A Universal System for Cough Detection in Domestic Acoustic Environments

Nikonas Simou*,Nikolaos Stefanakis* ,†, Panagiotis Zervas†
* FORTH-ICS, Heraklion, Crete, Greece, GR-70013
† Hellenic Mediterranean University, Department of Music Technology
and Acoustics, Rethymno, Greece, GR-74100

***Abstract*—Automated cough detection may provide valuable clinical information for monitoring a patient's health condition. In this paper, we present a cough detection system that utilises an acoustic onset detector as a pre-processing step, aiming to detect impulsive patterns in the audio stream. In a subsequent step, discrimination of coughing events from other impulsive sounds is handled as a binary classification task. In contrast to existing works, the proposed cough discrimination models are trained and tested with heterogeneous data uploaded by different users to online audio repositories. In that way, our system achieves robust performance to a wide range of audio recording devices and to varying noise and/or reverberation conditions. Our evaluation results showed that a sensitivity in the order of 90% and a specificity in the order of 99% can be achieved in a domestic environment with the utilization of Long-Short-Term-Memory deep neural network architecture.**

I. Introduction

Cough is a symptom related to many diseases of the respiratory system. For patients suffering from such conditions, monitoring of the cough activity may provide important information regarding the severity of their health condition or their response to different medical treatments [1], [2]. However, manual cough monitoring can be a slow and tedious process. Therefore, automatic cough detection algorithms have been proposed as the means to reduce the data to be manually analyzed or even eliminate the need for manual analysis.

In the recent years, several techniques that rely on the audio signal for monitoring of the cough activity have emerged. In the pioneering work of Matos et al [3] an automatic cough detection system was proposed following a keyword-spotting approach using Hidden Markov Models (HMMs) trained with mel-frequency cepstral coefficients (MFFCs). Since then, several other works [4]–[6] have proved the suitability of MFCCs for constructing a classifier that discriminates cough from others types of sounds. Drugman et al used tools from information theory to assess the relevance of several well known audio features for this problem [7]. Apart from MFCCs, their list with most relevant acoustic features included zero-crossing rate, spectral flatness, fundamental pitch and harmonic to noise ratio. In a similar approach, Pham proposed combining Entropy features together with MFCC and zero-crossing [8]. It is also worth mentioning the work from Larson et al who used dimensionality reduction techniques in order the extract features directly from the audio spectrogram [9].

Various classification techniques have been proposed as the means for recognizing the cough pattern; Among the most successful ones are Gaussian Mixture Models [4], [8], [10], Artificial Neural Networks [6], [7] and Support Vector Machines (SVM) [10]. In [9] Random Forests were also utilized with good results. Recently, with the introduction of deep neural network (DNN) architectures additional works showed their efficiency in the task of cough classification. Liu et al proposed the use of a pretrained DNN in combination with HMM [5] and MFCCs as acoustic features, while in [11] the authors proposed the use of a Convolutional Neural Network (CNN) that accepts a raw audio spectrogram as input. CNNs demonstrated a significant gain in the classification performance when compared to an HMM as well as to an SVM approach with MFCC features.

Segmentation of the audio stream into frames is an essential part of the cough detection process since it affects the final acoustic features presented to the classifier. Yet, in many of the works previously mentioned, this part of the process is not efficiently explained. Excluding the approaches in [3] and [9], most of the works rely on simple energetic criteria (e.g. the Root Mean Square Energy) for deciding whether an audio segment contains an interesting event that should be passed to the classifier or not [6], [8], [11], [12]. To our opinion, a better criterion for audio segmentation can be realized by combining energy criteria with onset detection criteria. This relies on the observation that the the explosive phase of the cough signal (see [10] as a reference for the three phases of cough) produces a rapid increase in acoustic energy that would very likely trigger an onset detection algorithm [13]. This represents the first of the two main contributions of this paper.

