Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus

Cedric Fayet[1,2], Arnaud Delhay[1,2], Damien Lolive[1,2], Pierre-François Marteau[1,3]

[1]IRISA, EXPRESSION Team, France
[2] Université de Rennes 1, [3] Université de Bretagne Sud
{cedric.fayet,arnaud.delhay,damien.lolive,pierre-francois.marteau}@irisa.fr

Abstract

This paper presents an attempt to evaluate three different sets of features extracted from prosodic descriptors and Big Five traits for building an anomaly detector. The Big Five model enables to capture personality information. Big Five traits are extracted from a manual annotation while Prosodic features are extracted directly from the speech signal. Two different anomaly detection methods are evaluated: Gaussian Mixture Model (GMM) and One-Class SVM (OC-SVM), each one combined with a threshold classification to decide the "normality" of a sample. The different combinations of models and feature sets are evaluated on the SSPNET-Personality corpus which has already been used in several experiments, including a previous work on separating two types of personality profiles in a supervised way. In this work, we propose the above mentioned unsupervised or semi-supervised methods, and discuss their performance, to detect particular audio-clips produced by a speaker with an abnormal personality. Results show that using automatically extracted prosodic features competes with the Big Five traits. The overall detection performance achieved by the best model is around 0.8 (F1-measure).

Index Terms: Anomaly detection, Gaussian Mixture Model, One Class-Support Vector Machine, Threshold Classification, Social Signal, Big Five, Prosody, SSPNET-Personality.

1. Introduction

According to [1], "an anomaly is defined as a pattern that does not conform to an expected normal behavior". The main objective of any anomaly detection system is to identify abnormal states from normal state distributions. The feature sets that describe these states are related to the nature of the input data (continuous, categorical, spatial, or spatio-temporal data), but also to the nature of the anomalies that we aim to track (pointwise, contextual or collective anomalies) [1]. Finally, to choose an adequate anomaly detection model, we also need to consider the desired output (score or label) and the availability of labeled data (supervised or unsupervised techniques).

In the literature, anomaly detection techniques are applied in intrusion detection [2], fraud detection [3], sensor network [4], monitoring flight safety [5], *etc.* As far as we know, the use of anomaly detection with the speech signal is more focused on speech pathology or disorder [6], or on the deduction of another type of pathology or disease, as cancer for instance, from the speech signal [7]. The speech signal is also used for detecting stress or depression [8, 9], that could be seen as an anomaly detection in the way that this state of mind can be considered as an abnormal mental state. Handling audio, video or biological signal to infer social information such as personality is part of a field called *Personality Computing*. This field is focused on the recognition of self perceived personality (Automatic Personality Recognition), the prediction of the personality perceived by the others (Automatic Personality Perception), and the generation of artificial personalities (Automatic Personality Synthesis) [10].

The Big Five model has been proposed to describe the speaker personality using five personality traits: openness, conscientiousness, extraversion, agreeableness and neuroticism [11]. Different tests with adaptation to a local context have been conducted on different languages and cultures, and the big five model seems to be generalizable to them [12, 13].

SSPNET-Personality corpus was built to experiment the prediction of the big five scale over audio features [14]. It is composed of french audio clips extracted from the "Radio Suisse Romande" and recorded by either professional or guest speakers. [15] offers a good overview of the different systems using this corpus. A lot of experiments have already been conducted on this corpus and most of them try to predict the value of the big five scales using different audio representations. Adding to this, some experiments in [14] suggest that the big five representation could be a good predictor of the role of a speaker. The authors propose a supervised SVM method that learns to recognize if a speaker is a professional one or a guest. The feature set used in this work is the psychological evaluation of a sample given by annotators using Big five traits. Their predictor reaches an accuracy of about 75%.

The purpose of our work is to compare the use of the big five features to the use of prosodic features to design an anomaly detector able to separate a normal personality class from an abnormal one. To this end, we consider a professional speaker as belonging to a normal personality class and a guest to an abnormal one. Contrary to [14], we propose to work in the unsupervised anomaly detection framework and we propose to evaluate two unsupervised strategies. Three sets of features are evaluated. Results show that prosodic features perform well and are less costly compared to manually annotated Big Five traits, even if those features may provide better detection performance.

The remainder of the paper is structured as follows. The anomaly detection method used is described in section 2. In section 3, the feature sets we compare are detailed before describing the experimental setup in section 4. Finally, the results are presented and discussed in section 5.

2. Method

This work aims at determining if a sample is normal (professional speaker) or abnormal (guest speaker) based on a train set containing only normal audio clips. Each audio clip is described by a feature vector summarizing the time evolution of the features as described in section 3.

