E-ISSN: 2676-6167 | P-ISSN: 2538-5100

Journal of Applied Research on Industrial Engineering

www.journal-aprie.com

J. Appl. Res. Ind. Eng. Vol. 9, No. 2 (2022) 197–211.

Paper Type: Research Paper

Automatic Personality Recognition and Perception Using Deep Learning and Supervised Evaluation Method

Effat Jalaeian Zaferani[1], Mohammad Teshnehlab[1,*], Mansour Vali[1]

[1] Department of Electrical and Computer Engineering, Faculty of K. N. Toosi University of Technology, Tehran, Iran; Jalaeian@ee.kntu.ac.ir; teshnehlab@eetd.kntu.ac.ir; mansour.vali@eetd.kntu.ac.ir.

Citation:

Jalaeian Zaferani, E., Teshnehlab, M., & Vali, M. (2022). Automatic personality recognition and perception using deep learning and supervised evaluation method. *Journal of applied research on industrial engineering, 9*(2), 197-211.

Received: 24/04/2021 Reviewed: 28/05/2021 Revised: 02/10/2021 Accepted: 06/10/2021

Abstract

The personality in the present world plays a critical role in social interactions, the use of modern technologies, and individuals' success. Therefore, in the last two decades, the study of Automatic Personality Perception (APP) and Automatic Personality Recognition (APR) has become more prevalent than speech processing. These studies have shown that personality traits affect acoustic features. However, the intrinsic imbalanced distribution of personality classes across the dataset is an issue mentioned in most previous studies and the classification results suffer from it. In this paper, an innovative supervised k-fold Cross-Validation (CV) method was proposed to cope with the problem of affecting the imbalanced distribution of data across different classes. The classification outcomes showed better performance in comparison with three traditional data balancing methods. Moreover, the obtained results of the proposed evaluation method indicated that the proposed method acts as a k-fold CV method if the data distribution is balanced; otherwise, it will improve the classification results.

Keywords: Stacked Auto-Encoder, Cross fold validation, Big Five Inventory, Personality trait, Speech processing.

1 | Introduction

With the technological advances and the demand for making smart cyberspace in recent decades, Automatic Personality Perception (APP) and Automatic Personality Recognition (APR) studies from handwriting, speech, and facial expression have been becoming more popular [1]. Personality is one of the most important issues in psychology that plays an essential role in human life and social interaction [2]. Among all personality theories, the Big Five Inventory (BFI) is frequently used in computer science research [3] and [4]. The BFI is a five-factor model that measures five personality traits, including Openness to experience (Ope.), Conscientiousness (Con.), Extraversion (Ext.), Agreeableness (Agr.), and Neuroticism (Neu.) [5]. These five traits exist in a person simultaneously but with different degrees. One can obtain these measurable traits via self-assessment BFI-questionnaires (personality recognition) or other assessments (personality perception) [2] and [6].

Corresponding Author: teshnehlab@eetd.kntu.ac.ir

http://dx.doi.org/10.22105/jarie.2021.282961.1314

Despite APP and APR studies' success, some challenges restrict the research. A critical challenge frequently mentioned in articles is acquiring a public, suitable, and annotated dataset for training Machine Learning (ML) models [7] and [8]. On the one hand, collecting such a dataset is time-consuming, and the ground truth annotation process is costly [9]. On the other hand, the analysis of existing datasets specifies that the binary classes of these datasets are imbalanced, which means one class instance is larger than the other [10]. The main factor behind this imbalance is the variety in culture, language, geography, and religion, which cause differences in personality distribution across societies [11]. The imbalanced distribution may cause biased (skewed) and incomplete learning during the ML algorithms training process. It means the ML parameters fitted on the class of the majority samples increase the misclassification rate [12] and invalidate the model classification results [13]. This problem becomes acute when using Cross-Validation (CV) algorithms to evaluate the model's generalization (predictive performance estimation). In this regard, data-level and algorithm-level techniques are two ways for addressing the class imbalance problem in traditional ML summarized below.

Data-level techniques. These techniques create data balanced distribution artificially. Random OverSampling (ROS) of the minority class is one of these techniques [14]. Its drawback is that data replication leads the model to over-fitting (good outcomes on the training data and poor performance on testing data). Moreover, by increasing the size of training data, the training time is increased. Synthetic Minority Oversampling Technique (SMOTE) was created to decrease oversampling weaknesses by generating synthetic data [15]. Despite the success of SMOTE in several applications, the generated samples have correlations with some original minority samples. Also, SMOTE can decrease between-class scattering by increasing overlapping between classes. Both issues reduce the ML performance [16]. The other technique is Random UnderSampling (RUS) of the majority class, which in this technique, some useful information is lost [14] and [17].

Algorithm-level methods. These techniques combine several weak learners (models) simultaneously and make strong decisions to increase the ability of the model estimation output (ensemble methods). This technique can reduce classifier bias toward the majority class. Boosting and bagging methods are examples of ensemble methods [17]. In the boosting method, the next learner parameters depend on the previous learner parameters, and in the bagging method, each learner training process is separate. The drawback of this technique is that few learners can be used [15].

The above techniques have been applied in various domains. However, for evaluating Deep Neural Networks (DNNs), some other issues should be considered.

DNNs are ML algorithms with good representation learning and feature engineering ability, especially in high input dimensions. They map raw input to new representations through multiple nonlinear functions and extract high-level features automatically that no experts can easily do. Despite the feature engineering ability of DNNs, their good performance extremely depends on the input data. In imbalanced data, the gradient value of a DNN in the minority class is much smaller than in the majority class. The error of the majority class reduces rapidly (in early epochs), and the error of the minority class is ignored caused incomplete network training [18] and [19].

On the other hand, since they are fine-tuned by gradient descent back-propagation, the use of the classifier is confined to the gradient-based one [17]. As a result, algorithm-level techniques can not use easily [20]. Also, the performance of these networks is sensitive to over-fitting, and data-level techniques should be used consciously.

The above explanations provide the impetus for proposing our novel algorithm on DNNs called the Supervised K-fold CV (SK-fold CV). Hence, this study aims to perceive and recognize personality based on speech signals with a novel evaluation method.

The rest of our study is organized as follows. In the next section, the literature review of related works is expressed. In Section 3, the summary of using the two datasets is explained. The materials and methods, including feature engineering using auto-encoders, proposed evaluation method, and evaluation criteria, are described in Section 4. Finally, Section 5 provides the results of the simulations and discussion.