An additional important remark is that all the cough detection systems mentioned so far are trained and tested on specific acoustic sensing equipment while in many cases, a specific placement of the sensing device with respect to the patient [3], [5], [9], [11] is also required. In contrast to such restrictions, we present in this paper a universal system for

Work presented in the paper was funded by a the TDI State Aid Action "Research – Create – Innovate" grant (MyHealthWatcher - T1EDK-03090), which is co-financed by Greece and the European Union (European Regional Development Fund) as part of the Operational Program "Competitiveness, Entrepreneurship and Innovation" of the National Strategic Reference Framework (NSRF) 2014-2020.

cough detection in domestic environments that does not require a specific audio equipment or placement of the audio device with respect to the user. Key for achieving an acceptable discrimination performance in heterogeneous conditions is the fact that our system is trained not only on recordings that we made ourselves (using multiple recording devices), but also on data extracted from audio repositories that can be found on the internet such as Freesound, Google Audioset and others. It should be noted that users are free to contribute to these repositories by uploading content captured with any type of audio device, at different audio formats (e.g. .wav or .mp3) and/or at acoustic conditions that may significantly vary from one case to the other (e.g. in terms of noise and reverberation). By using such a diverse collection of audio samples to train a classification model, it can be claimed that the model is subjected to a type of multi-condition training [14]. Results in this paper are shown in terms of two classifiers, a DNN accepting log-mel energies [15] as input and a SVM classifier operating on a selection of hand-crafted acoustic features that have already been proposed in the context of cough detection.

II. Data collection

Obviously, the main issue facing researchers initiating a study on automatic audio event classification is the availability of appropriate data. In the context of this work, recordings were collected from various online repositories, such as Freesound [16], Soundsnap [17] and Google AudioSet [18]. Since cough detection was handled as a binary classification problem, the non-cough category had to be populated with a wide array of audio events that can occur in a real monitoring scenario. Thus, in addition to the aforementioned repositories, we collected data from CHiME HOME [19] and CHiME-5 [20] audio databases. Table I tabulates some details of our database.

In that way we were able to capture high variability related, (a) to user gender and age and (b) to capturing quality (i.e. background noise, reverberation). Although the utilized audio equipment in each recording is unknown, it is reasonable to assume that they came from various recording devices (from medium quality smartphone microphones to professional audio equipment).

In enriching our database with more data, we carried out two additional recording sessions, Session 1 and Session 2, In Session1, sound was recorded simultaneously with two different smartphone devices, the built in microphone in a laptop and a portable digital recorder. The devices were placed at different angles and distances with respect to the subject in order to capture cough and other sounds from three male and one female subjects in an office environment. In Session2, sound was captured using two smartphones and an analogue microphone connected to a professional sound card. The main recorded audio content was speech captured from 15 different individuals. During Session2, 27 cough events were also recorded.

Dataset	Duration (hours)	Cough instants
Audioset-”cough”	2.50	1820
Soundsnap-”cough”	0.8	339
Freesound-”cough”	2.50	1560
Session 1	0.93	343
Session 2	1.65	27
CHiME-5	2.70	–
CHiME HOME	6.50	–
Freesound-”kitchen”	16.8	–
Freesound-”laughter”	5.91	–
Freesound-”giggle”	1.80	–
Freesound-”living room”	1.65	–

TABLE I
Information regarding the datasets that were used for training and testing the system.

III. Methodology

Onset detection is used for spotting impulsive events in the audio stream. For each detection, a short signal segment is extracted around the onset which is subsequently passed as input to the feature extraction step. The feature representation of the audio segment is then passed to the binary classifier that decides whether a cough or a non-cough event occurred.

A. Onset detection

At the explosive phase of the cough, the sudden release of air from the lungs leads to a rapid increase of the acoustic energy. This variation can be efficiently detected with the utilization of onset detection methods [13]. Moreover, the fact that the direct sound precedes the reflected sound components in a reverberant environment makes acoustic onsets excellent features for various detection problems, as for example in Direction of Arrival estimation techniques [21].

