Prediction of Personal Characteristics and Emotional State based on Voice Signals using Machine Learning Techniques

Marta Babel Guerreiro[1], Catia Cepeda[1], Joana Sousa[2,3], Carolina Maio[2], João Ferreira[2] and Hugo Gamboa[1]

[1]*LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal*

[2]*NOS Inovação, Lisboa, Portugal*

[3]*Bold International, Lisboa, Portugal*

Keywords: Gender, Age, Emotion, Machine Learning, Voice Signal.

Abstract: Voice signals are a rich source of personal information, leading to the main objective of the present work: study the possibility of predicting gender, age, and emotional valence through short voice interactions with a mobile device (a smartphone or remote control), using machine learning algorithms. For that, data acquisition was carried out to create a Portuguese dataset (consisting in 156 samples). Testing Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) classifiers and using features extracted from the audio, the gender recognition model achieved an accuracy of 87.8%, the age group recognition model achieved an accuracy of 67.6%, and an accuracy of 94.6% was reached for the emotion model. The SVM algorithm produced the best results for all models. The results show that it is possible to predict not only someone's specific personal characteristics but also its emotional state from voice signals. Future work should be done in order to improve these models by increasing the dataset.

1 INTRODUCTION

Over the past few years, great advances have been made in the technological field to obtain a more human-centered approach to its development. Furthermore, this technological evolution is not only bringing the resources to study and develop innovations in many different areas but also changing people's life habits and routines by being a part of human's daily life (Mamyrbayev et al., 2020).

Some personal characteristics could be combined to identify an individual, such as gender, age, race, ethnicity, physical attributes, personality, and so on (Satterfield, 2017). Voice signals are a rich source of personal data, and they have been increasingly used in many applications, particularly in recognition, as long as voice characteristics are unique, varying from person to person (Mamyrbayev et al., 2020; Aizat et al., 2020). Moreover, with the current worldwide pandemic situation, the study of emotions has become clearer, once mental health is being affected in different communities - states of depression and anxiety are gaining weight within confinement conditions (Bueno-Notivol et al., 2021; Salari et al., 2020; Tang et al., 2021). The most common ways to identify emotions are through facial expressions, electrophysiological signals or voice signals. As stated by Michael W. Kraus, identifying someone's emotions by its facial expressions is less accurate than by its voice signals so developing tools to identify emotions by voice has become even more relevant (Kraus, 2017).

Thereby, the aim of the present study is to identify user's gender and age group and its emotional state through a simple and short sentence captured by the mobile phone microphone, using machine learning algorithms.

The developed models have applicability in real context, in situations such as teleconsultations to monitor the patient's mental health status, in where the emotional state of the patient can be extracted from its voice and help the doctor understanding how it is; and it can also be implemented as a tool to understand if a patient is emotionally affected in a postoperative situation, once doctors usually cannot detect.

In addition, models could also be applied in telecommunications contexts: recognizing gender,

Guerreiro, M., Cepeda, C., Sousa, J., Maio, C., Ferreira, J. and Gamboa, H.
Prediction of Personal Characteristics and Emotional State based on Voice Signals using Machine Learning Techniques.
DOI: 10.5220/0010802700003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 4: BIOSIGNALS*, pages 142-149
ISBN: 978-989-758-552-4; ISSN: 2184-4305

age and emotional state of the user behind the mobile phone and/or the television, could be useful to profile the individual and, therefore, to prepare and present a more personalized content and design.

1.1 Related Work

Being gender, age and emotion recognition topics of interest for some years, there is already some literature on how to automatically recognize these characteristics based on voice signals, using machine learning models. Even so, the studies that most resemble to the present work are presented below.

