델	Accuracy	Precision	Specificity	Sensitivity
CNN1	0.811	**0.804**	0.655	**0.911**
CNN2	0.797	0.800	0.655	**0.889**
CNN3	0.743	0.741	0.517	**0.889**
CNN4	0.784	0.796	0.655	0.867
CNN5	0.770	0.780	0.621	0.867
CNN-vote	**0.808**	**0.932**	**0.857**	0.788
SVM	0.784	0.774	0.586	**0.911**
DT2	0.716	0.750	0.586	0.800
DT9	0.757	0.776	0.621	0.844
KNN	0.703	0.702	0.414	**0.889**
RF	**0.851**	**0.870**	**0.793**	**0.889**
Ada	0.703	0.756	0.621	0.756
Bagging	0.797	0.788	0.621	**0.911**
3 model	**0.811**	**0.830**	**0.724**	0.867

표 3. 파라미터별 좋은 식별기 순위
Table 3. List of best identifiers per parameter

순위	Accuracy	Precision	Specificity	Sensitivity
1	RF	CNN+vote	CNN+vote	CNN1
2	CNN1	RF	RF	Bagging
3	3 model	3 model	3 model	SVM
4	CNN–vote	CNN1	CNN1, 2, 4	RF

RF, random forest; CNN, convolutional neural network; DT2, decision tree depth=2; KNN, Knearest neighbor; Ada, adaboost; 3 model: SVM, DT, KNN의 voting에 의한 앙상블.

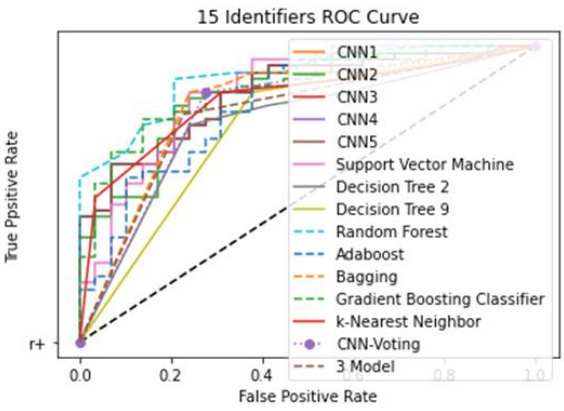


그림 7. 15개 식별기의 ROC 곡선
Figure 7. ROC curves of 15 identifiers

이번 실험의 결과를 통해 후두 질환, 그중에서도 악성 종양과 양성 및 정상 음성을 식별하는 방법으로 기존의 기계학습 방법 외에 앙상블 방법에 의해 성능을 개량할 수 있다는 가능성을 확인할 수 있었다.

5. 결론 및 향후 계획

본 논문에서는 여러 가지 CNN 모델을 결합하여 음성 신호로부터 후두 질환 음성을 정상 및 양성 질환과 악성 질환으로 구분하는 실험에 voting에 의한 앙상블 및 다양한 기계학습 방법을 적용하고 결과를 검토하였다. 또한 기존의 방법 중 널리 사용되고 있는 기계학습 방법들을 결합하여 적용하였을 때 성능을 개선할 가능성을 확인하였다. 또 후두 질환 음성과 같이 데이터 수가 작을 수밖에 없는 경우 소량의 데이터에 의한 과적합을 해소하면서 식별기의 성능을 개선할 수 있는 방안으로 앙상블 방법이 효과가 있음을 확인하였다. 앙상블 모델은 정확도뿐만 아니라 정밀도, 민감도, 특이도에서도 개선된 값을 보여주었다. 비록 향상된 수치값은 적었지만 앙상블 방법에 의해 다양한 파라미터의 개선이 이루어질 수 있다는 것을 확인한 실험이었다.