2 | Related Works

The literature of imbalances dataset studies in APP/APR is reviewed below. According to the type of dataset, the way of studies addressing class imbalance is different.

Data augmentation [9] and transfer learning [7] methods were typically used to deal with the class imbalance problem in APP/APR based on video and image datasets.

In the handwriting datasets, both data-level and algorithm-level techniques were used. Personality identification from handwriting feature extraction and AdaBoost classification was examined in [21]. Zarnoufi and Abik [22] proposed a combination of RUS and ROS techniques to solve severe class imbalance distribution. Cyber-violence detection by personality analysis of online users is one of the applications of personality traits identification—such an interesting study mentioned in [22]. An ensemble classifier including random forest, XGBoost, and AdaBoost to identify BFI traits of online users were applied to the feature set. To tackle the imbalance in data distribution, the article [4] used SMOTE technique. In [23], several supervised classifiers were examined in the ensemble method. Horvath et al. [24] provide an interesting study on the association between BFI personality traits and suicidal attempts. It dealt with the small sample size of the dataset, which affected classifiers results. Therefore, an ensemble classifier comprising gradient tree boosting, random forest, decision tree, logistic regression, and linear regression was applied to overcome this problem.

Studies based on speech signals datasets indicate that APP/APR from the speech was more difficult than other datasets. Their focus was mostly on feature extraction and classification methods, and to the best of our knowledge, none of the balancing techniques was applied [25]. However, in order to consider the effect of class imbalance, the Unweighted Average (UA) recall criterion has been utilized [26]. The UA recall is the average recall of binary classes used for highly imbalanced distribution classes [27], and its value illustrates the classifier's ability to classify imbalanced data correctly [28]. It has been used since the INTERSPEECH conference in 2009 [29]. In this regard, An et al. [30] worked on acoustic and lexical features classified by Support Vector Machine (SVM). In order to feature selection, the spearman correlation was applied to deceptive and non-deceptive speech from native speakers. Both regression and sequential-minimal-optimization classification were considered [30]. Personality traits classification using Stacked Auto-Encoder (SAE) and Long Short Term Memory (LSTM) was performed in [31]. The purpose of Jothilakshmi et al. [31] was to predict the speaker's behavior by non-verbal features. The authors deliberated the relationship between the speech signal and personality traits using spectral properties. They investigated and compared the K-Nearest Neighbors (KNN) clustering and the SVM approaches.

Mohammadi et al. [32] classified the personality traits by nonverbal features and two classifiers (the SVM and logistic regression). The feature sets contained low-level description, lexical and word embedding features. Koutsombogera et al. [33] reported that acoustics features are more important than the content (lexical) features for classifying personality traits. Results also demonstrated that there is no specific model or feature set for predicting a trait across personality. However, different models and feature sets outperform for every trait. Deep learning and LSTM capabilities were examined in [11] for feature engineering. Kampman et al. [34] suggested a virtual agent based on the online user personality speech and Convolutional Neural Network (CNN). Liu et al. [35] discussed that APR and APP need various speech features in size and type to outperform. Nevertheless, with diverse features, the designed model comes into trouble. So a feature filtering and hierarchical clustering algorithm were applied.

3 | Dataset

As the aim of our study, APP and APR from speech signals, we used two datasets, the first for assessing personality perception and the second for determining speech recognition personality. The labelling route of both datasets was the same as below.

The personality traits scores in the BFI can be described as continuous numbers in the range of [0,100] by using a psychological questionnaire [36]. A score of 50 is indicated a normal person. Hence, five personality characteristics were classified into two classes: less than normal and greater than normal. Scores more than 50 got "high" labels, and the scores below 50 got "low" tags [35]. Accordingly, APP and APR become binary classification projects.

3.1 | The SSPNet Speaker Personality Corpus (SPC)

The SSPNet SPC is a common and available dataset in APP research, including 640 speech clips recorded by 322 French people. The emotionally neutral short clip (10-second clip) was recorded. The 11 assessors evaluated each clip based on the BFI-10 questionnaire. The average of the 11 questionnaire scores was reported for each clip [37]. *Table 1* indicates the clip's number of the SPC dataset at the high and low levels in each trait.

Table 1. The number of short audio clips at high and low levels in each trait at the SPC dataset. The imbalanced distribution is apparent.

Traits	The Number of Samples		
	Low	High	Total sample
Neu.	337	303	640
Ext.	370	270	640
Ope.	227	413	640
Agr.	334	306	640
Con.	418	222	640

3.2 | Unemotional Text-Reading Speech (UTS) Corpus

The UTS corpus was collected from 139 participants via recording a Persian reading text. On average, each reading time per participant was about 1.25 minutes, whereas the total signal recording time was equal to 3 hours and 31 minutes. Participants were native speakers, and there were no presuppositions to the text. After recording, the participants filled the BFI questionnaire, and the scores were obtained [38].

Table 2. The number of short audio clips at high and low levels in each trait at the UTS dataset. The imbalanced distribution is obvious

Traits	The Number of Samples		
	Low	High	Total sample
Neu.	948	173	1,121
Ext.	339	782	1,121
Ope.	531	590	1,121
Agr.	442	679	1,121
Con.	507	614	1,121

For applying the same process on both datasets, the same length of clips was needed. Therefore, UTS clips were divided into 10-second utterances (short clips) to build 10-second clips similar to SPC recordings. Extraction of 10-second utterances increased the number of samples to 1,121. *Table 2* shows the samples number of five personality traits at low and high levels.

4 | Materials and Methods

The framework of our study is exposed in *Fig. 1*. This figure indicates the process of feature extraction, classification, and model evaluation.

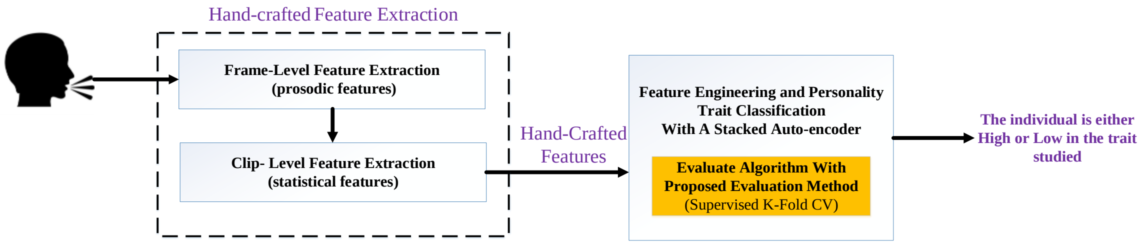

Hand-crafted Feature Extraction
Frame-Level Feature Extraction (prosodic features)
Clip- Level Feature Extraction (statistical features)
Hand-Crafted Features
Feature Engineering and Personality Trait Classification With A Stacked Auto-encoder
Evaluate Algorithm With Proposed Evaluation Method (Supervised K-Fold CV)
The individual is either High or Low in the trait studied

Fig. 1. Network with eight vertices.