A study was conducted in 2015, in which the author have used features such as mel frequency cepstral coefficients (MFCC) and formant frequencies to perform gender and age recognition from voice signals. The selected algorithms were Support Vector Machine (SVM) to perform gender recognition and K-Nearest Neighbors (KNN) to age recognition. From a Kurdish dataset, 40 samples of males and 42 of females were used to perform gender recognition, and the best accuracy achieved was 96%. From the same Kurdish dataset, samples from 114 speakers were used, divided into 7 classes ranging from age 5 to 65, and an accuracy of 81.44% was reached (Faek, 2015).

In 2017, Júnior presented a system capable of recognizing emotions through voice signals from the Berlin Database of Emotional Speech (Emo-DB) - recordings of voice signals, with acted emotions, from 5 female and 5 male actors, with a total of 530 audio signals - and from a Brazilian Portuguese one - built by extracting excerpts from audios of films and videos from YouTube, with 4 emotions (anger, happiness, sadness and neutral) in Brazilian Portuguese, consisting of 110 audio signals in the end. A predefined list of only 4 basic emotions (neutral, anger, happiness, and sadness) was chosen to get the best accuracy. Some features such as MFCC, energy, and pitch were extracted, and KNN and SVM classifiers were tested. In the end, an accuracy of 86.79% was obtained for the German database and 70.83% for the Brazilian Portuguese, and both results were reached while using the SVM classifier (Júnior, 2017).

In a study from 2018, some features such as the pitch, energy, and MFCC, were extracted and used to train and test a SVM classifier, in order to perform gender recognition. These features achieved an accuracy of 96.45% with a database made out of 280 English speech files (140 files of males and 140 files of females) (Chaudhary and Sharma, 2018).

In 2019, a short review of age and gender recognition based on speech was done and some conclusions are taken about the most common extracted features, and selected classifiers. According to the studies investigated in this review, and concerning gender recognition, the most common extracted features are MFCC, spectral features, and energy; regarding age recognition, short-time energy, zero-crossing rate, and energy entropy are the most commonly extracted features. The most common classifiers are SVM, Random Forest, and Gaussian Mixture (Zhao and Wang, 2019).

In 2021, a study was presented in which the authors have used the DEMos database - an Italian database of audio signals, consisting of 9365 spoken utterances produced by 68 native speakers (23 females, 45 males) in a variety of emotional states - to recognize happiness, guilt, fear, anger, surprise, and sadness. The selected features were related to magnitude spectrum, voicing features (such as pitch and energy), MFCC, and RelAtive SpecTrAl (RASTA) coefficients, and the selected classifier was SVM. The experiments were carried out for females and males separately and, in the end, the best result had an accuracy of 81.7% for male voices and 84.1% for female voices (Costantini et al., 2021).

As seen above, there are already some studies focused on recognizing gender, age, and emotions through voice signals, using different methods. However, by the literature review that was presented, only one study was found using a data set with Brazilian Portuguese audio signals, which can lead to different results when comparing to Portuguese audio signals - machine learning models developed in this topic are trained in a specific language and are not applicable to other languages.

2 DATA ACQUISITION

To the best of our knowledge, there is no dataset of audio files in Portuguese so, to have one to use in this article, a process of data acquisition was made.

The whole acquisition process was conducted during 12 weeks, making use of a free online platform to build and create personalized forms called JotForm, which is also compliant with General Data Protection Regulation (JotForm, 2021). To acquire data, a personalized form was created and, to be eligible to fill out the form, all that was needed was for the person to be native Portuguese. A device capable of accessing the form with internet connection was required to record the audio.

At the beginning of the form, it was always shown an animated GIF referring to one question below that the person filling out the form was asked to read

and answer by voice recording. The questions were changed throughout the acquisition process (weekly) and were never related to confidential information but always to current issues, like "Hi! What is your favorite Olympic sport?" or "Hi! Which team do you think will win UEFA Euro 2020?".