Our method for onset detection relies on measures of spectral energy across successive time-frames. We form frames by windowing the signal with a short-length Hanning window of length L, moving on a continuous time-grid with hop-size h. At each frame, the short-time Fourier transform (STFT) is calculated and the frequency bins $k \in [k_{min}, k_{max}]$ corresponding to a specified spectral range are used for further processing. Following the approach proposed by Tan et al. [22], we perform measure of ”percussiveness”, which relies on the ratio of the magnitude of each frequency bin between the current frame and the previous frame. In particular, percussiveness is measured at the kth frequency bin as

$$p[k] = \log_2(|X_n[k]| \, / \, |X_{n-1}[k]|), \tag{1}$$

where $X_n[k]$ and $X_{n-1}[k]$ is the Fourier coefficient at the current and the previous time-frame respectively. A binary variable $b[k]$ is then calculated at each frequency bin as follows; if $p[k]$ is greater than a threshold T_p then $p[k] = 1$, otherwise $p[k] = 0$. The sum of all binary variables across $k \in [k_{min}, k_{max}]$ should then be greater than threshold T_b in order for a time-frame to be considered as an onset. An additional condition that needs to be fulfilled is that $E_n > rT_E$

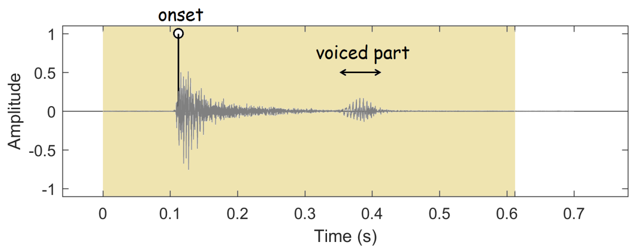

Fig. 1. A cough instance in the time domain. The yellow shadowed region illustrates the extend of the time window used for extracting the acoustic features.

where r is a fixed parameter, E_n is the L1 norm of the vector comprised of the STFT amplitudes across the considered frequency range at the nth time frame and T_E is the maximum of all the E_n measures obtained in the recording. Observe that this last criterion is the only one associated to signal energy.

In order to facilitate onset detection further, we assume finally that there is a minimum amount of time Δt between two successive onsets. This criterion is used in order to prevent the appearance of a '‘double" onset due to ambiguities in the sound in the neighbourhood of a strong attack.

B. Data annotation

Using the method described in the previous section, an annotation tool based on onset detection was developed in Matlab. This tool extracts a segment starting a little before and extending a little after each detected onset and plays back the segment to the annotator, who may assign a label to it with the press of a single button on the keyboard.

Since in our scenario cough detection is a binary classification problem, two different labels are considered. A "cough" label for assigning the audio segment to the True category and an "other" label for the False category. Additionally, the user can select to omit assigning a label to an audio segment, if he/she believes that it's context is ambiguous. The data annotation process was applied on the first five datasets of Table I (the last column presents the number of annotated cough instances). It should be noted that while in most cases the onset was located at the explosive part of the cough, in some cases onsets appeared also at the voiced part of the cough (see Fig. 1). We chose to admit voiced onsets to the "other" class and not the "cough" class. This was because cough instances very often occur without a voice part and therefore, voiced onsets are unreliable for cough detection.

C. Acoustic feature representation

We exploit the outcome of onset detection in order to extract acoustic data along a short temporal segment that starts 0.125 s before the onset and ends 0.5 s after it (see Fig. 1). We then use log-mel energy in the time frequency domain for representing the data input to the DNNs. The mel spectrograms are created in Python using the librosa library, with a window length of 0.04 s and hop-size of 0.02 s. 128 frequency zones are used to cover the full spectrum from 0 Hz to the Nyquist frequency (8kHz). In the end, the final data used as input to the DNN is a matrix of size 29×128, where each bin carries the energy in each mel bin in logarithmic scale.