4.1 | Hand-Crafted Features

The following principle was applied on every 10-second utterance of both datasets to extract hand-crafted features (features extracted manually).

Firstly, the six prosodic features were extracted from each clip by the open SMILE 2.3 toolkit. For extracting time-domain features, each clip was divided into 60ms frames with a 20 ms overlap. In the frequency domain, framing was performed with a 20 ms window and a 10 ms overlap. These six frame-level features were pitch, first and second formants, energy, and the length of voiced and unvoiced segments frequently used in the related studies [6] and [32]. Secondly, twenty-four statistical features were acquired by applying the mean, standard deviation, maximum and minimum functions to the six prosodic features over 10 seconds (4 statistical functions*6 prosodic features=24 statistical features). Although these features have valuable information in speech processing, previous studies have proved they cannot significantly identify personality. Thus, we used the feature engineering ability of DNNs.

4.2 | Feature Engineering and Classification with the SAE

Our prior work implemented two DNN approaches for personality trait classification, a one-dimensional (1-D) CNN and the SAE [39]. We concluded that despite the feature engineering ability of CNN in most fields, it does not provide good performance compare with SAE in our field of study. It has two reasons: 1) losing some useful information through the pooling layer to obtain downsampling features. 2) Local feature engineering (learning) in the convolutional layer based on the kernel size that does not take some hand-crafted features part at the feature engineering stage. For instance, the relation between the first hand-crafted feature and the last one does not consider. For better illustration, the architecture of 1-D CNN is shown in *Fig. 2*.

Although two stated reasons are the feature extraction process in CNN, the datasets challenge, as noted earlier, has forced researchers to extract hand-crafted features. Therefore, hand-crafted features cause CNN poor performance, which in comparison, it does not occur in the fully connected SAE. Please see [39] for more detail. Therefore, some nonlinear features were extracted by the SAE in the current study as below.

A SAE is created by putting some auto-encoder together. An auto-encoder is a neural network designed to reconstruct its input at its output, as shown in *Fig. 3*. It has two layers called the encoder (hidden) and the decoder (output) layers [40].

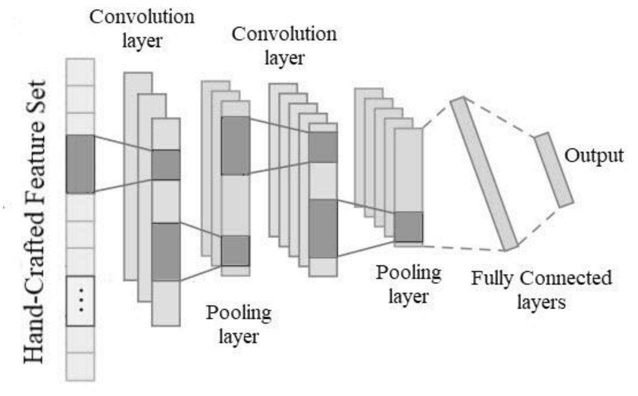


Fig. 2. The architecture of 1-D CNN.

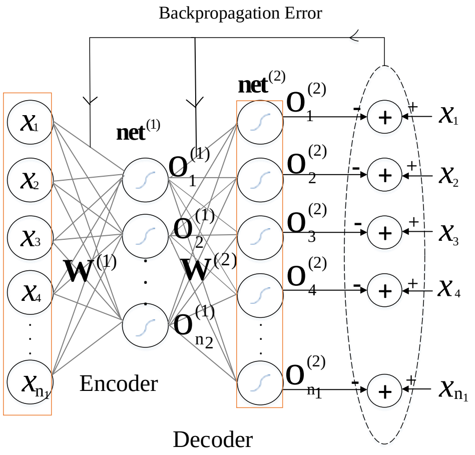


Fig. 3. Schematic of an auto-encoder with n₁ input neurons and n₂ encoder neurons.

In the following, the feed-forward and back-propagate equations of an auto-encoder are described. The feed-forward equations:

$$\mathrm{net}^{(1)} = \mathrm{W}^{(1)}\mathrm{X}. \quad (1)$$

$$\mathrm{O}^{(1)} = \mathrm{f}\left(\mathrm{net}^{(1)}\right). \quad (2)$$

$$\mathrm{net}^{(2)} = \mathrm{W}^{(2)}\mathrm{O}^{(2)} \qquad \text{where} \qquad \mathrm{W}^{(2)} = \mathrm{W}^{(1)\mathrm{T}}. \quad (3)$$

$$\mathrm{O}^{(2)} = \mathrm{f}\left(\mathrm{net}^{(2)}\right). \quad (4)$$

X is the input matrix of the auto-encoder. The weight matrixes of encoder and decoder layers are $W^{(1)}$ and $W^{(2)}$, respectively, which $W^{(1)} = W^{(2)T}$. The matrix $O^{(1)}$ is named encoder output. The decoder's output matrix is $O^{(2)}$. The desired matrix and activation function are indicated by X and f [41]. n_1 and n_2 are the numbers of input and encoder neurons, respectively.

The SAE feed-forward equations are the same as the conventional neural network (*Eq. (1)* to *(4)*). Though, the desired output is equal to the input, which means the output estimates the input and the encoder output is the compression and representation of the input.

The back-propagation equations:

The lost function is the Mean Square Error (MSE) expressed in *Eq. (5)*.

$$E_{MSE} = \frac{1}{2}\sum_{i=1}^{n}\left(x_i - o_i^{(2)}\right)^2. \tag{5}$$

Where x_i and $o_i^{(2)}$ are the i-th vector of the input and decoder output matrix, respectively.

Eq. (6) is the update of the weight matrix:

$$W^{(2)} = W^{(2)} - \eta.\nabla W^{(2)}, \tag{6}$$

where $\nabla W^{(2)}$ is the gradient of $W^{(2)}$ and η is the learning rate.