After that, it was asked to choose from a list of four emotions - angry, happy, neutral, and sad - which one best suits your state of mind when filling out the form. Then, the person was asked to select:

- Age group: less than 21 years old, 21-30, 31-40, 41-50, 51-60, more than 60 years old;
- Gender: male, female;
- Hometown: Porto and north, center, Metropolitan Area of Lisbon, south, Azores, Madeira, other or "rather not answer".

Finally, the participant must give the last four digits of its mobile phone, or landline number, to be used as unique identifier of the generated audio.

The audio files were recorded in a *.wav* format of two-channel, with a sample rate of 22050 Hz.

2.1 Sample Description

The dataset consists of 156 samples from only 88 different people. The audio signals have durations between 1 second and 54 seconds, the median is 7 seconds, and the mean is 10 seconds.

According to the information acquired through the form, the description of the sample is presented in the next tables (Tables 1, 2, and 3). Since certain age groups and emotions have few samples, it was decided to group these categories into new ones in order to get better results, given that a small number of samples of data could be not enough for machine learning algorithms. These new groups are also found in the above-mentioned tables.

Table 1: Number of samples per gender.

Gender	Number of Samples
Female	64
Male	92

Table 2: Number of samples per age group.

Age Group	Number of Samples	Label	Number of Samples
Teens	7	A	72
Twenties	65		
Thirties	48	B	64
Forties	16		
Fifties	19	C	20
Olds	1		

Table 3: Number of samples per emotion.

Emotion	Number of Samples	Label	Number of Samples
Happy	71	Positive	136
Neutral	65		
Angry	6	Negative	20
Sad	14		

3 METHODS

To analyze and process the audio signals, and build the machine learning models (SVM, KNN, and Random Forest), some different libraries were used, such as Pandas, NumPy, LibROSA, pyAudioAnalysis; to extract the audio features, the libraries TSFEL and Python Speech Features were used; in the data preprocessing, the library imblearn was used to conduct the oversampling process; when it comes to feature selection, both libraries TSFEL and scikit-learn were used; regarding machine learning algorithms, scikit-learn was the chosen library to develop the models (Barandas et al., 2020; Lyons, 2013; Lemaître et al., 2017; Pedregosa et al., 2011).

It was also developed a platform that shows in real-time the results obtained with the developed models. For that, Dash was used to build the web application (Dash, 2021).

In the next sections, the tools used for feature extraction and selection will be presented, as well as the algorithms used to perform the classification, and even the metric used to evaluate the model.

3.1 Data Pre-processing

The first step of the pre-processing was to convert the audio files from the data acquisition with two-channel to one-channel.

After that, a data cleaning was conducted manually. Since the conditions under which the acquisition was carried out were not in a controlled environment, a data pre-selection process was performed with the intention of eliminating all corrupted data and also those files that did not meet the acquisition requirements. Consequently, files that could not be opened, files that were just silent or background noise, and files recorded by people who were not Portuguese were deleted from the database.

Posteriorly to the data cleaning, the dataset had to be organized in order to have all the audio files labeled according to the information that was asked in the data acquisition and, therefore, ready to be used in the models. From the raw files with random names, a method was applied to rename the file to include the

four digits and a code for gender, age, emotion, hometown. An example for a female, in the second (21-30) age group, feeling happy and living in Metropolian Area of Lisbon with 0000 as last four digits is: 0000_f_2_h_AML.

The last step of the pre-processing was to remove the silence periods from the acquired files so that, in the end, the files were only made out of speech samples. For that, the energy of the acquired audio files was calculated and used to remove this silent samples, as described below.

Each audio file was divided into windows and the energy of each window was calculated. When all windows were covered, the maximum energy value was calculated from the energy values of all windows. After that, the signal was divided into windows once again, in order to eliminate some windows according to the following formula:

$$E_w < \alpha * max(E_s) \quad (1)$$

where:

- E_w: energy calculated in each window of the signal;
- α: constant;
- $max(E_s)$: maximum energy value from the calculated energy values in each window.

