EXPLORING AUDITORY ACOUSTIC FEATURES FOR THE DIAGNOSIS OF COVID-19

Madhu R. Kamble, Jose Patino, Maria A. Zuluaga and Massimiliano Todisco

EURECOM, Sophia Antipolis, France

ABSTRACT

The current outbreak of a coronavirus, has quickly escalated to become a serious global problem that has now been declared a Public Health Emergency of International Concern by the World Health Organization. Infectious diseases know no borders, so when it comes to controlling outbreaks, timing is absolutely essential. It is so important to detect threats as early as possible, before they spread. After a first successful DiCOVA challenge, the organisers released second DiCOVA challenge with the aim of diagnosing COVID-19 through the use of breath, cough and speech audio samples. This work presents the details of the automatic system for COVID-19 detection using breath, cough and speech recordings. We developed different front-end auditory acoustic features along with a bidirectional Long Short-Term Memory (bi-LSTM) as classifier. The results are promising and have demonstrated the high complementary behaviour among the auditory acoustic features in the Breathing, Cough and Speech tracks giving an AUC of 86.60% on the test set.

***Index Terms*—** COVID-19, auditory acoustic features, bi-LSTM, respiratory sounds.

1. INTRODUCTION

Coronavirus disease, so-called COVID-19, is an infectious disease caused by the recently discovered coronavirus, the SARS-CoV-2. This disease has spread rapidly worldwide over the past year, causing a global crisis with serious health, social and economic consequences. To put an end to this pandemic, various initiatives are being carried out worldwide, including the development of new systems for rapid diagnosis of the disease.

Recently, the DiCOVA 2021 Challenge [1] was carried out to promote research in development of systems for the detection of COVID-19 through recordings of respiratory sounds. Several systems have been proposed to detect the COVID-19 signature within acoustic indicators [2, 3, 4, 5, 6, 7]. Only a few of them focused on the study of acoustic clues, giving more emphasis to classifiers. The study reported in [8] explores the Autoregressive Predictive Coding (APC) to pre-train a unidirectional LSTM and spectral augmentation. In [9], authors used ComParE 2016 feature set, and two classical machine learning models, namely Random Forests, and Support Vector Machines (SVMs). The use of breathing patterns for the diagnosis of COVID-19 is studied in [10]. COVID-19 detection by means of a Contextual Attention Convolutional Neural Networks and gender information is studied in [11].

The first COVID-19 challenge consisted on 2 tracks: Track-1 focused on diagnosing COVID-19 using cough sounds, while Track-2 focused on a collection of breath, sustained vowel phonation, and a number of counting speech recordings. As a follow up of the first successful DiCOVA 2021 challenge, the second DiCOVA challenge has been organised [12]. The second challenge aimed at 4 different tracks, namely, breathing, cough, speech and fusion. The organisers provided a baseline system for the second challenge based on Log Mel Spectrogram front-end and Bidirectional Long Short-Term Memory (bi-LSTM) back-end.

In this paper, we describe and propose a system for automatic COVID-19 detection presented on all the four different tracks of the second DiCOVA challenge. Our system focuses more on features than classifiers by using 4 perceptually-motivated acoustic features at front-end. The features we explored are Teager energy operator cepstral coefficients (TECCs), Instantaneous Amplitude Cepstral Coefficients (IACCs), Constant Q-Cepstral Coefficients (CQCCs) and Filterbank Constant Q Transform (FBCQT) [13, 14, 15], along with the bi-LSTM classifier.

The remainder of this paper is organized as follows. Section 2 presents the technical details of auditory acoustics features used for the detection of COVID-19. Section 3 describes the second DiCOVA challenge database. Experimental setup and results are presented in Section 4 and Section 5, respectively. Finally, the main conclusions of this work and future research lines are drawn in Section 6.

2. AUDITORY ACOUSTICS FEATURES

In this section we discuss the acoustic features used to diagnose COVID-19 from breath, cough and speech.

2.1. MelSPEC

Studies have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in

The first author is supported by the RESPECT project funded by the French Agence Nationale de la Recherche (ANR).

lower frequencies than higher frequencies [16]. The Mel spectrum contains a short-time Fourier transform (STFT) for each frame of the spectrum (energy/amplitude spectrum), from the linear frequency scale to the logarithmic Mel-scale, and then goes through the filter bank to get the eigen vector, these eigenvalues can be roughly expressed as the distribution of signal energy on the Mel-scale frequency.