4.3 | Proposed Supervised K-Fold CV

The CVs are statistical approaches to evaluate ML models' prediction skills [42] and are dependent on the data distribution [43]. Among all types of CVs, the K-Fold CV is popular in APP/APR due to the data's equal chance to participate in the training process [44].

Let us introduce our innovative and simple method to prevent stated obstacles in the introduction. The SK-Fold CV steps are as follows. The bold statements are those that are different from the conventional ROS-based K-Fold CV.

I. Consider the 15% of the original dataset as a testing dataset and others as the training dataset.
II. Choose K (the number of Folds) based on the number of speakers.
III. Split the training dataset into K segments based on the speaker ID. Since there are different numbers of clips per speaker ID, It causes varied instance numbers in each Fold.
IV. Balance data with the ROS method in each Fold. The number of repetitions of the samples is equal to the maximum number of majority class instances in the Foldes.
V. Pick (k – 1) Folds as the new training dataset and the hold-out Fold as development dataset.
VI. Train the model parameters with the new training dataset until the maximum epoch reach. If evaluating the model with the development dataset shows an over-fitting problem, the training process will be stopped before the maximum epoch.
VII. Transfer the parameters of the trained model into the next step.
VIII. Repeat steps 5-7, K times.
IX. Test the model with the final model parameters and testing dataset.

As shown in Steps 1 and 5, the original dataset was divided into three groups (training, development, and testing). Each time, the designed model was trained by the training dataset and evaluated by the development dataset. After the training process was finished, the final model valuation was done by the testing dataset for evaluating the ability to predict unseen data. *Fig. 4* illustrates the proposed K-Fold CV.

In the SK-Fold CV method, the number of speakers has divided into K equal subdivisions. The maximum majority samples in all Folds was considered as repeat level. Then ROS was applied to each fold individually. The training process on each training subset continued to the maximum epoch. Nevertheless, data replication increases the likelihood of over-fitting. The error loss of evaluating the development dataset was calculated in each epoch. If this error loss were ascending in ten consecutive epochs, the training process would be stopped in the training subset (early stopping). The trained model was trained again by the next training subset. This speeded the convergence rate up. The process was repeated K times, where K=10 according to previous studies [35] and [45].

Someone may ask why the data balancing method was applied in Step 4 and not Step 1. The reason is that each speaker had multiple audio clips in the SPC and UTS datasets. The number of audio clips of one speaker was not equal to another, which was the second challenge of these datasets. We had to place all clips of one speaker in one Fold to prevent the placement of some clips of one speaker in both the training and development datasets. Therefore, the random placement of audio clips in the Folds of CV caused the second class imbalance. On the other hand, data balancing methods perform these placements without considering data labels, while artificially increasing or decreasing the data without considering the speaker ID will cause an imbalanced distribution of data in the Folds.

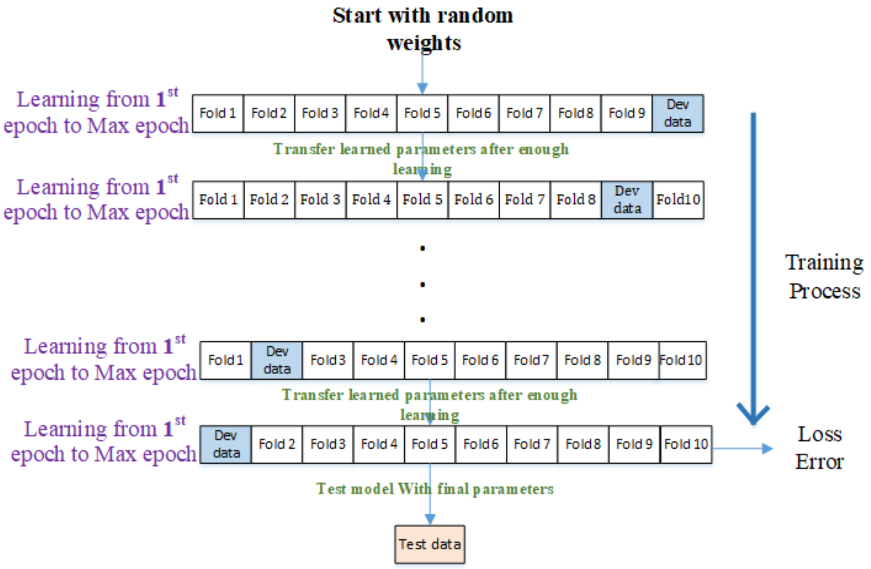

Start with random weights
Learning from 1st epoch to Max epoch
Fold 1 | Fold2 | Fold 3 | Fold4 | Fold 5 | Fold6 | Fold 7 | Fold8 | Fold 9 | Dev data
Transfer learned parameters after enough learning
Learning from 1st epoch to Max epoch
Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Dev data | Fold10
Learning from 1st epoch to Max epoch
Fold1 | Dev data | Fold3 | Fold 4 | Fold 5 | Fold 6 | Fold7 | Fold 8 | Fold9 | Fold10
Transfer learned parameters after enough learning
Learning from 1st epoch to Max epoch
Dev data | Fold 2 | Fold3 | Fold 4 | Fold 5 | Fold6 | Fold 7 | Fold8 | Fold 9 | Fold 10
Training Process
Loss Error
Test model With final parameters
Test data

Fig. 4. Schematic illustration of the SK-Fold CV method for the model evaluation.

4.4 | Evaluation Criteria

The accuracy criterion is a misleading metric for imbalanced dataset problems due to the domination of the majority class [46]. As mentioned previously, the APP and APR datasets are class imbalances, and the UA recall criterion presented in INTERSPEECH 2009 is an appropriate evaluation criterion of models designed on these datasets [31]. It is because the recall of class *low* is not affected by class high and vice versa.

This article uses the accuracy and UA recall expressed as *Eq. (7)* and *(8)* as evaluation criteria.

$$\mathrm{Acc}=\frac{\mathrm{T_{LL}}+\mathrm{T_{HL}}}{\mathrm{T_{LL}}+\mathrm{T_{HL}}+\mathrm{F_{HL}}+\mathrm{F_{LL}}}. \tag{7}$$

T_{LL} is the number of true Low Level (LL). T_{HL} is the number of true High Level (HL). F_{LL} is the number of false low level. F_{HL} is the number of false high level.

$$\mathrm{recall(X)}=\frac{\mathrm{T_X}}{\mathrm{T_X}+\mathrm{F_{NX}}}. \tag{8}$$

The recall is the fraction of the total amount of relevant instances that were retrieved. X indicates relevant instances, and NX is the non-relevant instance.

