Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-022-01142-1

ORIGINAL RESEARCH

Person identification based on voice biometric using deep neural network

Noor D. AL-Shakarchy[1] · **Hadab Khalid Obayes**[2] · **Zahraa Najm Abdullah**[1]

Received: 6 July 2022 / Accepted: 13 December 2022

Abstract Nowadays in all everyday transactions, technological progress has become an intrinsic characteristic that depends on such electronic applications as financial and banking transfers, health care, project management, and other crucial life aspects. The core of these applications is person Identification and/or verification steps which can be considered one of the complicated limitations. Accordingly, the employment of biometric attributes can yield promising outcomes in these fields. A One's voice is a unique bio-feature whereby people can be authenticated and precludes others from assuming a one's identity without their previous knowing or assent. This work proposes a model with a new architecture to identify the person by exploiting the unique individual characteristics available in one's voice based on deep learning. An augmentation method is utilized to increase the samples in the available dataset. The available temporal information at an input audio file is analysed then feature maps from this information are extracted which represent the salient temporal feature (time-domain features). The decision is made based on tracking these voice features over time. Successful and promising results are achieved through this work, the accuracy is close to 99.81% (± 1.78%) and the values of loss function are close to 0.009 over VoxCeleb1 dataset for identifying 40 subjects.

Keywords One Dimensional Convolutional Neural Networks (1DCNN) · Voice biometric · Person identification PI · Rectified linear units (ReLU) · Pooling layer (Pool) · Batch normalization layer (BN)

1 Introduction

Nowadays with the progress in technology, biometrics has become essential for one's authentication techniques in all everyday transactions [1] Among the topics of high popularity in speech research is voice recognition which is included biometric identification besides facial, retinal, and fingerprint recognition. The human voice became an essential part of systems for verifying and identifying speakers. These systems are used to identify crime suspects, improve human-machine interactions, and adapt music when waiting in line [2]. Despite the numerous studies conducted to extract traits and develop enhanced classifiers, the accuracy of classification remains below satisfaction levels [3]. System identification mainly deals with addresses the issue of creating mathematical models for dynamic systems in light of the system data that is observed [1]. That is, these systems try to find a mathematical model based on measured relationships and approximate the unknown function [4]. According to the information that is available, there are two approaches to describe the behaviour of system identifications, namely the State-Space and the Black-Box approaches. The first approach is adopted when the dynamic equation is available and involves the description of the system's internal state [5]. As for the Black-Box, it is used whenever the only information provided is its inputs and outputs and it involves an input-output description. Automatic person identification is an interesting field in research aimed at the ability of autonomous machines to identify a person

✉ Noor D. AL-Shakarchy
noor.d@uokerbala.edu.iq

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, Kerbala University, Kerbala, Iraq

[2] College of Education for Humanities Studies, University of Babylon, Babylon, Iraq

Published online: 28 December 2022

through biological metrics [6]. By detecting the person's identity, the decoded speech could be used as an input for a variety of applications, such as managing calls, identity security, processing client requests, and computer dictations. Among the factors which influence how reliable systems are when it comes to identifying or authenticating speech are the spectral density and segments of speech, context sensitivity, stress, and pronunciation [3, 7, 8].

2 Related work

For speaker recognition, the authors of [9] proposed a model using a two-dimensional convolutional neural network (2-D CNN) and gated recurrent unit (GRU). The convolutional layer is employed in the network model design to extract voiceprint features and decrease dimensionality in temporal and frequency- related domains. This allows for GRU layers to be computed more quicker. The stacked GRU recurrent network layers could also pick up a speaker's auditory characteristics. The DNN model (deep GRU) attained an accuracy of 98.96% in the experiments.

The authors present a block diagram for identifying system users through the individual voice featuresstated in the research work [10]. The latter presents a model of elementary speech units by combined of the DNN method and the I-vector. It has a remarkable increase in the rate of security againstdifferent types of attacks that biometric identification systems suffer from. This allows users to be identified by their voice characteristics. They investigated the vulnerable aspects of the identification systems based on voice biometrics and developed a structural framework that identifies users through voice using advanced security.