In this section, we describe the two methods that we have

used to perform the anomaly detection: a gaussian mixture-based approach and a One-Class Support Vector Machine approach (referred to as OC-SVM in the rest of the paper). The OC-SVM has been used to ensure a comparison baseline with the SVM method used in [14].

2.1. Gaussian-Mixture Model approach (GMM)

The Gaussian-mixture approach for detecting anomaly can be decomposed into two steps:

1. learning a Gaussian Mixture Model (GMM) to model the feature space distribution for the normal samples,
2. choosing a threshold based on the likelihood for a sample to decide if it has been generated or not by the learned distribution, and thus if the sample could be considered as normal.

Considering the set $\mathbf{X}$ of feature vectors $\mathbf{x}$, each representing an audio clip, a GMM $\mathcal{M}_{\mathbf{X}}$ with M Gaussians is chosen to model the dataset $\mathbf{X}$. Its probability distribution is given by

$$P(\mathbf{x}|\mathbf{\Theta}) \quad = \quad \sum_{m=1}^{M} \alpha_m P(\mathbf{x}|\theta_m)$$

where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_M)$ is the set of parameters. α_m is the mixing coefficient associated to the m^{th} Gaussian with parameters $\theta_m = (\mu_m, \Sigma_m)$ that represent respectively the mean vectors and the covariance matrices and distribution $P(\mathbf{x}|\theta_m)$. The *Expectation-Maximization (EM)* algorithm is used to learn the GMM parameters from unlabeled data and maximizes the loglikelihood of the data and the model [16].

By using the distribution learned at the first step, considering a new sample that we want to label either as normal or abnormal, its likelihood is evaluated and compared to a threshold value. If the likelihood of the sample is above the threshold, the sample is considered as normal. Otherwise, it is considered to be abnormal.

The structure of covariance matrix used for each Gaussian component is considered to be an hyper-parameter of the model. Consequently, we have three hyper-parameters : the covariance matrix type, the number of components (both are related to the Gaussian distribution) and the threshold value.

2.2. One-Class SVM approach (OC-SVM)

In short, a SVM classifier learns a boundary which maximizes the margin between classes. This well-known approach has been shown to be very effective on many classification problems. OC-SVM is an adaptation of SVM to the one-class problem. After transforming the feature via a kernel, OC-SVM considers as a starting point, all the available data as member of a single class $C_{inliers}$ and the origin as the only member of a class $C_{outliers}$ [17].

During the training, the hyper-parameter ν corresponds to a penalizing term which represents a trade-off between inliers and outliers. With the SVM approaches, the choice of the kernel is important to improve the results. The most widely used kernels are linear (inner product), polynomial, RBF and sigmoid. Related to the kernel used (RBF, polynomial, sigmoid) we have to determine parametric coefficients related to the kernel. For example, with the sigmoid kernel :

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x}^T \mathbf{y} + c_0)$$

where $\mathbf{x}$ and $\mathbf{y}$ are input vectors and γ (slope) and c_0 (intercept) are parametric values related to the kernel.

To maintain a similar approach to the one that we have adopted for the GMM, we choose to use the distance between a new sample and the learned boundary as an anomaly score, instead of directly deciding which class the samples belong to. We then use a classification threshold based on this anomaly score to decide if a sample is normal or abnormal. This two steps strategy introduces some adjustable fuzziness around the boundary.

3. Materials

The experiments are conducted over the SSPNET-Personality Corpus[14]. The corpus contains 640 audio clips divided into 307 audio clips of professional speakers and 333 audio clips of guests in French language. The duration of each audio clip is about 10 seconds, based on the assumption that it takes short time to get an opinion about others personality. For each sample, 11 non-native french speakers evaluate the BFI-10 Questionnaire [18] from which a score is computed for each Big Five's scale.

3.1. Big Five features

For each sample, 11 evaluations of the BFI-10 Questionnaire (so 11 evaluations of the big five features) are available. For our experiment, we consider two sets of features based on the big five model:

- *BigFive5*: for each sample and for each Big Five scale, we compute the mean of the 11 evaluations which leads to 5 features.
- *BigFive55*: for each sample, we concatenate the 11 evaluations given for each sample which leads to 55 features.

The purpose of this last set of features is to verify if the information contained in several distinct annotations is complementary or can simply be aggregated in a lower dimension feature vector, as in *BigFive5* feature vector. The main drawback of these two feature sets is that they are the result of manual annotation, which is furthermore difficult to predict from the speech signal.