The α constant was found by simultaneously observing the signal defined as being the noisiest of the entire dataset and its energy spectrum, in order to obtain the maximum value of this energy spectrum and the minimum energy while there was speech and calculating the ratio between them.

Running through all the signal windows, the energy of each one was calculated and if it was less than a threshold (defined by a constant multiplied by the maximum energy value found in that signal), then that signal window was removed from the signal.

Finally, to avoid biased results, since the data is imbalanced and it could comprise the models' learning process, an oversampling was carried out in all models to generate synthetic samples from minority classes, based on Adaptive Synthetic Sampling (ADASYN), using the library imblearn (Lemaître et al., 2017; He et al., 2008). Thus, the total number of samples per class is found in tables 4, 5, and 6:

Table 4: Number of samples per gender after oversampling.

Gender	Number of Samples
Female	87
Male	92

Table 5: Number of samples per age group after oversampling.

Age Group	Number of Samples
A	72
B	64
C	74

Table 6: Number of samples per emotion valence after oversampling.

Emotion Valence	Number of Samples
Positive	136
Negative	132

3.2 Feature Extraction and Selection

The set of features was extracted using two different libraries: Python Speech Features and TSFEL (Lyons, 2013; Barandas et al., 2020).

From Python Speech Features library, log filterbank energies (26 coefficients) and spectral subband centroids (26 coefficients) were extracted (Lyons, 2013).

TSFEL was used in order to obtain a very complete feature set since this library encompasses the extraction of features from time series in three different domains at once: spectral, statistical, and temporal (Barandas et al., 2020).

In the spectral domain, there are a total of 26 different features, such as the fundamental frequency of the signal, maximum power spectrum density, MFCC, and linear prediction cepstral coefficients (LPCC) (Barandas et al., 2020).

In the statistical domain, there are 16 features, such as the histogram of the signal, kurtosis, skewness, maximum, minimum, mean and median values, root mean square, standard deviation, and variance (Barandas et al., 2020).

In the temporal domain, there are 18 features. Examples of these features are the auto-correlation of the signal, the number of negative and positive turning points of the signal, mean and median absolute differences, zero-crossing rate, and total energy (Barandas et al., 2020).

In the end, we have a list of 442 different values counting as features (390 from TSFEL and 52 from Python Speech Features). From this list, a feature selection was done before each classification. This process was taken in two steps: first, if there were constant features, these features were removed, using the library scikit-learn, and, likewise, if there were highly correlated features, these features were also removed, using the library TSFEL. In the end, all the features were normalized using also the library scikit-learn (Barandas et al., 2020; Lyons, 2013; Pedregosa et al., 2011).

3.3 Classification

To perform classification, three different machine learning algorithms were used. The choice of these algorithms were based on the literature review results on other languages. Therefore, the chosen algorithms were SVM, KNN, and Random Forest.

The best parameters for each algorithm were selected based on grid-search for each experiment performed (gender, age, and emotion recognition).

3.3.1 Support Vector Machine

The basic principle of a SVM algorithm is to create lines or hyperplanes in a N-dimensional space (being N the number of features) called the decision boundaries, separating the labeled training data into classes (Witten and Frank, 2005). The classification of the new unknown sample is done according to the region of the hyperplane where the sample fits.

3.3.2 K-Nearest Neighbors

KNN is a simple algorithm, which is based on a distance approach. This type of algorithm assumes that observations with similar characteristics exist in close proximity and will tend to have similar outcomes, meaning that the value of a specific data point is determined by the values of the data points around it (Kramer, 2013).

There are several ways to calculate the mentioned distance between examples but the ones used in this work are the Euclidean and the Manhattan distances (Black, 2004; Black, 2019).

3.3.3 Random Forest

As the name implies, the basic principle of a random forest algorithm is having a large set of different individual decision trees, each one performing the classification of a new sample and, in the end, the most common prediction becomes the random forest's classification. The key for the algorithm to work well and reach good results is the low correlation between different decision trees, making ensemble predictions more accurate than each individual tree's prediction (Liu et al., 2012).