Zhipeng et al. [11] focus on the speaker identification scene and artistic design. The researchers used ANNas it can distinguish between complex classification boundaries. Speech recognition was done by studying multi-layer perceptual networks for improving classification accuracy, using the back propagation method (BP algorithm). The identification system is able to detected the test objects (speakers) even in case those speakers are pronounce the same separated words by through the identification of their different voice fingerprints. This design does not only consider BP based recognition but also its variants, and explores voice fingerprint recognition according to a convolutional neural network (CNN).

In the paper of [12], the authors created a recognition system for the identification of human speakersby means ofCNN. The methods deployed in theirstudy are MFCC-CNN and RW-CNN. The first involves the standard method whereby MFCC is used for extracting audio features. These features enter the CNN for processing. The training CNN takes input as a picture, after which the training process starts using the proposed CNN. Moving to the second method (RW-CNN), it follows a similar procedure as with MFCC-CNN, except that these initial phases are skipped, starting at the entering of CNN. They obtained an accuracy rate of (96%) for both methods.

The system designed in [13] identifies text dependent and independent speakers in English. An audio wave recorder is used for recording speech, after which the given speech signals are pre-processed. UMRT is adopted for image compressing, and it is combined with MFCC to extract features. Through the MLP and BP algorithms, the features are classified, and accuracyis obtained by means of a confusion matrix. The accuracy rate for speech dependent and independent systems reached (97.91%)and (94.44%), respectively. The related work can be summarized in Table 1 below.

3 Proposed method

After the great developments observed in computer science and artificial intelligence, deep learning has evolved from traditional neural network technology to represent a multi-functional and layered neural network [14]. Deep learning has extensive uses in exploring speech and signals with an

Table 1 The related work summarization

Reference	Brief description	Accuracy
[9]	Using deep neural network (DNN) model using a two-dimensional convolution neural network (2-D CNN) and gated recurrent unit (GRU).	98.96
[10]	DNNs are initialized using the restricted Boltzmann machines (RBM).	Test error rate reached 0.56%
[11]	**Method(1)**:BP neural network using Mel frequency cepstrum coefficient (MFCC) to extract features;	0.740
	Method(2): a BP neural network that directly uses audio signals as the input;	0.996
	Method(3): CNN neural network	1.00
[12]	The methods deployed are MFCC-CNN and RW-CNN.	0.96, 0.96
[13]	Combination of MFCC and UMRT for feature extraction, Multi layer perceptron and Back propagation for classification	97.91% and 94.44%

efficiency enhancement of over 30% [15, 16].CNN is a type of machine learning algorithm that employs deep learning depending on a special case of the feed-forward neural network [17]. CNN can be implemented [18, 19] with different dimensions according to the specific applications. One-dimension (1D-CNN) is used to analyze signal data over a fixed duration (such as audio signals) [20, 21] which is useful for speech processing that deals with just one-Dimension.

This work proposes a model to identify the person via voice biological metrics by employing 1D CNN with a new architecture to find the model to find a mathematical model based on measured relationships and approximating the unknown function through the use of a number of input and output data. The main aim of the proposed architecture is to build a usable model to prophesy the identity of a specific person by their voice with a low complexity requirement. .The general block diagram of the proposed system stages are illustrated in Fig. 1.

3.1 Dataset

VoxCeleb is an audio-visual data set that consists of short audio and video fragments of human speech. These are extractions from YouTube interview videos . VoxCeleb speakers having various ethnicities, accents, professions, gender and ages. Each of the speaking face-tracks is recorded "in the wild", including background chatter, laughter, overlapping speech, pose variation and different lighting conditions. All segment have a length of minimally three seconds. VoxCeleb1 contains more than 100,000 utterances for 1,251 celebrities, taken from videos found on YouTube [22].

3.2 Network architecture

To provide immunity to overfitting when working on big data with any alarming scales; the best and strong models are the CNNs which provide easier control and training. The PIDNN performs two main stages which employed different functions by some layers respectively. These stages are the pre-processing stage which does all the preparation work on the input signal (audio file) then the identification predictor stage which extracts the salient features of the audio file for each person and used these feature maps in decision-making to decide the person's identity. The summary illustration of the proposed architecture is shown in Table 2 .

