Circuits, Systems, and Signal Processing (2023) 42:1707–1722
https://doi.org/10.1007/s00034-022-02189-y

Detection of Common Cold from Speech Signals using Deep Neural Network

Suman Deb[1] · Pankaj Warule[1] · Amrita Nair[1] · Haider Sultan[1] · Rahul Dash[1] · Jarek Krajewski[2]

Received: 4 February 2022 / Revised: 12 September 2022 / Accepted: 12 September 2022 /
Published online: 3 October 2022

Abstract

This paper presents a deep learning-based analysis and classification of cold speech observed when a person is diagnosed with the common cold. The common cold is a viral infectious disease that affects the throat and the nose. Since speech is produced by the vocal tract after linear filtering of excitation source information, during a common cold, its attributes are impacted by the throat and the nose. The proposed study attempts to develop a deep learning-based classification model that can accurately predict whether a person has a cold or not based on their speech. The common cold-related information is captured using Mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) from the speech signal. The data imbalance is handled using the sampling strategy, SMOTE–Tomek links. Then, utilizing MFCC and LPC features, a deep learning-based model is trained and then used to categorize cold speech. The performance of a deep learning-based method is compared to logistic regression, random forest, and gradient boosted tree classifiers. The proposed model is less complex and uses a smaller feature set while giving comparable results to other

✉ Suman Deb
sumandeb@eced.svnit.ac.in

Pankaj Warule
d20ec007@eced.svnit.ac.in

Amrita Nair
amrita.nair1999@gmail.com

Haider Sultan
haider.sultan406@gmail.com

Rahul Dash
rdash242@gmail.com

Jarek Krajewski
jarek.krajewsk@rfh-koeln.de

1 Sardar Vallabhbhai National Institute of Technology, Surat 395007, India

2 Rhenish University of Applied Sciences, 50678 Cologne, Germany

state-of-the-art methods. The proposed method gives an UAR of 67.71%, higher than the benchmark OpenSMILE SVM result of 64%. The study's success will yield a noninvasive method for cold detection, which can further be extended to detect other speech-affecting pathologies.

Keywords Cold speech · MFCC · LPC · Gradient boosted trees · Random forest · Deep neural network

1 Introduction

Research on speech-based health assessment is gaining interest due to its noninvasive nature and ease of transmission. The speech-based health assessment system may be embedded into smart devices with microphones and wearable technologies, which will be highly helpful for remote and early diagnosis of various medical disorders. The key benefit of a speech-based health assessment system is that it may only diagnose and monitor patient health from speech signals, enabling patients to avoid repeated hospital visits for medical examinations.

The common cold is a viral infectious disease of the upper respiratory tract. A person suffering from a common cold may show various symptoms such as coughing, sore throat, runny nose, sneezing, and hoarseness [39]. Human speech is produced when the air from the lungs passes through the vocal cords, which may vibrate or let it pass, and then is shaped by the mouth and the nose. Since common cold symptoms affect the nose, throat, and hence the vocal tract, the characteristics of the speech of a person suffering from the common cold are different compared to normal speech. Cold speech, or speech of a person suffering from the common cold, is observed to have a lower pitch, increase in noise due to hoarseness or coughing, and change in the timbre of the voice [41].

Upper Respiratory Tract Infections, such as the common cold and influenza (flu), are a major public health concern, causing 3 to 5 million cases of serious infection per year [13]. Detecting common cold and related illnesses through speech may be useful in preventing the spread of these viral infections. It may be beneficial to monitor a patient's health remotely. Recognizing whether a person has a common cold from their speech can be beneficial for speech signal processing tasks. Speech recognition [27], speaker verification [38], and emotion detection systems [14, 24] are trained on normal speech, so their performance is negatively affected when tested on cold speech. Analysis of cold speech can make these systems more robust.

This study uses the Upper Respiratory Tract Infection Corpus (URTIC) dataset. The URTIC dataset consists of recordings of 630 male and female subjects with a mean age of 29.5 years, sampled at 16KHz [33]. These 630 recordings are split into 28,652 segments of 3–10 s duration. Out of the total 28,652 segments, 2876 belong to the positive 'Cold' class, while the rest are from the negative 'No cold' class. Our URTIC corpus provides important complementary data within the challenge of capturing Covid-19 from speech. Without the common cold as a reference, false alarm errors might occur, misclassifying the common cold as Covid-19. Analysis of common cold can further enable differential diagnosis of Covid-19 [11, 16].