2.2. TECC and ESA-IACC

The Teager Energy Operator (TEO) ($\psi\{\cdot\}$) track the running estimate of instantaneous energy fluctuations of the narrow-band signal. The Teager energy profile obtained from the bandpass filter is further given to the Energy Separation Algorithm (ESA) to isolate the Instantaneous Amplitude (IA) ($a_i[n]$) and Instantaneous Frequency (IF) ($\Omega_i[n]$) and is given as [17, 18, 19]:

$$\Psi_d\{x_i[n]\} = x_i^2[n] - x_i[n-1]x_i[n+1] \approx a_i^2[n]\Omega_i^2[n], \quad (1)$$

where $x_i[n]$ is i^{th} bandpass filtered signal.

$$a_i[n] \approx \frac{2\Psi_d\{x[n]\}}{\sqrt{\Psi_d\{x[n+1] - x[n-1]\}}}, \quad (2)$$

where $x_i[n]$ is i^{th} bandpass filtered signal. The block diagram of Teager Energy Cepstral Coefficients (TECC) and Energy Separation Algorithm Instantaneous Amplitude Cepstral Coefficients (ESA-IACC) feature set is shown in Figure 1. The TECC feature set is computed as per our earlier studies in [13, 20] and ESA-IACC feature set according to [14, 21].

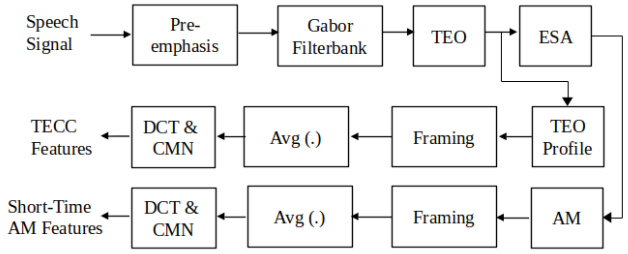

Fig. 1. Block diagram of TECC, and ESA-IACC feature sets.

2.3. Filterbank CQT and CQCC

The constant Q transform (CQT) is a perceptually motivated approach to time-frequency analysis introduced by Youngberg and Boll [22] in 1978 and refined over the last few decades by Brown [23]. In contrast to Fourier-based approaches, the CQT gives a greater frequency resolution for lower frequencies and a greater temporal resolution for higher frequencies, which emulate the human auditory perception. The CQT of a discrete signal $x(n)$ is defined by:

$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j)a_k^*(j-n+N_k/2) \quad (3)$$

where $k = 1, 2, ..., K$ is the frequency bin index, $a_k(n)$ are the basis functions, $*$ is the complex conjugate and N_k is a variable window length. The center frequencies f_k are defined according to $f_k = 2^{(k-1)/(B)} f_1$, where f_k is the center frequency of bin k, f_1 is the center frequency of the lowest frequency bin and B is the number of bins per octave.

The filter selectivity Q which reflects the ratio between center frequency and bandwidth is constant and defined as:

$$Q = \frac{f_k}{f_{k+1} - f_k} = (2^{1/B} - 1)^{-1} \quad (4)$$

In practice, B determines time-frequency resolution trade-off.

2.3.1. FBCQT

Similar to MelSPEC, FBCQT is calculated filtering $X^{CQ}(k,n)$ with a filterbank composed of n_{fb} triangular filters equally spaced along the linear-scale, and then calculating the logarithm of the energy in each band.

2.3.2. CQCC

Constant Q cepstral coefficients (CQCCs) were introduced recently and successfully in the context of fake audio detection [15]. CQCC features are based on a combination of the constant Q transformation (CQT) and cepstral analysis. The cepstral coefficients are calculated from the transformation at constant Q, which imply a re-sampling in frequency domain from the geometric scale to a linear scale, according to:

$$CQCC(p) = \sum_{l=0}^{L-1} \log \left|\bar{X}^{CQ}(l)\right|^2 \cos\left[\frac{p\left(l-\frac{1}{2}\right)\pi}{L}\right] \quad (5)$$

where $\bar{X}^{CQ}$ is the linearised CQT-derived spectrum, l is the linear-scale index and $p = 0...L-1$. The full CQCC extraction algorithm is described in [15].

3. SECOND DICOVA CHALLENGE DATASET

After the first successful DiCOVA Challenge, the organisers released the second DiCOVA challenge focusing on four different tracks, namely, breathing, cough, speech, and fusion [12].

The training/validation set for the all the tracks contains 965 audio files stored in .FLAC format at 44.1 kHz sampling frequency. Each audio file corresponds to a single subject. This set comprises an audio recording from 172 COVID-19 positive subjects and 793 COVID-19 negative subjects. Gender and age of the subjects is also provided as extra metadata. Validation set are performed using 5-fold cross validation from train lists. The test set consists of 471 audio files with the same format as the training/validation set, but with the COVID-19 status hidden from the participants. The detailed information of the respective tracks are reported below:

Track 1: Breathing - The goal of this track is to use the key differences and analyze the breathing signal from COVID-19 positive and negative subjects that can contribute towards the detection of the disease. In total, the samples provided by the organizers include the data of 965 subjects that is further split into train and validation set. The dataset contains lists corresponding to a 5-fold cross validation split.