- Extraction Step

Thos step consists of many convolutional blocks (3–7 blocks as shown in Table 2) to debrief and track the temporal (salient) features of the input signal (audio file) over time and construct higher-level feature patterns for each person's voice. Each block consists of many layers which are 1D

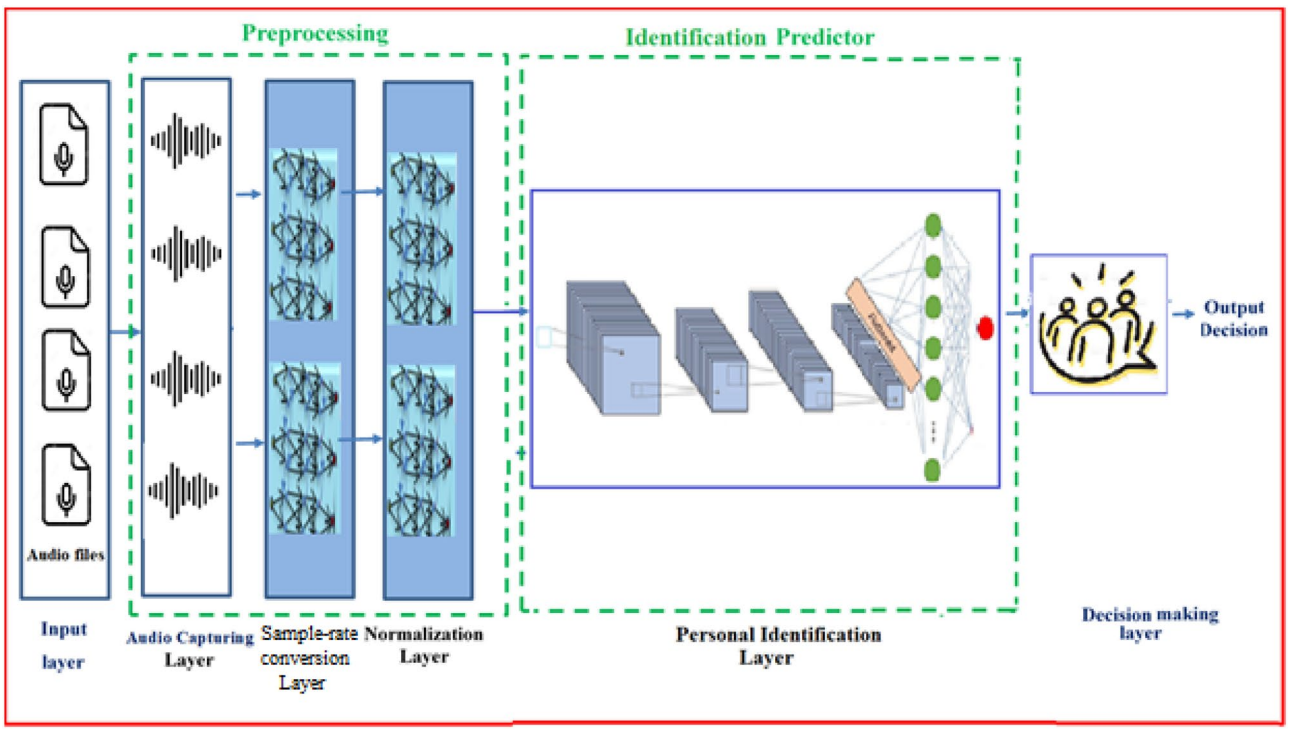

Fig. 1 The suggested PIDNN system stages

Table 2 The suggested system's overview illustration

Block no.	Layer (type)	Output shape	Parames no.
1	reshape_1 (Reshape)	(None, 60,000, 1)	0
2	batch_normalization_1	(Batch (None, 60,000, 1)	4
3	Co1d_1 (Conv1D)	(None, 60,000, 8)	96
	(M_P)1d_1	(MaxPooling1 (None, 30,000, 8)	0
	Dr_1 (Dropout)	(None, 30,000, 8)	0
4	Co1d_2 (Conv1D)	(None, 30,000, 16)	912
	batch_normalization_2	(Batch (None, 30,000, 16)	64
	M_P1d_2	(MaxPooling1 (None, 15,000, 16)	0
	Dr_2 (Dropout)	(None, 15,000, 16)	0
5	Co_3 (Conv1D)	(None, 15,000, 32)	2592
	M_P1d_3	(MaxPooling1 (None, 7500, 32)	0
	Dr_3 (Dropout)	(None, 7500, 32)	0
6	Co_4 (Conv1D)	(None, 7500, 64)	10,304
	M_P1d_4	(MaxPooling1 (None, 1500, 64)	0
7	Co1d_5 (Conv1D)	(None, 1500, 128)	24,704
	M_P 1d_5	(M_P)1 (None, 750, 128)	0
8	flatten_1 (Flatten)	(None, 96,000)	0
	D_1 (Dense)	(None, 500)	48,000,500
9	D_2 (Dense)	(None, 100)	50,100
	Dr_4 (Dropout)	(None, 100)	0
10	D_3 (Dense)	(None, 8)	808