D. DNN architecture

Various DNN architectures were tested, among which the most promising appeared to be the one based on Long Short Term Memory (LSTM) units. Specifically, two LSTM layers of 256 units each, were used, followed by one fully connected layer of 64 units and finally a dropout layer with 0.3 probability and an output softmax layer. Between each consecutive hidden layer we included a batch normalization layer which is known for regularizing the network [23]. On the other hand, CNN based architectures of varying depth and/or width that were tested were unable to produce satisfactory results. In order to optimize training, we also apply a learning rate decay of a 0.2 factor and tolerance of 5 epochs while using early stopping with a tolerance of 10 epochs, to avoid overfitting. Both are implemented according to the validation loss which is calculated over a random 5% split of the available training data.

IV. Evaluation

A. Handcrafted feature description

In order to be able to compare the proposed method to more traditional ones, we compute a set of handcrafted features, mostly based on the work reported in [7]. These features include the Mel Frequency Cepstrum Coefficients (MFCC) mean, variance, delta and delta-deltas, using a total of 13 mel frequency bins. Additionally we compute the mean and variances for the zero-crossing rate, spectral flatness, spectral bandwidth and spectral contrast features. In every case we use an STFT of 512 samples with a hop size of 256 samples. These features are calculated (a) on the entire length of each audio segment extracted from each onset (described in III-C) and (b) on a portion of each segment that begins exactly at the onset and ends 64 ms after it. Following this approach, we have at the same time a global acoustic representation of the entire cough instant and an additional one which is specific to the acoustic content around the onset. Both actions (a) and (b) result to 64×1 feature vectors. We will refer to the 64×1 acoustic feature that results from action (a) as the simple feature vector, and to the 128×1 acoustic feature that results by the concatenation of both features as the enriched feature vector. All features were standardized so that every feature in the training set has zero-mean and unit-variance. The same standardization is applied on the testing set using the statistics of the available training set.

In order to improve robustness and to avoid the curse of dimensionality, we additionally perform feature selection using Recursive Feature Elimination (RFE). The RFE technique is employed using random forest classifiers with 500 estimators, while gini importance [24] is used for feature ranking. The approach involves using RFE in a cross validation scheme in order to first define the optimal number of features and then applying it on the entire training set in order to decide

parameter	Value
Sampling rate	16 kHz
L and also length of STFT	672 samples
h	512 samples
k_{min} corresponding frequency	120 Hz
k_{max} corresponding frequency	6 kHz
T_p	1.9
T_b	77
r	0.08
Δt	0.12 s

TABLE II
ONSET DETECTION PARAMETERS

specifically which features to keep. In the end, 36 out of the 64 features were preserved from the simple feature vector and 50 out of the 128 from the enriched feature vector. For the evaluation that follows, a Support Vector Machine (SVM) classifier with a radial basis function kernel was used.

B. Onset detection performance

It is obvious that if a cough instant does not trigger an onset, it will be missed by the system. Therefore, one has to tune the detection parameters of Section III-A so that the onset detection step is neither too loose nor too strict. The onset parameters proposed in this paper are shown in Table II. These parameters were empirically tuned after several trials performed on a subset of the available audio recordings. More particularly, 58 recordings were randomly selected from the first four databases of Table I. A subject listened to these recordings in Praat and for each cough instant, marked the time interval starting a little before and ending a little after the cough onset. This way, 284 cough onsets were identified. Running the onset detection algorithm, we then counted the number of detected onsets that were located within these intervals. Using the parameters of Table II, the onset detection step spotted 275 out of the 285 cough onsets, achieving a detection rate of 96.8 %. At the same time, 575 onsets that were detected by the algorithm were outside from the marked time intervals and were in most cases triggered by events that did not belong to the cough class.

C. Test cases

In order to test the classification performance, we chose to split the data in a recording-wise fashion rather than a sample-wise fashion. This means that audio samples from the same audio recording are used either for testing, or for training, and it is not allowed for these portions to be distributed in both categories. Doing so, we manage to minimize the probability that cough samples from the same individual (and thus from the same audio recording device) are at the same time in the training and in the testing set. Doing so, we are able to better assess the system's robustness to "unseen" coughers as well as to "unseen" recording devices.