Track 2: Cough - The goal of this track is to use cough sound recordings from COVID-19 positive and negative subjects. The validation set is composed of cough audio data from 965 subjects. The dataset also contains lists corresponding to a 5-fold cross validation split.

Track 3: Speech - Similar to the Track 1, the goal of this track also aims to detect the COVID-19 disease using the speech signals from positive and negative COVID-19 subjects. The data distribution and 5-fold cross validation is also similar to the previous tracks.

Track 4: Fusion - In this track, all the scores from the breath, speech and cough track of the corresponding folds are used to have the simple arithmetic mean of those particular folds. The validations scores are the concatenation of all the folds.

3.1. Baseline and evaluation metrics

The organisers of the challenge provide a baseline system for 4 different tracks based on log Mel spectrogram. The back-end classifier used is a bidirectional Long Short-Term Memory (bi-LSTM) Classification performance evaluation is measured using traditional detection metrics, namely, true positive rate (TPR) and false positive rate (FPR) over a range of decision thresholds. From these metrics, the probability scores for each audio file are used to compute the receiver operating characteristic (ROC) curve, and the area under the curve (AUC) metric to quantify the model performance [24].

4. EXPERIMENTAL SETUP

We have performed the experiments on the second DiCOVA Challenge database. Five acoustic features discussed in Section 2 have been used along with a cascade of two bi-directional long-short term memory (bi-LSTM) and a fully connected neural network with an encoder-decoder style network. The encoder consists of two bi-LSTM layers with 128 units in both the forward and back-ward direction. This is fully connected neural network comprising of 256 nodes in the first layer and 64 nodes and a tanh$(\cdot)$ non-linearity in the second layer. Finally, a single node output, passed through a sigmoid non-linearity is obtained as the COVID-19 probability score for the input feature matrix.

Parameters used to extract acoustic features are detailed hereafter.

MelSPEC: The MelSPEC feature set was extracted, similarly for the baseline system, using 64-dimensional log Mel spectrogram with Δ and $\Delta\Delta$ resulting in total 192-D feature vector.

TECC: The TECC feature set was extracted using 40 Mel-spaced Gabor filterbank with f_{min}=10 Hz, and f_{max}=fs/2 Hz [13]. For each subband filtered signals, we obtain 40-D static features appended along with their Δ and $\Delta\Delta$ coefficients resulting in 120-D feature vector.

ESA-IACC: The ESA-IFCC feature set was extracted using same parameters as used for TECC feature set expect the frequency scale in Gabor filterbank, here we used linearly-spaced Gabor filterbank. However, ESA-IACC feature set is computed with the pre-processing technique and cepstral mean normalization (CMN) technique for COVID-19 classification task.

CQCC: The CQCC features are extracted with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 44.1kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq 43$Hz (9 being the number of octaves). The number of bins per octave B is set to 96. Only 20 static coefficients (with log-energy) were considered, resulting in total 60-dimensional (D) feature vector (including *20*-Δ and *20*-$\Delta\Delta$).

FBCQT: The Filterbank CQT features set were extracted using 63-dimensional log linearised CQT with with a maximum frequency of $F_{max} = F_{NYQ}/2$ and a minimum frequency of $F_{min} = F_{max}/2^10 \simeq 43$Hz, with Δ and $\Delta\Delta$ resulting in total 189-D feature vector.

5. EXPERIMENTAL RESULTS

The results in terms of AUCs obtained on the validation folds for Breathing and Cough tracks are reported in Table 1 and for Speech and Fusion tracks reported in Table 2. For each fold the classifier is trained using the training data and evaluated on the validation data. The average validation AUC denotes the average over the AUCs for the 5 folds. The acoustic features considered have their strengths and weaknesses and therefore the AUC for some folds and tracks are better compared to other folds and tracks. For all the tracks of validation set, FBCQT gave the higher AUC compared to other features. In particular, for breathing track FBCQT gave an average AUC of 80.52%. For Cough, Speech and Track Fusion it yielded an average AUC of 79.60%, 81.04, and 84.18%, respectively. IACC and FBCQT feature obtained the same AUC of 81.04% for the speech track.