The UA recall is calculated by *Eq. (9)*.

$$\mathrm{UA}=\frac{\mathrm{recall}\left(\mathrm{LL}\right)+\mathrm{recall}\left(\mathrm{HL}\right)}{2}. \tag{9}$$

The recall (HL) indicates the recall of high-level instances, and recall (LL) is the recall of low-level samples.

5 | Results and Discussion

Researchers have used various kinds of features to improve classification outcomes of APP/APR based on speech signals. They have shown that each personality trait has different effects on different features [2], [18] and [49]. Since our study focuses on improving the evaluation method, not feature extraction, we used twenty-four prevalent statistical features. Although these features were not appropriate for all five personality traits classification, network inputs had to be considered constant among all five traits to compare the proposed evaluation method's effectiveness. Hence it was expected that classification results were not satisfactory for some personality traits in both datasets.

The following principle was utilized in the four simulations: RUS-based K-Fold CV, ROS-based K-Fold CV, SMOTE-based K-Fold CV, and SK-Fold CV. In four CVs, the obtained twenty-four statistical features were given to the five separate SAEs. The motive of using separate networks is that research has shown that personality traits affect speech characteristics differently, and it outperforms if implemented separately [46]. Each SAE contained fully connected auto-encoders with 15 and 8 neurons in their encoder layers. The number of neurons and layers was chosen by grid search. The activation function in each layer was tanh. This function has positive, negative, linear, and nonlinear ranges and is suitable for feature exploring. The initial weights were set the same for four CVs so that the simulation results were not affected by initial weight values. The batch normalization method was used in each layer for normalizing layer input. The dropout technique was utilized for preventing over-fitting in each layer. Since the SAE is a neural network, it can be used as the feature extractor and classifier together. Therefore, a layer with one neuron and the sigmoid activation function was used as the output layer to classify each trait's low level (lower than 0.5) or high level (higher than 0.5). The loss function was the MSE, and the loss error was back-propagated to the layers for fine-tuning.

Since the SK-Fold CV is designed on the ROS method, we compare loss diagrams of ROS-based K-Fold CV and our CV below. *Fig. 5* and *8* illustrate the training and development error diagrams of implementing the 10-Fold CV and the S10-Fold CV for Conscientiousness trait in both datasets. In *Fig. 6*, the loss error is approximately equal to 8×10^{-2} in epoch 180, while in *Fig. 5* (the conventional CV), the same amount of error is found in epoch 1000 for the UTS dataset. *Fig. 7* and *Fig. 8* compare conventional CV and SK-fold CV for the SPC dataset and show an increase in convergence speed. As mentioned earlier, one drawback of the oversampling methods is that the model training time increases by increasing the size of training data. Therefore the increase in convergence speed is the first benefit of our proposed method.

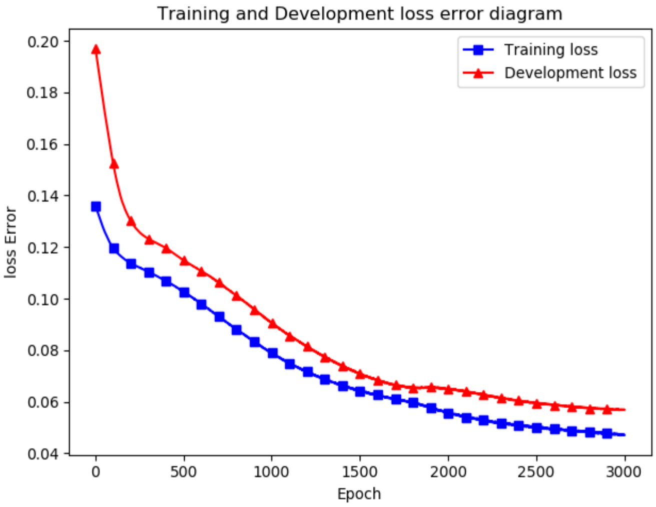


Fig. 5. Training and development loss error diagrams for conscientiousness trait classification in the UTS dataset evaluated by conventional 10-Fold CV.

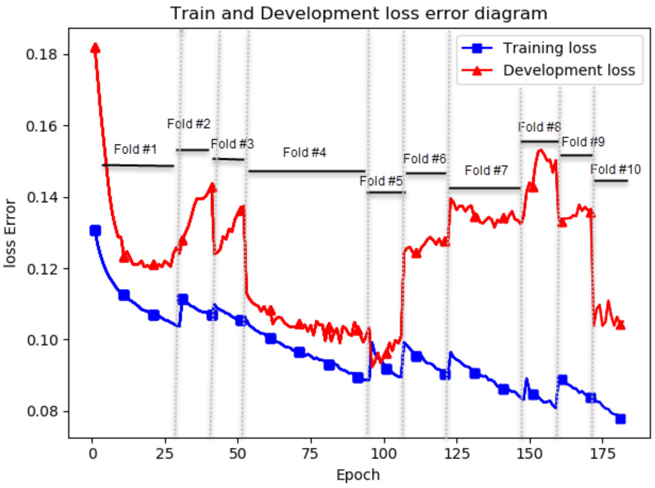


Fig. 6. Training and development loss error diagrams for conscientiousness trait classification in the UTS dataset evaluated by the S10-Fold CV method.

It can worth mentioning that *Fig. 5* and *Fig. 7* are the average diagrams of K Folds training, but *Fig. 6* and *Fig. 8* illustrate all K Fold training in one plot. It is another benefit of our method that the over-fitting in all Folds can be traced. Contrary to the CV method that the over-fitting in the average development loss in K iterations is considered ($Error = \frac{1}{K}\sum_{i=1}^{K} E_i$) and if over-fitting occurs in one subset of the training dataset, it will be ignored until the average of K faces over-fitting.