There have been several attempts at detecting cold from speech using the Upper Respiratory Tract Infection Corpus (URTIC) dataset [33]. The URTIC dataset has been utilized in the INTERSPEECH 2017 Computational Paralinguistics Cold Subchallenge ComParE 2017 [33]. The baseline of the ComParE-2017 challenge uses an end-to-end learning technique with a convolutional neural network (CNN) and long short-term memory (LSTM), ComParE-2013 feature set, and bag-of-audio-words (BoAW) features. ComParE-2013 has 6373 static features computed from various functionals over low-level descriptors (LLD). For BoAW, from the ComParE feature set, one codebook for 65 LLD and 65 deltas of these LLD was revealed. Gosztolya et al. [18] applied two classification approaches using SVM classifier, one trained on frame-level features and the other one on utterance-level features. The training instances were randomly down-sampled, and the average of a 100 posterior score was used during classification. Kaya et al. [26] utilized a weighted partial least square regression classifier and combination of RASTA-PLP and MFCC features encoded by Fisher vector for classification. The class imbalance problem in the dataset was handled by a weighting scheme in the kernel classifier. Tavarez et al. [36] proposed a fusion of the Gaussian mixture model (GMM) and cosine distance classifiers using features such as magnitude spectrum, relative phase shift, and suprasegmental features (MFCC, LPCC). The fusion of classifier decisions improved the performance of standalone detectors. Wagner et al. [41] investigated the effects of cold on the phonetic level of speech. They derived an automatic speech recognizer (ASR)-based phonetic transcription, and based on phonetic transcription, and separate classifiers were trained using features like MFCC, invariant-integration features (IIF), and constrained maximum likelihood linear regression (CMLLR). However, they concluded that phoneme-level classification was not worth the effort. Huckvale and Beke [21] examined the best features for the detection of cold, including voice quality features (energy, entropy), voice spectral features (MFCC), modulation spectrogram, and spectral envelope features. They concluded that the DCT-coded modulation spectrogram gave the best results. Cai et al. [7] utilized perception aware MFCC and constant Q cepstral coefficients (CQCC) features for the detection of common cold. Suresh et al. [35] analyzed the effect of the common cold on sound quality using phoneme state posteriorgram (PSP) characteristics. The Gaussian mixture model (GMM) was used to construct 120-dimensional PSP features for each frame from 120 states of 40 three-state phonetic hidden Markov models (HMM). In our previous work [15], we used variational mode decomposition to extract various statistical amplitude, entropy, and energy features from the speech signal. A mutual information-based weight assignment strategy and the SVM classifier were used for classification. Kao et al. [25] proposed a method for automatic detection of cold speech using discriminative autoencoders and strength modeling with multiple sub-directory generations. Teixeira et al. [37] proposed fully homomorphic encryption (FHE)-based encrypted neural network (ENN) using an extended Geneva minimalistic acoustic parameter set (eGeMAPS) for cold speech classification without any performance degradation compared to the unencrypted neural network (NN). Albes et al. [1] used a combination of pruning and quantization to reduce network size by up to 95% while maintaining accuracy for common cold detection on the URTIC database. This will be helpful for deploying the speech-based artificial intelligent system on device which has memory constraints. Vicente et al. [17] derived Fisher vector

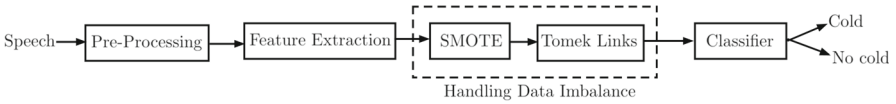

Fig. 1 Block diagram of proposed method

using MFCC features and generative Gaussian mixture model for identification of cold speech. Power normalization and principal component analysis on Fisher vectors further improved the classifier performance. Warule et al. [43] classified cold speech using only vowel-like region segments of speech. They concluded that MFCC features extracted from only vowel-like region segments of speech reduces time and computational complexity without overly affecting the recognition performance.

In biomedical diagnosis problems, the datasets are often highly imbalanced, with only a small percentage of samples belonging to the positive, i.e., anomalous class. In the URTIC dataset, only 2876 samples belong to the positive 'Cold' class, whereas the 'No cold' category contains 25776 samples. The classifiers are biased toward the majority class as they are trained to minimize classification error. This can lead to most of the samples being labeled as the negative class and failure of the classifier to detect the disease. This paper proposes the use of a combination of sampling strategies, a weighting scheme, and a modified binary cross-entropy loss function for training the classifiers to overcome the dataset imbalance. Thus, the classifiers favor the minority class, leading to an improved average recall.

The paper is organized as follows. Section 2 is an in-depth explanation of the pre-processing, feature extraction and sampling method. Section 3 explores the classifiers. The results are shown in Sect. 4 and results are discussed in Sect. 5. Finally, the conclusion is drawn in Sect. 6.

2 Method

The proposed method for the detection of common cold from speech signals is illustrated in Fig. 1. The method involves pre-processing of speech data, feature extraction, sampling strategies and classification using logistic regression, random forest, gradient boosted trees, and deep neural network classifiers.

2.1 Pre-Processing

The pre-processing steps include normalization and removal of silence regions from the speech samples. The speech samples are normalized such that they have zero mean and unit variance. Then, the regions of silence are removed from the speech sample based on short-term energy and spectral centroid thresholds. After this, we extract the desired features from the speech samples.

2.2 Feature Extraction

Feature extraction is the process of extracting the relevant information from the speech signal. For the proposed work, we have used Mel-frequency cepstral coefficients (MFCCs) and linear prediction coefficients (LPCs).

2.2.1 Mel-Frequency Cepstral Coefficients

Human speech is filtered by the shape of the vocal tract, which manifests itself in the envelope of the short time power spectrum. Mel-frequency cepstral coefficients capture this envelope. In this work, we extract 13 Mel-frequency cepstral coefficients (MFCCs), followed by the delta and delta-delta coefficients from the speech samples, giving a 39 sized feature vector. Mel scale is used to divide the frequency into sub-bands, and then, DCT is used to extract the MFCCs [22].