3.3.4 Grid Search

All the machine learning algorithms have associated to it some hyperparameters that must be defined and should be tuned to improve the models' performance (Feurer and Hutter, 2019).

The chosen method to tune the three different algorithms' hyperparameters was grid search. This method works based on the Cartesian product of values from a finite set of values defined by the user for each parameter (Feurer and Hutter, 2019).

To perform grid search, the scikit-learn library was used, and in Table 7 are presented the available values for the hyperparameters that were tuned per algorithm (Pedregosa et al., 2011).

3.4 Model Evaluation

To evaluate the performance of a specific model, it has to go through the training process to, later, be tested and evaluated before being applied to real situations. Before the training and testing processes, the labeled dataset must be randomly divided into different groups of samples - train-test split -, giving rise to the train and test subsets (Tsukerman, 2019). It is important to mention that the data was split in a stratified way so that, in the end, the test population could be distributed over the different classes.

The model's performance is evaluated by comparing the predicted classes of the samples of the test set to their real classes. Since we are dealing with classification problems that take into account a maximum of 3 different classes, while classifying an unknown sample, there is always a positive class and the remaining ones are considered the negative classes. Thus the possible outcomes are True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

In this study, the evaluation was done through accuracy (Equation 2), that returns the ratio between the number of correctly predicted classes and the number of all predicted classes within the test set (Nighania, 2018):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

4 RESULTS

From the data cleaning process, a total of 19 samples were removed from the dataset.

After that, to remove periods of silence from the audio signals, we used the algorithm explained in the section 3.1. Making a balance between keeping all the information of the less noisy signals and cutting the maximum amount of silence periods of the noisier signals, the value found for constant α in Equation 1 was 0.034. Using this value, the whole list of audio files passed through this algorithm.

Table 7: Possible values for each parameter when performing grid search.

	Parameters	Values
SVM	C	0.001, 0.01, 0.025, 0.05, 0.1, 0.5, 1.0, 10.0
	γ	0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 10.0
	kernel	linear, radial basis function
KNN	n_neighbors	1, 3, 5, 7, 9
	weights	uniform, distance
	metric	euclidean, manhattan
Random Forest	max_depth	1, 2, 3, 4, 5
	max_features	1, 2, 3, 4, 5, 6
	n_estimators	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Given that the size of the dataset is not big and that some of the audios are from the same person, the results of the models vary according to the random state defined on the train-test split. Consequently, it was decided to perform the same experiments for 5 different random states so that, in the end, the evaluation could be done taking into account different training sets.

The results were obtained using a test size of 40%. The experiments were conducted for those 5 different random states, randomly chosen, and performing grid-search for each algorithm and random state in order to get the best accuracy.

From the accuracy results of each random state applied to each algorithm, a statistical analysis was done for the three models: gender (Table 8). age group (Table 9), and emotion recognition (Table 10). In this analysis, and for each algorithm, it is included the minimum accuracy, the maximum, the mean of the accuracy values, and the standard deviation for all those 5 random states.

5 DISCUSSION

Firstly, it is important to note that the fact that accuracy values change depending on the random state has to do with the fact that, within all the samples in the dataset, there are only 88 different people. What happens when doing the train-test split is that the train set can contain all samples from the same person and the test set is from different people, resulting in worse results, or instead it can happen to have samples from the same person spread across the train and the test set and therefore the results are better. To overcome this limitation, the models were run for different random states to be considered, in the end, the statistics of the results of each model.

From the results obtained, and considering the maximum value within the mean of accuracies for each model, it is possible to conclude that the emotion valence recognition model reached the best accuracies and the worst accuracies come from the age recognition model. From among the 3 different models (gender, age group and emotion valence recognition), the algorithm that produces the best result is unanimous and this algorithm is the SVM. Regarding gender recognition model, the best accuracy was 87.8%. For the age group recognition model, it was reached an accuracy of 67.6%. In the emotion valence recognition model, was achieved an accuracy of 94.6%.