3.2. Prosodic Features

Prosodic features are commonly used to capture affect cues in a speech signal. Contrary to the Big Five traits, a large number of prosodic features are much easier to extract automatically from the speech signal. In this study, we adopt the 6 dimensional prosodic feature set as described in [14]. Using the Praat software [19], we extract the pitch, the first two formants, the energy and the duration of voiced/unvoiced segments with a sliding analysis window size of 40 ms with a step of 10 ms. From these low-level features, resulting from the extraction, we derive the final features that summarize their time evolution by computing mean, maximum, minimum and entropy values for each of the 6 features. Consequently, the final *Prosodic* feature set is composed with 24 features for each audio clip.

3.3. Pre-processing

Normalizing the features is an important preprocessing step before using the OC-SVM approach. Features that do not share the same range of values and the same variations could affect the quality of the OC-SVM model. Therefore this step is really important in the case of *Prosodic* features, which are composed of different types of features. We choose to perform a standardization (zero mean and unit variance) on all types of features.

4. Experimental Setup

The different experiments described in this section are carried out by following the same procedure. From the corpus, we build three sets (train, test, and validation sets), as described below. To increase the statistical confidence of our results, we run each experiment 60 times by distributing randomly the samples on the three sets. For each experiment, we compute the mean and the standard-deviation divided by the mean of the different runs.

The data is thus split into three folds as follows:

- Train set: 207 clips of professional speakers and a variation from 0 to 103 guest clips.
- Test set: 50 clips of professional speakers and 50 clips of guests.
- Validation set: 50 clips of professional speakers and a variation from 0 to 50 of guest clips.

We use python 3.5 and scikit-learn [20] to conduct the different experiments.

4.1. Hyper-parameters tuning

According to the available data, since we have 207 samples in the training set, we need to be careful about the number of parameters in our model to avoid over-fitting.

In the case of the Gaussian Mixture approach, we need to determine two hyper-parameters: the co-variance matrices and the number of components. The choice to use a diagonal covariance matrix seems to be a good compromise between complexity of the model and size of training set. Once the structure of the co-variance matrix has been determined, we need to choose the number of considered components (referred to as #cp) . One of the most used method to make this choice is to balance the complexity of the model (number of components) and the quality of the model on the training set. The AIC and BIC scores [21] are well-known methods to reach this goal. In our case, the number of features is low (see in section 3), so the BIC score is more likely to be reliable [22].

In the case of the OC-SVM approach, after testing different types of kernel in our experiments, we have chosen to present the results with the sigmoid kernel which gives the best results. With the sigmoid kernel, we need to determine three hyper-parameters ν, γ and c_0. For the rest of the paper, we choose to fix c_0 at 0. Without considering the fuzzy boundary, the OC-SVM gives a first classification of the sample. By using it, we compute a classification score [23] which can be used to evaluate the classification quality and consecutively the quality of the hyper-parameters.

For each set of features, we considered a training set with all the available normal samples, to determine the hyper-parameters (table 1) with respect to the aforementioned methods.

Table 1: *Hyper-parameters chosen for each AD*

	GMM		OC-SVM		
	Co-var	#cp	Kernel	γ	ν
BF5	diagonal	4	sigmoid	0.016	0.7
BF55	diagonal	14	sigmoid	0.007	0.6
PROSODIC	diagonal	10	sigmoid	0.009	0.57

4.2. Experiments

4.2.1. Comparing two AD models on the different feature sets

To compare the different sets of features according to the chosen approach, we carry out the following steps: for each set of values (*BigFive5*, *BigFive55*, *Prosodic*), we test each anomaly detection model after tuning hyper-parameters as explained before (4.1). Then, we compare the different models by using a ROC (Receiver Operating Characteristic) curve. It means that for a given FPR (False Positive Rate), we search the associated TPR (True Positive Rate). This sampled association is obtained by testing a range of possible threshold values. The ROC curve gives an information about the detector quality as a function of the threshold value.

A last step consists in estimating the robustness of each detector to a degradation of the training set. In this purpose, we introduce a certain percentage of abnormal samples into the training set (0% to 50%). We keep the different hyper-parameter values unchanged.

4.2.2. Influence of the validation set purity on the threshold choice

In the previous experiments, we use a range of possible values for the threshold which allows to separate the normal and abnormal classes. In this step, we want to find a way to get the best threshold value. In this purpose, we design a validation set which contains 50 normal samples and a variable percentage of abnormal samples (variation of the percentage of abnormal sample, starting from 0, ending at 50 with a step of 5 percent). To evaluate the quality of each model, we use the F1-score (harmonic mean between recall and precision) with normal instance as positive and abnormal as negative instance. For each step, we find the threshold value that maximizes a F1-score on the validation set. Then, we evaluate the F1-score on the test set with the threshold value obtained before.