Following this tactic, four testsets (namely test A, B, C and D) were created for assessing the detection performance in the following manner; several audio recordings were randomly chosen from the first 3 databases of Table I and the labelled audio segments (both "cough" and "other" class) associated to these recordings were used to populate the four test sets. Additionally, all the onsets associated with a particular individual from Session 1 were added, so that each each testset contained the segments from one only individual. This way, test A contained all the onsets from first individual of Session 1, test B those produced from the second individual of Session 1 and so on. A number of 200 cough samples was reached for each one of the four testsets, while the number of non-cough instants varied from 154 to 387. All other data that was not in the testset was allowed to be used for training. This included not only the labelled audio segments from the first five datasets of Table I, but also the onsets detected in all the other datasets. Although the audio samples originating from the last six datasets of Table I were not manually labelled, it was verified that only a trivial percentage of these onsets was actually produced by human cough and therefore, it was safe to use these onsets as representatives of the "other" class. The entire dataset consisted of 4110 cough samples and 50004 "other" samples.

Two additional test cases were generated, one by putting the samples of Session 2 (27 coughs/4510 non-cough instants) and one by putting the samples of CHiME HOME (5627 non-cough instants) as a test. These two tests are mostly representative of the system's specificity, since the vast majority of the onsets detected in those two datasets belong to the non-cough class.

D. Results

Various results are presented in this section for the DNN and the two SVM approaches where each time, the classifiers are trained and tested on exactly the same datasets. Moreover, results shown for DNN are obtained as the average from five repetitions of the training/testing procedure, which was done in order to increase the statistical significance of the results.

With respect to test A-D, the results obtained using DNN and SVM are shown in terms of average sensitivity and average specificity in the second and third column of Table III. The results indicate that the DNN approach achieved better sensitivity and specificity compared to both SVM approaches. While the two SVM models achieve inferior scores, it is worth noting that the SVM_{50} with the enriched feature set shows a significant improvement compared to the simple feature set (SVM_{36}). This demonstrates that the acoustic features extracted from the onset region add significant knowledge to the classification process. As an additional metric for comparing the different classifiers, the Area Under the Curve (AUC) was calculated for each testset. AUC is obtained from the Receiver Operating Characteristics (ROC) curve that resulted by plotting the true positive rate against the false positive rate, using as the free parameter (a) the decision threshold used for the cough class probability provided by the softmax layer of the DNN and (b) the decision threshold used for the cough class probability provided by the (SVM_{50}) and the (SVM_{36})

Case	A-D			Session 2		CHiME HOME	
	sens	spec	AUC	sens	spec	sens	spec
DNN	**87.8**	**98.9**	**98.6**	95.6	99.7	-	99.0
SVM_{50}	80.0	96.8	97.2	77.8	**99.9**	-	99.7
SVM_{36}	65.1	94.9	90.8	25.9	99.3	-	**99.9**

TABLE III
SENSITIVITY & SPECIFICITY PAIRS FOR VARIOUS METHODS AND TEST CASES.

classifier. The average AUC from all four tests is shown in the fourth column of Table III, where again one may observe the superiority of the DNN classifier (AUC=98.6) compared to the SVM classifiers (AUC=97.2 and 90.8).

With respect to Session 2, the DNN approach has again by far the best sensitivity, but is surpassed by the $SVM_{50/128}$ model in terms of specificity. The superiority of the enriched feature vector compared to the simpler one can again be seen. Finally, the DNN classifier achieves a specificity of 99.0% in CHiME HOME dataset and it is surpassed by both (SVM_{50}) and (SVM_{36}) classifiers. It should be noted that the data of Session 2 includes samples from four different audio recording devices while CHiME HOME only from one recording device. On the other hand, a much higher number of recording devices is involved in testset A-D, and therefore, this test is more indicative of each classifier's performance across heterogeneous conditions.

Overall, the results demonstrate that the DNN approach may achieve a sensitivity in the order of 90% and a specificity in the order of 99% in a domestic environment. Possibly, the performance may be improved by the use of additional training data, so that the variability of the non-cough class is better learned.