The speech samples are first divided into frames of 20ms with a shift of 10ms. Hamming window is used in each of the frames. This makes the ends of the frames continuous and prevents distortions while taking discrete Fourier transform (DFT). A 2048-point DFT is then applied on each frame, and the positive part of the spectrum is taken. The power spectrum of the speech sample is computed as

$$P = |\mathrm{DFT}(x_i)|^2/N \tag{1}$$

where N is taken as 2048 and x_i is the i^{th} frame of individual speech sample. Next, the Mel filter points are computed to obtain a Mel filter bank. The conversion from Hertz scale to Mel scale is given by

$$M(f) = 1125\ln(1 + f/700) \tag{2}$$

where f is frequency in Hertz. In the Mel scale, 40 equally spaced triangular band-pass filters are constructed with a response of 1 at the center frequency and linearly decreasing till 0 at the center frequencies of adjacent filters. The filter banks are converted back into Hertz scale and normalized. The log of the power spectrum is then filtered by the normalized Mel filter bank to obtain the filter bank coefficients.

The MFCCs are highly correlated, and so DCT-III is applied to decorrelate them. The inverse DCT is computed as

$$X_k = \frac{1}{\sqrt{2}}x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N}n\left(k + \frac{1}{2}\right)\right] \tag{3}$$

where $k = 0, 1, \ldots, N-1$, x_n is the n^{th} filter bank coefficient, and N is 2048. The first 13 coefficients are retained, and the rest are discarded.

In addition to these coefficients, the delta coefficients are also computed to represent the dynamics of the coefficients. The delta coefficients are computed as

$$dt = \sum_n n\,(C_{t+n} - C_{t-n}) \,/\, 2\sum_n n^2 \tag{4}$$

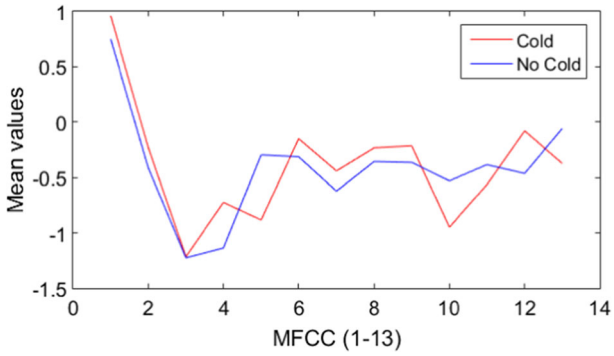

Fig. 2 Mean values of the MFCC feature for the Cold and No cold classes of speech

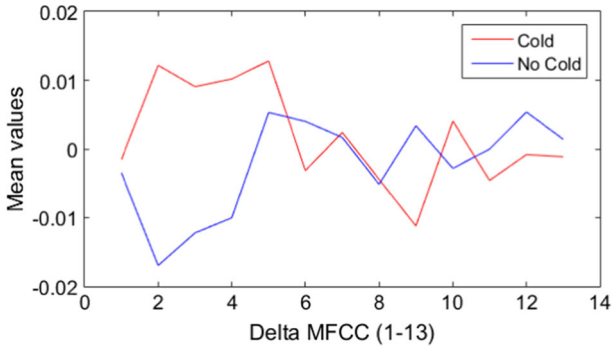

Fig. 3 Mean Values of the delta MFCC feature for the Cold and No cold classes of speech

where N is taken as 2 and t is the frame number. Delta-delta or acceleration coefficients are computed from the delta coefficients. Thus, we obtain a total 39 features from each speech frame, which are then used for the purpose of classification.

Figure 2 shows the mean values of 13 MFCC features. The mean values are calculated from first 1000 samples of training dataset. It is observed that variations of MFCC features are different for 'Cold' class and 'No cold' class. Similar variations are observed with delta MFCC features as shown in Fig. 3.

2.2.2 Linear Predictive Coding Coefficients

Linear predictive (LP) coding supplies us with an appropriate model for speech recognition. It helps estimate the spectral envelope of the speech signal and is extensively used for compressing the audio signal. It provides us a small number of speech parameters called the LP coefficients (LPCs) that relate to the characteristics of the vocal tract and thus the sound being uttered. These coefficients can be used and stored as templates for pattern recognition. LPC calculates the subsequent speech signal sample based on the linear combination of the prior speech sample. The method to calculate the LPC coefficients is shown in Fig. 4.

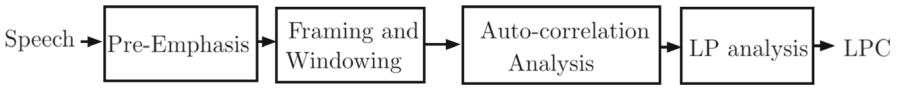

Fig. 4 LPC feature extraction

Pre-Emphasis After sampling of the speech signal, pre-emphasis is required. Filtering the speech signal gives us a smooth spectral shape of the same. Pre-emphasis filter based on the relation of input/output on time domain which is shown by the equation

$$y(n) = x(n) - ax(n-1) \tag{5}$$

a is a constant of pre-emphasis filter, ordinary have $0.9 < a < 1.0$.