All params:48,090,084

Params trainable: 48,090,050

Params non-trainable: 34

convolution layers; with filter numbers (8, 16, 32, 64, and 128) respectively, and kernels sizes of (11, 7, 5, 5, and 3); integrated with nonlinear, pooling (Max pooling), dropout, and some batch normalization layers.

The pooling or sub-sampling layer is responsible for resisting noise and blurring by making the features strong enough by reducing features resolution. The max-pooling function with window size (2) is used in the proposed model, which outputs the maximal value from some clusters of neurons (vectors of non-overlapping one-dimensional spaces produced by dividing the input audio files) at the prior layer.

A dropout layer is presented in the proposed model for avoiding the case of over-fitting and effectively controlling noise during the training process by selecting some neurons randomly (around 25%) of the specific layer, and their weights are set to zero. In order to accelerate the learning process and regulate the update in the model layers, some other layers are added which are Batch normalization layers.

- Identification Step

The identification step is a process that involves building a mathematical model based on measured relationships and approximating the unknown function; this is done through the final three fully connected block layers (Dense layers). Which have 500,100, and 8 units respectively with the "ReLU" activation function except final fully connected layer (output or decision-making layer) applies a sigmoid function to represent the binary identification of persons. Dropout layers are also added in this step with 25% randomly neurons selected.

The proposed model performance is evaluated with two metrics (accuracy and loss functions). A quick way to understand the behavior of the proposed model learning on a specific data set is by plotting the results for each epoch based on Implementing the training and validation data set, as can be shown in Fig. 2.

As mentioned, the input to the proposed model is an audio file, therefore the first pre-processing step in order to regularize the audio is the Sample-rate conversion step. It is the process of changing the sampling rate (the number of times the audio is re-sampled per second) of a discrete signal for obtaining a new discrete representation of the underlying continuous signal. This conversion is done based on the resampling (up-sampling and down-sampling) of each audio file. In the proposed system, all audio files are resampled to 60,000 Hz, so the input shape to the proposed model is (the number of all samples, input size which is set to 60,000, number of channels which set to 1).