We now focus on the results obtained on the blind test set that we submitted to the challenge. For evaluation on the test dataset, the COVID-19 positive likelihood score for each file

Table 1. Results on validation set for Breathing and Cough on Second DiCOVA Challenge database in terms of AUC (%).

Folds	Breathing					Cough				
	MelSPEC	TECC	IACC	CQCC	FBCQT	MelSPEC	TECC	IACC	CQCC	FBCQT
0	76.40	74.40	78.50	78.60	80.00	66.80	74.30	76.00	79.40	79.60
1	75.00	73.90	80.30	78.10	80.00	77.10	67.50	78.70	74.50	84.20
2	75.00	74.80	76.60	78.30	74.80	77.40	69.80	74.60	77.70	78.80
3	80.30	72.30	77.70	75.00	78.90	74.80	81.00	74.70	78.20	77.50
4	82.10	87.80	86.40	82.60	88.90	77.40	73.40	79.00	83.20	77.90
Avg	77.76	76.64	79.90	78.52	**80.52**	74.70	73.20	76.60	78.60	**79.60**

Table 2. Results on validation set for Speech and Fusion on Second DiCOVA Challenge database in terms of AUC (%).

Folds	Speech					Track Fusion				
	MelSPEC	TECC	IACC	CQCC	FBCQT	MelSPEC	TECC	IACC	CQCC	FBCQT
0	74.60	70.70	75.10	74.20	78.10	75.00	77.30	79.90	80.70	82.20
1	85.90	79.20	80.30	82.50	87.00	82.90	76.90	83.90	80.30	86.70
2	81.20	75.90	80.60	75.50	76.50	82.70	78.10	80.70	80.10	79.90
3	78.90	76.80	81.80	76.80	77.70	81.20	81.90	82.20	78.90	81.60
4	83.80	81.00	87.40	81.70	85.90	88.40	87.40	90.90	88.00	90.50
Avg	80.88	76.72	**81.04**	78.14	**81.04**	82.04	80.20	83.52	81.60	**84.18**

was computed by taking the average over the score outputs from the 5 validation fold models. As discussed earlier on validation set, for all the tracks FBCQT gave high AUC compared to all the features considered here. However, on test set the results were contradictory to the validation set. The FBCQT feature did not perform well on test set and results in lower AUC for all the tracks compared to all the other features taken into consideration.

The best and second best system on the test set includes MelSPEC and CQCC features giving an AUC of 84.50% and 82.21% on breathing track (with sensitivity of 31.67 % and 36.67 % at specificity of 95.13 %) and 74.89% and 76.98% on cough track (with sensitivity of 36.67 % and 25.00 % at specificity of 95.13 %), respectively. On speech track, MelSPEC and IACC feature set are the best and second best giving an AUC of 84.70% and 80.92%, respectively, (with sensitivity of 43.33 % and 45.00 % at specificity of 95.13 %) as can be viewed from Table 3.

Table 3. Single System results on test set of the Second DiCOVA Challenge Database

Subset	Single System				
	MelSPEC	TECC	IACC	CQCC	FBCQT
Breathing	**84.50**	67.92	79.02	82.21	55.10
Cough	74.89	68.31	71.55	**76.98**	52.20
Speech	**84.26**	77.41	80.92	76.90	51.90
Fusion	**84.70**	77.75	83.26	82.01	53.00

Last but not least, we also report fusion experiments to understand the complementary information that is present in each acoustic feature under investigation. Systems were selected each time by adding the next worst in terms of AUC according to the Track Fusion results reported in Tables 2. Fusion results are shown in Table 4. Unexpectedly, the best combination results in MelSPEC + IACC + FBCQT feature set giving an AUC of 86.60%. All the combinations outperform the single system based on MelSPEC except the IACC + FBCQT.

Table 4. System's Fusion on test set of the Second DiCOVA Challenge Database

System's Fusion						
MelSPEC	TECC	IACC	CQCC	FBCQT	AUC
×	×	✓	×	✓	82.70
✓	×	✓	×	✓	**86.60**
✓	×	✓	✓	✓	86.00
✓	✓	✓	✓	✓	85.80

6. SUMMARY AND CONCLUSIONS

This paper reports on the exploration of acoustic cues using different auditory-based features for the diagnosis of COVID-19. Particularly, the systems presented are based on 5 different acoustic features based on Mel frequency scale, Teager energy operator, speech demodulation, and constant Q transform. For a proper comparison of these features we use the same back-end consisting of a bi-LSTM network. The FBCQT system outperforms all the proposed systems for the validation set on all the tracks, however, for the test set, it demonstrated a low capacity for generalisation with limited accuracy. Fusion experiments showed that the features considered are highly complementary. The best fusion gave an AUC of 86.60% for the MelSPEC + IACC + FBCQT feature combination on the test set, which led us to a result between the challenge baseline system (84.70%) and the challenge winner system (88.44%).

