sensors

MDPI

Article

A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech

Cem Doğdu [1,2,3,*], Thomas Kessler [1], Dana Schneider [1,2,3,4], Maha Shadaydeh [2,5] and Stefan R. Schweinberger [2,3,6,7,*]

1 Department of Social Psychology, Institute of Psychology, Friedrich Schiller University Jena, Humboldtstraße 26, 07743 Jena, Germany
2 Michael Stifel Center Jena for Data-Driven and Simulation Science, Friedrich Schiller University Jena, 07743 Jena, Germany
3 Social Potential in Autism Research Unit, Friedrich Schiller University Jena, 07743 Jena, Germany
4 DFG Scientific Network "Understanding Others", 10117 Berlin, Germany
5 Computer Vision Group, Department of Mathematics and Computer Science, Friedrich Schiller University Jena, 07743 Jena, Germany
6 Department of General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/Haus 1, 07743 Jena, Germany
7 German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, 07743 Jena, Germany
* Correspondence: cem.dogdu@uni-jena.de or cemdogdupsk@gmail.com or cem.dogdu@dzne.de (C.D.); stefan.schweinberger@uni-jena.de (S.R.S.)

Abstract: Vocal emotion recognition (VER) in natural speech, often referred to as speech emotion recognition (SER), remains challenging for both humans and computers. Applied fields including clinical diagnosis and intervention, social interaction research or Human Computer Interaction (HCI) increasingly benefit from efficient VER algorithms. Several feature sets were used with machine-learning (ML) algorithms for discrete emotion classification. However, there is no consensus for which low-level-descriptors and classifiers are optimal. Therefore, we aimed to compare the performance of machine-learning algorithms with several different feature sets. Concretely, seven ML algorithms were compared on the Berlin Database of Emotional Speech: Multilayer Perceptron Neural Network (MLP), J48 Decision Tree (DT), Support Vector Machine with Sequential Minimal Optimization (SMO), Random Forest (RF), k-Nearest Neighbor (KNN), Simple Logistic Regression (LOG) and Multinomial Logistic Regression (MLR) with 10-fold cross validation using four openSMILE feature sets (i.e., IS-09, emobase, GeMAPS and eGeMAPS). Results indicated that SMO, MLP and LOG show better performance (reaching to 87.85%, 84.00% and 83.74% accuracies, respectively) compared to RF, DT, MLR and KNN (with minimum 73.46%, 53.08%, 70.65% and 58.69% accuracies, respectively). Overall, the emobase feature set performed best. We discuss the implications of these findings for applications in diagnosis, intervention or HCI.

Keywords: machine learning; vocal emotion recognition; speech; emotional speech database; feature set

Citation: Doğdu, C.; Kessler, T.; Schneider, D.; Shadaydeh, M.; Schweinberger, S.R. A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech. *Sensors* **2022**, *22*, 7561. https://doi.org/10.3390/s22197561

Academic Editors: Soo-Hyung Kim and Gueesang Lee

Received: 31 August 2022
Accepted: 2 October 2022
Published: 6 October 2022

1. Introduction

Vocal emotion recognition (VER) is a crucial part of the socio-emotional skill set that allows humans to understand others' affective states within multimodal emotion processing. Together with visual signals and the verbal content of speech, nonverbal signals are also crucial for us to understand the emotional situation in a social context and therefore to adjust our behaviors to react appropriately. For the past two decades, areas within computer science such as Human Computer Interaction (HCI) have aimed to provide tools for automatic emotion recognition in natural speech [1]. Automatic emotion recognition tools are increasingly important in a growing range of applications. For instance, these methods can offer benefits to clinical diagnosis and intervention, such as in the fields of

autism [2,3], prognosis and prevention (e.g., in the detection of depression or suicidal tendencies [4,5], or even in the context of therapies and human conflict resolution [6–8]. Such methods may also support better understanding of temporal causality of nonverbal behavior in dyadic human social interactions [9] or help to provide a non-invasive yet objective assessment of imitation behavior [10,11]. In the context of robotics and HCI, and especially when combined with high-quality speech synthesis technology [12], efficient emotion recognition algorithms can also contribute to experiencing interactions with robot or other technological devices as rewarding, seamless and trustworthy. At the same time, there is a relative gap in systematic research regarding the relative efficiency of these methods to classify specific emotions. Accordingly, the main aim of the present paper—a comparison of such algorithms—seems to be relevant for a range of applications.

Several classical machine-learning (ML) and recent deep-learning (DL) algorithms have been used with databases from different languages on supervised and unsupervised classification tasks [13]. However, automatization of emotion recognition is still challenging considering the significant number of complex parameters determining the accuracy and generalizability of computational methods.

Research suggests that the success of supervised ML and DL algorithms is contingent on the variation of emotional models (i.e., discrete vs. dimensional), dataset types (i.e., acted, elicited, natural or semi-natural), data pre-processing (i.e., framing, windowing, voice activity detection, normalization, noise reduction, feature selection) and supporting modalities (e.g., visual and physiological signals) [13,14]. Crucially, the extracted features constitute the core of model training process. These features are composed of physical acoustic parameters (i.e., Low Level Descriptors (LLDs)) such as prosodic and spectral features. In the past, several acoustic parameter feature sets were developed, which are now being used with linear/non-linear ML classifiers and complex Deep Neural Networks. For instance, Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), Decision Tree (DT) (i.e., "C4.5", see [15]), Linear Regression and Naïve-Bayes [16–20] are among the most popular classical machine-learning algorithms usually showing good overall accuracy within the databases with relatively low computational costs. These algorithms contribute to the aim of real-time raw emotion recognition in speech. Furthermore, recent developments provide us with complex artificial and deep neural network algorithms such as the Multilayer Perceptron Neural Network (MLP), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [21]. These complex algorithms are capable of handling challenging tasks although they require very high computing power, expensive hardware, and a significant amount of energy. Finally, we note that Hidden Markov Models (HMMs) are also successfully used for automatic emotion recognition [22] and were also recently used in combination with MFCCs to address other challenging tasks of automatic classification of complex non-vocal human sounds [23]. However, HMMs also require substantial resources in terms of computing power and time, which can be a limitation to their feasibility [24]. Overall, the work towards determining the best methods for VER remains ongoing.