V. CONCLUSION

A universal system for cough detection from audio recordings was presented in this paper. It was shown that cough detection can be greatly assisted by the use of onset detection as a pre-processing step, with aim to detect impulsive patterns in the audio stream. In a subsequent step, discrimination of coughing events from other impulsive sounds was possible by using a binary classifier. Results indicate that DNN and particularly, the LSTM architecture, may achieve a sensitivity in the order of 90% and a specificity in the order of 99% in a domestic environment. The presented system was trained and tested with data extracted mainly from online audio repositories. Due to the high diversity characterizing these recordings, the presented system can theoretically be deployed on any audio device equipped with a microphone and is also robust to a variety of acoustic conditions in terms of noise and reverberation.

REFERENCES

[1] S Subburaj, L Parvez, and T. Rajagopalan, "Methods of recording and analysing cough sounds," *Pulmonary pharmacology*, vol. 9, no. 5-6, pp. 269–279, 1996.

[2] KF Chung, "Assessment and measurement of cough: the value of new tools," *Pulmonary pharmacology & therapeutics*, vol. 15, no. 3, pp. 267–272, 2002.

[3] S. Matos, S. Birring, I. Pavord, and H. Evans, "Detection of cough signals in continuous audio recordings using hidden markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1078–1083, 2006.

[4] Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, I. Lane, and P. Zhang, "Coughloc: Location-aware indoor acoustic sensing for non-intrusive cough detection," in *International Workshop on Emerging Mobile Sensing Technologies, Systems, and Applications*. Citeseer, 2011.

[5] J. Liu, M. You, Z. Wang, G. Li, X. Xu, and Z. Qiu, "Cough event classification by pretrained deep neural network," *BMC medical informatics and decision making*, vol. 15, no. S4, pp. S2, 2015.

[6] Samantha J Barry, Adrie D Dane, Alyn H Morice, and Anthony D Walmsley, "The automatic recognition and counting of cough," *Cough*, vol. 2, no. 1, pp. 8, 2006.

[7] T. Drugman, J. Urbain, and T. Dutoit, "Assessment of audio features for automatic cough detection," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1289–1293.

[8] C. Pham, "Mobicough: real-time cough detection and monitoring using low-cost mobile devices," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2016, pp. 300–309.

[9] E. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, 2011, pp. 375–384.

[10] T. Drugman, J. Urbain, N. Bauwens, R. Chessini, A. Aubriot, P. Lebecque, and T. Dutoit, "Audio and contact microphones for cough detection," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] J. Amoh and K. Odame, "Deepcough: A deep convolutional neural network in a wearable cough detection system," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2015, pp. 1–4.

[12] H. Lu, W. Pan, N. Lane, T. Choudhury, and A. Campbell, "Soundsense: scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*, 2009, pp. 165–178.

[13] J. Bello, L. Daudet, S. Abdallah, M. Duxbury, C.and Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on speech and audio processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[14] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[15] J. Volkmann, S. Stevens, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 208–208, 1937.

[16] "Freesound," https://freesound.org/.

[17] "Soundsnap," https://www.soundsnap.com/.

[18] "Google audioset," https://research.google.com/audioset/.

[19] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[20] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth'chime'speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[21] A. Moore, C. Evers, P. Naylor, D. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct path dominance test," in *Proc. of 23rd International Signal Processing Conference (EUSIPCO)*, 2014, pp. 2341–2345.

[22] H. Tan, Y. Zhu, L. Chainsorn, and S. Rahardja, "Audio onset detection using energy-based and pitch-based processing," in *Proc. 2010 IEEE Int. Symposium on Circuits and Systems*, Paris, 2010, pp. 3689–3692.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[24] L Breiman, JH Friedman, CJ Stone, RA Olshen, and CJ Stone, "Classification and regression trees (wadsworth, belmont, ca, 1984)," in *Proceedings of the Thirteenth International Conference, Bari, Italy*, 1996, p. 148.