Framing and windowing Here, the speech signal become some frame which overlaps to avoid any signal loss. Windowing functions are applied to minimize discontinue signal at the frame edge. Generally, we use hamming window with the form:

$$w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \tag{6}$$

where $0 \leq n \leq N-1$.

Autocorrelation analysis The next step is autocorrelation analysis toward each frame result by windowing. This process calculates the similarity between a piece of current and a prior speech frame.

LPC Analysis and coefficient In this work, we have used the Burg algorithm for LPC feature extraction. The Burg algorithm [28] is a constrained least-squares estimation procedure. It uses the summation of the backward and forward linear prediction error energies. Assuming an all-pole stationary stochastic process, the forward linear prediction error $f_{M,k}$ is given by

$$f_{M,k} = x_{k+M} + \sum_{i=1}^{M} a_{M,i} x_{k+M-i} = \sum_{i=0}^{M} a_{M,i} x_{k+M-i} \tag{7}$$

for 1 ≤ k ≤ N - M.

The backward linear prediction error $b_{M,k}$ is given by

$$b_{M,k} = \sum_{i=0}^{M} a^*_{M,i} x_{k+i} \tag{8}$$

also for 1 ≤ k ≤ N - M. The sum of the forward and backward prediction error energies are minimized to obtain the estimates and subject to the constraint that Levinson recursion is satisfied, for all orders from 1 to M. We have also experimented with covariance and the Yule–Walker algorithm for the LPC feature extraction. The best performance is obtained utilizing the Burg algorithm. Therefore, we have reported all results based on the LPC feature extracted using the Burg algorithm.

2.3 Handling Data Imbalance

The URTIC dataset is highly imbalanced, with only 2876 samples having 'Cold.' On the other hand, 'No cold' class contains 25776 samples. This can cause the classifier to overfit on the 'Cold' samples and negatively affect its performance. To overcome this, re-sampling techniques are employed on the training data to obtain a more balanced dataset. First, the MFCC features are up-sampling using the synthetic minority over-sampling technique (SMOTE) [8], which randomly generates synthetic samples from the minority class samples. It is preferred to random over-sampling as we obtain new samples, instead of replicating the 'Cold' samples. This is followed by under-sampling using Tomek links.

Tomek links are closely related pairs of instances belonging to opposite classes. Applying SMOTE may lead to the creation of many such pairs, making it difficult for the classifier to separate into opposite classes [20]. We can remove either the majority instance from the Tomek links or all such instances. We examine the effect of different sampling ratios and the removal of majority class instances or all such instances on classifier performance.

3 Classifier

For the classification of speech as 'Cold' and 'No cold,' different models have been examined, namely logistic regression, random forest, gradient boosted trees, and deep neural network. This section explains the architectures of these classifiers. The hyper-parameters for the logistic regression, random forest gradient boosted trees classifiers were set using Bayesian optimization, which uses the Bayes theorem to search for the optimum hyperparameters which give the best UAR score. The *HyperOpt* toolkit [5] has been used to implement the hyperparameter tuning of all classifiers.

3.1 Logistic Regression Classifier

Logistic regression (LR)-based classifier attempts to describe the relationship between a discrete response variable and one or more independent variables [6]. The logistic distribution function proposed by Cox [12] is given as follows:

$$\pi(x) = \frac{e^x}{1 + e^x} \tag{9}$$

The range of the logistic function is between 0 and 1, which can be interpreted as probabilities of a sample belonging to a particular class. In this work, feature vector is initially sampled using random under-sampling, SMOTE, or a combination of SMOTE and Tomek links, and is then classified with a cost-sensitive logistic regression classifier. Additionally, L2 regularization has been used to avoid overfitting. During training, value of C was 0.5 used for linear regression.

3.2 Random Forest Classifier

Random forest (RF) classifier is an ensemble of a set of decision trees built in randomly selected subspaces of the feature space [19]. Since decision trees are vulnerable to overfitting on the training data, random forest classifier is preferred. For classification problems, the class selected by majority of the trees in the forest is taken as the output class. To overcome the imbalanced data problem in our work, the feature vector is first re-sampled and then classified using a cost-sensitive random forest classifier.

3.3 Gradient Boosted Trees

Gradient boosted trees or XGBoost comprise of an ensemble of weaker decision trees. It optimizes the performance of decision trees by training them in a gradual, additive manner. In this work, the open-source *XGBoost* package has been used, which speeds up learning of gradient boosted trees using parallel and distributed computing. In addition, classes have been weighted to combat the dataset imbalance. During training, maximum tree depth was set to 5.0 and the $L1$ and $L2$ regularization rates were considered as 102.0 and 0.09 with the complexity control of 2.7 and 6.0 minimum child weight.