Taking into account the literature review, it was expected that the three final models would produce different results. Once female and male voices are quite distinct, it was expected that gender recognition model would produce the best results. On the other hand, it was expected that both age group and emotion recognition models would produce worse results when comparing to the first one.

Contrary to what was expected, the model that produced the best results among the three models was the emotion valence recognition model. This may have to do with the sampling size, which is higher than the size in the other models (268 samples for emotion valence model, 164 samples for gender and 210 samples for age), after the oversampling process. With more samples, the learning process was probably improved and achieved better results.

In accordance with the literature review, the gender recognition model produced better results than the age model. With regard to age group recognition, and taking into account that 3 different categories were considered, it becomes much more complicated to classify the samples when there are age groups in which the voice remains practically unchanged. For this reason, it is more complicated to classify age compared to other characteristics.

The differences between our results and the ones shown in 1.1 could arise from the difference between the used datasets - our dataset consists of few samples acquired under uncontrolled conditions when comparing to the ones used in the literature.

With regard to the choice of algorithms, we can

Table 8: Statistical analysis of the accuracies values obtained for gender recognition.

Algorithm	Min of accuracy (%)	Max of accuracy (%)	Mean of accuracies (%)	Standard Deviation (%)
SVM	83.3	91.7	87.8	3.8
KNN	69.4	76.4	72.8	3.1
Random Forest	63.8	86.1	76.7	7.5

Table 9: Statistical analysis of the accuracies values obtained for age group recognition.

Algorithm	Min of accuracy (%)	Max of accuracy (%)	Mean of accuracies (%)	Standard Deviation (%)
SVM	60.7	77.4	67.6	6.0
KNN	60.7	70.2	66.0	3.2
Random Forest	57.1	72.6	64.8	4.9

Table 10: Statistical analysis of the accuracies values obtained for emotion valence recognition.

Algorithm	Min of accuracy (%)	Max of accuracy (%)	Mean of accuracies (%)	Standard Deviation (%)
SVM	91.7	98.1	94.6	2.3
KNN	76.9	85.2	80.4	3.3
Random Forest	82.4	90.7	86.1	3.2

conclude that the chosen ones produce promising results for these models.

The results were good, but the study still has some limitations.

6 CONCLUSIONS

As a first approach, the work developed in this study allowed to conclude that is possible to extract some personal characteristics and also the emotional state from Portuguese voice signals. Nevertheless, there are still some limitations that can be overcome in further studies.

In the end, an accuracy of 87.8% was achieved in gender recognition, using a SVM classifier; also SVM produced the best accuracy of 67.6% for age group recognition; and 94.6% was the best reached accuracy for emotion recognition when using the same algorithm. These are very promising results, considering that the dataset in which the experiments were made have a quite small sample size, with short sentences.

Some improvements should be done in order to achieve better results. First of all, more samples should be acquired in order to have a larger dataset and, therefore, oversampling methods would not be necessary. Moreover, for gender and age recognition models, each participant should have just one sample to avoid repetitive voices that could affect the results. Each participant should have just one sample per question for the emotion recognition model, as long as different questions could arise to different emotional states. In addition, it should be developed an algorithm to automate the pre-selection process of samples.

Another improvement that should be performed is related to the removal of noise from the acquired audio samples. Assuming that the data acquisition method remains the same, the conditions of the data acquisition are not controlled and the problem related to background noise arises: it becomes impossible to guarantee that the person participates in the acquisition in a quiet environment and sometimes there is a lot of background noise. The biggest problem is that there are various types of background noise from sample to sample, which makes the removal of these background noises a complex task.