본 연구의 후속 연구는 다음과 같은 방향으로 진행되어야 한다. 앞으로 실제 진단에 의미가 있는 식별기로 개선하기 위해서는 근본적으로 지속적인 데이터의 확보가 필요하며 훈련에 사용된 데이터가 아닌 외부 데이터에 대해 어떤 식으로 식별률을 개선할 수 있는지에 연구의 초점이 모아져야 한다. 또한 추가적으로 수집된 데이터를 활용하여 기존의 모델을 효과적으로 개선시킬 수 있는 식별 모델 적응화 방법에 대한 연구가 필요하다. 그밖에 임상 전문가가 환자들로부터 데이터를 수집하는 과정 및 정제 기준에 대한 절차의 표준화가 무엇보다 필요하다. 이를 위해서는 유사 연구에 관심있는 의료기관들 간의 협력 등 여러 가지 연구 외적인 문제의 해결도 필요하다.

감사의 글

이 논문은 창원대학교 2021-2022년도 창원대학교 자율연구과제 연구비 지원으로 수행된 연구 결과임.

References

Aicha, A. B. (2018). Noninvasive detection of potentially precancerous lesions of vocal fold based on glottal wave signal and SVM approaches. *Procedia Computer Science*, *126*, 586-595.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., Malki, K. H., ... Bencherif, M. A. (2017). An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, *31*(1), 113.e9-113.e18.

Bezdek, J. C., Keller, J., Krisnapuram, R., Pal, N. R. (2005). *Fuzzy models and algorithms for pattern recognition and image processing*. (pp. 442-490). New York, NY: Springer.

Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C.,

& Wang, C. T. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, *33*(5), 634-641.

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, *33*(6), 947.e11-947.e33.

Jeon, B. U., Kang, J. S., & Chung, K. (2021). AutoLM and CNN-based soft-voting ensemble classification model for road traffic emerging risk detection. *Journal of Convergence for Information Technology*, *11*(7), 14-20.

Jo, C., Kim, K., Kim, D., & Wang, S. (2001, September). Screening of pathological voice from ARS using neural networks. *Proceedings of the Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) 2nd International Workshop* (pp. 241-245). Florence, Italy.

Jung, H., Choi, M. K., Kim, J., Kwon, S., & Jung, W. (2020). CNN-based weighted ensemble technique for ImageNet classification. *IEMEK Journal of Embedded Systems and Applications, 15*(4), 197-204.

Kim, H. B., Jeon, J., Han, Y. J., Joo, Y. H., Lee, J., Lee, S., & Im, S. (2020). Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. *Journal of Clinical Medicine, 9*(11), 3415.

Ko, H., Ha, H., Cho, H., Seo, K., & Lee, J. (2019, May). Pneumonia detection with weighted voting ensemble of CNN models. *Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 306-310). Chengdu, China.

Lee, J. Y. (2021). Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database. *Applied Sciences, 11*(15), 7149.

Librosa. (2021). Librosa: Audio and music processing in Python. Retrieved from http://librosa.org/

Liu, F., Liu, Y., & Sang, H. (2020). Multi-classifier decision-level fusion classification of workpiece surface defects based on a convolutional neural network. *Symmetry*, *12*(5), 867.

Lv, X., Ming, D., Lu, T., Zhou, K., Wang, M., & Bao, H. (2018). A new method for region-based majority voting CNNs for very high resolution image classification. *Remote Sensing*, *10*(12), 1946.

Massachusetts Eye and Ear Infirmary. (1994). *Voice disorders database, version.1.03 (CD-ROM)*. Lincoln Park, NJ: Kay Elemetrics.

Morvant, E., Habrard, A., & Ayache, S. (2014, August). Majority vote of diverse classifiers for late fusion. *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition* (*SSPR*) (p. 20). Joensuu, Finland.

Roy, S., Sayim, M. I., & Akhand, M. A. H. (2019, May). Pathological voice classification using deep learning. *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT).* Dhaka, Bangladesh.

Ruta, D., & Gabrys, B. (2000). An overview of classifier fusion methods. *Computing and Information Systems*, *7*(1), 1-10.