5. Results

5.1. Comparing two AD models on the different feature sets

Our first experiment (table 2) consists in evaluating the quality of each feature set. On our data, the *BigFive55* feature vector, regardless of the method used, performs better than the others. By considering, the OC-SVM approach, the *Prosodic* and *BigFive5* feature vectors get comparable results. For the GMM approach, *BigFive5* feature vector achieves a poor performance. A possible explanation is the fact that we consider a very simple model with few parameters. Moreover, the *BigFive5* feature vector has the lowest results for both methods, thus showing that the aggregation of the individual annotations induces a significant information loss.

Our second experiment (figure 1) consists in evaluating the quality of each AD model when the training set is degraded, *i.e.* by including a certain percentage of abnormal samples in it. The results show that the GMM and OC-SVM approaches have

Table 2: *Mean area (and his standard-deviation divided by the mean) under ROC curve (ROC-AUC score) for the two approaches combined with the three feature sets.*

	BF5	BF55	PROSODIC
GMM	0.757 (*0.053*)	0.936 (*0.025*)	0.892 (*0.032*)
OC-SVM	0.857 (*0.036*)	0.918 (*0.026*)	0.876 (*0.041*)

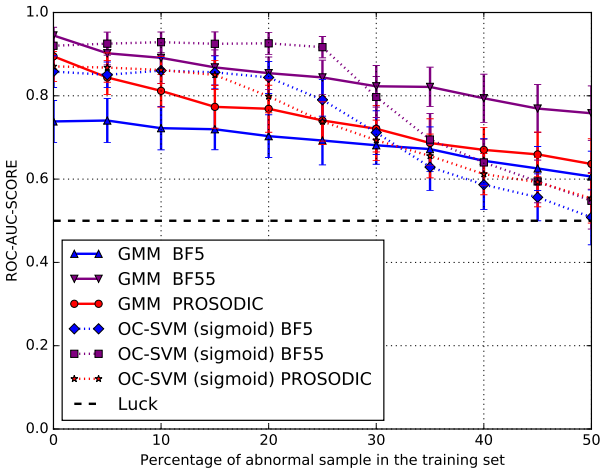

Figure 1: *Robustness of the models to the introduction of abnormal samples in the training set.*

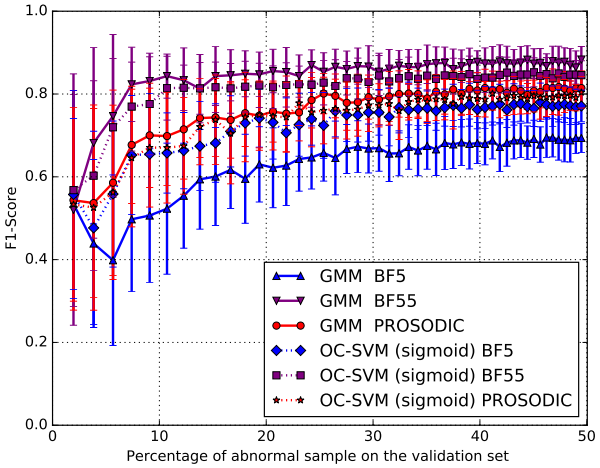

Figure 2: *Percentage of annotated abnormal sample in the validation set.*

different responses to degradation. The OC-SVM approaches have a degradation in two steps: before a certain percentage of contamination they are robust to degradation and after this limit their scores decrease down to around 0.5 (random detector) for 50% of degradation. The GMM approaches are more sensible to small degradation compared to the OC-SVM approaches but they have a kind of linear degradation along the degradation axis, and their scores for 50% of contamination depend on the feature set but, however, they are still better than a random detector.

A possible explanation for the OC-SVM approach, is that the number of abnormal samples included into the training set became, after a certain percentage, too important compared to the number of slacks of variables considered to make a trade-off between inliers and outliers. This phenomenon drives the OC-SVM to increasingly consider the abnormal samples as inliers.

For the OC-SVM approach, the *BigFive55* feature vector seems to keep stable results for approximately 25 percents of degradation. The other two feature vectors start to have a decrease of their quality for less than 15 percent of degradation. For the GMM approach, the *prosodic* feature vector seems to be more sensible to degradation than the two feature sets composed with Big Five information. The *BigFive55* feature vector gives a more robust separation between normal and abnormal samples.