As future work, it would be interesting to test if results improve with the aggregation by gender when training and testing both the age group recognition model and the emotion recognition model.

REFERENCES

Aizat, K., Mohamed, O., Orken, M., Ainur, A., Zhumazhanov, B., and Pham, D. (2020). Identification and authentication of user voice using DNN features and i -vector. *Cogent Engineering*, 7(1):1751557.

Barandas, M., Folgado, D., Fernandes, L., Santos, S.,

Abreu, M., Bota, P., Liu, H., Schultz, T., and Gamboa, H. (2020). Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456.

Black, P. E. (2004). Euclidean distance. https://www.nist.gov/dads/HTML/euclidndstnc.html. Accessed: 2021-08-28.

Black, P. E. (2019). Manhattan distance. https://www.nist.gov/dads/HTML/manhattanDistance.html. Accessed: 2021-09-07.

Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., and Santabárbara, J. (2021). Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, 21(1):100196.

Chaudhary, S. and Sharma, D. K. (2018). Gender Identification based on Voice Signal Characteristics. *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, pages 869–874.

Costantini, G., Parada-Cabaleiro, E., and Casali, D. (2021). Automatic Emotion Recognition from DEMoS Corpus by Machine Learning Analysis of Selected Vocal Features. In *14th International Joint Conference on Biomedical Engineering Systems and Technoloies*.

Dash (2021). Dash for Python Documentation — Plotly. https://dash.plotly.com/introduction. Accessed: 2021-02-09.

Faek, F. (2015). Objective Gender and Age Recognition from Speech Sentences. *Aro, The Scientific Journal of Koya University*, 3(2):24–29.

Feurer, M. and Hutter, F. (2019). *Hyperparameter Optimization*, pages 3–33. Springer International Publishing, Cham.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 1322–1328.

JotForm (2021). Free online form builder & form creator — jotform. https://www.jotform.com/. Accessed: 2021-08-17.

Júnior, J. D. R. (2017). *Reconhecimento automático de emoções através da voz*. PhD thesis, Universidade Federal de Santa Catarina.

Kramer, O. (2013). K-Nearest Neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*, volume 51, pages 13–23. Springer, Berlin, Heidelberg.

Kraus, M. W. (2017). Supplemental Material for Voice-Only Communication Enhances Empathic Accuracy. *American Psychologist*, 72(7):644–654.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Liu, Y., Wang, Y., and Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. *LNCS*, 7473:246–252.

Lyons, J. (2013). Welcome to python_speech_features's documentation! — python_speech_features 0.1.0 documentation. https://python-speech-features.readthedocs.io/en/latest/#. Accessed: 2021-09-08.

Mamyrbayev, O., Mekebayev, N., Turdalyuly, M., Oshanova, N., Ihsan Medeni, T., and Yessentay, A. (2020). Voice Identification Using Classification Algorithms. In *Intelligent System and Computing*, chapter Voice Iden, pages 1–13. IntechOpen.

Nighania, K. (2018). Various ways to evaluate a machine learning model's performance.

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

Salari, N., Hosseinian-Far, A., Jalali, R., Vaisi-Raygani, A., Rasoulpoor, S., Mohammadi, M., Rasoulpoor, S., and Khaledi-Paveh, B. (2020). Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. *Globalization and Health*, 16(1):57.

Satterfield, J. M. (2017). The Iceberg—Visible and Hidden Identity. https://www.thegreatcoursesdaily.com/visible-and-hidden-identity/. Accessed: 2021-08-24.

Tang, F., Liang, J., Zhang, H., Kelifa, M. M., He, Q., and Wang, P. (2021). COVID-19 related depression and anxiety among quarantined respondents. *Psychology & Health*, 36(2):164–178.

Tsukerman, E. (2019). *Machine Learning for Cybersecurity Cookbook*. Packt.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, second edition.

Zhao, H. and Wang, P. (2019). A short review of age and gender recognition based on speech. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 183–185.