Saarbruecken Voice Database. (2020). Saarbruecken Voice Database. Retrieved from http://www.stimmdatenbank.coli.uni-saarland.de/

Saldanha, J. C., Ananthakrishna, T., & Pinto, R. (2014). Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. *Journal of Medical Imaging and Health Informatics*, *4*(2), 168-173.

Scikit learn. (2022). Ensemble methods. Retrieved from https://scikit-learn.org/stable/modules/ensemble.html

Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors*, *19*(7), 1733.

Szmurlo, R., & Osowski, S. (2021, September). Deep CNN ensemble for recognition of face images. *Proceedings of the 2021 22nd International Conference on Computational Problems of Electrical Engineering (CPEE)* (pp. 1-4). Hrádek u Sušice, Czech Republic.

Tensorflow. (2021). Retrieved from http://www.tensorflow.org/

Wu, H., Soraghan, J., Lowit, A., & Di Caterina, G. (2018, July). Convolutional neural networks for pathological voice detection. *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1-4). Honolulu, HI.

• **조철우 (Cheolwoo Jo)** 교신저자
창원대학교 전기전자제어공학부 교수
경남 창원시 의창구 창원대학로 42
Tel: 055-213-3662
Email: cwjo@changwon.ac.kr
관심분야: 디지털 신호처리, 머신러닝, 음성신호처리

• **왕수건 (Soo-Geun Wang)**
부산대학교 의과대학 이비인후과학교실 명예교수
경남 양산시 물금읍 부산대학로 49 부산대학교 의과대학
Tel: 051-553-4089
Email: entwangsg@daum.net
관심분야: 음성장애, 후두암, 두경부종양

• **권익환 (Ickhwan Kwon)**
부산대학교 대학원 IT응용공학과 박사과정
경상남도 밀양시 삼랑진읍 삼랑진로 1268-50
Tel: 051-895-1713
Email: kanikwon@pusan.ac.kr
관심분야: 인공지능, 영상처리, 데이터사이언스

기계학습에 의한 후두 장애음성 식별기의 성능 비교*

조 철 우[1] · 왕 수 건[2] · 권 익 환[3]

[1]창원대학교 전기전자제어공학부, [2]부산대학교 의과대학 이비인후과, [3]부산대학교 대학원 IT응용공학과

국문초록

본 논문은 후두 장애음성 데이터의 식별률을 CNN과 기계학습 앙상블 학습 방법에 의해 개선하는 방법에 대한 연구이다. 일반적으로 후두 장애음성 데이터는 그 수가 적으므로 통계적 방법에 의해 식별기가 구성되더라도, 훈련 방식에 따라 과적합으로 인해 일어나는 현상으로 인해 외부 데이터에 노출될 시 식별률의 저하가 발생할 수 있다. 본 연구에서는 다양한 정확도를 갖도록 훈련된 CNN 모델과 기계학습 모델로부터 도출된 결과를 다중 투표 방식으로 결합하여 원래의 훈련된 모델에 비해 향상된 분류 효율을 갖도록 하는 방법과 함께, 기존의 기계학습 중 앙상블 방법을 적용해 보고 그 결과를 확인하였다. 알고리즘을 훈련하고 검증하기 위해 PNUH(Pusan National University Hospital) 데이터셋을 이용하였다. 데이터셋에는 정상음성과 양성종양 및 악성 종양의 음성 데이터가 포함되어 있다. 실험에서는 정상 및 양성 종양과 악성종양을 구분하는 시도를 하였다. 실험결과 random forest 방법이 가장 우수한 앙상블 방법으로 나타났으며 85%의 식별률을 보였다.

핵심어: 진단, 후두암, 후두 장애, 기계 학습, convolutional neural network (CNN)

참고문헌

전병욱, 강지수, 정경용(2021). 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델. *융합정보논문지*, *11*(7), 14-20.

* 본 연구는 창원대학교 2021-2022년도 연구지원을 받아 수행되었음.