5.2. Influence of the validation set purity on the threshold choice

To conduct this experiment, we considered the threshold values obtained by computing the ROC curve as explained in section 4.2.2. For the two approaches, the curves obtained (figure 2) for a feature vector seem to share a same tendency.

The *BigFive55* feature vector curves seems to reach an asymptotic value for at least 10 percents of abnormal samples on the validation set (around 6 samples), and at least 30 percents (21 samples) for the two other feature vectors.

The asymptotic F1-score reached with *BigFive55* feature vector and GMM approach is above 0.8, and the F1-score obtained with $Prosodic$ feature vector and OC-SVM approach corresponds to the one obtained in a supervised way with SVM [14]. The variation of the result in the figure 2 shows the difficulty to reach a good and stable threshold value.

6. Conclusion

The main objective of this paper was to compare the use of prosodic cues and the Big Five annotation traits as feature sets for an anomaly detection. We have conducted some experiments with the SSPNET-Personality Corpus using professional speaker as normal samples and guest as abnormal samples. This choice was motivated by Mohammadi *et al.* work [14] that demonstrates the effectiveness of using the Big Five features to train a classifier able to separate these two categories of speakers. We built three sets of features (*BigFive5*, *BigFive55* and *Prosodic*) based on the speech signal and a psychological evaluation (Big Five model) available on the dataset. We have used two different machine learning methods (GMM, OC-SVM) to build our anomaly predictors. Based on the results, the *BigFive55* feature set seems to outperform the two other sets of features. The *Prosodic* feature set seem to have an advantage over the *BigFive5* feature set. The good performance of the *BigFive55* feature set over the *Prosodic* feature set indicates that features based on psychological information can bring more information than audio features only. However, the prosodic features are easy to extract from the speech signal and thus seem to be the best compromise between ease of extraction and performance.

A natural follow up is to test other feature sets: for instance one can increase the number of features to reach the same order of dimension as the *BigFive55* feature vector. Finally, conducting these experiments on other audio corpora, or trying to generalize our results on multimedia data are part of future work.

7. Acknowledgments

This research has been financially supported by the French Ministry of Defense - Direction Générale pour l'Armement and the région Bretagne (ARED) under the MAVOFA project.

8. References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[2] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," Technical report, Tech. Rep., 2000.

[3] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[4] K. Park, Y. Lin, V. Metsis, Z. Le, and F. Makedon, "Abnormal human behavioral pattern detection in assisted living environments," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2010, p. 9.

[5] L. Li, M. Gariel, R. J. Hansman, and R. Palacios, "Anomaly detection in onboard-recorded flight data using cluster analysis," in *Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th*. IEEE, 2011, pp. 4A4–1.

[6] J. B. Alonso, F. Daz-de Mara, C. M. Travieso, and M. A. Ferrer, "Using nonlinear features for voice disorder detection," in *ISCA tutorial and research workshop (ITRW) on non-linear speech processing*, 2005.

[7] R. P. Clapham, L. v. d. Molen, R. J. J. H. v. Son, M. W. M. v. d. Brekel, and F. J. M. Hilgers, "NKI-CCRT Corpus - Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy," in *LREC*, 2012.

[8] L. He, M. Lech, N. C. Maddage, and N. Allen, "Stress detection using speech spectrograms and sigma-pi neuron units," in *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, vol. 2. IEEE, 2009, pp. 260–264.

[9] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[10] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[11] O. P. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[12] L. R. Goldberg, "Language and individual differences: The search for universals in personality lexicons," *Review of personality and social psychology*, vol. 2, no. 1, pp. 141–165, 1981.

[13] M. Gurven, C. Von Rueden, M. Massenkoff, H. Kaplan, and M. Lero Vie, "How universal is the Big Five? Testing the five-factor model of personality variation among foragerfarmers in the Bolivian Amazon." *Journal of personality and social psychology*, vol. 104, no. 2, p. 354, 2013.

[14] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[15] B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, and others, "A survey on perceived speaker traits: personality, likability, pathology, and the first challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100–131, 2015.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[17] B. Schlkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[18] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Dec. 2016. [Online]. Available: http://www.praat.org/

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] R. J. Steele and A. E. Raftery, "Performance of Bayesian model selection criteria for Gaussian mixture models," *Frontiers of Statistical Decision Making and Bayesian Analysis*, vol. 2, pp. 113–130, 2010.

[22] G. Schwarz and others, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[23] T. Caliski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.