In experimental studies on ML algorithm comparisons, there is still no consensus for which physical acoustic parameters are optimal for the highest recognition accuracy [14]. From the broadest perspective, extracted LLDs are listed under four main categories [13,25]: (a) Prosodic Features (i.e., Fundamental frequency/pitch (F_0), energy-volume/intensity, duration/speech rate), (b) Spectral Features (e.g., Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC)), (c) Voice Quality Features: (e.g., Jitter, Shimmer and Harmonics-to-Noise Ratio-HNR) and d) Teager Energy Operator (TEO) Based Features. Among these features, a wide range of extraction methods were applied such as calculating global statistics (e.g., mean and range) or voiced region local level analyses of the F_0 separately or in combination [26]. This variation in the feature extraction can also be seen in the MFCC features, which are among the most frequently used spectral features. For instance, the number of used MFCCs and applied statistical functionals varies across established default feature sets [27,28] (e.g., 12 MFCCs/3 Inter-Quartile Ranges–

emobase or 1–4 MFCCs/20th to 80th percentile–*eGeMAPS*). With regard to these variations in feature extraction, further exploratory comparisons are needed.

During the past decades, several feature sets were developed combining some of these LLDs. Popular examples are emobase [27], GeMAPS, eGeMAPS [28] and IS-09 [29]. The feature extraction tool openSMILE may be the most popular software deriving the above listed acoustic parameters—involving long-term (i.e., global) features and short-term (i.e., local) stationary features. In particular, the latter allows defining temporal changes along chunks of speech signal [27]. Although feature sets have commonalities, they provide different levels of performances depending on the LLDs, statistical derivations but also the number of the extracted LLDs. In this manner, the effects of extracted features to the performance of ML classifiers needs to be investigated and cross-validated.

Considering these parameters in VER, the aim of this methodological study is to compare existing algorithms in order to identify relatively more accurate ML classifiers and physical acoustic parameters among some of the existing materials. It would also be crucial to identify the level of performance on each emotion class separately. Overall, the results would indicate the crucial parameters determining discrete emotion prediction accuracy.

2. Related Works

In the past, several studies conducted comparisons between combinations of several features and classifiers on the widely used Berlin Database of Emotional Speech (EMO-DB) [30] and more recent databases such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [31].

In the comparison of classical ML algorithms, the SVM and the KNN algorithms are frequently used. For example, SVM, KNN, Linear Discriminant Analysis (LDA) and Regularized Discriminant Analysis (RDA) were compared with a fusion of spectral and prosodic features on the EMO-DB and a Spanish emotional database [32]. Overall, results indicated accuracies between 62% and 81%, with the RDA performing best among these classifiers. In another study, the SVM outperformed KNN and Naïve-Bayes algorithms, showing up to 86% accuracy on the EMO-DB with a gain-ratio feature selection approach [19]. In a more recent study, Logistic Regression was reported to give better (in fact, ceiling, at 100%) performance compared to MLP (84.62%) and SVM (91,67%) algorithms on an unstandardized own database with MFCCs [33]. In these experimental designs, it is apparent that researchers used different kinds of feature extraction methods but also only a single set of features across the algorithms. In relation to this limitation, Sugan and colleagues [34] compared the SVM and the Feedforward Backpropagation Artificial Neural Network (FF-BP-ANN) across three different sets of cepstral features on the EMO-DB. However, this study contains no prosodic features as some other studies do (e.g., [35]), despite the distinctive characteristic effect of intonation and rhythm on vocal emotion expression [36].

Considering this comparison approach and its limitations in the literature, the aim of this study is to evaluate performance of ML algorithms across different feature sets containing different types of physical features (e.g., prosodic, spectral, voice quality) on the most frequently used EMO-DB database. This approach would provide a broader perspective to evaluate the performance of ML algorithms and their large variations of performance. In addition, we conduct statistical significance tests on the algorithm comparisons which are not found in previous research. Finally, another aim of this study is to discuss classification performances on each emotion class level in more detail, rather than focusing only on overall accuracy percentages.

3. Materials and Methods

3.1. Database

The Berlin Database of Emotional Speech (EMO-DB) [30] was used for the classification of the emotional states anger, boredom, disgust, fear, happiness, sadness and neutral. The EMO-DB comprises 10 German everyday life sentences (i.e., 5 short and 5 long sentences) that were recorded from 10 actors (5 females, 5 males) for each emotional state. In total,

535 utterances (Table 1) are used in the final form of the database (i.e., 127 anger, 69 fear, 46 disgust, 62 sadness, 71 happiness, 79 neutral, 81 boredom).

Table 1. Number of instances for each emotion class of the EMO-DB.

Emotions	Number of Instances
Anger	127
Fear	69
Disgust	46
Sadness	62
Happiness	71
Boredom	81
Neutral	79
Total	**535**

3.2. Feature Extraction

In total, 4 distinct feature sets of the openSMILE (Version 3.0) [27] were used for the extraction of physical acoustic parameters (also called Low-Level-Descriptors (LLDs)) and their statistical derivations: The openSMILE/openEAR "emobase" Feature Set (emobase), The INTERSPEECH 2009 Emotion Challenge Feature Set (IS-09) [29], The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version (eGeMAPS) [28].

The emobase feature set is composed of 988 attributes derived from 26 LLDs and their delta coefficients by implementing several statistical functions (Table 2) such as standard deviation, skewness, kurtosis, range and arithmetic mean. Included LLDs are Fundamental Frequency (F_0), 12 MFCCs, Zero-Crossing Rate (ZCR), Probability of Voicing, Intensity, Loudness, F_0 Envelope and 8 Line Spectral Frequencies.

The IS-09 contains 16 LLDs (Table 2) and 384 attributes. Fundamental Frequency (F_0), 12 MFCCs, Zero-Crossing Rate (ZCR) and Probability of Voicing are common LLDs just as emobase. In addition, it also contains Root-Mean-Square (RMS) Energy parameters. As in emobase, the delta coefficients are computed and 12 statistical functionals are implemented (i.e., mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, range and two linear regression coefficients with their mean square error).

Table 2. Low-Level-Descriptors and Functionals of Each Feature Set. (*) indicates features of emobase and IS-09 distinct from each other. (**) indicates the features of eGeMAPS in addition to GeMAPS features.

Feature Sets	Low-Level-Descriptors	Functionals
emobase and IS-09 (Common Features)	F_0, 12 MFCCs, ZCR, Probability of Voicing	Mean, Standard Deviation, Skewness, Kurtosis, Minimum and Maximum Value, Range, Slope and Offset of Linear Approximation with Quadratic Error
emobase	* Intensity, Loudness, F_0 Envelope, 8 Line Spectral Frequencies	* 3 Inter-Quartile Ranges, Quartile 1–3
IS-09	* (RMS) Energy	-
GeMAPS	F_0, H1-H2 Harmonic Difference F_0, H1-A3 Harmonic Difference (F_0 − A3), Jitter, Formant 1-2-3 Frequency, Formant 1, Shimmer, Loudness, HNR, Alpha Ratio, Hammarberg Index, Spectral Slope 0–500 Hz and 500–100 Hz, Formant 1-2-3 Relative Energy	Mean, Coefficient of Variation; (For loudness and F_0): *20th, 50th and 80th Percentile, the Range of 20th to 80th percentile, Mean and Standard Deviation of the Slope of rising/falling signal parts*; (6 Additional Temporal Features): *Rate of Loudness Peaks, Mean Length and Standard Deviation on the Regions $F_0 > 0$ and $F_0 = 0$, Pseudo Syllable Rate*
eGeMAPS	** MFCCs 1–4, Spectral Flux and Formant 2–3 Bandwidth	* Equivalent Sound Level. Voiced and unvoiced region inclusions vary among some LLDs.

The more recently developed GeMAPS feature set contains 18 LLDs (Table 2) including Fundamental Frequency (F_0), H1-H2 Harmonic Difference (F_0), H1-A3 Harmonic Difference (F_0 − A3), Jitter, Formant 1-2-3 Frequency, Formant 1, Shimmer, Loudness, Harmonics to Noise Ratio (HNR), Alpha Ratio, Hammarberg Index, Spectral Slope 0–500 Hz and 500–100 Hz and Formant 1-2-3 Relative Energy. Mean and Coefficient of Variation are calculated on all smoothed LLDs. The 20th, 50th and 80th Percentile, the Range of 20th to 80th percentile, Mean and Standard Deviation of the Slope of rising/falling signal parts are applied only to pitch and loudness. In addition, several temporal features are added such as rate of loudness peaks and mean length of unvoiced regions (for more details see [25]). The extended version (eGeMAPS) contains additional spectral parameters MFCCs 1–4, Spectral flux and Formant 2–3 Bandwidth. In addition, Equivalent Sound Level, Voiced and unvoiced region inclusions are added to the eGeMAPS (Table 2).

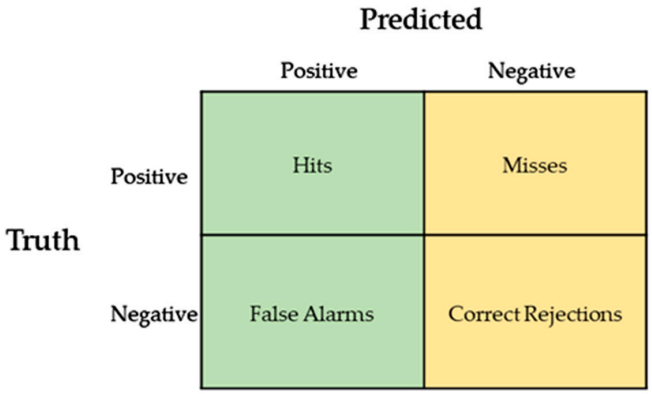
3.3. Classifiers

In terms of the determining emotional classifiers, classifications were conducted with the Python wrapper package [37] of the WEKA [38] ML classifiers: Multilayer Perceptron Neural Network (MLP), Support Vector Machine with Sequential Minimal Optimization (SMO), J48 Decision Tree (DT), Random Forest (RF), k-Nearest Neighbor (KNN), Simple Logistic Regression (LOG) and Multinomial Logistic Regression (MLR). Classifier configurations were set to default values of the WEKA except the MLP.

The MLP contained 3 hidden layers. The number of nodes at the first layer varied depending on the number of input attributes (i.e., sum of the number of attributes and number of classes is divided by 2). The second and third hidden layers contained 32 and 16 nodes, respectively. Further neural network configurations were default values of the WEKA algorithm (i.e., learning rate = 0.3, momentum rate of the backpropagation = 0.2, number of epochs = 500).

3.4. Statistical Analyses

The overall performance matrix was based on overall accuracy (ACC) and F-measures (F) that calculated via precision and recall scores for each classifier and feature set combination (Figure 1).



Recall = Hits / (Hits + Misses)

Precision = Hits / (Hits + False Alarms)

Specificity = Correct Rejections / (Hits + False Alarms)

False Negative Ratio = Misses / (Hits + False Alarms)

False Positive Ratio = False Alarms / (False Alarms + Correct Rejections)

Accuracy = Hits + Correct Rejections / (All)

F-measure = (2 * Precision * Recall) / (Precision + Recall)

Figure 1. Computations for the prediction performance evaluations.

Several model fit measures and weighted (w) averages along imbalanced classes of the database were calculated for the comparison of the models such as precision, recall, Area Under Precision Recall Curve (AUPRC), F-measure, Matthews Correlation Coefficient (MCC) [39], Area Under Curve (AUC—calculated via Receiver Operating Characteristics) [40], Cohen´s Kappa (κ) [41], and Root Mean Squared Error (RMSE).

In addition, significance tests were conducted on the accuracy percentages of the classifiers taking the best classifier as a test base and implementing 10-fold cross-validated paired T-tests. Comparisons were conducted on the final average accuracy of all 10-fold.

4. Results

4.1. 10-Fold Cross-Validation

In total, 4 feature sets and 7 classifiers were included in the 10-fold cross-validation studies on the EMO-DB (see also Figure S1 for all prediction performance evaluations in the supplementary materials). Overall, the most accurate model and feature set combination was the SMO classifier and the emobase feature set (ACC = 87.85%; $AUPRC_w$ = 0.84; F_w = 0.88; MCC = 0.86; AUC_w = 0.97; κ = 75; RMSE = 0.31; see Table 3 and Figure 2). On all feature sets, $SMO_{acc\ range}$ [78.32%: 87.85%], $MLP_{acc\ range}$ [76.63%: 84.00%] and $LOG_{acc\ range}$ [79.44%: 83.74%] showed relatively better performance compared to the decision tree-based DT, RF and KNN classifiers. Additional classification performance measures were also consistent in regard to this difference between two group of algorithms (Figure 2). Crucially, the MLR classifier reached up to 84.70% accuracy with the emobase feature set ($AUPRC_w$ = 0.92; F_w = 0.85; MCC = 0.82; AUC_w = 0.98; κ = 0.64; RMSE = 0.29) while showing relatively low performance on the GeMAPS (ACC = 70.65%; $AUPRC_w$ = 0.75; F_w = 0.71; MCC = 0.66; AUC_w = 0.94; κ = 0.65; RMSE = 0.29) and eGeMAPS (ACC = 71.96%; $AUPRC_w$ = 0.78; F_w = 0.72; MCC = 0.67; AUC_w = 0.94; κ = 0.67; RMSE = 0.27) feature sets.

Table 3. Accuracy percentages of each classifier on each feature set. Classifier and feature set names and abbreviations are written bold.

%	MLP	SMO	DT	RF	KNN	LOG	MLR
emobase	84.00	87.85	54.95	75.70	63.93	83.74	84.70
IS-09	80.37	83.74	53.08	73.46	58.69	79.44	78.51
GeMAPS	76.63	78.32	56.10	75.00	66.00	79.63	70.65
eGeMAPS	79.25	79.63	55.14	74.77	68.22	79.81	71.96

Overall, the worst accuracy performances were detected on $DT_{acc\ range}$ [53.08%: 56.10%] and $KNN_{acc\ range}$ [58.69%: 68.22%] classifiers. Although $RF_{acc\ range}$ [73.46%: 75.70%] provided a substantial improvement compared to DT, it showed inferior performance compared to MLP, LOG and MLR, especially with emobase and IS-09 feature sets.

The feature sets emobase and IS-09 showed relatively better performance compared to the GeMAPS and eGeMAPS (Table 3). However, GeMAPS and eGeMAPS reached around 76% to 79% accuracy with the MLP, SMO and LOG classifiers. Specifically, with the MLR classifier, the performance difference was biggest between emobase (ACC = 84.70%; $AUPRC_w$ = 0.92; F_w = 0.85; MCC = 0.82; AUC_w = 0.98; κ = 0.64; RMSE = 0.29) and GeMAPS (ACC = 70.65%; $AUPRC_w$ = 0.75; F_w = 0.71; MCC = 0.66; AUC_w = 0.94; κ = 0.65; RMSE = 0.29) or eGeMAPS (ACC = 71.96%; $AUPRC_w$ = 0.78; F_w = 0.72; MCC = 0.67; AUC_w = 0.94; κ = 0.67; RMSE = 0.27) feature sets.

Feature Set	Algorithm	Precision	Recall	AUPRC	F-measure	MCC	AUC	Kappa	RMSE
emobase	*SMO*	0.88	0.88	0.84	0.88	0.86	0.97	0.75	0.31
	MLP	0.84	0.84	0.87	0.84	0.81	0.97	0.71	0.25
	RF	0.78	0.76	0.84	0.74	0.71	0.96	0.71	0.14
	DT	0.56	0.55	0.41	0.55	0.47	0.75	0.51	0.33
	MLR	0.85	0.85	0.92	0.85	0.82	0.98	0.64	0.29
	KNN	0.65	0.64	0.49	0.64	0.58	0.78	0.61	0.31
	LOG	0.84	0.84	0.90	0.84	0.81	0.98	0.86	0.30
IS-09	*SMO*	0.84	0.84	0.78	0.84	0.81	0.96	0.81	0.31
	MLP	0.81	0.80	0.83	0.80	0.77	0.96	0.77	0.22
	RF	0.76	0.74	0.81	0.72	0.69	0.96	0.68	0.26
	DT	0.53	0.53	0.40	0.53	0.45	0.75	0.44	0.35
	MLR	0.79	0.79	0.86	0.79	0.75	0.96	0.75	0.23
	KNN	0.59	0.59	0.42	0.59	0.51	0.75	0.51	0.34
	LOG	0.79	0.79	0.87	0.79	0.76	0.97	0.76	0.21
GeMAPS	*SMO*	0.78	0.78	0.70	0.78	0.74	0.94	0.74	0.31
	MLP	0.77	0.77	0.75	0.77	0.73	0.93	0.72	0.25
	RF	0.76	0.75	0.82	0.74	0.70	0.96	0.70	0.24
	DT	0.56	0.56	0.44	0.56	0.48	0.76	0.48	0.34
	MLR	0.71	0.71	0.75	0.71	0.66	0.94	0.65	0.29
	KNN	0.66	0.66	0.51	0.66	0.60	0.80	0.60	0.31
	LOG	0.79	0.80	0.86	0.79	0.76	0.97	0.76	0.21
eGeMAPS	*SMO*	0.80	0.80	0.73	0.79	0.76	0.95	0.76	0.31
	MLP	0.79	0.79	0.81	0.79	0.75	0.94	0.75	0.23
	RF	0.76	0.75	0.82	0.74	0.70	0.96	0.70	0.24
	DT	0.55	0.55	0.44	0.55	0.47	0.77	0.47	0.34
	MLR	0.72	0.72	0.78	0.72	0.67	0.94	0.67	0.27
	KNN	0.68	0.68	0.54	0.68	0.62	0.81	0.62	0.30
	LOG	0.80	0.80	0.85	0.80	0.76	0.96	0.76	0.21

Figure 2. Classification performance measures among feature sets. Precision, Recall, AUPRC and AUC values are weighted averages among number of instances of each class in the database. Data bars represent values between 0 and 1. Length of the data bars are determined by the number in each cell.

Cross-validated paired *t*-tests also confirmed that the SMO classifier showed an overall better performance compared to most of the classifiers (Table 4). Particularly, SMO significantly outperformed DT, RF and KNN algorithms among all comparisons (although some of these differences were less prominent with the GeMAPS and eGeMAPS feature sets). Interestingly, the SMO classifier was significantly better than the LOG only on the emobase feature set ($t = 4.56$, $p = 0.001$). This indicates that the largest feature set emobase created a more robust difference between SMO and LOG.

Table 4. Cross-Validated Paired T-Test Comparison (two-tailed) with the Test Base SMO Classifier. *: $p < 0.05$. **: $p \leq 0.001$. Note that positive t-values indicate better performance of the SMO classifier. Classifier and feature set names and abbreviations are written bold.

Feature Set	MLP	DT	RF	KNN	LOG	MLR
emobase	$t = 4.79$ $p = 0.001$ **	$t = 14.78$ $p < 0.001$ **	$t = 12.21$ $p < 0.001$ **	$t = 10.93$ $p < 0.001$ **	$t = 4.56$ $p = 0.001$ **	$t = 3.58$ $p = 0.005$ *
IS-09	$t = 3.05$ $p = 0.001$ **	$t = 10.00$ $p < 0.001$ **	$t = 7.20$ $p < 0.001$ **	$t = 10.77$ $p < 0.001$ **	$t = 1.40$ $p = 0.198$	$t = 1.74$ $p = 0.116$
GeMAPS	$t = 2.45$ $p = 0.014$ *	$t = 17.24$ $p < 0.001$ **	$t = 2.58$ $p = 0.030$ *	$t = 5.04$ $p = 0.001$ **	$t = -0.49$ $p = 0.638$	$t = 3.81$ $p = 0.004$ *
eGeMAPS	$t = 1.12$ $p = 0.292$	$t = 12.50$ $p < 0.001$ **	$t = 1.58$ $p = 0.150$	$t = 5.36$ $p = 0.001$ **	$t = 1.04$ $p = 0.328$	$t = 1.22$ $p = 0.254$

Finally, comparisons of the SMO with the other two best classifiers MLR and MLP indicated that SMO slightly outperformed these. However, the difference to MLR was statistically significant only with the emobase (t = 3.58, $p = 0.005$) and GeMAPS (t = 3.81, $p = 0.004$) feature sets. With the similar pattern, SMO performed better than MLP only with the emobase (t = 4.79, $p = 0.001$), GeMAPS (t = 2.45, $p = 0.014$) and IS-09 feature sets (t = 3.05, $p = 0.001$).

4.2. F-Measures of Class Predictions and Confusion Matrices

F-measures of each emotion class are displayed across classifiers and feature sets (Figure 3). Overall, the emotion class with the best prediction performance was "Sadness" $_{\text{Fmean}}$[0.79:0.88] while the lowest performance was detected for "Happiness" $_{\text{Fmean}}$[0.54:0.64].

Among the least accurate classifiers, it is apparent that F-measures of the "Sadness" class are still acceptable such as for RF [emobase = 0.86, IS-09 = 0.80, GeMAPS = 0.91, eGeMAPS = 0.93] and KNN [emobase = 0.87, IS-09 = 0.76, GeMAPS = 0.92, eGeMAPS = 0.95]. Crucially, this performance increase can be detected more correctly in GeMAPS and eGeMAPS sets when compared to emobase and IS-09 feature sets. Moreover, the least accurate classifier DT showed even good performance on the "Sadness" emotion, especially on GeMAPS (0.80) and eGeMAPS (0.86) feature sets. However, this increase in performance was not as prominent on emobase (0.66) and IS-09 (0.65) feature sets.

Conversely, even the most accurate classifiers showed lowest F-measures for the "Happiness" emotion such as MLP [emobase = 0.66, IS-09 = 0.65, GeMAPS = 0.59, eGeMAPS = 0.63], LOG [emobase = 0.74, IS-09 = 0.70, GeMAPS = 0.67, eGeMAPS = 0.51], SMO [emobase = 0.82, IS-09 = 0.75, GeMAPS = 0.67, eGeMAPS = 0.63] and MLR [emobase = 0.77, IS-09 = 0.72, GeMAPS = 0.54, eGeMAPS = 0.54]. However, it is crucial to note that this performance decrease of "Happiness" class detection among SMO and MLR algorithms was detected more prominently on GeMAPS and eGeMAPS feature sets compared to emobase and IS-09.

Furthermore, confusion matrices for each classifier and feature set combinations were extracted (Figure 4 and see Figure S2 for all in the supplementary materials). As a complementary finding, some of the confusion matrices seem problematic in regard to the low performance of the class "Happiness" predictions. For instance, 20% (14 out of 71) of "Happiness" voices were classified as "Anger" and 12% (15 out of 127) of "Anger" voices were classified as "Happiness" by the MLP classifier with the emobase feature set (Figure 4a). Similarly, these false classifications were 21% (15/71) "Anger" and 6% (8/127) "Happiness" on the SMO classifier with IS-09 feature set (Figure 4b). Furthermore, GeMAPS and eGeMAPS feature sets were also problematic in terms of the differentiation of "Happiness" and "Anger", for instance with MLR and LOG classifiers (Figure 4c,d). Moreover, this problem is more apparent among the classifiers with overall low performance such as RF

with IS-09 and KNN with eGeMAPS (Figure 4e,f), showing 52% (37/71) and 35% (25/71) of confusion, respectively.

Feature Set	Emotions	MLP	SMO	DT	RF	KNN	LOG	MLR	MEAN
emobase	Anger	0.86	0.90	0.69	0.80	0.78	0.88	0.89	0.83
	Fear	0.82	0.83	0.45	0.68	0.65	0.84	0.80	0.72
	Disgust	0.84	0.85	0.43	0.67	0.65	0.87	0.84	0.74
	Sadness	0.86	0.89	0.66	0.86	0.87	0.85	0.87	0.84
	Happiness	0.66	0.82	0.49	0.49	0.54	0.74	0.77	0.64
	Neutral	0.91	0.92	0.54	0.81	0.47	0.84	0.85	0.76
	Boredom	0.90	0.89	0.49	0.81	0.49	0.83	0.86	0.75
	Weighted Average	0.84	0.88	0.55	0.74	0.64	0.84	0.85	
IS-09	Anger	0.83	0.88	0.67	0.80	0.70	0.87	0.84	0.80
	Fear	0.75	0.84	0.43	0.69	0.49	0.76	0.75	0.67
	Disgust	0.79	0.82	0.40	0.74	0.53	0.81	0.77	0.69
	Sadness	0.87	0.83	0.65	0.80	0.76	0.85	0.79	0.79
	Happiness	0.65	0.75	0.35	0.52	0.50	0.70	0.72	0.60
	Neutral	0.86	0.88	0.57	0.77	0.57	0.78	0.76	0.74
	Boredom	0.84	0.81	0.52	0.69	0.49	0.77	0.81	0.70
	Weighted Average	0.84	0.84	0.53	0.72	0.59	0.79	0.79	
GeMAPS	Anger	0.83	0.84	0.70	0.80	0.78	0.86	0.78	0.80
	Fear	0.73	0.73	0.41	0.68	0.59	0.73	0.68	0.65
	Disgust	0.67	0.63	0.30	0.58	0.51	0.71	0.64	0.58
	Sadness	0.86	0.90	0.80	0.91	0.92	0.94	0.82	0.88
	Happiness	0.59	0.67	0.40	0.59	0.46	0.67	0.54	0.56
	Neutral	0.79	0.77	0.50	0.75	0.63	0.80	0.69	0.70
	Boredom	0.82	0.81	0.64	0.79	0.60	0.80	0.75	0.74
	Weighted Average	0.77	0.78	0.56	0.74	0.66	0.79	0.71	
eGeMAPS	Anger	0.83	0.84	0.69	0.81	0.76	0.69	0.76	0.77
	Fear	0.76	0.73	0.37	0.67	0.68	0.64	0.67	0.65
	Disgust	0.70	0.74	0.39	0.56	0.55	0.66	0.72	0.62
	Sadness	0.94	0.93	0.86	0.93	0.95	0.65	0.85	0.87
	Happiness	0.63	0.63	0.43	0.57	0.47	0.51	0.54	0.54
	Neutral	0.79	0.77	0.47	0.74	0.65	0.68	0.66	0.68
	Boredom	0.84	0.85	0.53	0.78	0.63	0.76	0.81	0.74
	Weighted Average	0.79	0.79	0.55	0.74	0.68	0.66	0.72	

Figure 3. F-measures for each emotion. Color coding indicates performance, with dark green indicating best and dark red indicating poorest performance, and with yellow indicating intermediate classification performance, as shown in the color bar.

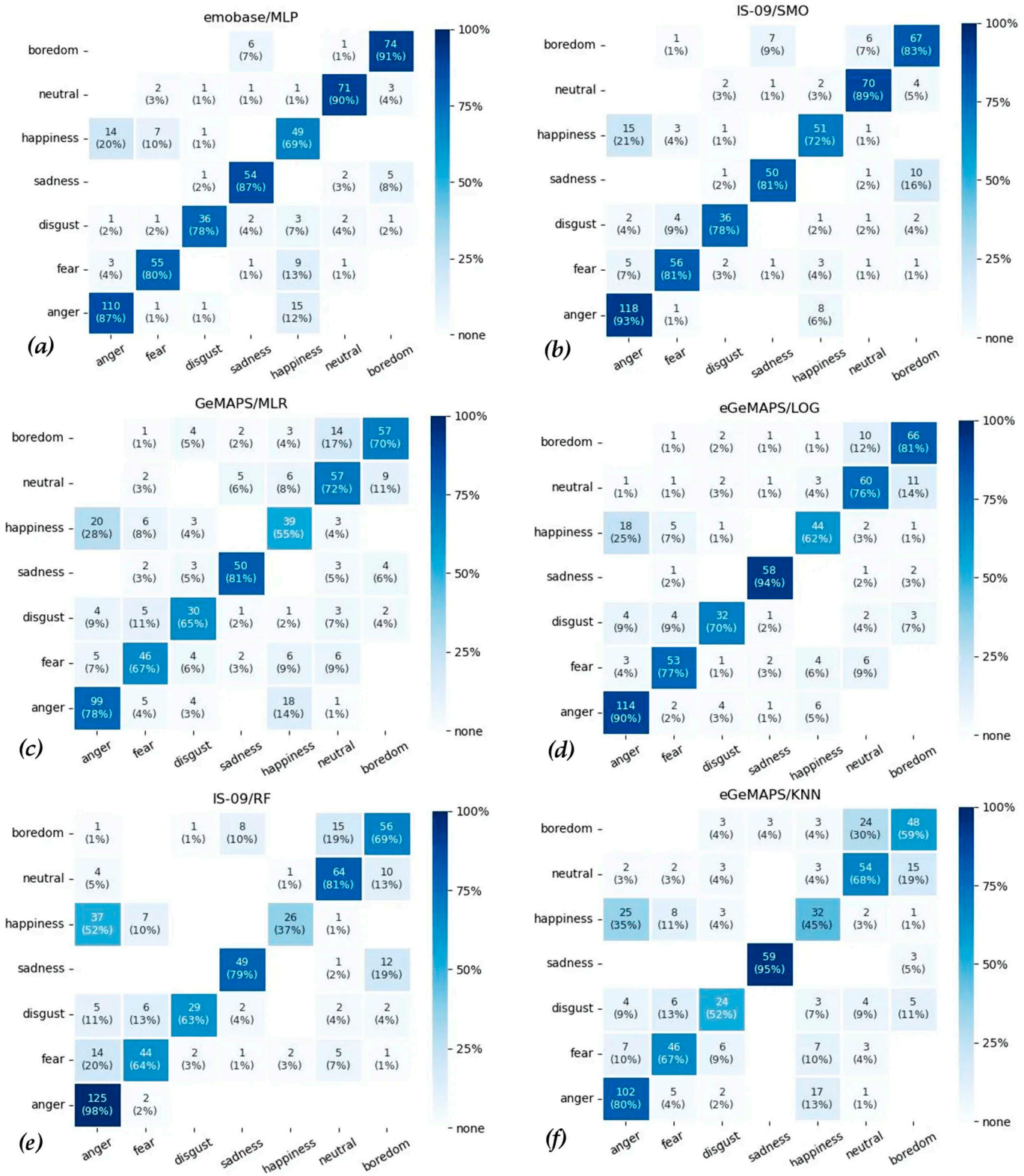


Figure 4. Confusion Matrices of the Predictions With (**a**) emobase/MLP, (**b**) IS-09/SMO, (**c**) GeMAPS/MLR, (**d**) eGeMAPS/LOG, (**e**) IS-09/RF, (**f**) eGeMAPS/KNN. The *x*-axis represents the ground truth labels and the *y*-axis represents predicted labels. Note: Figures give percentages determining the color map but also provide absolute numbers in parentheses to transparently indicate different base frequencies of the predicted emotions. Note also that percentages and numbers are omitted for empty cells to enhance readability.

Finally, some confusion matrices also indicate a differentiation problem between the "Boredom" and "Neutral" classes. These cases were observed especially among the classifications with low overall performance such as RF with IS-09 (Figure 4e), KNN with eGeMAPS (Figure 4f), KNN with emobase, DT with GeMAPS and eGeMAPS (please see supplementary materials Figure S2).

5. Discussion

Conducting successful automatic vocal emotion recognition (VER) in speech is challenging and depends on numerous complex factors. With a set of analyses, we aimed to identify the ML algorithms and feature sets with best performances on the EMO-DB emotional speech database. Performance analyses on 10-fold cross-validation indicated that SMO, MLP and LOG showed more accurate predictions compared to DT, RF and KNN. Overall, the classifications were more accurate with the emobase and IS-09 feature sets compared to GeMAPS and eGeMAPS. It is crucial to better understand these differences.

The SMO algorithm showed the best overall accuracy percentage, achieving 87.9% with the emobase feature set in our analyses and comparisons. Further, our SMO results indicate relatively better overall accuracy compared to classical SVM, which has been widely used in the vocal emotion recognition literature [26,42,43]. The Sequential Minimal Optimization algorithm in our analyses provided significant improvements, specifically with the implemented default feature sets.

Considering the classifications with low overall accuracy in our analyses, it is important to note that decision-tree-based algorithms (i.e., DT and RF) are not ideal for the used emotional dataset and acoustic parameter sets, at least when it comes to the default configurations of WEKA. Even though accuracy was low in our C4.5 classical DT algorithm, it should be noted that the decision-tree concept seems to be successfully implemented in several other algorithms. For instance, Lee et al. [18] managed to reach up to 89.6% accuracy by implementing the Extreme Learning Machine (ELM) technique and the SVM binary decision-tree (DTSVM) algorithm with a correlational feature selection approach. Moreover, in a study with a new approach called DNN-decision tree, SVM algorithm reached 75.8% accuracy on the EMO-DB database [44]. However, the feature extraction methods are not identical among these studies, which makes the performance comparison difficult.

KNN was another algorithm with relatively low overall accuracies in our study. This outcome stands in contrast to previous research that provided very high overall accuracies with the KNN algorithm on VER. This discrepancy could potentially also relate to differences in the applied feature extraction method. For instance, Khan et al. [45] reported reaching 91.7% overall accuracy on an own database using the Forward Selection (FS) feature selection method. Using the EMO-DB database, Zhu and Ahamd [46] showed that the KNN algorithm reached to up to 78.8% accuracy by using Gamma Frequency Cepstral Frequencies (GTCCs) in addition to prosodic features and spectral frequencies.

All this suggests that the performance of the implemented algorithms is highly dependent on the acoustic parameters used as input set to the models. Accordingly, further refinements in this regard seem a promising way to enhance accuracy.

Interestingly, more contemporary default feature sets as implemented in openSMILE (i.e., GeMAPS and eGeMAPS) failed to outperform the older and more extensive feature sets. This difference might be mainly based on the size of the feature sets. Note that the GeMAPS feature set actually could have an advantage in the real-life task as the larger feature sets could have a problem of overfitting to the training set [28]. In addition, the smaller feature set of the GeMAPS should be advantageous in real-time applications. In this manner, it would be even more informative to conduct feature set and algorithm comparisons on unseen data from other databases or real life, rather than just from untrained data within the same dataset.

As a limitation of the present study, note that we conducted analyses only on the EMO-DB database. Although this is one of the most popular databases in the literature, it has a relatively small number of samples (i.e., 535) compared to other widely used databases

such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS; 2496 samples) [31] or the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP; 1150 samples) [47]. It is important to appreciate differences between these databases in terms of producing emotional expressions (e.g., posing, acting to script, mood induction, interaction vs. solo expression) or variability of utterances (e.g., two sentences in RAVDESS, 10 sentences in EMO-DB and open spontaneous material in IEMOCAP). In addition, databases differ with respect to the underlying emotion model, number of emotion categories or dimensions, and language. In this manner, future research in this area could reduce database dependency by using multiple databases with different characteristics. At the same time, reviewing databases with different characteristics could help to select the best material for a particular research question. For instance, databases that use mood induction could have advantages over databases that use enacted/posed emotions when the aim is to assess algorithms' performance in the context of real-life emotions. Moreover, whereas existing databases tend to focus on salient and basic emotions, the increasing trend in emotion research to study more subtle emotions (e.g., pride, compassion, gratitude, admiration, desire; for review see [48]) could call for the development of databases that permit research on the automatic classification of subtle emotions.

Finally, the confusion matrices provide an informative pattern of misclassifications regarding specific emotion. Most prominently, nearly all classifications had a tendency for confusing "Happiness" and "Anger" categories, especially with the GeMAPS and eGeMAPS feature sets. This confusion could be related to the fact that both emotions are expressed with high arousal that results in physically high pitch and volume/intensity on the vocal samples [49]. In addition, the second most confused emotion categories were "Boredom" and "Neutral". Interestingly, "Boredom" is the only emotion that is not one of the Ekman´s six basic emotion categories [50]. At the same time, in a 2D emotional model, specifically "indifferent boredom" has a rather neutral position in terms of valence [51] exhibiting low volume/intensity in relation to lower arousal level.

6. Conclusions

This paper has provided a comparison of classical ML algorithms for automatic vocal emotion recognition in speech, which recently gained importance in a growing range of applications. The present analyses were performed on existing recordings and feature sets of vocal emotions, and the findings therefore will be of interest for researchers who use these algorithms for off-line automatic emotion analysis. Human emotions are inherently multimodal in nature [52], such that sensor integration and simultaneous consideration of facial, vocal, and bodily data (where available) can be expected to enhance automatic emotion recognition. Where applications depend on a real-time automatic analysis of vocal emotions, more comparisons between efficient algorithms for real-time analyses will be warranted. At the same time, such work could benefit from considering the tight link between the perception and expression of human vocal emotions [53]. In the past two decades, substantial progress has been made in methods for automatic emotion recognition. To achieve large-scale usability in applied fields of diagnosis, intervention, communication research or human–robot interaction, both systematic evaluation of available algorithms and ongoing efforts at refining these are indispensable.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/s22197561/s1, Figure S1: Prediction Performance Evaluations; Figure S2: Confusion Matrices.

Author Contributions: Conceptualization, C.D.; methodology, C.D.; software, C.D.; validation, M.S.; formal analysis, C.D.; investigation, C.D., D.S. and M.S; writing—original draft preparation, C.D. and S.R.S.; writing—review and editing, C.D., S.R.S., D.S., T.K. and M.S.; visualization, C.D.; supervision, T.K., S.R.S. and D.S.; funding acquisition, S.R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Carl Zeiss Foundation within the scope of the program line "Breakthroughs: Exploring Intelligent Systems" for "Digitization—explore the basics (No P2017-01-003), use applications" and the Competence Center for Interdisciplinary Prevention at Friedrich Schiller University, (No 1.2.14.22).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The implemented database (http://emodb.bilderbar.info/start.html accessed on 5 June 2022) and feature sets (https://github.com/audeering/opensmile/releases/tag/v3.0.0 accessed on 1 July 2022) are publicly available. The scripts in Python programming language, extracted data and raw outputs can be found via Open-Science-Framework (https://osf.io/k9usg/?view_only=d4b0ede657544849916e79a204956074).

Acknowledgments: The authors would like to thank Okan Ertürk and Oliver Kresin for their advisory support within the scope of data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
2. Drimalla, H.; Scheffer, T.; Landwehr, N.; Baskow, I.; Roepke, S.; Behnia, B.; Dziobek, I. Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (SIT). *Npj Digit. Med.* **2020**, *3*, 25. [CrossRef] [PubMed]
3. Kowallik, A.E.; Schweinberger, S.R. Sensor-Based Technology for Social Information Processing in Autism: A Review. *Sensors* **2019**, *19*, 4787. [CrossRef] [PubMed]
4. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]
5. Dong, Y.Z.; Yang, X.Y. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* **2021**, *441*, 279–290. [CrossRef]
6. Longobardi, T.; Sperandeo, R.; Albano, F.; Tedesco, Y.; Moretto, E.; Di Sarno, A.D.; Dell'Orco, S.; Maldonato, N.M. Co-regulation of the voice between patient and therapist in psychotherapy: Machine learning for enhancing the synchronization of the experience of anger emotion: An experimental study proposal. In Proceedings of the 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 22–24 August 2018; pp. 113–116.
7. Tanana, M.J.; Soma, C.S.; Kuo, P.B.; Bertagnolli, N.M.; Dembe, A.; Pace, B.T.; Srikumar, V.; Atkins, D.C.; Imel, Z.E. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behav. Res. Methods* **2021**, *53*, 2069–2082. [CrossRef]
8. Halperin, E.; Schori-Eyal, N. Towards a new framework of personalized psychological interventions to improve intergroup relations and promote peace. *Soc. Personal. Psychol. Compass* **2020**, *14*, 255–270. [CrossRef]
9. Shadaydeh, M.; Muller, L.; Schneider, D.; Thummel, M.; Kessler, T.; Denzler, J. Analyzing the Direction of Emotional Influence in Nonverbal Dyadic Communication: A Facial-Expression Study. *IEEE Access* **2021**, *9*, 73780–73790. [CrossRef]
10. Kowallik, A.E.; Pohl, M.; Schweinberger, S.R. Facial Imitation Improves Emotion Recognition in Adults with Different Levels of Sub-Clinical Autistic Traits. *J. Intell.* **2021**, *9*, 4. [CrossRef]
11. Shaham, G.; Mortillaro, M.; Aviezer, H. Automatic facial reactions to facial, body, and vocal expressions: A stimulus-response compatibility study. *Psychophysiology* **2020**, *57*, e13684. [CrossRef]
12. Yamagishi, J.; Veaux, C.; King, S.; Renals, S. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoust. Sci. Technol.* **2012**, *33*, 1–5. [CrossRef]
13. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
14. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [CrossRef]
15. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240. [CrossRef]
16. Casale, S.; Russo, A.; Scebba, G.; Serrano, S. Speech emotion classification using machine learning algorithms. In Proceedings of the 2008 IEEE International Conference on Semantic Computing, Santa Monica, CA, USA, 4–7 August 2008; pp. 158–165. [CrossRef]
17. Chavhan, Y.; Dhore, M.L.; Yesaware, P. Speech emotion recognition using support vector machine. *Int. J. Comput. Appl.* **2010**, *1*, 6–9. [CrossRef]

18. Lee, C.C.; Mower, E.; Busso, C.; Lee, S.; Narayanan, S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* **2011**, *53*, 1162–1171. [CrossRef]
19. Gjoreski, M.; Gjoreski, H.; Kulakov, A. Machine learning approach for emotion recognition in speech. *Informatica* **2014**, *38*, 377–384.
20. Wang, S.; Wang, W.; Zhao, J.; Chen, S.; Jin, Q.; Zhang, S.; Qin, Y. Emotion recognition with multimodal features and temporal models. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 598–602. [CrossRef]
21. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* **2021**, *21*, 1249. [CrossRef]
22. Meng, H.Y.; Bianchi-Berthouze, N. Affective State Level Recognition in Naturalistic Facial and Vocal Expressions. *IEEE Trans. Cybern.* **2014**, *44*, 315–328. [CrossRef]
23. Sitaula, C.; He, J.; Priyadarshi, A.; Tracy, M.; Kavehei, O.; Hinder, M.; Withana, A.; McEwan, A.; Marzbanrad, F. Neonatal bowel sound detection using convolutional neural network and Laplace hidden semi-Markov model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1853–1864. [CrossRef]
24. Er, M.B. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. *IEEE Access* **2020**, *8*, 221640–221653. [CrossRef]
25. Nordström, H. Emotional communication in the human voice. Doctoral Dissertation, Department of Psychology, Stockholm University, Stockholm, Sweden, 2019.
26. Rao, K.S.; Koolagudi, S.G.; Vempada, R.R. Emotion recognition from speech using global and local prosodic features. *Int. J. Speech Technol.* **2013**, *16*, 143–160. [CrossRef]
27. Eyben, F.; Wöllmer, M.; Schuller, B. openSMILE: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM 2010 International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462. [CrossRef]
28. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.J.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
29. Schuller, B.; Steidl, S.; Batliner, A. The Interspeech 2009 Emotion Challenge. In Proceedings of the Interspeech 2009 Emotion Challenge, Brighton, UK, 6–10 September 2009; pp. 312–315.
30. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the 2005 Interspeech Conference, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
31. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
32. Kuchibhotla, S.; Vankayalapati, H.D.; Vaddi, R.S.; Anne, K.R. A comparative analysis of classifiers in emotion recognition through acoustic features. *Int. J. Speech Technol.* **2014**, *17*, 401–408. [CrossRef]
33. Rumagit, R.Y.; Alexander, G.; Saputra, I.F. Model comparison in speech emotion recognition for Indonesian language. *Procedia Comput. Sci.* **2021**, *179*, 789–797. [CrossRef]
34. Sugan, N.; Srinivas, N.S.; Kar, N.; Kumar, L.S.; Nath, M.K.; Kanhe, A. Performance comparison of different cepstral features for speech emotion recognition. In Proceedings of the 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 5–7 July 2018; pp. 266–271. [CrossRef]
35. Palo, H.K.; Sagar, S. Comparison of neural network models for speech emotion recognition. In Proceedings of the 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 21–23 September 2018; pp. 127–131. [CrossRef]
36. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Patt. Analy. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef]
37. GitHub. Available online: https://github.com/fracpete/python-weka-wrapper3 (accessed on 10 September 2022).
38. Frank, E.; Hall, M.A.; Witten, I.H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed.; Morgan Kaufmann Publishers: San Fransisco, CA, USA, 2016.
39. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
40. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
41. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
42. Shen, P.; Changjun, Z.; Chen, X. Automatic speech emotion recognition using support vector machine. In Proceedings of the 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, Harbin, China, 12–14 August 2011; pp. 621–625. [CrossRef]
43. Bitouk, D.; Verma, R.; Nenkova, A. Class-level spectral features for emotion recognition. *Speech Commun.* **2010**, *52*, 613–625. [CrossRef] [PubMed]
44. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37. [CrossRef]

45. Khan, M.; Goskula, T.; Nasiruddin, M.; Quazi, R. Comparison between k-nn and svm method for speech emotion recognition. *Int. J. Comput. Sci. Eng.* **2011**, *3*, 607–611.
46. Zhu, C.; Ahmad, W. Emotion recognition from speech to improve human-robot interaction. In Proceedings of the 2019 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 5–8 August 2019; pp. 370–375. [CrossRef]
47. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
48. Sauter, D.A. The nonverbal communication of positive emotions: An emotion family approach. *Emot. Rev.* **2017**, *9*, 222–234. [CrossRef] [PubMed]
49. Banse, R.; Scherer, K.R. Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psychol.* **1996**, *70*, 614. [CrossRef]
50. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef]
51. Goetz, T.; Frenzel, A.C.; Hall, N.C.; Nett, U.E.; Pekrun, R.; Lipnevich, A.A. Types of boredom: An experience sampling approach. *Motiv. Emot.* **2014**, *38*, 401–419. [CrossRef]
52. Young, A.W.; Frühholz, S.; Schweinberger, S.R. Face and voice perception: Understanding commonalities and differences. *Trends Cogn. Sci.* **2020**, *24*, 398–410. [CrossRef]
53. Frühholz, S.; Schweinberger, S.R. Nonverbal auditory communication–evidence for integrated neural systems for voice signal production and perception. *Prog. Neurobiol.* **2021**, *199*, 101948. [CrossRef]