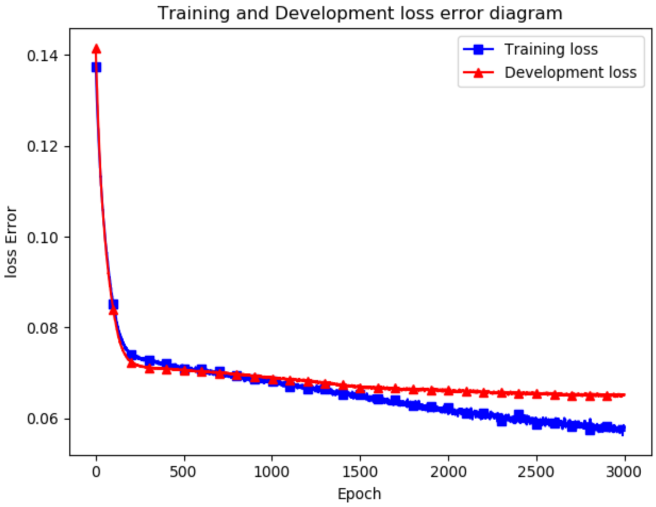


Fig. 7. Training and development loss error diagrams for conscientiousness trait classification in the SPC dataset was evaluated using the conventional 10-Fold CV method.

Table 3 and *Table 4* indicate the personality traits classification results in terms of accuracy and UA for the UTS and SPC datasets. In both tables, the outcomes under column SK-Fold CV show the simulation results with our method. The three extra implementations based on RUS, ROS and SMOTE methods were done to better show the proposed method's outstanding results.

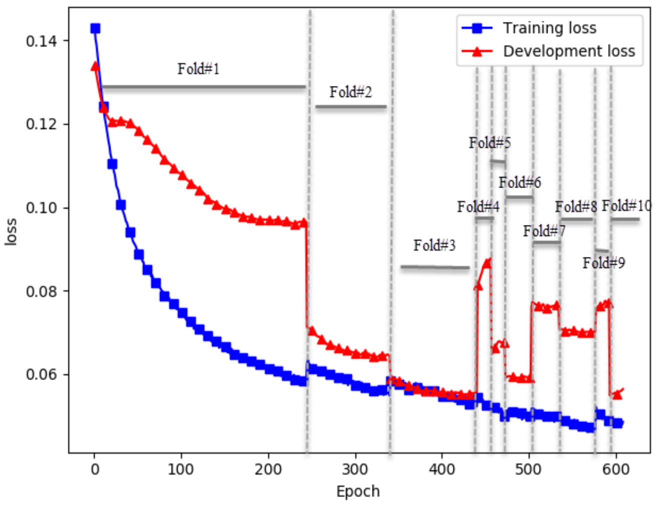


Fig. 8. Training and development loss error diagrams for conscientiousness trait classification in the SPC dataset evaluated by the S10-Fold CV method.

Table 3. Comparison of personality classification by K-Fold CV and SK-Fold CV methods in SPC dataset.

Traits	K-Fold CV			SK-Fold CV
	RUS	ROS	SMOTE	
Neu.	63.4 (60.7)	65.7 (68.3)	64.4 (62.5)	66.1 (72.1)
Ext.	55.3 (54.7)	61.9 (63.9)	59.1 (57.3)	62.5 (66.7)
Ope.	55.9 (60.1)	60.8 (73.4)	61.5 (72.1)	65.1 (77.5)
Agr.	59.9 (55.6)	60.5 (53.8)	60.1 (64.3)	61.3 (68.1)
Con.	53.1 (54.6)	61.6 (77.1)	63.1 (66.3)	66.9 (72.6)
Average	57.5 (57.1)	62.1 (67.3)	61.6 (64.5)	64.5 (71.4)

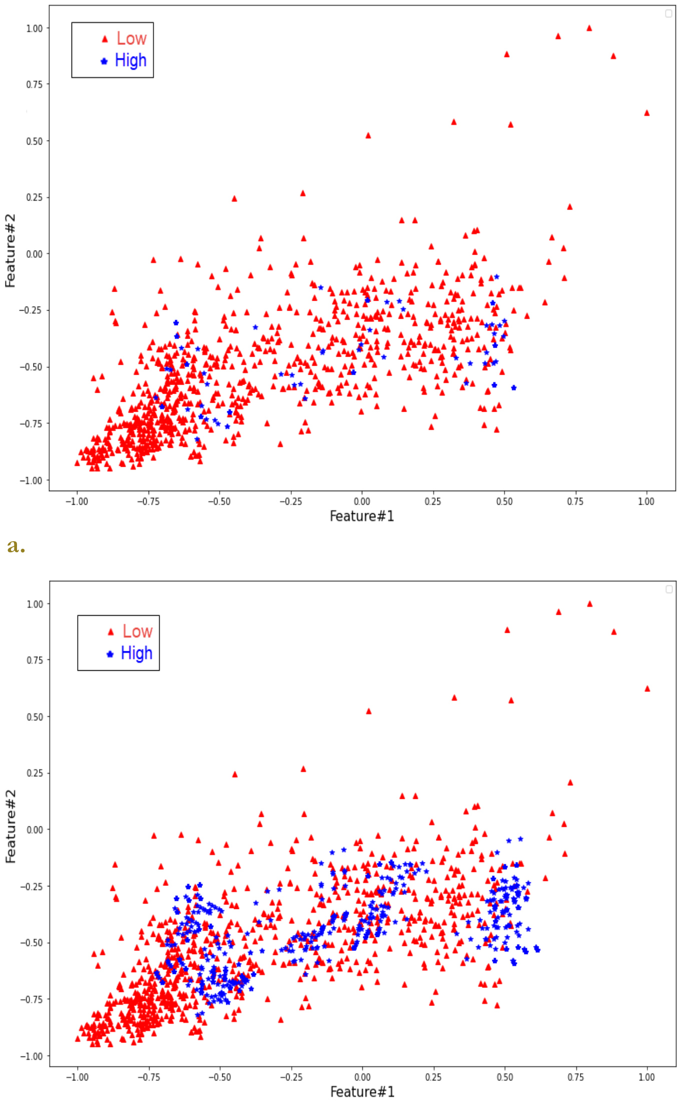
Table 4. Comparison of personality classification by K-Fold CV and SK-Fold CV methods in UTS dataset.