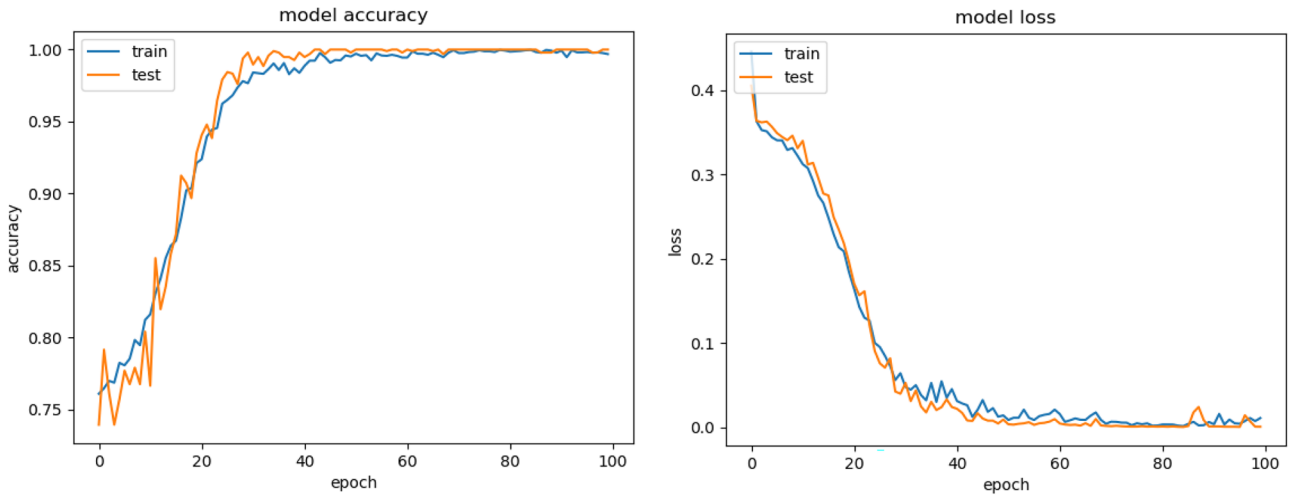

Fig. 2 Accuracy and loss function of proposed model

To speed up learning and lead to faster convergence; DNN models have to deal with small weight values as well as input values close to 0 during processing. The learning process can be retarded or slowed down due to large integer value inputs. The normalization process is dividing each value by the largest one in order to specify the intensity range of input values between 0 and 1.

4 Evaluation of the proposed architecture

Given the fact that DNNs have many parameters, these might over-fit the training data throughout the learning procedure [21]. This implies that the model's performance is excellent with the training data, but leads to failure when generalizing unseen data. This leads to an inferior performance on new data (usually the test set).

There are a number of regularizing approaches for avoiding such issues, mainly through certain intuitive ideas [23]:

- introducing stochastic behaviour towards neural activations (e.g., dropout and drop connect);
- normalizing batch statistics in feature activations (e.g., batch normalization);

using guidance from a validation set to stop the learning process (e.g., early stopping).

Deep learning models are stochastic. This means these models have variable processes at any time. So the outcome leads to some randomness and some uncertainty by differing predictions with differing overall skills although it uses the same data. The model can be evaluated in two skills: Model Skill Estimation (Model Variance Controlling); K-fold Cross-validation [24]is used; whereby different results are obtained when different data trained on the same model. The second evaluation is Stochastic Model's Skill Estemation (Model Stability Controlling). It obtains differing results when training the same model onto the same data and repeates evaluating of a non-stochastic model multiple times then obtain the mean value.

4.1 K-Folds cross validation

The deep learning model performance on unseen data can be estimated using cross-validation process; which based on resampling data sets on a single parameter (k) and evaluating the model on all specific resampling. In other words, the general expectancy estimation of predicting unused data during the model training is done based on a limited sample.

Table 3.

Table 3 Fold CV for proposed model

Fold number	Fold accuracy (%)	Accuracy of model
1	99.07	99.81% (± 1.78%)
2	99.35	
3	99.74	
4	100.00	
5	100.00	
6	100.00	
7	100.00	
8	100.00	
9	100.00	
10	100.00	

5 Conclusion

There are various conclusions established during this work. The deep neural networks-based classification system is considered to be a superior approach than other traditional methods in accuracy and loss functions considerations. On it, the proposed model achieved successful results to identify individuals through their voices based on using 1-D CNN. The time and effort required for the annoying pre-processing of the audio signal are economized by dealing directly with raw data and thus lead at the same time to making the model useable in many organizations. The proposed model provided the possibility to handle the noise associated with samples via the behavior of the ReLU activation function which extracts the salient features and neglects the weak ones. All noise elements from the sequence are dropped and only those carrying a positive value are kept. The pose variation problem in the sample was decreased by obeying the convolutional property, this was done by adding some Batch Normalization layers after the convolutional layers. This leads to diverse coefficients of the same feature map, at diverse locations, being normalized in the same manner independently of its direct spacial circumference. The computation complexity and the memory requirements were reduced due to the 1D structure of the proposed model which exploits all audio information. Taking successive frames of the audio file as one set directly led to using the same coefficients across all spaces.

6 Future works

The face recognition performance degrades considerably for images of identical twins. And therefore the performance of the face aging model has degraded accordingly. Upon that, proposing the face aging model that can distinguish the twin people is an active improvement to the presented work. The future work solves the challenges for face aging twin people. Baby faces have not uniformly transitioned between baby and adult and aging cases. The other future work may be presented as a face-aging model that can deal with these not uniform transitions and solve the challenges for baby images.