7. REFERENCES

[1] Ananya Muguli, Lancelot Pinto, Nirmala R, Neeraj Sharma, Prashant Krishnan, Prasanta Kumar Ghosh, Rohit Kumar, Shrirama Bhat, Srikanth Raj Chetupalli, Sriram Ganapathy, Shreyas Ramoji, and Viral Nanda, "DiCOVA Challenge: Dataset, Task, and Baseline System for COVID-19 Diagnosis Using Acoustics," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 901–905.

[2] Madhu R. Kamble, Jose A. Gonzalez-Lopez, Teresa Grau, Juan M. Espin, et al., "PANACEA Cough Sound-Based Diagnosis of COVID-19 for the DiCOVA 2021 Challenge," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 906–910.

[3] Rohan Kumar Das, Maulik Madhavi, and Haizhou Li, "Diagnosis of COVID-19 Using Auditory Acoustic Cues," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 921–925.

[4] Swapnil Bhosale, Upasana Tiwari, Rupayan Chakraborty, and Sunil Kumar Kopparapu, "Contrastive Learning of Cough Descriptors for Automatic COVID-19 Preliminary Diagnosis," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 946–950.

[5] Flavio Avila, Amir H. Poorjam, Deepak Mittal, Charles Dognin, Ananya Muguli, Rohit Kumar, Srikanth Raj Chetupalli, Sriram Ganapathy, and Maneesh Singh, "Investigating Feature Selection and Explainability for COVID-19 Diagnostics from Cough Sounds," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 951–955.

[6] Kotra Venkata Sai Ritwik, Shareef Babu Kalluri, and Deepu Vijayasenan, "COVID-19 Detection from Spectral Features on the DiCOVA Dataset," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 936–940.

[7] Vincent Karas and Björn W. Schuller, "Recognising Covid-19 from Coughing Using Ensembles of SVMs and LSTMs with Handcrafted and Deep Audio Features," in *INTERSPEECH*, 2021, pp. 911–915.

[8] John Harvill, Yash R. Wani, Mark Hasegawa-Johnson, Narendra Ahuja, David Beiser, and David Chestek, "Classification of COVID-19 from Cough Using Autoregressive Predictive Coding Pretraining and Spectral Data Augmentation," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 926–930.

[9] Isabella Södergren, Maryam Pahlavan Nodeh, Prakash Chandra Chhipa, Konstantina Nikolaidou, and György Kovács, "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 916–920.

[10] Gauri Deshpande and Björn W. Schuller, "The DiCOVA 2021 Challenge — An Encoder-Decoder Approach for COVID-19 Recognition from Coughing Audio," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 931–935.

[11] Adria Mallol-Ragolta, Helena Cuesta, Emilia Gómez, and Björn W. Schuller, "Cough-Based COVID-19 Detection with Contextual Attention Convolutional Neural Networks and Gender Information," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 941–945.

[12] Debarpan Bhattacharya Debottam Dutta Pravin Mote Sriram Ganapathy Neeraj Kumar Sharma, Srikanth Raj Chetupalli, "The Second DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," in *submitted to IEEE Intl. Conference on Acoustics Speech Signal Processing (ICASSP)*, Singapore, 2022, pp. 1–5.

[13] Madhu R Kamble and Hemant A Patil, "Analysis of reverberation via teager energy features for replay spoof speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, London, UK, pp. 2607–2611.

[14] Madhu R. Kamble, Hemlata Tak, and Hemant A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 641–645.

[15] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[17] P. Maragos, J.F. Kaiser and T.H. Quatieri , "On separating amplitude from frequency modulations using energy operators," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, California, USA, 1992, vol. 2, pp. 1–4.

[18] Maragos, Petros and Kaiser, James F and Quatieri, Thomas F, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[19] Thomas F Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1^{st} edition, Pearson Education India, 2015.

[20] Madhu R. Kamble and Hemant A. Patil, "Detection of replay spoof speech using Teager energy feature cues," *Computer Speech Language*, vol. 65, pp. 101140, 2021.

[21] Madhu R. Kamble, Hemlata Tak, and Hemant A. Patil, "Amplitude and frequency modulation-based features for detection of replay spoof speech," *Speech Communication*, vol. 125, pp. 114–127, 2020.

[22] James Youngberg and Steven Boll, "Constant-q signal analysis and synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tulsa, Oklahoma, USA, 1978, vol. 3, pp. 375–378.

[23] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.

[24] Tom Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition.