

Advanced Analysis and Model Development on Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

21 CST H3Art



Angry Speech



Angry Song



Neutral Speech



Neutral Song



Fearful Speech



Fearful Song



Outline



- **Introduction**
- **Data Preprocessing**
- **Imbalanced Learning**
- **Classification**
- **Conclusion**

Introduction


RAVDESS Emotional speech audio

Emotional speech dataset

Data Card Code (342) Discussion (2) Suggestions (0)

About Dataset

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS. Full dataset of speech and song, audio and video (24.8 GB) available from [Zenodo](#). Construction and perceptual validation of the RAVDESS is described in our Open Access [paper in PLoS ONE](#). 

Check out our [Kaggle Song emotion dataset](#).

Files

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

The Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS) is a validated **multimodal dataset** developed by Livingston and Russo (2018) that contains audiovisual recordings of **24 professional actors** uttering **two lexically matched sentences** neutrally in a North American accent.

Available in [RAVDESS Emotional speech audio \(kaggle.com\)](https://www.kaggle.com/datasets/ryanbradford/ravdess-emotional-speech-audio)

Introduction: Data Source

The first part of the study used a modified version of the original RAVDESS dataset, where **numerical attributes** (The Second Table is at next page) were generated based on the **original categorical attributes** (The First Table is as follows) by **extracting quantitative statistics of the original audio signal**.

Name	Type	Description
modality	Nominal	Recording mode
vocal_channel	Nominal	Type of vocal communication
emotion	Nominal	Emotion expressed
emotional_intensity	Ordinal	Degree of emotional involvement
statement	Nominal	Statement uttered
repetition	Ordinal	Repetition of the statement
actor	Nominal	Actor's ID
sex	Nominal	Actor's sex
filename	Nominal	Record's ID

Introduction: Data Source

Name(s)	Type	Description
frame_count	Interval	Number of frames per sample
mean, std, min, max, skew, kur, q_01, q_05, q_25, q_50, q_75, q_95, q_99	Ratio	Statistics of original audio signal
lag1_sum, lag1_mean, lag1_std, lag1_min, lag1_max, lag1_kur, lag1_skew, lag1_q01, lag1_q05, lag1_q25, lag1_q50, lag1_q75, lag1_q95, lag1_q99	Ratio	Statistics of Lag (difference between each observation and the antecedent)
zc_sum, zc_mean, zc_std, zc_min, zc_max, zc_kur, zc_skew, zc_q01, zc_q05, zc_q25, zc_q50, zc_q75, zc_q95, zc_q99	Ratio	Statistics of Zero Crossing Rate
mfcc_sum, mfcc_mean, mfcc_std, mfcc_min, mfcc_max, mfcc_q01, mfcc_q05, mfcc_q25, lag1_q50, mfcc_q75, mfcc_q95, mfcc_q99, mfcc_kur	Ratio	Statistics of Mel-Frequency Cepstral Coefficients
sc_sum, sc_mean, sc_std, sc_min, sc_max, sc_kur, sc_skew, c_q01, sc_q05, sc_q25, sc_q50, sc_q75, sc_q95, sc_q99	Ratio	Statistics of Spectral Centroid
stft_sum, stft_mean, stft_std, stft_min, stft_max, stft_kur, stft_skew, stft_q01, stft_q05, stft_q25, stft_q50, stft_q75, stft_q95, stft_q99	Ratio	Statistics of Short-Time Fourier Transform

Introduction: Data Source

Further attributes have been created by **dividing each time series into 4 non overlapping windows** and computing all the quantitative statistics described in **previous table** at a local level. The names referring to such features can be easily derived from the expression:

NAME_wN

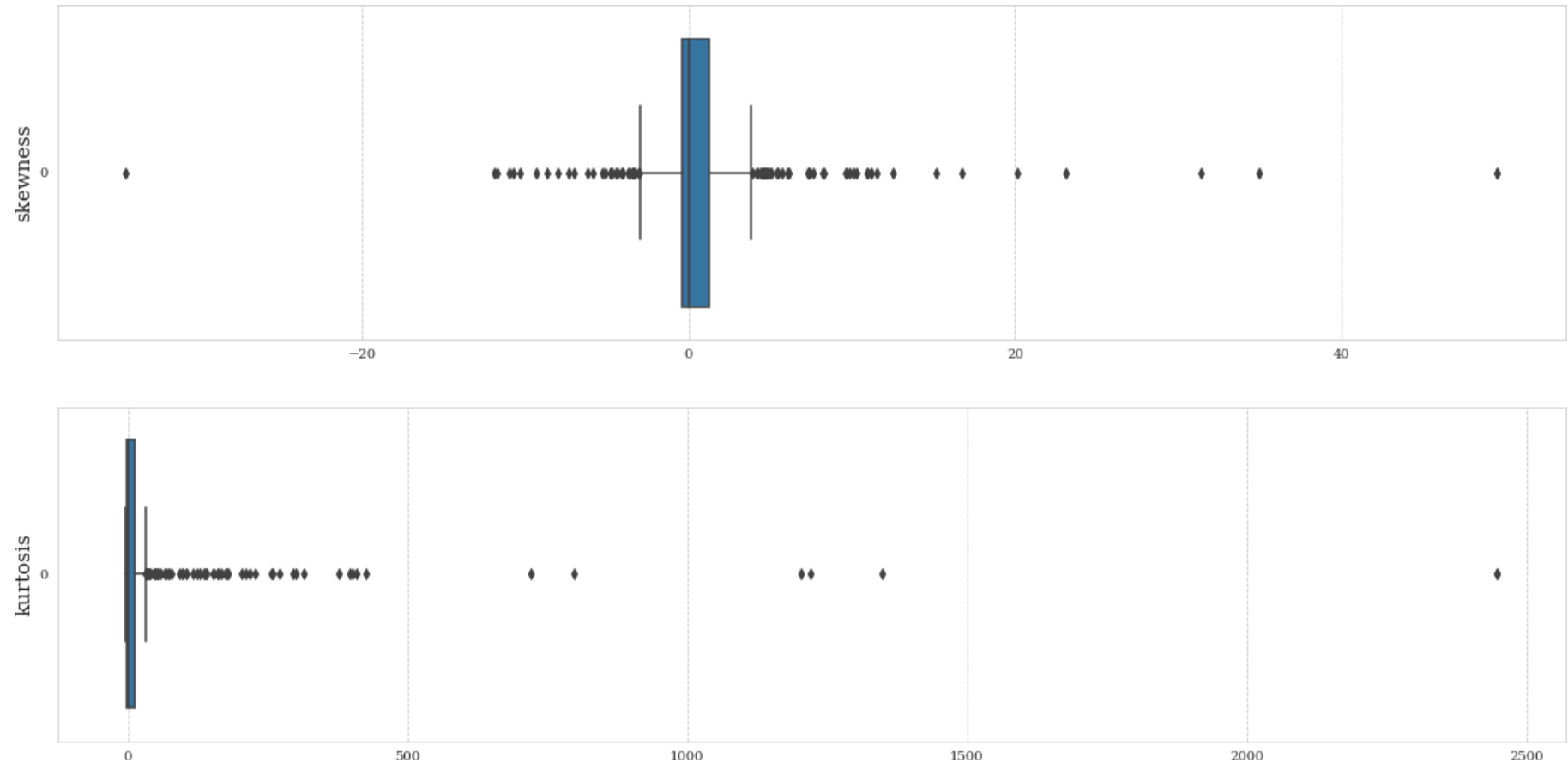
and by replacing ***NAME*** with the name of the numerical attribute and N with **the index (1, 2, 3 or 4) of the window considered**. The only exception is represented by the global-level feature ***frame_count***, which is replaced locally with the denomination ***length_wN***.

Data Preprocessing: Data Understanding

1. Data Partitioning

2. Distribution Analysis

3. Data Imbalance



- Data was divided into **training (TR)** and **test (TS)** sets, with **34.1% reserved for testing**.
- Numerical attributes were analyzed for **skewness** and **kurtosis**, revealing that only 4.5% were close to a normal distribution and 8.5% were sufficiently symmetric.
- Categorical attributes showed **slight imbalances** in **vocal channel** (+18.2% for speech) and **emotional intensity** (+8% for normal emotion), with **significant imbalances** in **emotions** (8% for neutral, disgust, surprise).

Data Preprocessing: Simple Preparation

1. Initial Observations

2. Data Cleaning

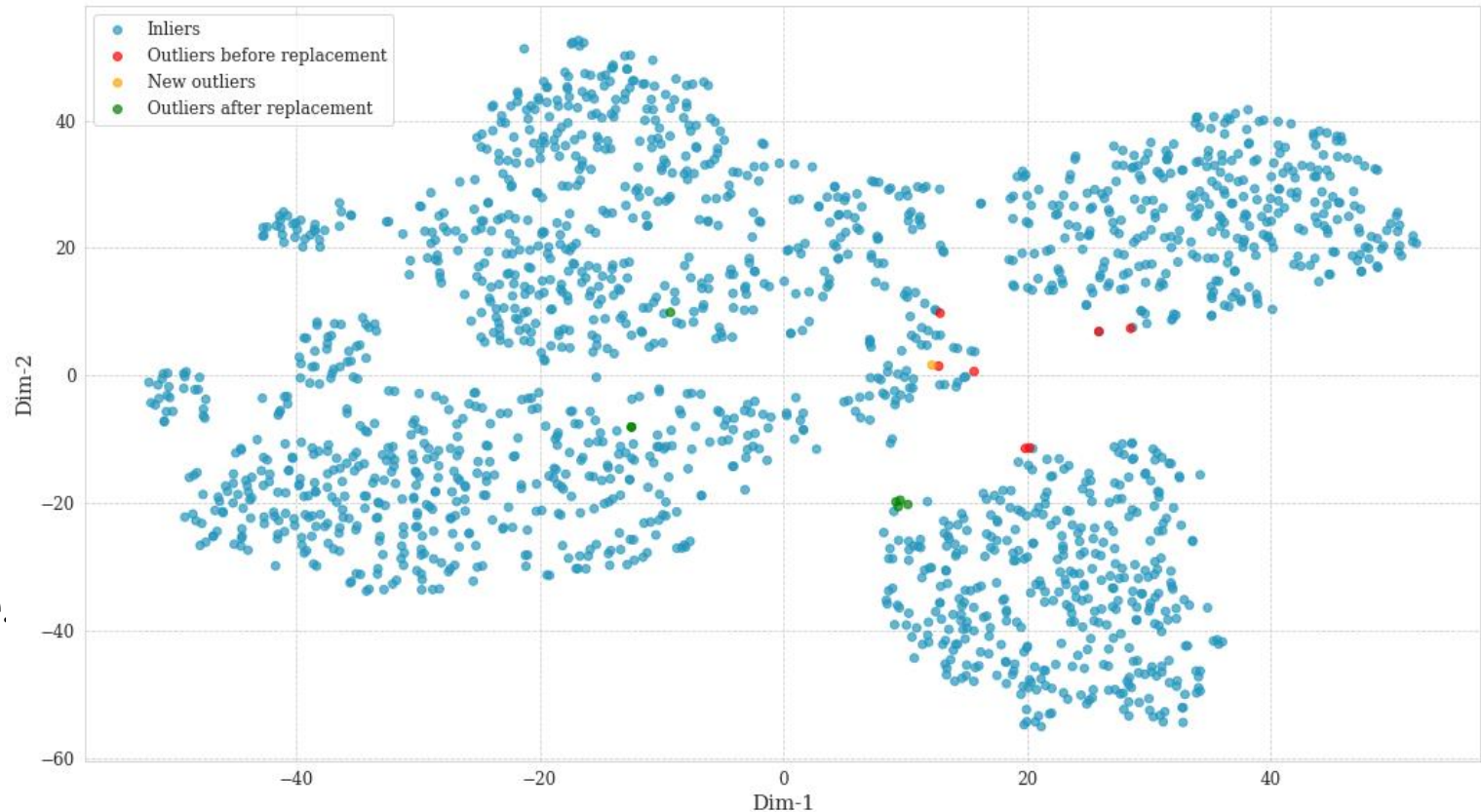
3. Simple Preprocessing

- No inconsistencies, missing values, or duplicates were found.
- Continuous attributes with **zero variance** and categorical attributes with **unique values** were **removed**, reducing the **dimensionality from 434 to 383**.
- **Standardization, one-hot encoding, and feature reduction** were applied as necessary.

Data Preprocessing: Anomaly Detection

I used several unsupervised **anomaly detection methods**, such as Histogram-Based Outlier Score (HBOS), Deviation-Based (DB), K-Nearest Neighbors (KNN), Local Outlier Factor (LOF) and Isolation Forest (IF). Preprocessing steps included standardization, one-hot encoding, and PCA for feature reduction. The **top 1% of data points with the highest anomaly scores** were identified as **outliers**.

Then, anomalous data points were managed by **replacing their values with the median of each continuous attribute**. After replacement, the data was re-evaluated to ensure that the new top 1% anomalies did not overlap with the previous set, validating the effectiveness of the anomaly management.



Dimensionality reduction has been performed with t-SNE

Research Topics



Imbalanced Learning:

- Many real-world datasets suffer from class imbalance, where certain classes are significantly underrepresented.
- The section will cover techniques for handling imbalanced data and their effects on classification tasks.

Classification:

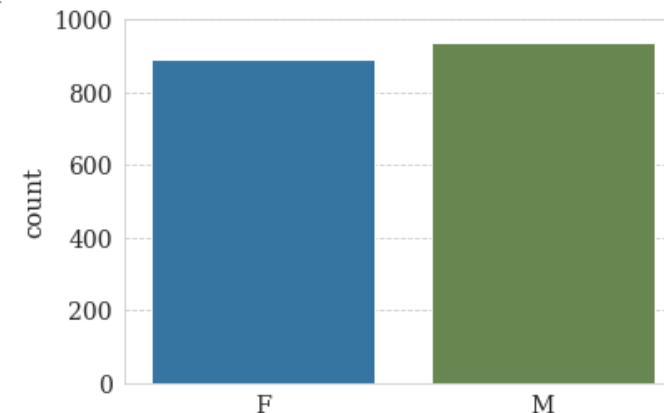
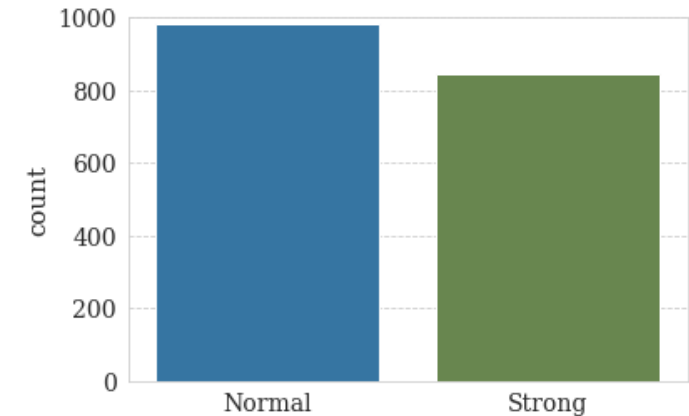
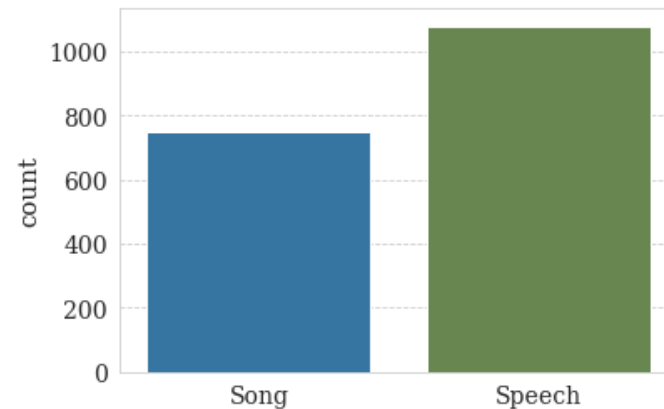
- Involves classification tasks for vocal channel, sex, emotional intensity, and emotion.
- Various classifiers like Logistic Regression, SVM, Neural Networks, etc., will be introduced and their performances compared.

Imbalanced Learning

In many real-world datasets, class imbalance is a common issue, where certain classes have significantly more samples than others. This imbalance causes classifiers to be biased towards the majority class, thereby affecting the recognition of minority classes.

This section aims to test the performance of three different techniques for handling a user-created imbalanced setting: **random undersampling**, **Syntetic Minority Oversampling Technique (SMOTE)** and **class weight adjustment**. I measure the performance of each technique by looking at the improvements in the **F1-scores** of two simple classifiers: **Decision Tree (DT)** and **K-Nearest Neighbors (K-NN)**³, before and after the re-balancing of data. The analysis of the performance of each classifier has been restricted to the binary classification of *vocal_channel*, *sex* and *emotional_intensity*.

The distributions of the target attributes before the imbalancing are reported as follows (From left to right: distributions of *vocal_channel*, *sex* and *emotional_intensity*):



Imbalanced Learning: Hyperparameters

Model selection has relied on a randomized search with repeated stratified 5-fold CV. The corresponding hyperparameters are as follows:

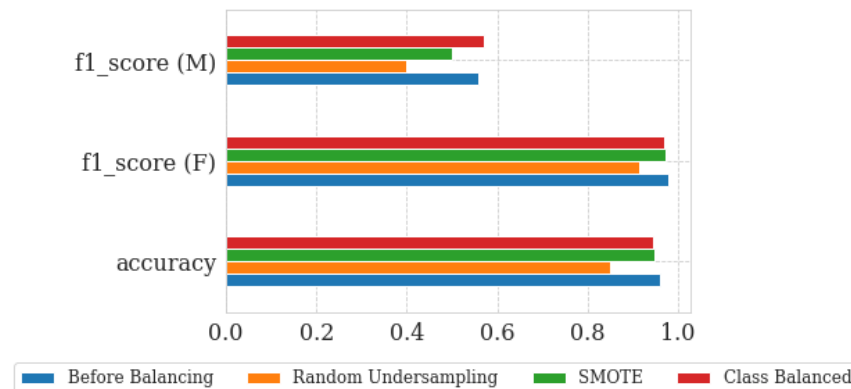
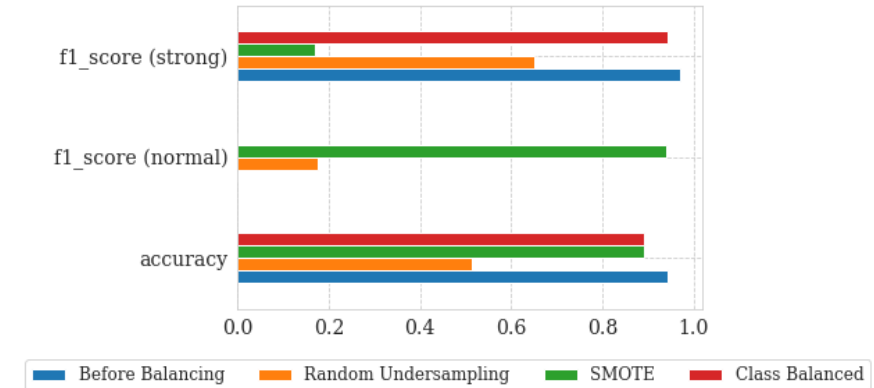
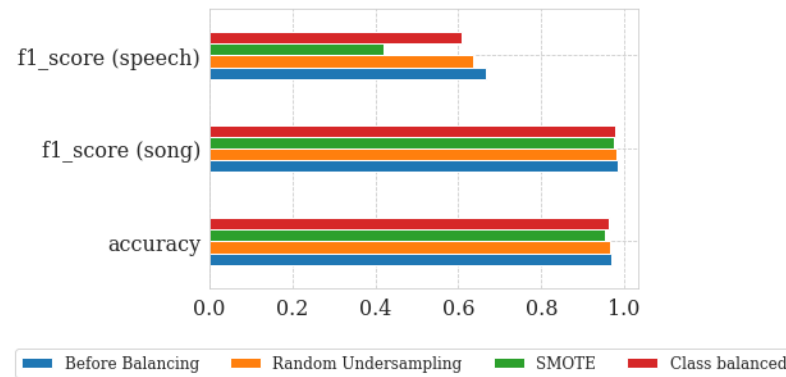
Decision Tree Tested Hyperparameters	Description	Tested Values
Criterion	Metric to choose the best split	Gini, Entropy, Log-Loss
Max Depth	Maximum depth of the tree	Discrete interval [2, 200]
Min Samples Split	Minimum number of samples for split	Log-uniform distribution between [0.01, 1]
Min Samples Leaf	Minimum number of samples in leaf	Uniform distribution between [0.001, 0.2]

K-NN Tested Hyperparameters	Description	Tested Values
K	Number of neighbors	Discrete interval [2, N/2]
Weights	Weight function in prediction	Uniform, Distance
Metric	Distance metric	City-Block, Euclidean, Cosine, Chebyshev

Imbalanced Learning: DT Results

Model evaluation has been performed with a simple hold-out on the previously separated TS data. Since random undersampling involves discarding potentially valuable data from the majority class, the performance of each classifier may be affected by the randomness of the sampling process. For this reason I have performed random undersampling for 10 different iterations and considered as performance indicator of DT and K-NN the mean of the respective F1-scores.

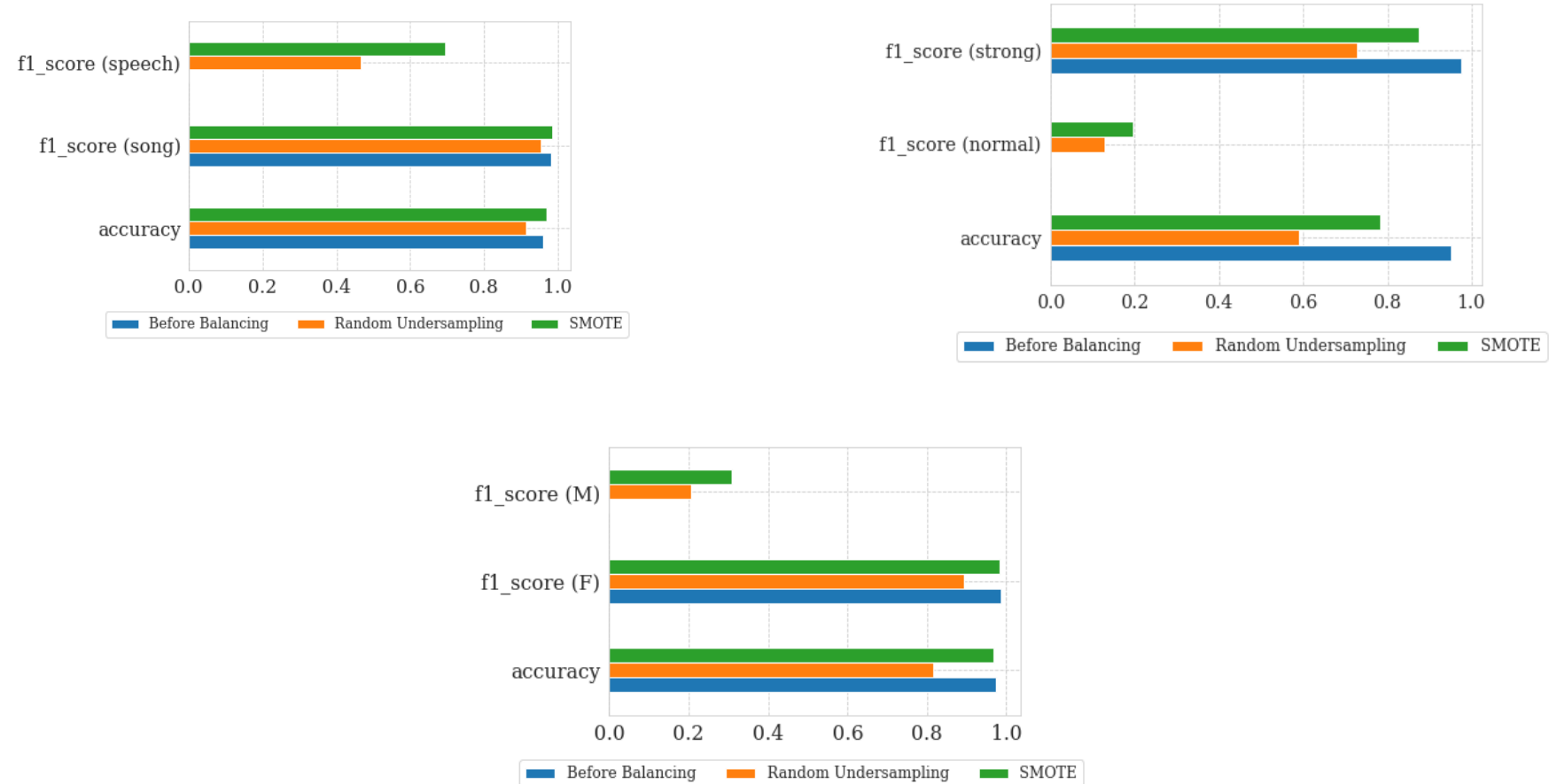
For Decision Trees, balancing methods seem to have a negligible effect on the classification of ***vocal_channel*** and ***sex***, but they play a significant role in the classification of ***emotional_intensity***, where they allow for an increase in the model capability in detecting the minority class normal (whose F1-score is null in the unbalanced setting), even if at the cost of a decrease in the F1-score of the majority class strong.



Imbalanced Learning: KNN Results

Model evaluation has been performed with a simple hold-out on the previously separated TS data. Since random undersampling involves discarding potentially valuable data from the majority class, the performance of each classifier may be affected by the randomness of the sampling process. For this reason I have performed random undersampling for 10 different iterations and considered as performance indicator of DT and K-NN the mean of the respective F1-scores.

As for K-NN, the F1-score of minority classes (speech, M, normal) in each imbalanced learning is null. In this case, both SMOTE and random undersampling seem to have a major role in improving the capability of the model in recognizing the minority class. Overall SMOTE seems to perform better than random undersampling, probably because of the relatively small size of the data; and also better than class weight adjustment, which fails in improving the Decision Tree capability in recognizing the minority class instances of ***emotional_intensity***.



Classification: Overview

The objective is to classify *vocal_channel*, *sex*, *emotional_intensity* (**binary classification**), and *emotion* (**multi-class classification**) using various classifiers. For each model and task, pre-processing operation have concerned, in the order: the **standardization** of numerical attributes with **min-max scaler**; the **one-hot encoding** of categorical attributes (target excluded); and the **label encoding** of the target attribute. Final model evaluation is performed with a hold-out on the already provided TS data.

Classifiers	
Logistic Regression (LR)	Support Vector Machines (SVM)
Neural Networks (NN)	Decision Tree Bagging (DTB)
Random Forests (RF)	AdaBoost (AB)
Gradient Boosting (GB)	

Classification: Logistic Regression

Model selection for LG has consisted of a grid search with 5-fold CV. Tested hyperparameters and correspondent values are as follows. Since some **penalties are not compatible** with some solvers, I set **L2** for all the candidates. Furthermore, I set to **800 the maximum number of iterations** required for the solver to converge.

Hyperparameter	Description	Tested Values
C	Inverse of regularization strength	Log-uniform distribution between $[10^{-4}, 10^3]$
Solver	Algorithm to use in the optimization problem	L-BFGS, LIBLINEAR, Newton-CG, Newton-Cholesky, SAG, SAGA

Here is the best result of LG:

Target	C	Solver	Weighted F1	Accuracy
Vocal Channel	1	L-BFGS	0.98	0.98
Sex	1	L-BFGS	0.85	0.85
Emotional Intensity	1	L-BFGS	0.77	0.77
Emotion	1	Newton-Cholesky	0.43	0.49

Classification: Support Vector Machines

Support Vector Machines with various kernels were tested, with the RBF kernel performing best for most tasks.

Hyperparameter	Description	Tested Values
C	Regularization parameter	Log-uniform distribution between $[10^{-4}, 10^4]$
γ	Kernel coefficient	Log-uniform distribution between $[10^{-4}, 10^4]$
Kernel	Kernel function	Linear, Polynomial, RBF

Here is the best result of SVM:

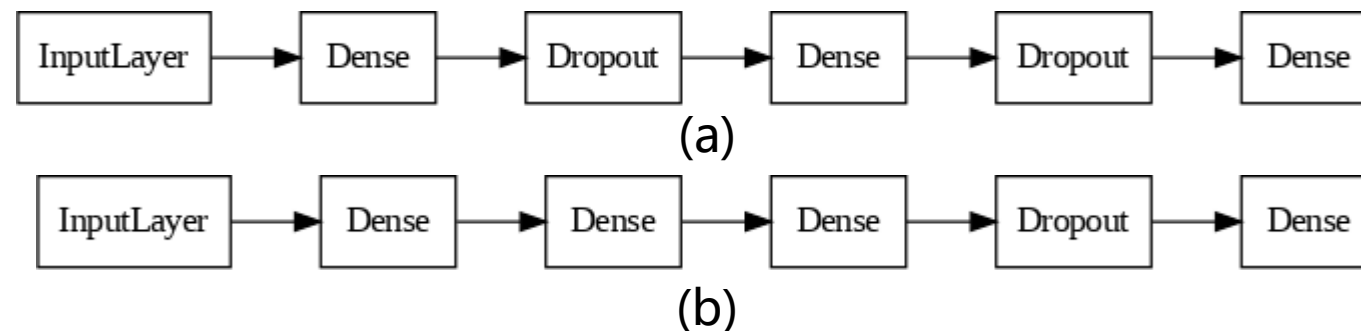
Target	C	γ	Kernel	Weighted F1	Accuracy
Vocal Channel	95.454	0.0006	RBF	0.98	0.98
Sex	0.7	0.1	RBF	0.90	0.90
Emotional Intensity	1072.26	0.0003	RBF	0.76	0.76
Emotion	0.018	0.097	Polynomial	0.47	0.50

Classification: Neural Networks

Neural Networks with different architectures and regularizations showed strong performance, especially for vocal channel and sex classification. I have tested two architectures for binary classification and multi-class classification. All the tested hyperparameters are as follows:

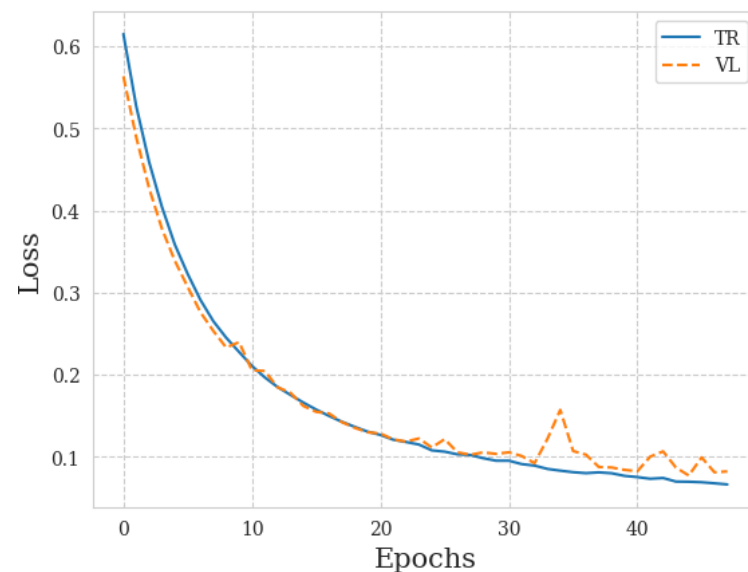
Hyperparameter	Description	Tested Values
Size	Number of units in hidden layers	Powers of 2 within [2, 28]
Epochs	Number of epochs	10, 20, 50, 100, 200
η	Learning rate	Log-uniform distribution between $[10^{-3}, 1]$
p	Dropout rate	0.0, 0.2, 0.4, 0.6

Architectures of the NNs for the binary classification targets (a) and multi-class classification target (b)

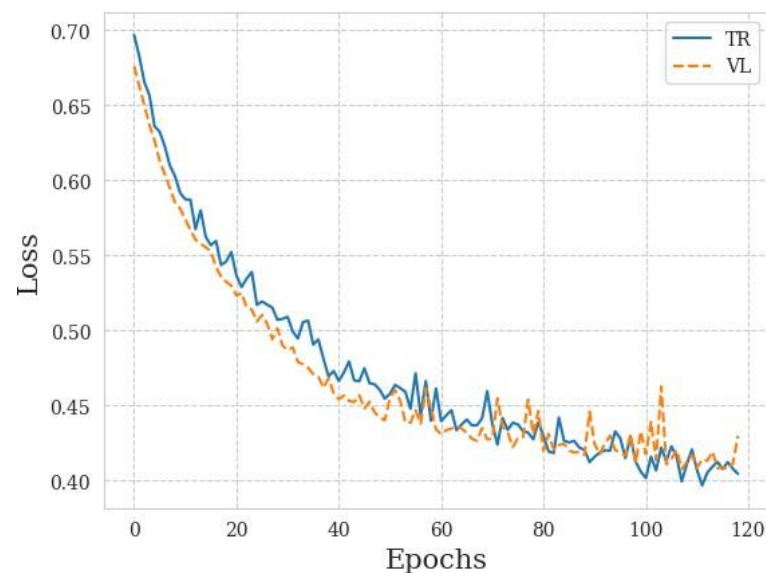


Classification: Neural Networks

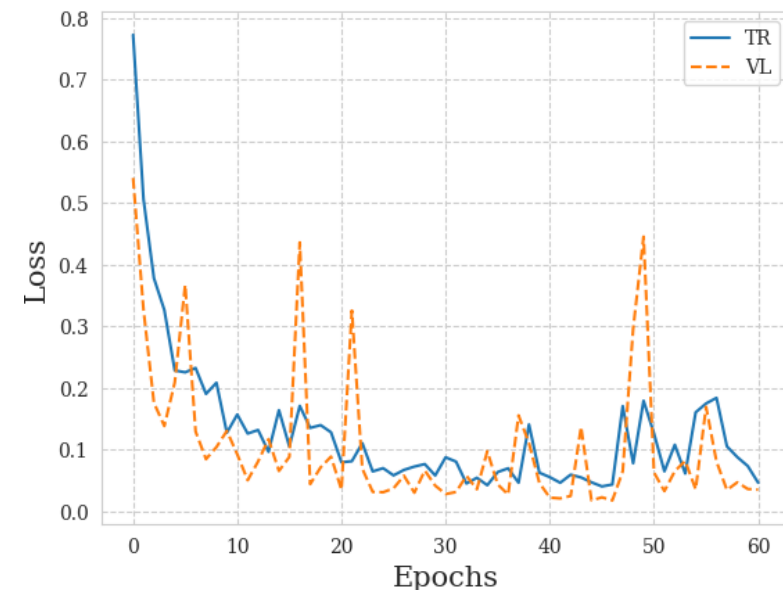
Learning
curves of best
models for
each target



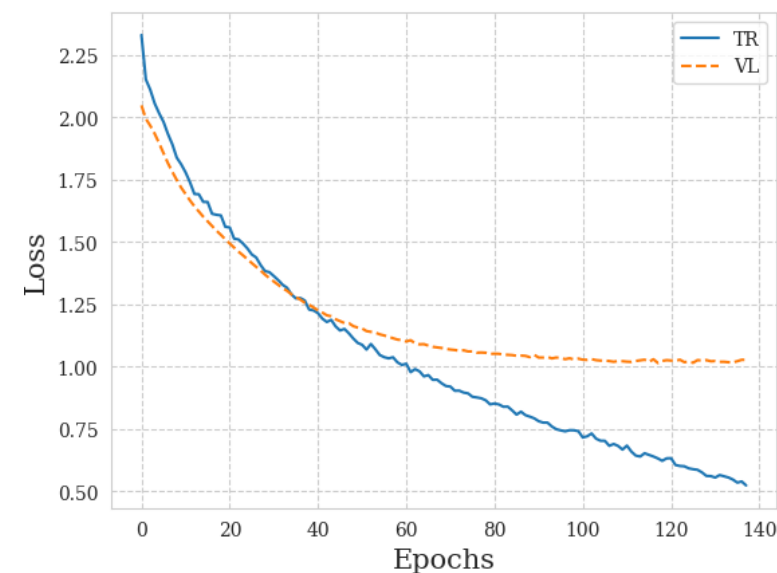
vocal_channel



emotional_intensity



sex



emotion

Classification: Neural Networks

Best results of NN and average and standard Deviation of performance metrics

Target	Size(s)	Epochs	η	p	Weighted F1	Accuracy
Vocal Channel	2	35	0.1	0.0	0.98	0.98
Sex	128	61	1	0.0	0.95	0.95
Emotional Intensity	16, 8	118	0.001	0.2	0.77	0.78
Emotion	256, 256, 256	290	0.0001	0.4	0.52	0.53

Target	Weighted F1	Accuracy
Vocal Channel	0.96± 0.02	0.96± 0.02
Sex	0.95± 0.00	0.95± 0.00
Emotional Intensity	0.77± 0.01	0.77± 0.01
Emotion	0.51± 0.01	0.52± 0.01

Classification: Tree-Based Models and Ensembles

I have analyzed the performance of four ensemble classifiers: **Bagging with Decision Tree** (DTB), **Random Forest** (RF), **AdaBoost** (AB) and **Light Gradient Boosting Machine** (GB). For DTB, RF and AB model selection has been carried out with a randomized search with **3-fold CV**. For both models the hyperparameter space includes the hyperparameters of the **base estimator in Imbalanced Learning Section**. For AB, I assume that the base estimator is a **decision stump**. Tested hyperparameters and

DTB Hyperparameter	Description	Tested Values
Max Samples	Maximum number of samples to train each base estimator	0.5, 0.6, 0.7, 0.8
Max Features	Maximum number of features to train each base estimator	Discrete interval [2, N]
RF Hyperparameter	Description	Tested Values
Max Features	Maximum number of features to choose best split	\sqrt{N} , $\log_2(N)$, N
AB Hyperparameter	Description	Tested Values
Learning Rate	Weight applied to each classifier at each boosting iteration	Log-uniform distribution between $[10^{-4}, 1]$

Best Results of Bagging with Decision Tree

Target	Criteri on	Max Depth	Min Split	Min Leaf	Max Sampl es	Max Featur es	Weigh ted F1	Accur acy
Vocal Chann el	Entro py	58	0.022	0.007	0.7	330	0.95	0.95
Sex	Log- Loss	51	0.010	0.002 8	0.7	78	0.88	0.88
Emoti onal Intens ity	Gini	16	0.013	0.016	0.6	381	0.75	0.75
Emoti on	Entro py	23	0.027	0.001 7	0.7	126	0.39	0.42

Best Results of Random Forest

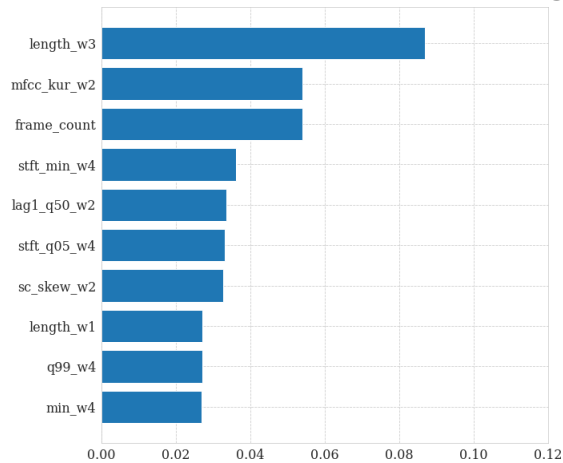
Target	Criterion	Max Depth	Min Split	Min Leaf	Max Features	Weighted F1	Accuracy
Vocal Channel	Gini	75	0.043	0.0044	\sqrt{N}	0.96	0.96
Sex	Entropy	78	0.011	0.005	N	0.85	0.85
Emotional Intensity	Gini	38	0.018	0.0040	N	0.77	0.77
Emotion	Log-Loss	81	0.021	0.010	N	0.41	0.43

Best Results of AdaBoost

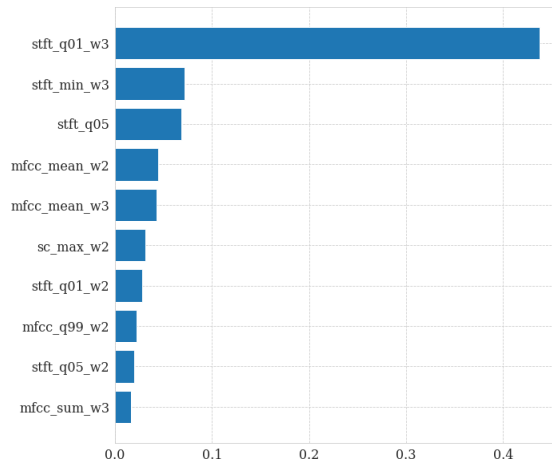
Target	Learning Rate	Weighted F1	Accuracy
Vocal Channel	0.0001	0.89	0.89
Sex	0.0001	0.83	0.83
Emotional Intensity	0.0001	0.64	0.64
Emotion	0.0001	0.29	0.30

Classification: Tree-Based Models and Ensembles

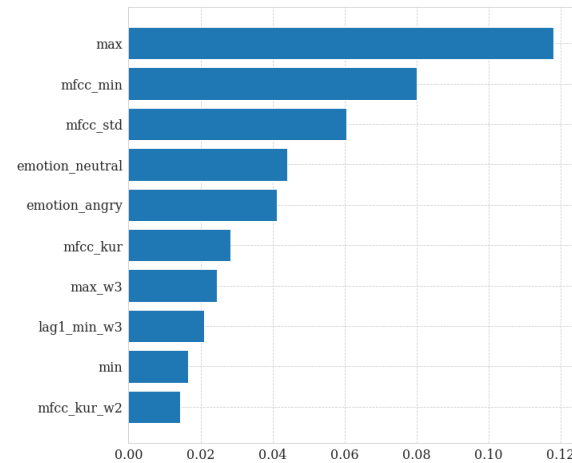
A further analysis has concerned the computation of **the importance of each input feature** in **RF's** predictive performance, which allows to gain a more in-depth understanding of the information used by the model to discriminate between classes of the target variable.



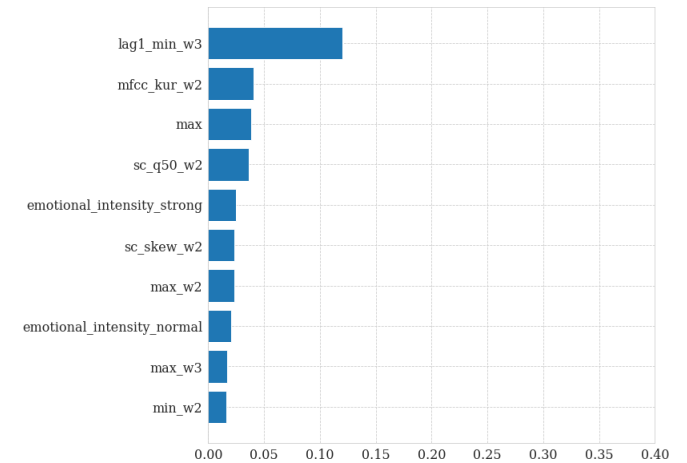
(a) *vocal_channel*



(b) *sex*



(c) *emotional_intensity*



(d) *emotion*

As evidenced, emotion and *emotional_intensity* share a significant number of their top 10 most important features: **max**, **max_w3**, **lag1_min_w3**. Furthermore, some emotions like neutral and angry seem to be significant in the classification of *emotional_intensity*, and both of *emotional_intensity*'s values like strong and normal seem to play a significant role in the recognition of emotions. On the other hand, *vocal_channel* shares only one attribute with *emotional_intensity* and *emotion* (**mfcc_kur_w2**), whereas none of the most important features for the recognition of sex are present in the other targets.

Classification: Tree-Based Models and Ensembles

Model selection for **GB** has consisted of a randomized search with **5-fold CV**. Additionally, a model selection has been performed by exploiting automatic model transformation of categorical attributes in place of one-hot-encoding, but nonetheless the former has yielded a better performance. The hyperparameters and best results are:

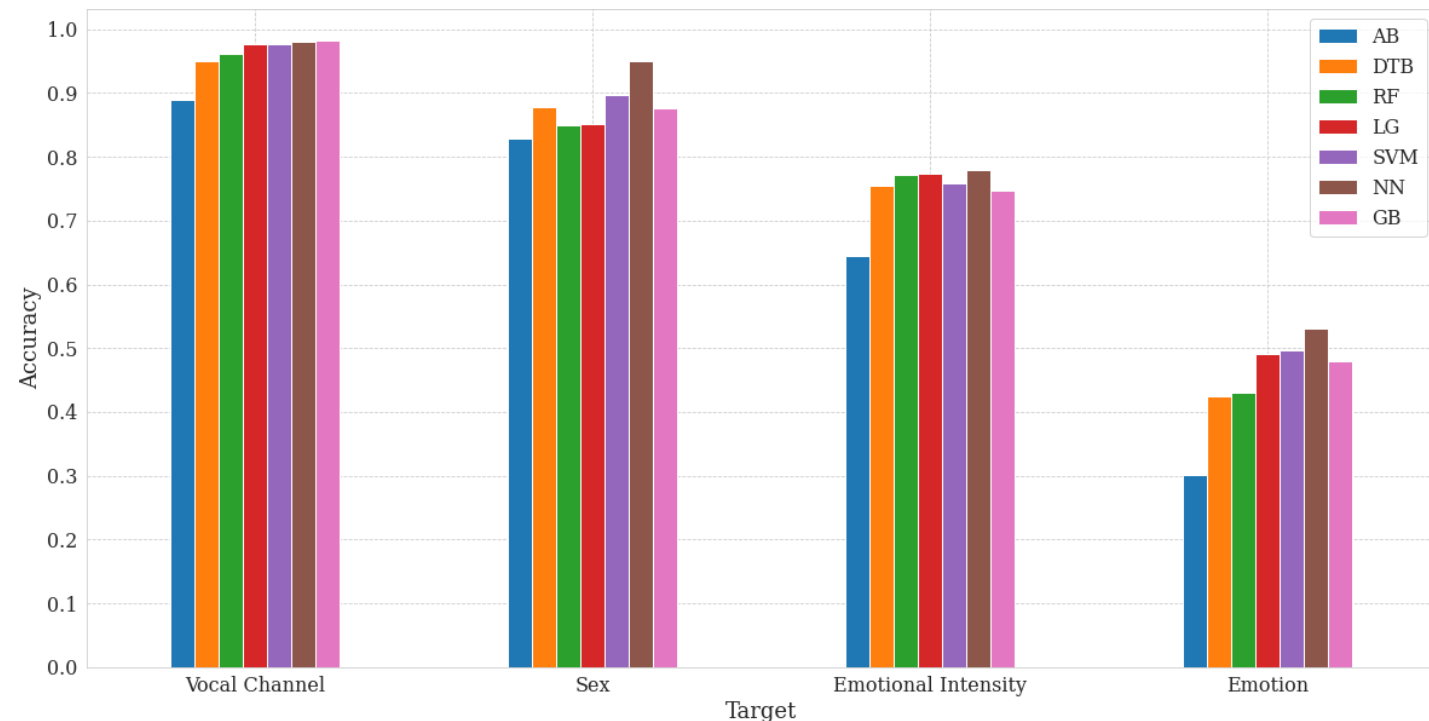
GB Hyperparameter	Description	Tested Values
Boosting type	Gradient boosting algorithm	GBDT, GOSS, DART
Estimators	Number of boosting trees	Uniform distribution between [50, 500]
η	Learning rate	Log-uniform distribution between [10^{-4} , 1]
Leaves	Maximum number of leaves per base learner	Uniform distribution between [5, 50]

Target	Boosting type	Estimators	η	Leaves	Max depth	Weighted F1	Accuracy
Vocal Channel	GOSS	395	0.1	20	191	0.98	0.98
Sex	GOSS	422	0.3	21	45	0.88	0.88
Emotional Intensity	GBDT	467	0.2	16	64	0.74	0.74
Emotion	GOSS	432	0.1	49	35	0.46	0.48

Classification: Results

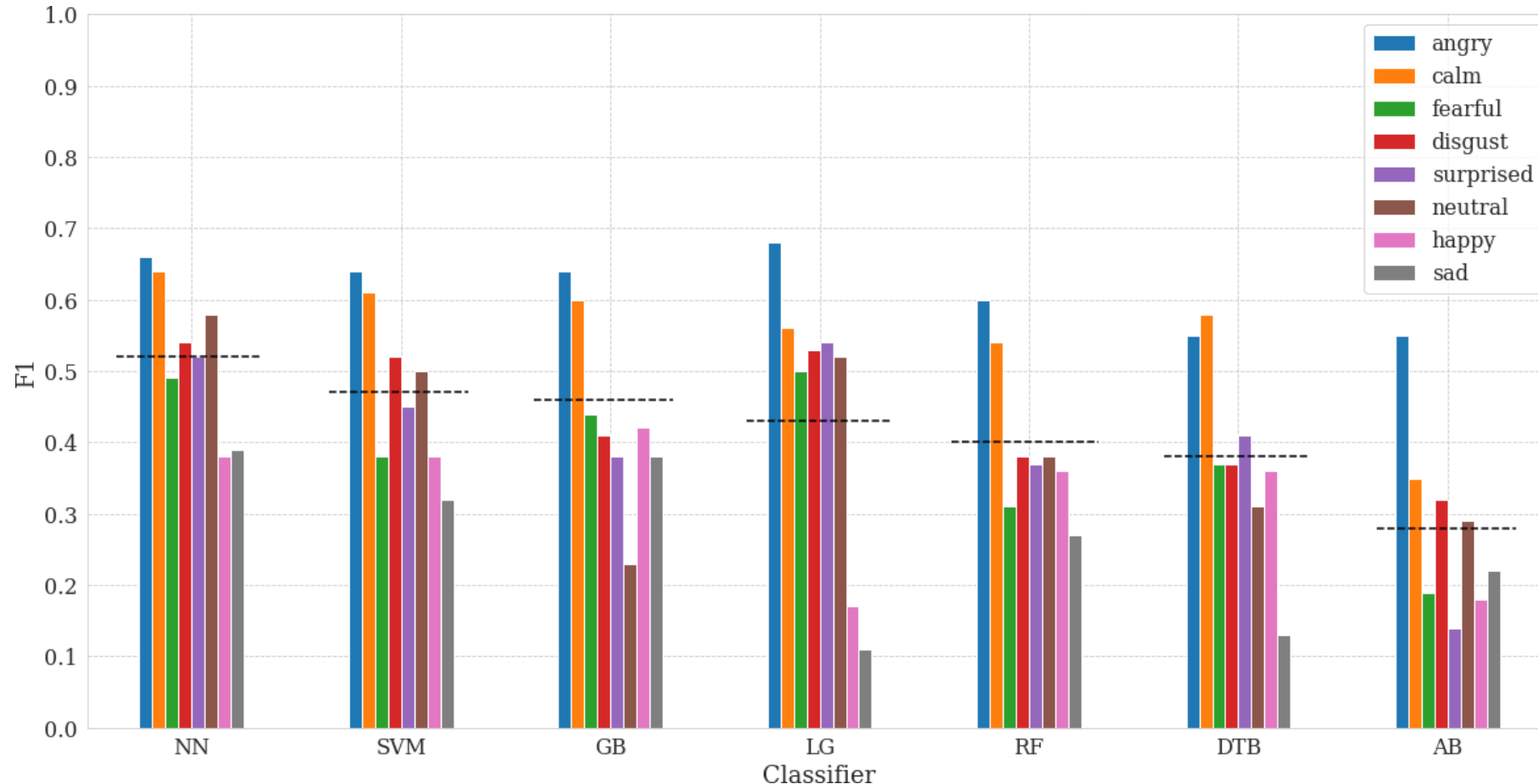
In this concluding section I make a comparative evaluation of the classification models considered for the binary tasks (***vocal_channel***, ***sex***, ***emotional_intensity***) and the multi-class task (emotion). In order to compare the relative performance of each classifier, I consider as a **baseline (alongside LG)** the performance of a random classifier with accuracy and weighted average F1 equal to the inverse of the number of target classes – **0.5 for the binary tasks and 0.125 for the multi-class task**.

All the models reach optimal accuracy results in the classification of ***vocal_channel***, with slight better values for GB and NN. For the classification of ***sex***, most of the models reach competitive results w.r.t. the baseline (LG), but a significantly higher accuracy is reached by NN. Differently, for the classification of ***emotional_intensity*** only NN is able to outperform LG, followed by RF. These results are confirmed by the visual inspection of ROC curves and the correspondent AUC. And the **multi-class problem** is more challenging for all classifiers: even if all of them outperform the random prediction, **only NN provides an accuracy above 0.5**.



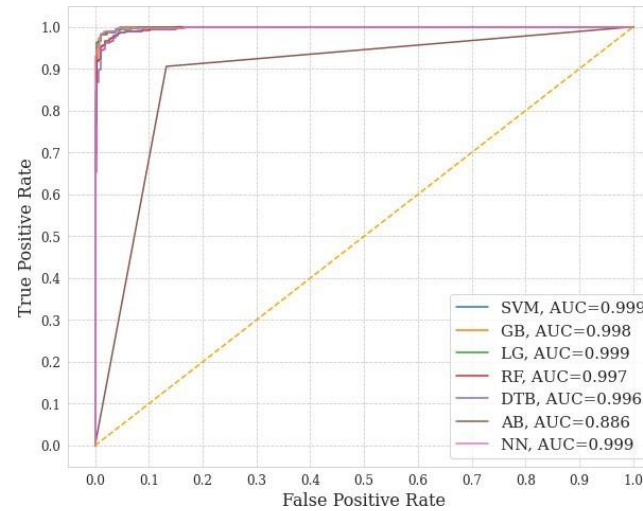
Classification: Results

A further comparison of the relative F1 scores can be useful to better appreciate the capability of each model to discriminate between emotions



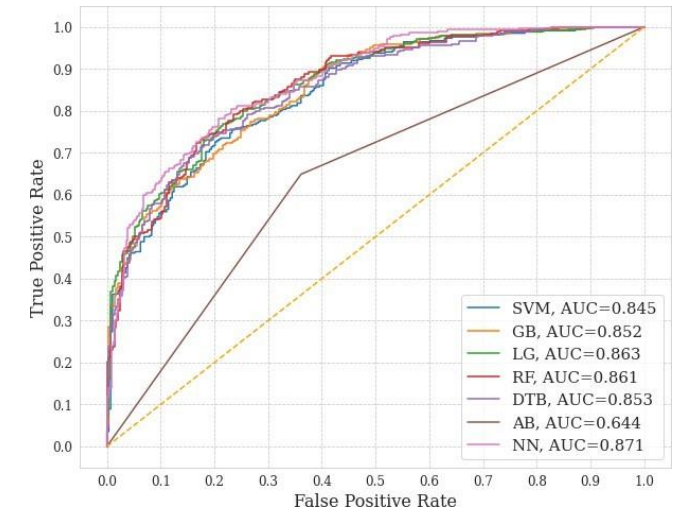
Classification: Results

ROC curves of classification models for each binary target

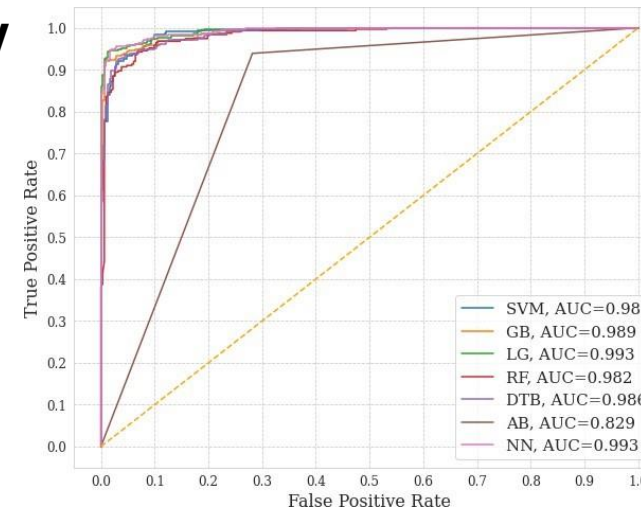


(a) *vocal_channel*

(b) *sex*



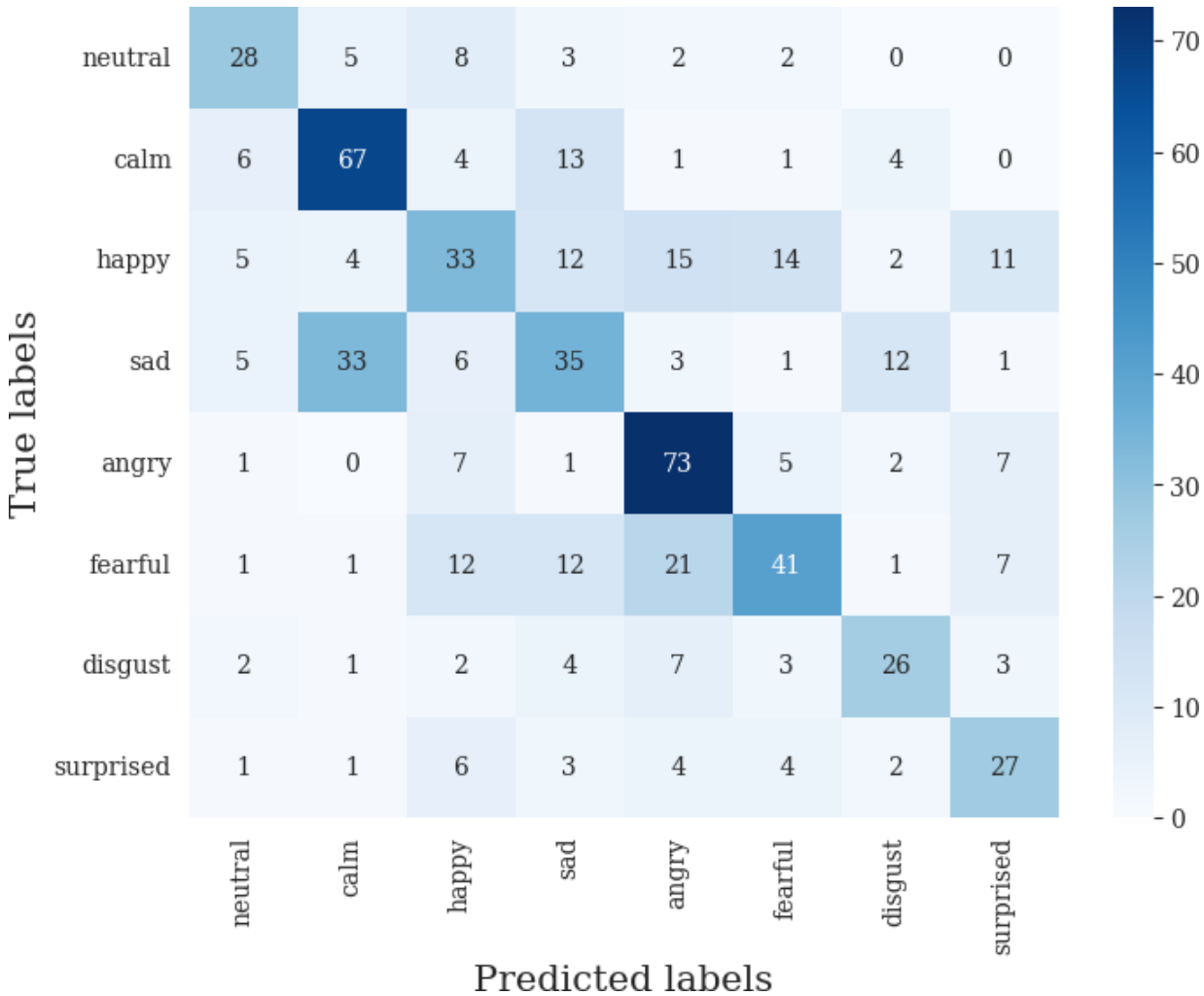
(c) *emotional_intensity*



Classification: Results

a visual inspection of the confusion matrix of the neural classification highlights similar mistakes committed by human classification: sad is frequently confused with calm; fearful is often confused with sad and angry; disgust is mainly confused with happy; and surprised is often confused with happy.

Confusion matrix of NN classification



Conclusion

Summary of Findings:

- **Imbalanced Learning:**
 - Applied techniques like random undersampling, SMOTE, and class weight adjustment to handle class imbalance.
 - Significant improvement in the classification of minority classes, especially in emotional intensity.
- **Classification:**
 - Evaluated classifiers including Logistic Regression, SVM, Neural Networks, Decision Trees, and ensemble methods.
 - Neural Networks and Random Forests demonstrated superior performance across tasks, with notable accuracy in sex and vocal channel classification.

Implications:

- **Imbalanced Learning:**
 - Techniques like SMOTE and class weight adjustment are effective in enhancing model performance on imbalanced datasets.
- **Classification:**
 - Combining multiple classifiers and preprocessing methods can significantly improve accuracy and robustness.



Thank you for listening