3.4 Deep Neural Network

Numerous studies have shown the usefulness of neural networks for medical diagnosis [2–4, 30, 31, 40, 42]. Deep neural network (DNN) consists of multiple neural network layers, i.e., hidden layers between the input and output layers. This work uses a deep neural network with multiple dense hidden layers for the detection of cold. Feature vector is scaled to zero mean and unit standard deviation. To handle the skew in dataset distribution, the feature vector was re-sampled using random under-sampling, SMOTE, and a combination of SMOTE and Tomek links. The hidden layers are followed by dropout to curb overfitting and improve generalization. The output layer has a sigmoid node to represent the final class probabilities.

Since the dataset is highly skewed toward the majority 'No cold' class, a custom loss function is used to optimize the classifier, allowing weights to be trained such that both classes are given equal importance. The backpropagation algorithm is used to train DNN, computing the gradient of the loss function with respect to each weight before iterating backward through the layers. The cross-entropy loss function measures the performance of a classification model whose output is between 0 and 1. The loss function separately calculates the binary cross-entropy loss for both classes.

Mean positive error (MPE) is the binary cross-entropy loss for the positive 'Cold' class and mean negative error (MNE) is the binary cross-entropy loss for the negative 'No cold' class. Mean positive error is calculated as:

$$\text{mpe} = \frac{-1}{N} \sum_{i=1}^{N} y_i \log(p(y_i)) + (1 - y_i)(1 - p(y_i)) \tag{10}$$

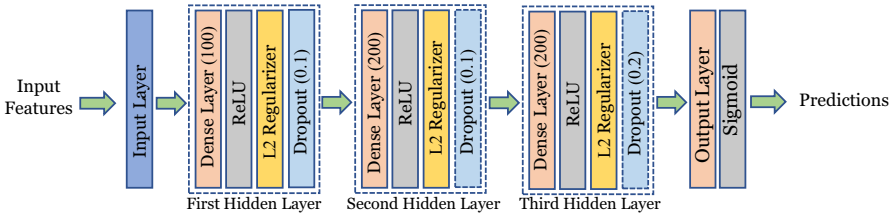

Fig. 5 The proposed architecture of DNN

and mean negative error is given by:

$$\text{mne} = \frac{-1}{M}\sum_{i=1}^{M} y_i \log(p(y_i)) + (1 - y_i)(1 - p(y_i)) \tag{11}$$

Here, N is the number of samples belonging to the positive 'Cold' class, M is the number of samples belonging to the negative 'No cold' class, y_i is the true class probability, and $\text{p}(y_i)$ is the predicted class probability. The total classification error is the sum of mean positive and negative errors. To further optimize the classifier to model the dataset better, we have used a custom loss function that separately calculates the binary cross-entropy loss for both classes. This allows the weights to be trained such that both majority and minority classes are given importance. In addition, threshold moving has been used to change the decision threshold of the classifier to counteract the dataset imbalance [10].

The architecture of the proposed DNN classifier used for classification is shown in Fig. 5. Three hidden layers are used, with the number of units per dense layer being 100, 100, and 200, respectively. A dropout of 0.1 or 0.2 is used after each dense layer to overcome overfitting. The network is trained up to 100 epochs, with early stopping used to stop the training once the number of true positives for the validation set stops increasing. The dense layers have been initialized using *He Initialization* and *L2* regularization has been used after each dense layer. In addition, the model is trained using *Adam* optimizer and learning rate of 0.00001. As mentioned, a custom loss function is used to ensure that both the classes are taken into equal consideration while training. Figure 6 shows the loss vs epoch curve during the training of DNN using a combination of MFCC and LPC features.

4 Results

The Upper Respiratory Tract Infection Corpus (URTIC) dataset used to evaluate the proposed method. The metric used for evaluation of the classifiers is unweighted average recall (UAR), which is the average of the individual recalls of the positive and negative classes. Additionally, the performance of the suggested method is evaluated in relation to the impacts of various re-sampling techniques.

The performance of the proposed deep learning-based technique using MFCC and LPC features is evaluated. The performance comparison of the proposed deep learning-

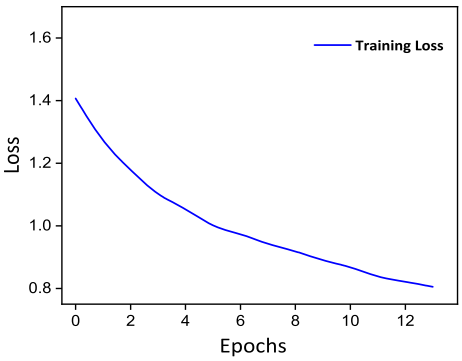

Fig. 6 The loss vs epoch curve during the training of DNN using a combination of MFCC and LPC features