Traits	K-Fold CV			SK-Fold CV
	RUS	ROS	SMOTE	
Neu.	34.6 (49.1)	60 (38.2)	42.7 (59.6)	67.2 (63.6)
Ext.	47.9 (53.1)	52.3 (64)	50.7 (57.3)	54.7 (70.6)
Ope.	64.1 (52.1)	64.2 (48)	64.2 (51.8)	64.4 (55.1)
Agr.	43.1 (46.8)	50.2 (40.1)	49.7 (59.1)	54.7 (62.8)
Con.	49.3 (50.1)	50.4 (53.4)	53.4 (56.8)	65.3 (59.6)
Average	47.8 (50.2)	49 (55.1)	52.5 (56.9)	61.2 (62.3)

From *Table 3* and *Table 4*, the four consequences can be concluded as follows.

I. The RUS method had a poor performance in our study. Because low class and high class are of equal importance, removing samples from each class reduces the classification results due to diminishing the sample size of the model learning.

II. Despite the success of SMOTE in several applications, it was not appropriate for some personality traits such as Neuroticism in the UTS dataset. Because generated samples have correlations with some original minority samples. Also, SMOTE can decrease between-class scattering by increasing overlapping between classes [17]. Suppose the two classes, despite having severely imbalanced data distribution, are easily separable in two groups. In that case, it becomes easy for any classifier to learn to discriminate between them. The challenges become profound when the classes are interspersed, and the generation of new samples in feature space leads the model into difficulty, as is often the case with real-world applications like our study. Since twenty-four features have low resolution, we have to use the auto-encoder feature engineering ability, but the sample generation by SMOTE reduced the separability even less than before and caused poor performance in traits classification. *Fig. 9* is prepared for better illustration. In this figure, Feature#1 and Feature#2 was statistical features obtained by applying the maximum function on the pitch and the first formant of the Neuroticism trait in the UTS dataset, respectively. As shown, these two features are not separatable and using SMOTE method decreased the between-class scatter of this trait severely. Although changing the model can get better performance of SMOTE, for a fair comparison, we had to keep the model of feature extraction and classification the same in all simulations. As the Neuroticism trait in the UTS dataset was severe imbalance class, we used it for *Fig. 9*.



a.

b.

Fig. 9. SMOTE effect on the Neuroticism trait in the UTS dataset; a. original data, b. data balancing with SMOTE.

III. As far as *Tables 3* and *4* show, it is perfectly understandable that oversampling is better due to keeping all the information in the training dataset. In this regard, SK-Fold CV and ROS-based CV results were near. However, by overcoming the over-fitting and time-consuming, the simulation outcomes of the proposed method are better than the traditional one. As *Table 3* shows, the UA recall is improved by an average of 2.46% compared to ROS in the SPC dataset. In *Table 4*, the proposed evaluation method could enhance the UA recall results compared to ROS by an average of 12.2% in the UTS dataset. UA recall enhancement is the third benefit of our method.

IV. As is clear, the UA recall of Openness in three K-Fold CVs and SK-Fold CV methods is approximately the same in the UTS dataset. Also, the Nerutisem and Agreeableness UA recalls are about the same as the conventional K-Fold CV in the SPC dataset. It is because the number of samples in the low and high levels of these traits is approximately equal (see *Tables 1* and *2*). It confirms that the proposed evaluation method performs the same as the traditional CV method when the two classes' samples are near equal. However, it improves the UA recall if there is an imbalance distribution in classes. This is the fourth advantage of SK-Fold CV.

6 | Conclusions

Research of APP and APR based on the big-five inventory needs more data for outperforming, but it is ideal. The challenge of data collection and data annotation restricts the research. Besides, imbalanced personality distribution across society is the reason for the low classifier performance. This study considered the pons and cons of some data-level techniques for addressing imbalanced class problems. We used the feature engineering ability of the auto-encoder to extract appropriate features from hand-crafted features and classify them by stacked autoencoder. Then a supervised K-Fold CV method was proposed to evaluate the model's generalisation ability, cope with affecting the imbalanced class and enhance the classification results. The consequences reveal that our proposed evaluation method enhances the UA results in all personality traits in both datasets.

Conflicts of Interest

All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

References

[1] Niebuhr, O. (2021). *Relations between speech sciences and industry*. Retrieved from http://www.letras.ufmg.br/padrao_cms/index.php?web=lbass&lang=1&page=3732&menu=&tipo=1

[2] Schultz, D. P., & Schultz, S. E. (2016). *Theories of personality*. Cengage Learning.

[3] Hagh-Shenas, H. (2017). *Big 5 personality*. Ravansanji. (**In Persian**). https://www.gisoom.com/book/11287274/

[4] Phan, L. V., & Rauthmann, J. F. (2021). Personality computing: new frontiers in personality assessment. *Social and personality psychology compass*, *15*(7), e12624. https://doi.org/10.1111/spc3.12624

[5] Komarraju, M., Karau, S. J., Schmeck, R. R., & Avdic, A. (2011). The Big Five personality traits, learning styles, and academic achievement. *Personality and individual differences*, *51*(4), 472-477.

[6] Mohammadi, G., & Vinciarelli, A. (2015, September). Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract. *2015 international conference on affective computing and intelligent interaction (ACII)* (pp. 484-490). IEEE.

[7] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... & Escalera, S. (2016, October). Chalearn lap 2016: first round challenge on first impressions-dataset and results. *European conference on computer vision* (pp. 400-418). Springer, Cham.

[8] Rosenberg, A. (2018, June). Speech, prosody, and machines: nine challenges for prosody research. *9th international conference on speech prosody* (pp. 784-793). Speech Prosody. DOI: 10.21437/SPEECHPROSODY.2018-159

[9] Junior, J. C. S. J., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., Baró, X., ... & Escalera, S. (2019). First impressions: a survey on vision-based apparent personality trait analysis. *IEEE transactions on affective computing*, *13*(1), 75-95. DOI: 10.1109/TAFFC.2019.2930058

[10] Vinciarelli, A., & Mohammadi, G. (2011). Towards a technology of nonverbal communication: vocal behavior in social and affective phenomena. In *Affective computing and interaction: psychological, cognitive and neuroscientific perspectives* (pp. 133-156). IGI Global.

[11] An, G., & Levitan, R. (2018, February). Lexical and acoustic deep learning model for personality recognition. In *Interspeech* (pp. 1761-1765). https://www.isca-speech.org/archive_v0/Interspeech_2018/pdfs/2263.pdf

[12] Zhao, S., Xu, Z., Liu, L., Guo, M., & Yun, J. (2018). Towards accurate deceptive opinions detection based on word order-preserving CNN. *Mathematical problems in engineering*, *2018*. https://doi.org/10.1155/2018/2410206

[13] Ewen, R. B. (2014). *An introduction to theories of personality*. Psychology Press.