References

1. Mohanty P, Nayak AK (2022) CNN based keyword spotting: an application for context based voiced Odia words. Int J Inf Technol. https://doi.org/10.1007/s41870-022-00992-z
2. Jain AK, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Trans Circuits Syst Video Technol 14(1):4–20. https://doi.org/10.1109/TCSVT.2003.818349
3. Farooq H, Naaz S (2020) Performance analysis of biometric recognition system based on palm print (2020). Int J Inf Technol 12:1281–1289
4. Rachad S, Nsiri B, Bensassi B (2015) System identification of inventory system using ARX and ARMAX models. Int J Control Autom 8(12):283–294. https://doi.org/10.14257/ijca.2015.8.12.26
5. Pappalardo CM, Guida D (2018) System identification algorithm for computing the modal parameters of linear mechanical systems. Machines. https://doi.org/10.3390/machines6020012
6. Mandalapu H et al (2021) Audio-visual biometric recognition and presentation attack detection: a comprehensive survey. IEEE Access 9:37431–37455. https://doi.org/10.1109/ACCESS.2021.3063031
7. Mamyrbayev OZ, Othman M, Akhmediyarova AT, Kydyrbekova AS, Mekebayev NO (2019) Voice verification using i-vectors and neural networks with limited training data. Bull Natl Acad Sci Repub Kaz 3(379):36–43. https://doi.org/10.32014/2019.2518-1467.66
8. Kumar A, Mittal VH (2021) Speech recognition in noisy environment using hybrid technique. Int J Inf Technol 13:483–492
9. Ye F, Yang J (2021) A deep neural network model for speaker identification. Appl Sci 11(8):1–18. https://doi.org/10.3390/app11083603
10. Aizat K, Mohamed O, Orken M, Ainur A, Zhumazhanov B (2020) Identification and authentication of user voice using DNN features and i-vector. Cogent Eng. https://doi.org/10.1080/23311916.2020.1751557
11. Zhipeng D, Jingcheng W, Yumin X, Qingmin M, Xiaoming W (2019) Voiceprint recognition based on BP neural network and CNN. J Phys Conf Ser. https://doi.org/10.1088/1742-6596/1237/3/032032
12. Khdier HY, Jasim WM, Aliesawi SA (2021) Deep learning algorithms based voiceprint recognition system in noisy environment. J Phys Conf Ser. https://doi.org/10.1088/1742-6596/1804/1/012042
13. Antony A, Gopikakumari R (2018) Speaker identification based on combination of MFCC and UMRT based features. Proced Comput Sci 143:250–257. https://doi.org/10.1016/j.procs.2018.10.393
14. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data. https://doi.org/10.1186/s40537-019-0192-5
15. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1–21. doi: https://doi.org/10.1186/s40537-014-0007-7
16. Obayes HK, Al-A'araji N, Al-Shamery E (2019) Examination and forecasting of drug consumption based on recurrent deep learning. Int J Recent Technol Eng 8(2):414–420. https://doi.org/10.35940/ijrte.B1069.0982S1019
17. Ravi D et al (2017) Deep learning for health informatics. IEEE J Biomed Heal Inform 21(1):4–21. https://doi.org/10.1109/JBHI.2016.2636665
18. Obayes HK, Al-Turaihi FS, Alhussayni KH (2021) Sentiment classification of user's reviews on drugs based on global vectors for word representation and bidirectional long short-term memory recurrent neural network. Indones J Electr Eng Comput Sci 23(1):345–353. doi: https://doi.org/10.11591/ijeecs.v23.i1.pp345-353
19. Al-Shakarchy ND, Ali IH (2019) Abnormal head movement classification using deep neural network DNN. AIP Conf Proc. https://doi.org/10.1063/1.5123123
20. Al-Shakarchy ND, Ali IH (2020) Detecting abnormal movement of driver's head based on spatial-temporal features of video using deep neural network DNN. Indones J Electr Eng Comput Sci 19(1):344–352. https://doi.org/10.11591/ijeecs.v19.i1.pp344-352

21. Fridman L et al (2017) MIT Autonomous vehicle technology study: large-scale deep learning based analysis of driver behavior and interaction with automation, pp 1–17. http://arxiv.org/abs/1711.06976
22. Nagrani; A, Chung JS, Zisserman A (2017) A large-scale speaker identification dataset. INTERSPEECH
23. Buduma N, Locascio N (2017) Fundamentals of deep learning: designing next-generation machine intelligence algorithms. Nikhil Buduma; with contributions by Nicholas Locascio
24. Jung Y (2018) Multiple predicting K-fold cross-validation for model selection. J Nonparametr Stat 30(1):197–215. https://doi.org/10.1080/10485252.2017.1404598