Table 1 UAR(%) scores of different methods on the URTIC dataset

Feature	LR	RF	XGBoost	DNN
MFCC	62.90	65.32	65.21	65.64
LPC	62.03	62.87	60.20	59.78
MFCC + LPC	64.80	64.67	66.49	64.97
MFCC + LPC with Under-sampling	65.11	64.95	65.51	63.64
MFCC + LPC with SMOTE + Tomek links	64.67	65.69	66.71	67.71

based method with other classification techniques such as logistic regression, random forest, and gradient boosted trees is also discussed in this section. Table 1 shows the classification results obtained with proposed deep learning-based techniques along with the other techniques using different re-sampling techniques on the URTIC dataset. The 39-dimensional MFCC features give the highest UAR of 65.64% using DNN. The 32-dimensional LPC features give the highest UAR of 62.87% using a random forest classifier. For the logistic regression classifier, a combination of MFCC and LPC features and under-sampling with 0.5 sampling ratio gave the best UAR score of 65.11%. The random forest classifier performs slightly better, giving a UAR of 65.69% for combined MFCC and LPC features with SMOTE ratio of 0.2 and removal of all Tomek links. The open-source XGBoost implementation of gradient boosted trees outperformed the other ML classifiers with a UAR of 66.49% for combined MFCC and LPC features without any re-sampling, indicating that gradient boosted trees can be preferred to model imbalanced data. Using custom loss function, threshold shifting and combination of SMOTE and Tomek links for re-sampling, the deep neural network gave the highest UAR score of 67.71%. The confusion matrix of DNN for best UAR using a combination of MFCC and LPC features is shown in Fig. 7. We have achieved 69.82% recall for the 'Cold' class (sensitivity) and 65.60% recalls for the 'No-cold' class (specificity). This shows that the proposed DNN-based framework with a combination of MFCC and LPC features is effective in classifying the 'Cold' and 'No-cold' classes. To validate our proposed framework for the classification of 'Cold' and 'No-cold' speech, we have performed the cross-validation into the training and testing set, which gives the UAR of 64.72%.

Fig. 7 The confusion matrix of DNN for classification of the URTIC dataset

	Predicted No cold	Predicted Cold
Actual No cold	65.60	34.4
Actual Cold	30.18	69.82

5 Discussion

This section discusses the performance comparison of the proposed method with other state-of-the-art methods. Table 2 shows the performance of the proposed method compared to the state-of-the-art methods using the URTIC dataset. The INTERSPEECH 2017 ComParE Challenge's Cold Sub-Challenge used the URTIC dataset, and the baseline results are reported in [33]. The UAR achieved using the proposed methods is greater than the baseline results. Here, we want to underline that the bag-of-audio-words (BoAW) features and the 6373-dimensional ComParE-2013 feature set were used to obtain the baseline results. From the ComParE feature set, one codebook for 65 LLD and 65 deltas of these LLD was obtained for BoAW [32]. The table shows that, with the exception of Huckvale and Beke's work (MOD + SVM) [21], the MFCC and LPC features with SMOTE–Tomek links and DNN classifier provide a greater UAR than other techniques. Huckvale and Beke used 288-dimensional MOD feature vector with an SVM classifier and they achieved UAR of 67.95%. In contrast, using only 71-dimensional MFCC and LPC feature vector with DNN classifier, we obtained a comparable UAR result. In our previous study, with 50-dimensional features based on variational mode decomposition (VMD) and an SVM classifier, we were able to get a UAR of 66.84% [15]. Kao et al. [25] utilized alternative neural network-based discriminative autoencoders and MFCC features to report 65.81%. Using 88-dimensional extended Geneva minimalistic acoustic parameter set (eGeMAPS), Teixeira et al. [37] reported 66.9% and 66.7% UAR using unencrypted neural networks (NN) and encrypted neural networks (ENN), respectively. Vicente et al. [17] used SVM with the Fisher vector (FV) and principal component analysis (PCA) technique and MFCC features to achieve 64.92%. Warule et al. [43] achieved 61.93% UAR using deep neural network and framewise MFCC features extracted from vowel-like regions of the speech signal.

The task of detection of cold from speech signals proved to possess unique difficulties. The available URTIC dataset was highly skewed in favor of the 'No cold' class. Initially, we obtained a very misleading value of prediction accuracy, which upon inspection, showed that our model was simply classifying all the samples as having 'No cold,' defeating the whole purpose of the study. Hence, the main challenge in the study is to overcome this imbalance in the dataset. In this study, we have considered various data re-sampling techniques for dealing with imbalanced data. We have employed an approach that combined the SMOTE with the Tomek links technique. This approach outperforms the other data balancing techniques. The SMOTE balancing is an over-sampling method in which the minority class is over-sampled by producing synthetic

Table 2 UAR(%) Performance comparison of the proposed method with the state-of-the-art methods on the URTIC dataset

Model	UAR(%)
ComParE functionals + SVM (Schuller et al.) [33]	64.00
ComParE BoAW + SVM (Schuller et al.) [33]	64.20
VOI + SVM (Huckvale and Beke) [21]	66.34
VOW + SVM (Huckvale and Beke) [21]	66.47
MOD + SVM (Huckvale and Beke) [21]	67.95
GPPS + SVM (Huckvale and Beke) [21]	66.07
MFCC + GMM (Cai et al.) [7]	64.80
CQCC + GMM (Cai et al.) [7]	65.40
PSP + SVM (Suresh et al.) [35]	64.00
VMD + SVM (Deb et al.) [15]	66.84
MFCC + Autoencoder (Kao et al.) [25]	65.81
eGeMAPS + NN (Teixeira et al.) [37]	66.90
MFCC + FV + PCA + SVM (Vicente et al.) [17]	64.92
Vowel-like regions MFCC + DNN [43]	61.93
Proposed (MFCC + LPC + SMOTE–Tomek links + DNN)	67.71

samples [8]. These observations are produced by placing synthetic points on the line linking any two observations. The close pair of instances of observation from opposite classes are known as Tomek links. The Tomek links are an under-sampling method in which the observation from the majority class is removed for each Tomek link to increase the boundary between the two classes. In SOMTE-Tomek links, SMOTE is used initially to extend the separation border and balance the majority and minority classes, followed by eliminating majority class observations from each Tomek link. The Tomek links were successfully used as a data cleaning strategy to eliminate samples produced around the categorization boundary by the SMOTE method. Hence, the highest UAR is obtained using the SMOTE–Tomek links for 'Cold' and 'No cold' classification compared to other data re-sampling methods.