[14] Ghosh, A., Hossain, A. A., & Raju, S. T. U. (2021, March). Classification of diabetic retinopathy using few-shot transfer learning from imbalanced data. *7th international conference on advanced computing and communication systems (ICACCS)* (Vol. 1, pp. 78-83). IEEE.

[15] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of big data*, *5*(1), 1-30.

[16] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

[17] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, *6*(1), 1-54.

[18] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[19] Yan, Y., Chen, M., Shyu, M. L., & Chen, S. C. (2015, December). Deep learning for imbalanced multimedia data classification. *IEEE international symposium on multimedia (ISM)* (pp. 483-488). IEEE.

[20] Tan, H. H., & Lim, K. H. (2019, June). Vanishing gradient mitigation with deep learning neural network optimization. *7th international conference on smart computing & communications (ICSCC)* (pp. 1-4). IEEE.

[21] Chen, Z., & Lin, T. (2017). Automatic personality identification using writing behaviours: an exploratory study. *Behaviour & information technology*, *36*(8), 839-845.

[22] Zarnoufi, R., & Abik, M. (2019, November). Big five personality traits and ensemble machine learning to detect cyber-violence in social media. *International conference Europe Middle East & North Africa information systems and technologies to support learning* (pp. 194-202). Springer, Cham.

[23] Khan, A. S., Ahmad, H., Asghar, M. Z., Saddozai, F. K., Arif, A., & Khalid, H. A. (2020). Personality classification from online text using machine learning approach. *International journal of advanced computer science and applications*, *11*(3), 460-476.

[24] Horvath, A., Dras, M., Lai, C. C., & Boag, S. (2021). Predicting suicidal behavior without asking about suicidal ideation: machine learning and the role of borderline personality disorder criteria. *Suicide and life-threatening behavior*, *51*(3), 455-466.

[25] Aylett, M. P., Vazquez-Alvarez, Y., & Butkute, S. (2020, March). Creating robot personality: effects of mixing speech and semantic free utterances. *Companion of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 110-112). Association for Computing Machinery, New York, NY, United States. https://doi.org/10.1145/3371382.3378330

[26] Maillard, P., Pellaton, J., & Kramer, U. (2019). Treating comorbid depression and avoidant personality disorder: the case of Andy. *Journal of clinical psychology*, *75*(5), 886-897.

[27] Pohjalainen, J., Kadioglu, S., & Räsänen, O. (2012). Feature selection for speaker traits. *Thirteenth annual conference of the international speech communication association* (pp. 270-273). ISCA Archive. https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2012/i12_0270.pdf

[28] Fayet, C., Delhay, A., Lolive, D., & Marteau, P. F. (2017, August). Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus. In *Interspeech* (pp. 3281-3285). Stockholm, Sweden. https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/1194.PDF

[29] Schuller, B., Steidl, S. and Batliner, A. (2009). *The interspeech 2009 emotion challenge* (No. 211486). Retrieved from https://d-nb.info/1218972491/34

[30] An, G., Levitan, S. I., Levitan, R., Rosenberg, A., Levine, M., & Hirschberg, J. (2016). Automatically Classifying Self-Rated Personality Scores from Speech. In *Interspeech* (pp. 1412-1416). San Francisco, USA. https://www.isca-speech.org/archive_v0/Interspeech_2016/pdfs/1328.PDF

[31] Jothilakshmi, S., Sangeetha, J., & Brindha, R. (2017). Speech based automatic personality perception using spectral features. *International journal of speech technology*, *20*(1), 43-50.

[32] Mohammadi, G., Origlia, A., Filippone, M., & Vinciarelli, A. (2012, October). From speech to personality: Mapping voice quality and intonation into personality differences. *Proceedings of the 20th ACM international conference on Multimedia* (pp. 789-792). Association for Computing Machinery, New York, NY, United States. https://doi.org/10.1145/2393347.2396313

[33] Koutsombogera, M., Sarthy, P., & Vogel, C. (2020, September). Acoustic features in dialogue dominate accurate personality trait classification. *IEEE international conference on human-machine systems (ICHMS)* (pp. 1-3). IEEE.

[34] Kampman, O., Siddique, F. B., Yang, Y., & Fung, P. (2019). Adapting a virtual agent to user personality. In *Advanced Social Interaction with Agents* (pp. 111-118). Springer, Cham.

[35] Liu, Z. T., Rehman, A., Wu, M., Cao, W. H., & Hao, M. (2020). Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features. *IEEE transactions on multimedia*, *23*, 3414-3426.

[36] Diener, E. and Lucas, R.E. (2019). *Personality traits*. Retrieved from https://nobaproject.com/modules/personality-traits

[37] Mohammadi, G., Vinciarelli, A., & Mortillaro, M. (2010, October). The voice of personality: mapping nonverbal vocal behavior into trait attributions. *Proceedings of the 2nd international workshop on Social signal processing* (pp. 17-20). Association for Computing Machinery, New York, NY, United States. https://doi.org/10.1145/1878116.1878123

[38] Fallahnezhad, M., Vali, M., & Khalili, M. (2017, May). Automatic Personality Recognition from reading text speech. In *2017 Iranian Conference on Electrical Engineering (ICEE)* (pp. 18-23). IEEE.

[39] Zaferani, E. J., Teshnehlab, M., & Vali, M. (2021). Automatic personality traits perception using asymmetric auto-encoder. *IEEE access*, *9*, 68595-68608.

[40] Le, Q. V. (2015). A tutorial on deep learning part 2: autoencoders, convolutional neural networks and recurrent neural networks. *Google brain*, *20*, 1-20.

[41] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), 1236-1246.

[42] Balasuriya, J., & Yang, Y. (2019). The role of personality traits in pension decisions: findings and policy recommendations. *Applied economics*, *51*(27), 2901-2920.

[43] Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE intelligent systems*, *32*(6), 74-80. https://sentic.net/sentiment-analysis-suitcase.pdf

[44] Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. *Journal of nonparametric statistics*, *30*(1), 197-215.

[45] Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2020). Recent trends in deep learning based personality detection. *Artificial intelligence review*, *53*(4), 2313-2339.

[46] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., ... & Weiss, B. (2015). A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Computer speech & language*, *29*(1), 100-131.