Though the proposed framework provides higher UAR than the baseline results of ComParE 2017, Cold Sub-challenge, the level of recognition is quite low. The recognition performance of the proposed framework can be improved by collecting more speech samples for the 'Cold' class as only 10% recordings belong to 'Cold' class. The URTIC dataset does not contain any information except the class labels. Hence, it is challenging to achieve higher performance by giving more attributes related to the speaker. More features that describe the properties of the vocal tract as common cold significantly affect the properties of the vocal tract can be utilized to increase the performance of the proposed framework. The proposed framework provides good results with the simple deep neural network and combination of 71-dimensional MFCC and LPC features. Hence, this system can be embedded in smart devices to detect and monitor common cold and related illnesses, which may be helpful in monitoring the patient's health remotely and avoiding spreading these diseases.

6 Conclusion

In this study, we have developed a deep learning-based classification model that can actually predict from speech whether a person has a cold or not. The Mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) are used to extract information about the common cold from the speech signal. It is difficult to meet the baseline performance standards because the URTIC dataset only contains class labels and dataset is highly imbalanced. Re-sampling of the dataset has been employed to balance the distribution of the dataset. A combination of SMOTE and Tomek links based down-sampling gave the best results, compared to their individual performance. The performances were evaluated using different classifiers. Deep neural network-based classifier gave a reasonable UAR score. The customized loss function further improved the performance. In this work, we have achieved a UAR of 67.71% which is higher than baseline results using the proposed simple DNN architecture and a combination of 39-dimensional MFCC and 32-dimensional LPC features. The proposed model is less complex and uses a smaller feature set while giving comparable results to other state-of-the-art methods. The proposed methodology can be used to automatically and remotely detect the common cold and related illnesses.

References

1. M. Albes, Z. Ren, B.W. Schuller, N. Cummins, Squeeze for sneeze: Compact neural networks for cold and flu recognition. INTERSPEECH **41**, 4546–4550 (2020)
2. E. Alickovic, A. Subasi, Effect of multiscale pca de-noising in ecg beat classification for diagnosis of cardiovascular diseases. Circuits Syst. Signal Process. **34**(2), 513–533 (2015)
3. S. Ayashm, M. Chehel Amirani, M. Valizadeh, Analysis of ecg signal by using an fcn network for automatic diagnosis of obstructive sleep apnea. Circuits Syst. Signal Process. **41**, 1–16 (2022)
4. M.M. Bassiouni, I. Hegazy, N. Rizk, E.-S.A. El-Dahshan, A.M. Salem (2022) Automated detection of covid-19 using deep learning approaches with paper-based ecg reports. Circuits Syst. Signal Process. pp. 1–43
5. J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D.D. Cox, Hyperopt: a python library for model selection and hyperparameter optimization. Comput. Sci. Discov. **8**(1), 014008 (2015)
6. J.R. Brzezinski, G.J. Knafl, Logistic regression modeling for context-based classification, in *Proceedings. Tenth International Workshop on Database and Expert Systems Applications*. DEXA 99 (IEEE, 1999), pp. 755–759
7. D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, M. Li, D. Cai, Z. Ni, W. Liu, W. Cai, End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum. INTERSPEECH (2017), pp. 3452–3456
8. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
9. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794
10. G. Collell, D. Prelec, K. Patil, Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. arXiv preprint arXiv:1606.08698 (2016)
11. H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, B.W. Schuller, End-2-end covid-19 detection from breath & cough audio. arXiv preprint arXiv:2102.08359 (2021)
12. D.R. Cox, E.J. Snell, *Analysis of binary data* (Routledge, New York, 2018)
13. N. Cummins, A. Baird, B.W. Schuller, Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods **151**, 41–54 (2018)
14. S. Deb, S. Dandapat, Classification of speech under stress using harmonic peak to energy ratio. Comput. Electr. Eng. **55**, 12–23 (2016)

15. S. Deb, S. Dandapat, J. Krajewski, Analysis and classification of cold speech using variational mode decomposition. IEEE Trans. Affect. Comput. **11**(2), 296–307 (2017)
16. G. Deshpande, A. Batliner, B.W. Schuller, Ai-based human audio processing for covid-19: a comprehensive overview. Pattern Recognit. **122**, 108289 (2022)
17. J.V. Egas-López, G. Gosztolya, Predicting a cold from speech using fisher vectors; svm and xgboost as classifiers, in *International Conference on Speech and Computer* (Springer, 2020), pp. 145–155
18. G. Gosztolya, R. Busa-Fekete, T. Grósz, L. Tóth, Dnn-based feature extraction and classifier combination for child-directed speech, cold and snoring identification Interspeech 3522–3526 (2017)
19. T.K. Ho, Random decision forests, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, (IEEE, 1995), pp. 278–282
20. T.R. Hoens, N.V. Chawla, Imbalanced datasets: from sampling to classifiers. Found. Algorithms Appl. Imbalanced Learn. (2013). https://doi.org/10.1002/9781118646106.ch3
21. M. Huckvale, A. Beke, It sounds like you have a cold! testing voice features for the interspeech 2017 computational paralinguistics cold challenge. Interspeech 3447–3451, (2017)
22. S. Imai, Cepstral analysis synthesis on the mel frequency scale, in *ICASSP'83 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, (IEEE, 1983) pp. 93–96
23. E.L. José Vicente, G. Gosztolya, Using the fisher vector approach for cold identification. Acta Cybern. **25**(2), 223–232 (2021)
24. S.R. Kadiri, P. Gangamohan, S.V. Gangashetty, P. Alku, B. Yegnanarayana, Excitation features of speech for emotion recognition using neutral speech as reference. Circuits Syst. Signal Process. **39**(9), 4459–4481 (2020)
25. Y.-Y. Kao,H.-P. Hsu, C.-F. Liao, Y. Tsao, H.-C. Yang, J.-L. Li, C.-C. Lee, H.-S. Lee, H.-M. Wang, Automatic detection of speech under cold using discriminative autoencoders and strength modeling with multiple sub-dictionary generation, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)* (IEEE, 2018), pp. 416–420
26. H. Kaya, A.A. Karpov, Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold. INTERSPEECH (2017), pp. 3527–3531
27. P.P. Kumar, G.T. Yadava, H.S. Jayanna, Continuous Kannada speech recognition system under degraded condition. Circuits Syst. Signal Process. **39**(1), 391–419 (2020)
28. L. Marple, A new autoregressive spectrum analysis algorithm. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 441–454 (1980)
29. T.L. Nwe, T.H. Dat, W.Z.T. Ng, B. Ma, An integrated solution for snoring sound classification using bhattacharyya distance based gmm supervectors with svm, feature selection with random forest and spectrogram with cnn. INTERSPEECH (2017), pp. 3467–3471
30. S. Poorani, P. Balasubramanie, Seizure detection based on eeg signals using asymmetrical back propagation neural network method. Circuits Syst. Signal Process. **40**(9), 4614–4632 (2021)
31. Z. Sabir, M.A.Z. Raja, H.A. Wahab, M. Shoaib, J.G. Aguilar, Integrated neuro-evolution heuristic with sequential quadratic programming for second-order prediction differential models. Numerical Methods for Partial Differential Equations. **19**(1), 663–687 (2022)
32. M. Schmitt, B. Schuller, Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit Interspeech, 3457–3461 (2017)
33. B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom et al., The interspeech 2017 computational paralinguistics challenge: addressee, cold & snoring, in *Computational Paralinguistics Challenge ComParE*. Interspeech (2017), pp. 3442–3446
34. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, S. Narayanan, Paralinguistics in speech and language-state-of-the-art and the challenge. Comput. Speech Lang. **27**(1), 4–39 (2013)
35. A.K. Suresh, S.R. KM, P.K. Ghosh, Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition. INTERSPEECH (2017), pp. 3462–3466
36. D. Tavarez, X. Sarasola, A. Alonso, J. Sanchez, L. Serrano, E. Navas, I. Hernáez, Exploring fusion methods and feature space for the classification of paralinguistic information. INTERSPEECH (2017), pp. 3517–3521
37. F. Teixeira, A. Abad, I. Trancoso, Patient privacy in paralinguistic tasks. INTERSPEECH (2018), pp. 3428–3432
38. R.G. Tull, J.C. Rutledge, C.R. Larson, Cepstral analysis of "cold-speech"for speaker recognition: a second look. PhD thesis, Acoustical Society of America (1996)

39. D. Tyrrell, S. Cohen, J. Schilarb, Signs and symptoms in common colds. Epidemiol. Infect. **111**(1), 143–156 (1993)
40. M. Umar, Z. Sabir, M.A.Z. Raja, J.G. Aguilar, F. Amin, M. Shoaib, Neuro-swarm intelligent computing paradigm for nonlinear hiv infection model with cd4+ t-cells. Math. Comput. Simul. **188**, 241–253 (2021)
41. J. Wagner, T. Fraga-Silva, Y. Josse, D. Schiller, A. Seiderer, E. Andre, Infected phonemes: how a cold impairs speech on a phonetic level (2017), pp. 3457–3461
42. B. Wang, Y. Wang, J. Gómez-Aguilar, Z. Sabir, M.A.Z. Raja, H. Jahanshahi, M.O. Alassafi, F.E. Alsaadi, Gudermannian neural networks to investigate the liénard differential model. Fractals **30**(3), 2250050–315 (2022)
43. P. Warule, S.P. Mishra, S. Deb, Classification of cold and non-cold speech using vowel-like region segments, in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)* (IEEE, 2022) pp. 1–5