healthcare

MDPI

Article

Predicting Emotional Valence of People Living with the Human Immunodeficiency Virus Using Daily Voice Clips: A Preliminary Study

Ray F. Lin [1,†], Shu-Hsing Cheng [2,3,†], Yung-Ping Liu [4,*], Cheng-Pin Chen [2,5], Yi-Jyun Wang [1] and Shu-Ying Chang [2]

[1] Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan 32003, Taiwan; juifeng@saturn.yzu.edu.tw (R.F.L.); puddingjyun1120@gmail.com (Y.-J.W.)
[2] Department of Infectious Diseases, Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan 33004, Taiwan; shcheng@mail.tygh.gov.tw (S.-H.C.); cpc.y@nycu.edu.tw (C.-P.C.); ingrid45640@gmail.com (S.-Y.C.)
[3] School of Public Health, Taipei Medical University, Taipei 110301, Taiwan
[4] Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung 413310, Taiwan
[5] Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei 11221, Taiwan
* Correspondence: ypliu@cyut.edu.tw
† These authors contributed equally to this work.

check for updates

Citation: Lin, R.F.; Cheng, S.-H.; Liu, Y.-P.; Chen, C.-P.; Wang, Y.-J.; Chang, S.-Y. Predicting Emotional Valence of People Living with the Human Immunodeficiency Virus Using Daily Voice Clips: A Preliminary Study. *Healthcare* **2021**, *9*, 1148. https://doi.org/10.3390/healthcare9091148

Academic Editors: Hsin-Chieh Wu and Min-Chi Chiu

Received: 9 June 2021
Accepted: 31 August 2021
Published: 2 September 2021

Abstract: To detect depression in people living with the human immunodeficiency virus (PLHIV), this preliminary study developed an artificial intelligence (AI) model aimed at discriminating the emotional valence of PLHIV. Sixteen PLHIV recruited from the Taoyuan General Hospital, Ministry of Health and Welfare, participated in this study from 2019 to 2020. A self-developed mobile application (app) was installed on sixteen participants' mobile phones and recorded their daily voice clips and emotional valence values. After data preprocessing of the collected voice clips was conducted, an open-source software, openSMILE, was applied to extract 384 voice features. These features were then tested with statistical methods to screen critical modeling features. Several decision-tree models were built based on various data combinations to test the effectiveness of feature selection methods. The developed model performed very well for individuals who reported an adequate amount of data with widely distributed valence values. The effectiveness of feature selection methods, limitations of collected data, and future research were discussed.

Keywords: HIV; speech emotion recognition; feature selection; artificial intelligence; clinical diagnosis

1. Introduction

1.1. PLHIV and Depression

Studies have found that people living with the human immunodeficiency virus (PLHIV) are more likely to be depressed than ordinary people due to negative emotions to initial diagnosis, the stress of living with chronic illness, perceived and internalized stigma, and the side effects of HIV drugs [1–5]. The prevalence of depression among PLHIV was between 18% and 81% [3,6–8]. Without proper interventions, depression in PLHIV can result in increased substance abuse [9,10], increased high-risk sexual behaviors [11], more rapid HIV progression [7,12], cognitive impairment [6,13,14], and increased risk of suicidality [5,15–19].

1.2. Depression Interventions for PLHIV

Recently, interventions that involve the use of consumer-grade hardware (i.e., mHealth [20]) were proposed to help PLHIV manage depression. For example, Swendeman et al. [21] designed a smartphone self-monitoring application to help PLHIV by asking PLHIV to fill surveys on medication adherence, mental health, substance use, and sexual risk behaviors, and brief text dairies on stressful events. van Luenen et al. [22] tested a Web-based cognitive behavioral therapy program to reduce PLHIV's depressive symptoms and anxiety. Li et al. [23] studied a WeChat-based intervention to reduce the suicide rate of PLHIV. Although the abovementioned interventions were found helpful in improving mental health outcomes and decreasing suicidal risks in PLHIV, their successes rely on PLHIV's willingness and adherence. Without PLHIV's cooperation, these interventions had difficulties tracking PLHIV's destructive emotions and preventing suicidal events effectively.

Integrating mHealth with artificial intelligence (AI) techniques could be a possible solution to track emotional states and prevent suicide. Schnall et al. [24] suggested that smartphones are a fantastic means to track people with chronic diseases. With sensing, computing, and data storage abilities, mobile devices could be an ideal tool to collect information from PLHIV, and the collected data can be tested for developing AI to track their emotional states. Among various data that mobile devices can collect, e.g., image, accelerometer, global positioning system, or temperature, voice can be easily collected from PLHIV in their verbal communications.

Studies have shown that emotion affects voice [24–26]. Many studies developed emotion recognition models using voice, e.g., [27–30]. However, there are limitations of these developed models for our use. First, most of these studies developed models to recognize categorical emotions, such as joy, guilt, contempt, anger, disgust, sadness, and fear [31–33]. While these models focus on the differences among these so-called "basic emotions," they performed ineffectively in the degree of emotional valence. Second, the data used to develop these models were the databases, e.g., [34,35], that used acted speech simulated by actors. As Ayadi et al. [36] questioned, these databases' naturalness is the biggest concern while using them for developing AI models. Lastly, the properties and the intentions of speak between these actors and PLHIV were different.

1.3. Research Objectives

While PLHIV have a high prevalence of depression and risks of suicidality, developing a means to track PLHIV's emotional states in real-time may reduce unfortunate outcomes for PLHIV's case managers or relatives prevent unfortunate events. In support of this hypothesis, this preliminary study aimed to test the integration of mHealth and AI to discriminate PLHIV's emotions. Specific objectives were:

1. To collect PLHIV's voice data using their mobile devices,
2. To screen critical voice features using statistic methods of correlation and analysis of variance (ANOVA),
3. To test AI modeling for discriminating PLHIV's emotional valence and compare the effectiveness of the two statistical methods.

2. Material and Methods

2.1. Study Design and Setting

This study was conducted in the Taoyuan General Hospital, a 900-bed tertiary referred hospital located in Northern Taiwan. The hospital was designated to provide HIV care for more than 3500 PLHIV who resided in Taoyuan City [37]. To provide healthcare services and prevent unfortunate events of PLHIV, the Taoyuan General Hospital has been making efforts to track PLHIV's emotional states. However, it is a considerable challenge for the HIV care team; the staff could merely investigate the emotional state through their regular visits per three months, restricted by the overloaded assignment. Hence, the hospital was seeking an effective means to overcome current limitations.

2.2. Participants and Sampling

Sixteen adult PLHIV, recruited from the Taoyuan General Hospital by convenience sampling, participated in this study from 2019 to 2020. The HIV care team recruited these participants from their PLHIV pool (approximately 3500 in total) via sending a flyer using

Line and via in-person invitation in their clinic visits. Inclusion criteria for the current study were HIV-seropositive status, at least 18 years of age, and agreeing to participate in the study. Their demographic data are shown in Table 1 that summarizes information on age, sex, drug usage, sexual activity, HIV transmission, period of diagnosed HIV, anxiety, depression, and cellphone usage. Before participating in the study, their states of anxiety and depression were assessed using the Chinese version of the Hospital Anxiety and Depression Scale (HADS) [38].

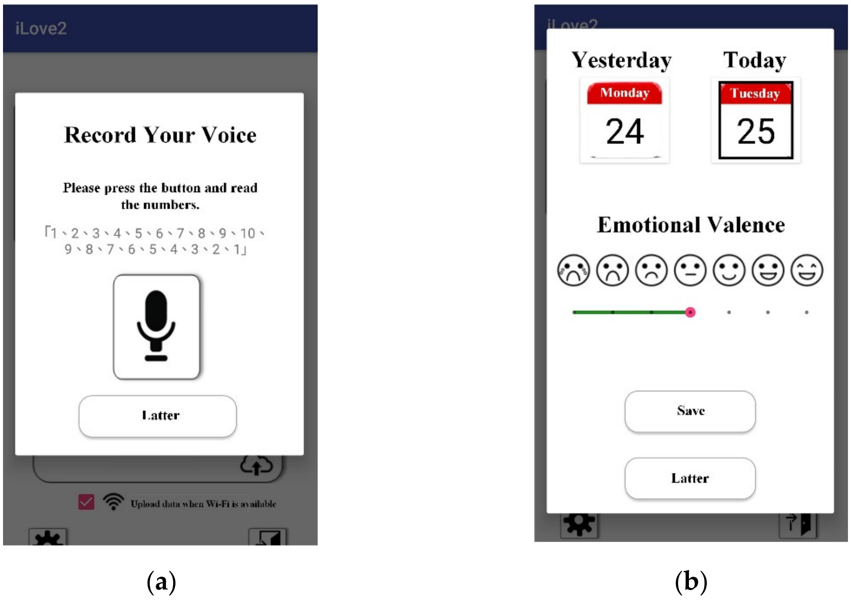
Table 1. Demographic data of the 16 participants.

Characteristics	*n*	%
Sample size	16	100
	Age	
Mean (years)	34.53	—
SD (years)	5.72	—
	Sex	
Male	16	100
	Occupational State	
Employed (stable)	6	37.5
Employed (unstable)	6	37.5
Self-employed	1	6.25
Unemployed	3	18.75
	Drugs even taken	
Amphetamine	7	43.75
Gamma-hydroxybutyrate	3	18.75
Rush	2	12.5
Took within 3 months	7	43.75
Never took	9	56.25
	Period between diagnosis of HIV infection	
Mean (years)	7.87	—
SD (years)	4.26	—
	Anxiety	
Definite (score 11–21)	5	31.25
Doubtful (score 8–10)	3	18.75
No (score 07)	8	50
Mean (score)	7.63	—
SD (score)	5.91	—
	Depression	
Definite (score 11–21)	5	31.25
Doubtful (score 8–10)	4	25
No (score 0–7)	7	43.75
Mean (score)	7.69	—
SD (score)	4.17	—

Ethical approval was obtained from the Institutional Review Board of Taoyuan General Hospital. All the participants signed the written consent and knew their right to withdraw from the study if they wished before participating in the study.

2.3. Data Collection

A self-developed mobile application (app), called iLove, was installed on the individuals' mobile phones to record their daily voice clips and report daily emotional valence. As shown in Figure 1a, the participant used the mobile app to record a daily voice clip while reading "1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1" in Chinese. Voice clips of numerals were selected to eliminate unwanted data variances that result from spoken content. After each voice recording, participants reported their daily emotional valence by selecting responding facial expressions shown in Figure 1b. The seven facial expressions from left to right were recorded as values from one to seven. The collected data were sent to cloud storage at midnight and waited for further use.



iLove2

Record Your Voice

Please press the button and read the numbers.

「1、2、3、4、5、6、7、8、9、10、9、8、7、6、5、4、3、2、1」

☑ Upload data when Wi-Fi is available

Emotional Valence

Figure 1. The self-developed iLove app. (**a**) Daily voice recording screenshot of iLove on which the participant record a daily voice clip while reading "1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1" in Chinese; (**b**) daily emotional valence reporting screenshot of iLove on which the participant reported their daily emotional valence by selecting responding facial expressions.

2.4. Data Processing and Feature Screening

The downloaded voice data first went through data-preprocessing of reducing noises and cutting silent clips, and then an open-source software, openSMILE, IS09_emotion [39], was applied to extracted 384 voice features. As shown in Table 2, the software, openSMILE, automatically produced 384 voice features (i.e., a table grid indicates a feature) from a voice clip. These features were calculated based on 16 descriptors (see the leftmost column of Table 2), comprising root mean square (RMS) frame energy, zero crossings (ZCR), pitch frequency (F0), harmonics-to-noise ratio (HNR), and 12 mel-frequency cepstrum coefficients (MFCC) in accordance to HTK (Hidden Markov Model Toolkit)-based computation [40]. These 16 descriptors were used to capture de-differentiated (partial differential) values (d') [35] to generate 16 non-personalized features [41]. Next, these 32 basic features were individually computed to obtain 12 statistical functionals, comprising mean, standard deviation (SD), skewness, kurtosis, maximum and minimum values, maximum and minimum positions, and range, as well as two linear regression coefficients (offset and slope) with their mean square error (MSE).

These 384 voice features were screened using two statistical methods: ANOVA and correlation. While using ANOVA, two p-values, 0.05 and 0.1, were set as criteria to screen two sets of critical features, whereas one p-value of 0.05 was set when using correlation.

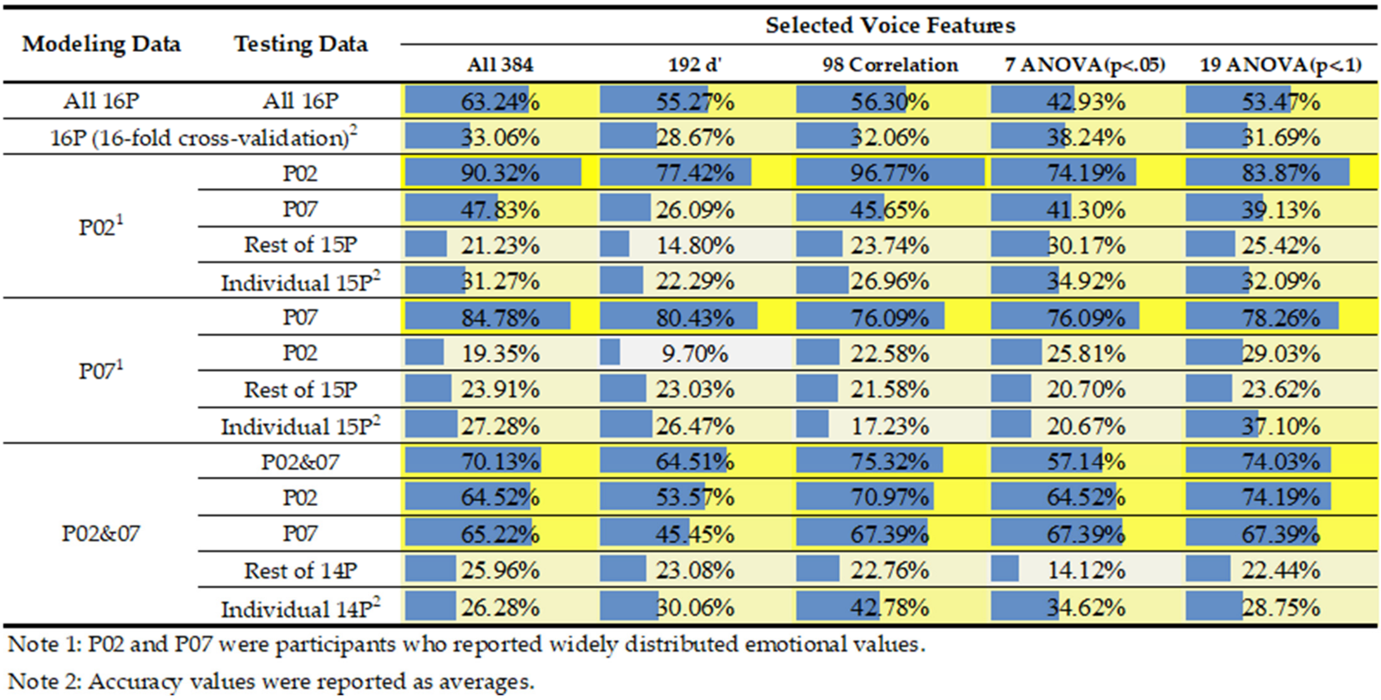
Table 2. Screened voice features using ANOVA and correlation from 384 voice features produced by openSMILE.

Effects		**Original Descriptors (*d*)**												**Delta Regression Coefficients of Descriptors (*d′*)**											
		Mean	**SD**	**Skewness**	**Kurtosis**	**Max Value**	**Min Value**	**Max Position**	**Min Position**	**Range**	**Offset**	**Slope**	**MSE**	**Mean**	**SD**	**Skewness**	**Kurtosis**	**Max Value**	**Min Value**	**Max Position**	**Min Position**	**Range**	**Offset**	**Slope**	**MSE**
RMS						🟨			^	🟨									🟨			🟨	^		
ZCR						🟨 ^	^	🟨 ^	🟨 *	🟨						*					🟨				
F0																🟨									
HNR					🟨			^	🟨																
MFCC	1		🟨										🟨	🟨	🟨			🟨		🟨	^	🟨	🟨	🟨	🟨
	2		🟨			🟨				🟨			🟨		🟨			🟨				🟨		^	🟨
	3				🟨		🟨			🟨	🟨									🟨 *					
	4				🟨		🟨	^								🟨				🟨				🟨	
	5		🟨			🟨	🟨			🟨			🟨		🟨				🟨 *		*	🟨			🟨
	6															^				^					
	7		🟨				🟨			🟨	🟨		🟨					🟨	🟨			🟨			
	8		🟨			🟨	🟨 ^			🟨			🟨		🟨	🟨	🟨	🟨	🟨			🟨			🟨
	9		🟨			🟨				🟨	🟨		🟨		🟨	🟨	🟨		🟨			🟨			🟨
	10		🟨							🟨		🟨	🟨		🟨		🟨	🟨	🟨			🟨	🟨	🟨	
	11		🟨	*								🟨	🟨		🟨										🟨
	12									🟨 *									🟨						

^ indicates $p < 0.1$ using ANOVA; * indicates $p < 0.05$ using ANOVA; a yellow shade indicates $p < 0.05$ using correlation; SD = Stadnard Deviation; MSE = Mean Square Error; RMS = Root Mean Square; ZCR = Zero Crossings; F0 = Pitch Frequency; HNR = Harmonics-To-Noise Ratio; MFCC = mel-frequency cepstrum coefficients.

2.5. Modeling

After extracting and screening voice features, the decision tree algorithm, tree.DecisionTree Regressor() through the scikit-learn library, was used to build various models based on different modeling data, testing data, and voice features. Because participants 2 and 7 were the only two participants who reported widely distributed emotional valence values, their data were specifically used when modeling and testing models. More explanations of the data will be detailed in the result section. As shown in Figure 2, four types of modeling data were used for modeling, comprising data of all 16 participants (All 16P), participant 2 (P02), participant 7 (P07), and participants 2 and 7 (P02&07). While developing models, the random state was set at 100, and the criterion was set as entropy. The max depth was tested with 4, 5, and 6, in which this range of values showed the best model performance. Only the best performance was reported in the result section.



Modeling Data	Testing Data	Selected Voice Features				
		All 384	192 d'	98 Correlation	7 ANOVA (p<.05)	19 ANOVA (p<.1)
All 16P	All 16P	63.24%	55.27%	56.30%	42.93%	53.47%
16P (16-fold cross-validation)[2]		33.06%	28.67%	32.06%	38.24%	31.69%
P02[1]	P02	90.32%	77.42%	96.77%	74.19%	83.87%
	P07	47.83%	26.09%	45.65%	41.30%	39.13%
	Rest of 15P	21.23%	14.80%	23.74%	30.17%	25.42%
	Individual 15P[2]	31.27%	22.29%	26.96%	34.92%	32.09%
P07[1]	P07	84.78%	80.43%	76.09%	76.09%	78.26%
	P02	19.35%	9.70%	22.58%	25.81%	29.03%
	Rest of 15P	23.91%	23.03%	21.58%	20.70%	23.62%
	Individual 15P[2]	27.28%	26.47%	17.23%	20.67%	37.10%
P02&07	P02&07	70.13%	64.51%	75.32%	57.14%	74.03%
	P02	64.52%	53.57%	70.97%	64.52%	74.19%
	P07	65.22%	45.45%	67.39%	67.39%	67.39%
	Rest of 14P	25.96%	23.08%	22.76%	14.12%	22.44%
	Individual 14P[2]	26.28%	30.06%	42.78%	34.62%	28.75%

Note 1: P02 and P07 were participants who reported widely distributed emotional values.

Note 2: Accuracy values were reported as averages.

Figure 2. Comparison of modeling accuracy of using a variety of data set. A longer blue bar indicates a relatively higher accuracy rate, and a brighter yellow shade indicates a better modeling performance.

The developed models predicted a variety of testing data. The model that used modeling data of 16P (Model_16P) tested the same data set of 16P (70% modeling and 30% testing). Sixteen-fold cross-validation (Models_16Folds) was performed using the data of 16P, in which 16 individual participant's data were predicted by the models developed using the rest of 15 participant data in turns. The models that used modeling data of P02 (Model_P02) and P07 (Model_P07) tested the data of P02, P07, the other 15 participants as a whole (Rest of 15P), and the rest of 15 individual participants (Individual 15P). The model that used modeling data of P02&07 (Model_P02&07) tested the data of P02&07 (70% modeling, 30% testing), P02, P07, the other 14 participants as a whole (Rest of 14P), and the rest of 14 individual participants (Individual 14P).

To test the effectiveness of the two statistical methods for model development, five sets of voice features were considered, comprising all extracted features (All 384), 192 de-differentiated features (192 d′), correlation-screened features with $p < 0.05$, ANOVA-screened features with $p < 0.05$, and ANOVA-screened features with $p < 0.1$. While performing correlation and ANOVA, only the data of participants 2 and 7 were used because of their widely distributed emotional values.

3. Results

3.1. Collected Data

Sixteen participants reported 1296 successful data records, in which a participant recorded and uploaded both the voice clip and emotional valence in a day. However, as shown in Figure 3, the participants reported a large amount of high emotional valence. They reported happy states (three facial expressions on the right in Figure 1b) in 82.41% of days. Furthermore, individual differences existed. As shown in Figure 4, participants 1, 3, 5, 6, 8, 9, 10, 11, and 12 tended to report happy states, whereas participant 4 tended to report unhappy states. Participants 14 and 15 reported few successful data records. Participants 2, 7, 12, 13, and 16 reported relatively normal distributed emotions. However, participants 2 and 7 were the only two who reported widely distributed emotional valence values.

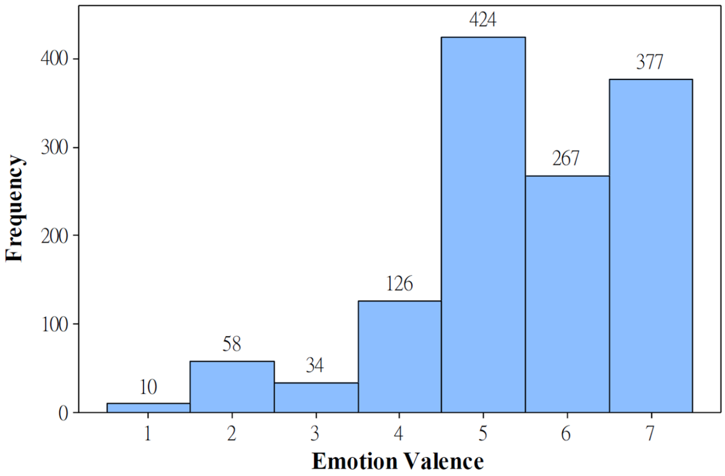


Figure 3. The distribution of data collection against the emotional valence of all 16 participants.

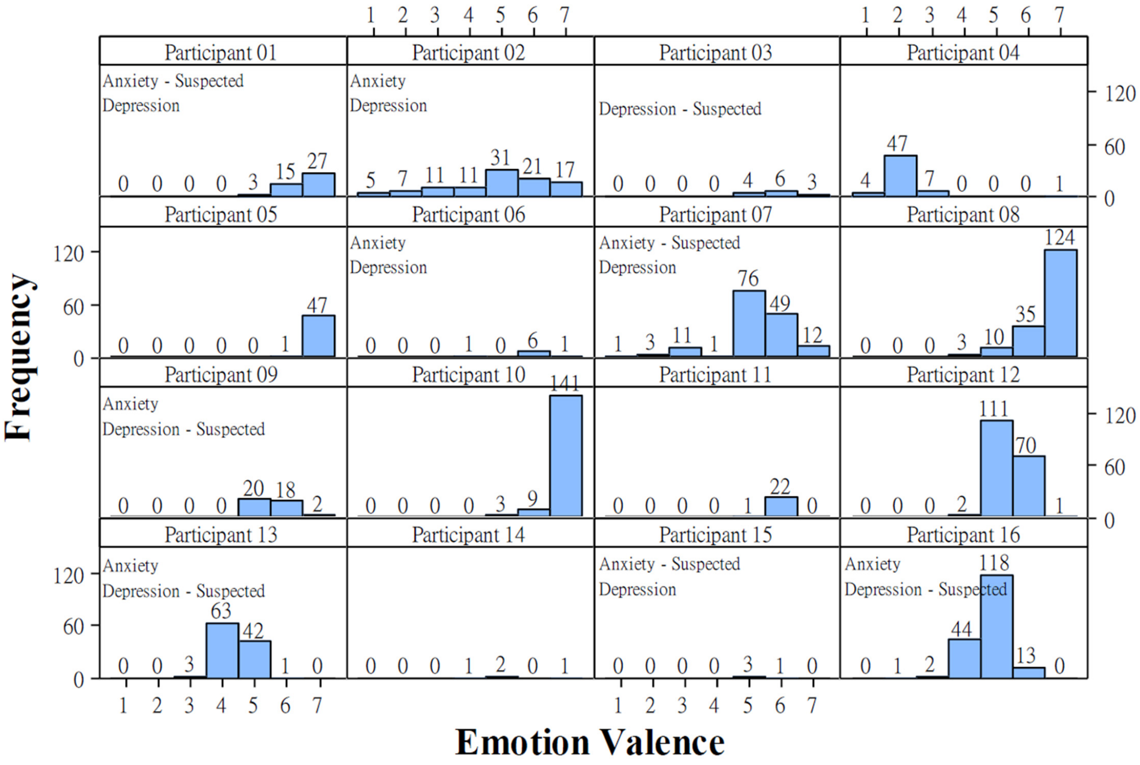


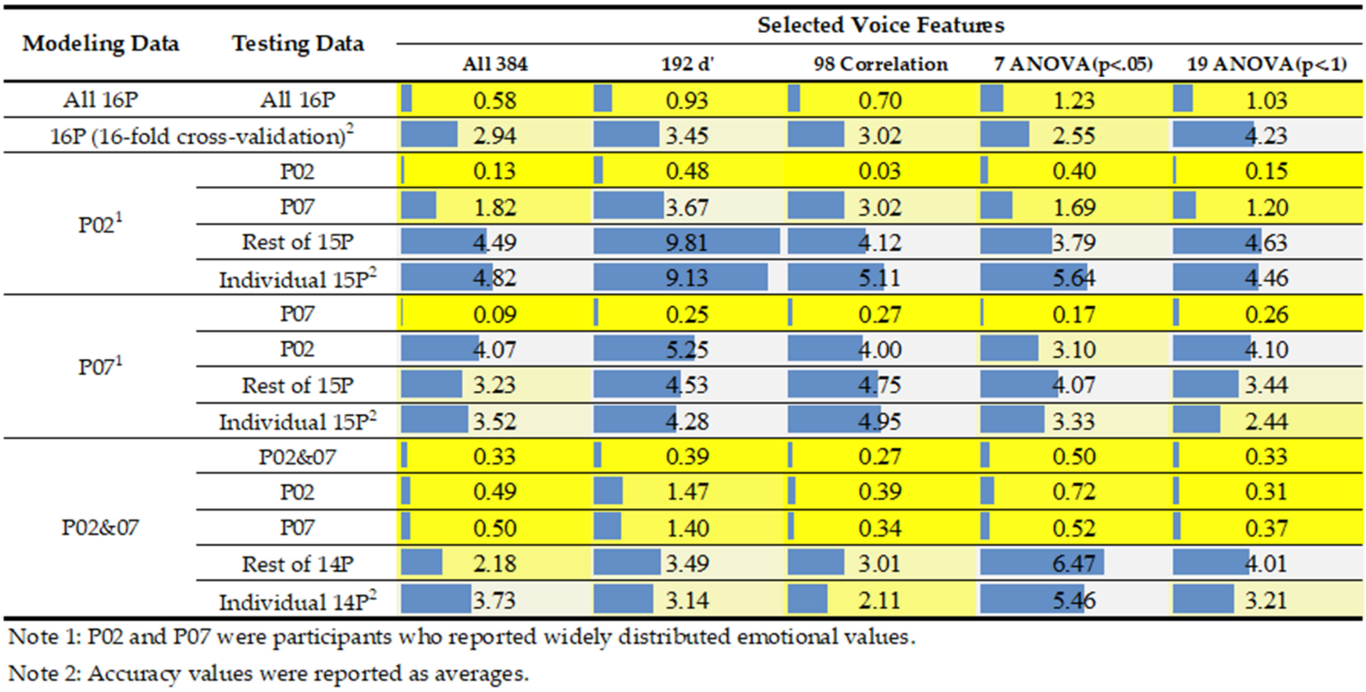
Figure 4. The distribution of data collection of 16 individual participants.

3.2. Critical Features Using Correlation and ANOVA

Table 2 shows critical features screened using correlation and ANOVA. There were 98 correlation-screened features (indicated with the yellow shading), seven ANOVA-screened features with $p < 0.05$ (indicated with *), and 19 ANOVA-screened features with $p < 0.1$ (indicated with ˆ and *). The table also shows all the 384 *d* features (all the combinations) and 192 *d′* features (the half combinations on the right side). These five sets of features were used to develop models.

3.3. Model Performance

The accuracy and mean squared error (MSE) of developed models are shown in Figures 2 and 5, respectively. First, regarding the use of all participants' data, Model_16P that used all 384 voice features obtained the best performance (63.24% accuracy rate and 0.58 MSE) compared to the models that used the other four feature sets (i.e., 192 *d′* features, 98 correlation-screened features, seven ANOVA-screened features, and 19 ANOVA-screened features). However, this is an overfitting result because the performance decreased dramatically (approximately 30%) when doing the 16-fold cross-validation. The results of 16-fold cross-validation showed that using 7 ANOVA-screened features resulted in the best performance (38.24% accuracy rate and 2.55 MSE).



Modeling Data	Testing Data	Selected Voice Features				
		All 384	192 d'	98 Correlation	7 ANOVA(p<.05)	19 ANOVA(p<1)
All 16P	All 16P	0.58	0.93	0.70	1.23	1.03
16P (16-fold cross-validation)[2]		2.94	3.45	3.02	2.55	4.23
P02[1]	P02	0.13	0.48	0.03	0.40	0.15
	P07	1.82	3.67	3.02	1.69	1.20
	Rest of 15P	4.49	9.81	4.12	3.79	4.63
	Individual 15P[2]	4.82	9.13	5.11	5.64	4.46
P07[1]	P07	0.09	0.25	0.27	0.17	0.26
	P02	4.07	5.25	4.00	3.10	4.10
	Rest of 15P	3.23	4.53	4.75	4.07	3.44
	Individual 15P[2]	3.52	4.28	4.95	3.33	2.44
P02&07	P02&07	0.33	0.39	0.27	0.50	0.33
	P02	0.49	1.47	0.39	0.72	0.31
	P07	0.50	1.40	0.34	0.52	0.37
	Rest of 14P	2.18	3.49	3.01	6.47	4.01
	Individual 14P[2]	3.73	3.14	2.11	5.46	3.21

Note 1: P02 and P07 were participants who reported widely distributed emotional values.

Note 2: Accuracy values were reported as averages.

Figure 5. The comparison of modeling MSE of various data set. A longer blue bar indicates a relatively higher value, and a brighter yellow shade indicates a better modeling performance.

Second, while using individual modeling data of P02 and P07, Model_P02 and Model_P07 had excellent performance. When predicting their own data sets, Model_P02 using 98 correlation-screened features had the best performance (96.77% accuracy rate and 0.03 MSE, see Figure 6a for a visual representation), whereas Model_P07 using the data of All 384 features had the best performance (84.78% accuracy rate and 0.09 MSE). When the two models predicted each other's data, Model_P02 using the data of All 384 features had the best performance to predict the data of P07 (47.83% accuracy rate and 1.82 MSE), whereas Model_P07 using 19 ANOVA features had the best performance to predict the data of P02 (29.03% accuracy rate and 4.1 MSE). When the data of the rest of 15 participants were tested, Model_P02 using 7 ANOVA-screened features obtained the best performance for predicting Rest of 15P (30.17% accuracy rate and 3.79 MSE) and Individual 15P (34.92%

accuracy rate and 5.64 MSE), whereas Model_P07 obtained the best performance for predicting Rest of 15P (23.91% accuracy rate and 3.23 MSE) using All 384 features and Individual 15P (37.1% accuracy rate and 2.44 MSE) using 19 ANOVA-screened features.

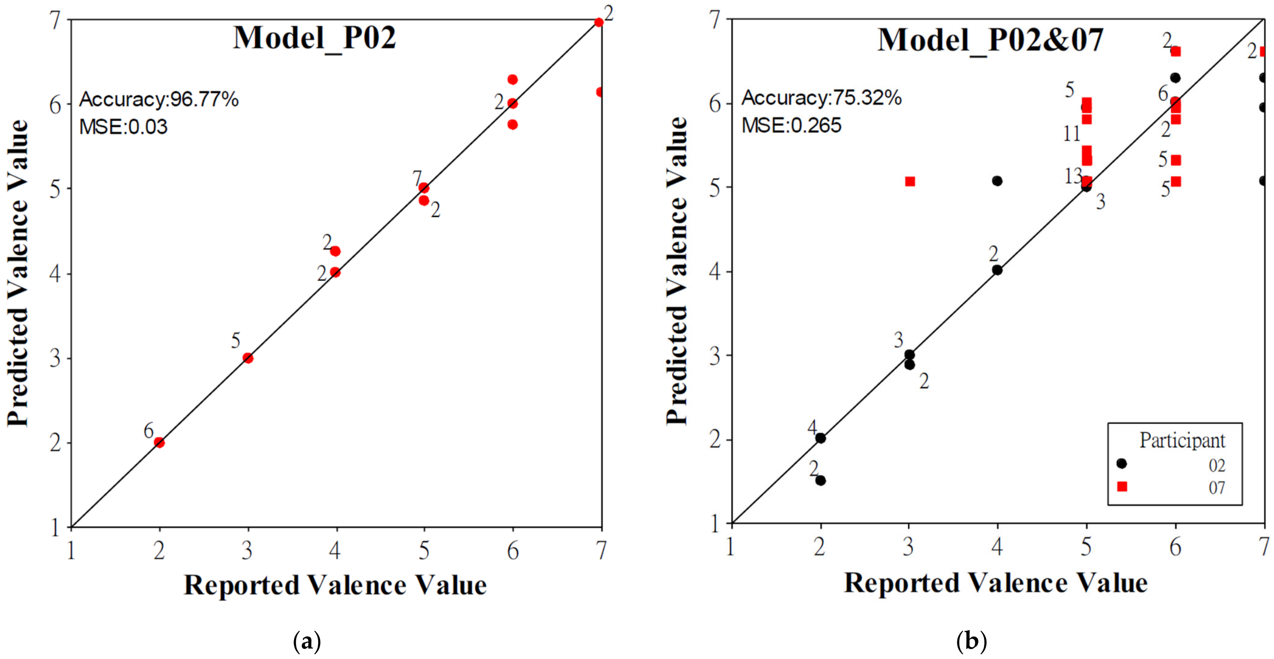
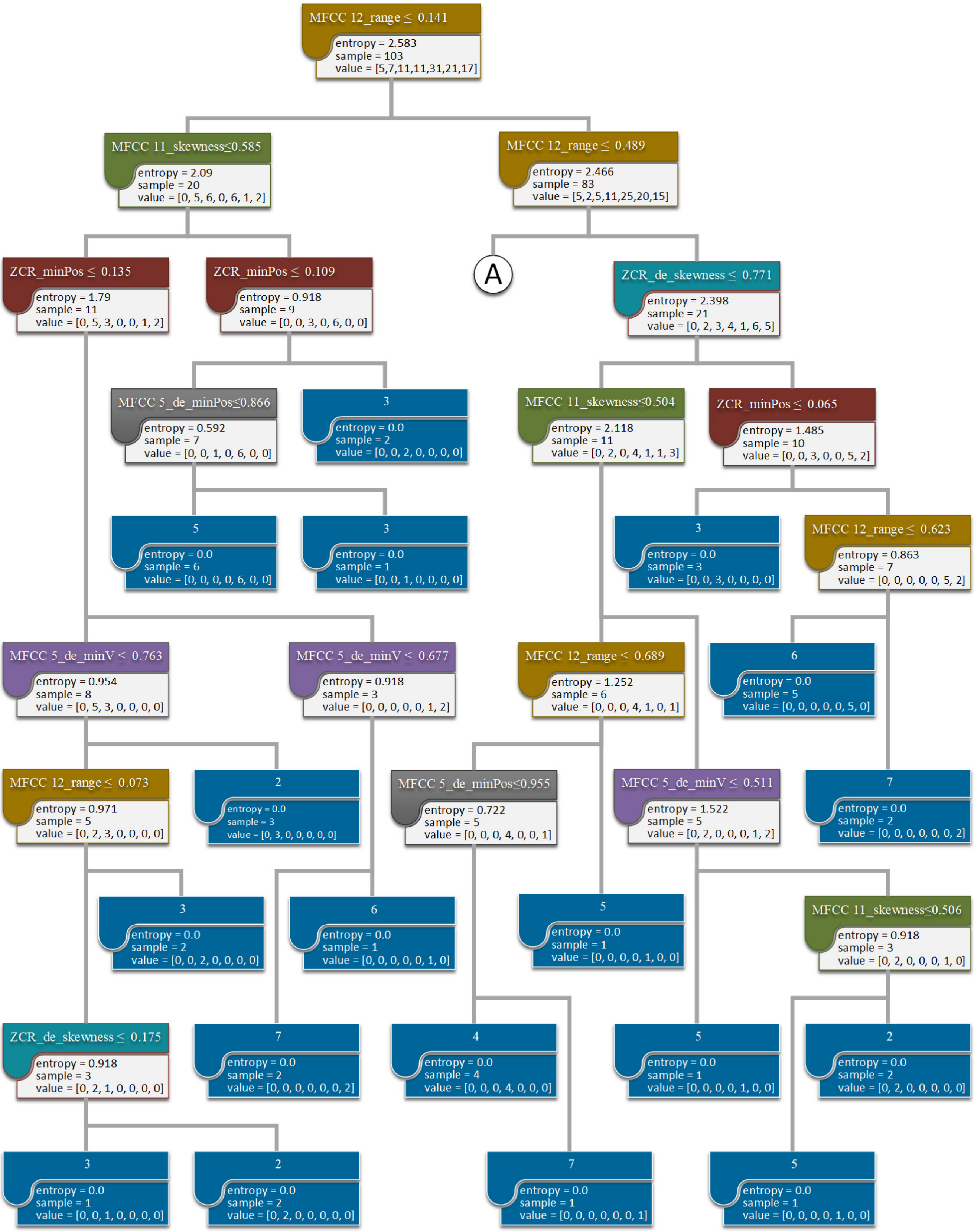


Figure 6. Visualization of model performance. (**a**) Performance of Model_P02 using 98 correlation-screened features (70% modeling, 30% testing); (**b**) performance of Model_P02&07 using 98 correlation-screened features (70% modeling, 30% testing). The number on the symbols represents the number of data with the same predictions.

Finally, while developing models using combined data of participants 2 and 7, Model_P02&07 using 98 correlation-screened features obtained the best performance for predicting the combined data sets (75.32% accuracy rate and 0.27 MSE, see Figure 6b for a visual representation). Compared to Model_P02 and Model_P07, Model_P02&07 had a much better model performance for predicting the testing data of P02, P07, Rest of 14P, and Individual 14P, although Model_P02&07 cannot compete with Model_P02 and Model_P07 when predicting their own data sets. Model_P02&07 had the best performance of the testing data of P02 (74.19% accuracy rate and 0.31 MSE) using 19 ANOVA-screened features, of P07 (67.39% accuracy rate and 0.34 MSE) using 98 correlation-screened features, of Rest of 14P (25.96% accuracy rate and 2.18 MSE) using all 384 features, and of Individual 14P (42.78% accuracy rate and 2.11 MSE) using 98 correlation-screened features.

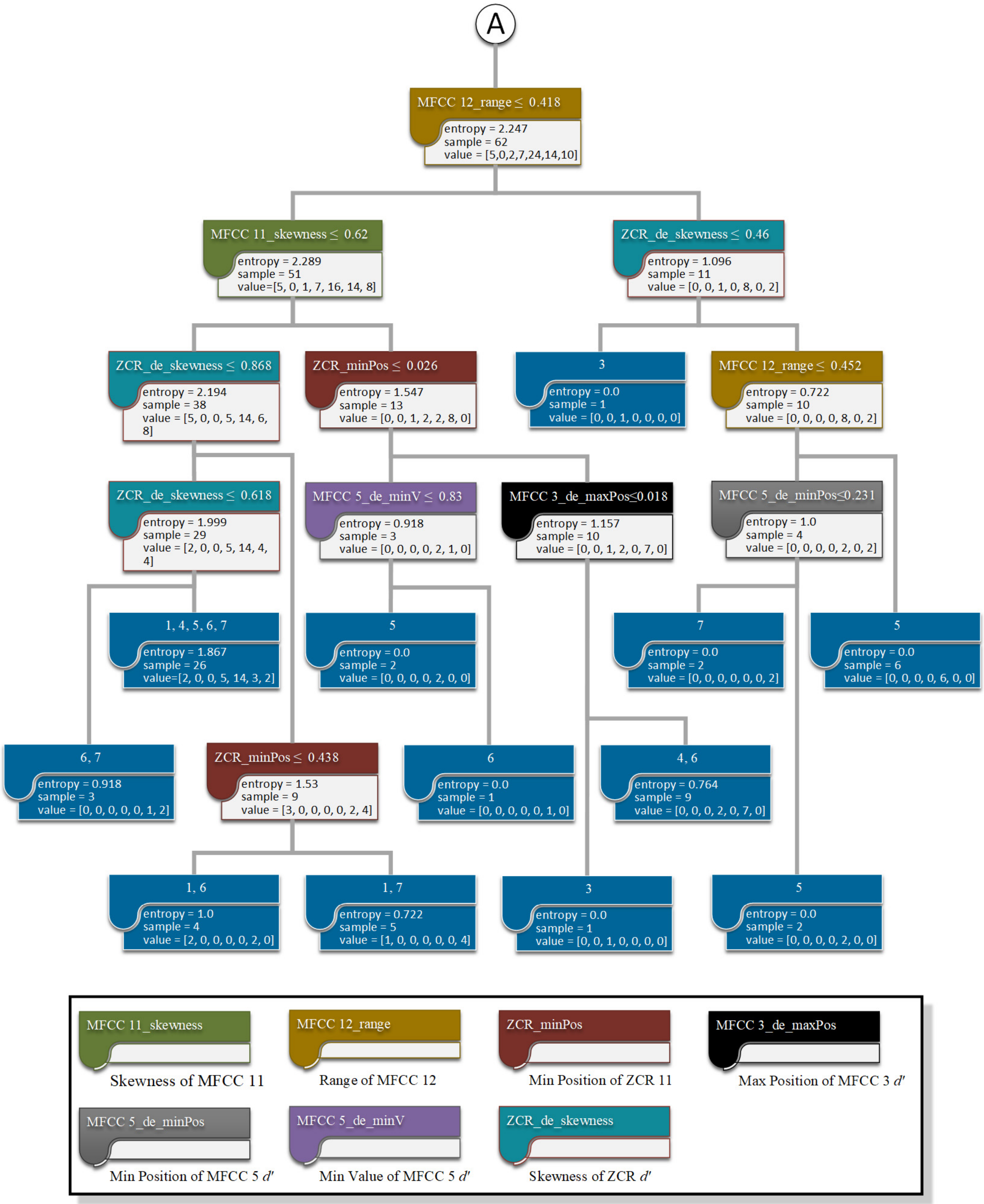
3.4. Decision Tree

All the developed models shown in Figure 2 could be visualized as trees showing decision rules. Due to limited space, only the decision tree of Model_P02 using 7 ANOVA-screened features is detailed here as an example. As shown in Figure 7a, the decision tree shows how the model determines the emotional valence value by using 30 decision rules. These rules are the paths from the first judgment node through the following judgment nodes (seven colored nodes indicate the seven critical features screened using ANOVA) to all the 30 end nodes (represented as solid blue ones). For example, the rightmost end node in the last level (Figure 7b) shows the prediction made with a rule, IF MFCC 12_range > 0.141 AND MFCC 12_range $\leq$ 0.489 AND MFCC 12_range $\leq$ 0.418 AND ZCR_de_skewness > 0.46 AND MFCC 12_range $\leq$ 0.452 AND MFCC 5_de_minPos > 0.231 THEN label = 5. This end node also shows two samples that matched the rule and had the assigned prediction of 5. Based on these decision rules, MFCC 12_range is the most critical feature to influence emotional valence. The feature was used as the first judgment node for all the 30 rules and applied the most frequently in the following decisions.



(**a**)

Figure 7. *Cont.*



(**b**)

Figure 7. Visualization of the developed decision tree of Model_P02 using 7 ANOVA-screened features. (**a**) Part one of Figure 7; (**b**) Part two of Figure 7.

4. Discussion

4.1. Limitations of Collected Data

Although it took nearly two years to collect data from 16 PLHIV, the data collection was not ideal. As shown in Figure 3, the collected emotional valence data concentrated on only a few high valence values. The unbalanced data would make the developed model tend to predict the valence as the values with great numbers of data (i.e., 3, 6, and 7). Furthermore, as shown in Figure 4, only two participants (i.e., participants 2 and 7) reported widely distributed valence values. The narrow ranges of emotional valence values reported by the other participants (participants 1, 3, 4, 5, 6, 9, 11, and 15) would result in an overfitting issue when applying all participants' data for developing AI models. That is, the model tended to recognize emotional valence using features related to individuals instead of using the features with emotional properties. The performance comparison of Model_P16 and Models_16Folds confirmed this overfitting issue. The modeling performance dramatically decreased when predicting the emotional valence of an additional patient whose data were not used in modeling. Furthermore, other than the limited amount of collected data, we should question the accuracy of emotional valence values reported by the participants. As shown in Figure 4, there were some participants (i.e., 5, 8, and 10) who consistently reported extremely high emotional valence.

Collecting sufficient and valid data from PLHIV for an extended period is challenging. Although the mobile app automatically showed a reminder to the participants every day, the participants could ignore the reminder or report unsuccessful information due to various reasons that include being busy, feeling troublesome, and forgetting. Strategies, such as providing monetary incentives and monitoring reported data daily by an experimenter, should increase the motivation and data validation for future data collections.

4.2. Model Performance When Using the Modeling Data of Participants 2 and 7

To overcome the limitations of collected data, we tested the data of participants 2 and 7 for developing models. As shown in Figures 2 and 5, Model_P02 and Model_P07, which used merely a single participant's data for modeling, could compete with Models_16Folds that used 15 participants' data when predicting the other participants' emotions. Furthermore, Model_P02&07 that used the two participants' data performed much better than Models_16Folds did when 98 correlation-screened features were used (42.78% vs. 32.06% accuracy rate and 2.11 vs. 3.02 MSE). Hence, widely distributed emotional valence data are critical for developing effective models for practical use.

It was surprising that Model_P02 and Model_P07 had excellent modeling performance when modeling data and testing data are from a single participant. As shown in Figure 6, the two models obtained very high accuracy rates and ideally predicted low emotional valence values. The most remarkable performance (96.77% accuracy rate and 0.03 MSE) was obtained by Model_P02 using 98 correlation-screened features. Compared to participant 7, participant 2 reported relatively wide and even valence values data (Figure 4). Hence, if a participant reported an adequate amount of widely distributed emotional values, a high-performance AI model could be expected for detecting his emotional valence.

4.3. Comparisons of Screened Voice Features

Except for manipulating modeling data, another method we used to overcome the overfitting issue was applying different voice features. As abovementioned, we stated that using all 384 features provided by openSMILE could result in the issue that the model applied the voice features related to individual differences instead of emotion-related features. Hence, the d' features and two statistical methods were applied to screen critical features. The d' features that use the derivative-based method [35] were proposed to produce non-personalized emotional characteristics and hence reduce individual differences [41]. Correlation and ANOVA tested the statistical relationships between emotional valence and voice features. According to our results, the d' features did not show their superiority in modeling over the two statistical methods. Perhaps one reason was that all participants were male, and they had fewer individual differences in voice. As shown in Table 2, the numbers of features screened by correlation, ANOVA with $p < 0.1$, and ANOVA with $p < 0.05$ were 98, 19, and 7, respectively. Although we did not have consistent results to conclude the best method for screening voice features yet, we see the benefits of using statistical methods to reduce the number of voice features effectively. Most importantly, the statistical method's application can overcome the overfitting issue using all the 384 *d* features for modeling. For example, using ANOVA features (either the numbers of 7 or 19), Model_P02 had better performance for predicting the other participants' emotional valence (both individuals and the group). Using 19 ANOVA features, Model_P07 had better performance for predicting the other individual participants' emotional valence. Using any of the three statistical features, Model_P02&07 had better performance for predicting the other individual participants' emotional valence.

4.4. Contribution and Future Research

While mHealth [20–23] has been promoted to help PLHIV manage depression, this preliminary study attempts to integrate mHealth with AI to track PLHIV's emotional valence. We show the potential to develop a low-emotional-valence detection model and statistical methods to use the limited data with currently available data efficiently. Due to the data limitations, we cannot develop an effective AI model for detecting PLHIV's low emotional valence yet. However, we have demonstrated that a general model (i.e., Model_P02&07) could be developed with widely distributed emotional data to provide reasonable performance. Additionally, excellent models (i.e., Model_P02 and Model_P07) could be developed for individuals if they provided widely distributed emotional data. Hence, our future research is to collect more data from participants, especially the participants who can report widely distributed emotional data. However, we do not expect that the data will increase dramatically due to the limited number of participants and the difficulties of reporting data every day. Therefore, the use of statistical methods is necessary to screen critical and meaningful voice features. With more data available, other AI methods could be tested while modeling. This study tested support vector machine, random forest, anomaly detection, and decision tree and found the decision tree was superior to the other three. Due to the characteristic of our collected data (fewer low emotional valence values), the method of anomaly detection may be suitable for developing models when more data are available. Besides testing these methods, the subsequent study is to develop a general model first and install this general model in a new PLHIV to start practical use. After a while, a transfer learning technique can use feedback to adjust the model for personal use.

Once data collection limitations are overcome, our ultimate goal is to develop an effective AI model installed in a PLHIV if he/she agrees. The AI model is expected to detect bad emotional states, and the participant's cellphone will automatically send a message to his/her case manager. The case manager then provides immediate care to the patient and hence reduces the unfortunate event.

5. Conclusions

This preliminary study shows the potential of using daily voice clips collected using a smartphone to develop an AI model for detecting the negative emotional states of PLHIV. With the collected data of the two specific participants, the developed decision-tree models predicted their emotional valence measured in a seven-scale range. The applications of correlation and ANOVA help screen voice features for developing AI models. However, these interpretations are limited due to the number of participants who reported adequate data. The following research is continuing to collect valuable data from more participants. While collecting data, emotional valence data reported by the participants need to be adequate and, most importantly, widely distributed on the emotional valence scale. The ultimate goal of this data collection is to improve the AI model detection of depression for case managers to effectively provide treatment and improve the mental health outcomes of PLHIV.

Author Contributions: Conceptualization, R.F.L., S.-H.C., Y.-P.L. and C.-P.C.; Data curation, R.F.L., Y.-P.L., Y.-J.W. and S.-Y.C.; Formal analysis, R.F.L. and Y.-J.W.; Funding acquisition, S.-H.C. and Y.-P.L.; Investigation, R.F.L., S.-H.C., Y.-P.L., Y.-J.W. and S.-Y.C.; Methodology, R.F.L., S.-H.C., C.-P.C. and Y.-J.W.; Project administration, R.F.L., S.-H.C. and S.-Y.C.; Resources, S.-H.C. and C.-P.C.; Software, Y.-P.L.; Supervision, R.F.L., S.-H.C. and C.-P.C.; Validation, R.F.L. and S.-H.C.; Visualization, R.F.L. and Y.-J.W.; Writing—original draft, R.F.L., Y.-P.L. and Y.-J.W.; Writing—review and editing, S.-H.C., Y.-P.L., C.-P.C. and S.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hospital and Social Welfare Organizations Administration Commission, Ministry of Health and Welfare, grant number 10807, 10908, and 11010 and Ministry of Science and Technology, MOST107-2221-E-155 -033 -MY3.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Taoyuan General Hospital (protocol code TYGH107028 and date of approval 30 October 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: We thank the team members of Comprehensive AIDS Care Center of Taoyuan General Hospital, Chien-Yu Cheng, Yi-Chun Lin, Mei-Hui Lin, Shin-Yen Ku, Na-Lee Sun, Wei-Wei Wu, and Hui-Ting Hsieh, for their dedication to patient care and data collection.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ruffieux, Y.; Lemsalu, L.; Aebi-Popp, K.; Calmy, A.; Cavassini, M.; Fux, C.A.; Günthard, H.F.; Marzolini, C.; Scherrer, A.; Vernazza, P.; et al. Mortality from suicide among people living with HIV and the general Swiss population: 1988–2017. *J. Int. AIDS Soc.* **2019**, *22*, e25339. [CrossRef] [PubMed]
2. Ciesla, J.A.; Roberts, J.E. Meta-analysis of the relationship between HIV infection and risk for depressive disorders. *Am. J. Psychiatry* **2001**, *158*, 725–730. [CrossRef]
3. Cook, J.A.; Burke-Miller, J.K.; Steigman, P.J.; Schwartz, R.M.; Hessol, N.A.; Milam, J.; Merenstein, D.J.; Anastos, K.; Golub, E.T.; Cohen, M.H. Prevalence, comorbidity, and correlates of psychiatric and substance use disorders and associations with HIV risk behaviors in a multisite cohort of women living with HIV. *AIDS Behav.* **2018**, *22*, 3141–3154. [CrossRef]
4. Grov, C.; Golub, S.A.; Parsons, J.T.; Brennan, M.; Karpiak, S.E. Loneliness and HIV-related stigma explain depression among older HIV-positive adults. *AIDS Care* **2010**, *22*, 630–639. [CrossRef] [PubMed]
5. Zeng, C.; Li, L.; Hong, Y.A.; Zhang, H.; Babbitt, A.W.; Liu, C.; Li, L.; Qiao, J.; Guo, Y.; Cai, W. A structural equation model of perceived and internalized stigma, depression, and suicidal status among people living with HIV/AIDS. *BMC Public Health* **2018**, *18*, 1–11. [CrossRef] [PubMed]
6. Rubin, L.H.; Maki, P.M. HIV, depression, and cognitive impairment in the era of effective antiretroviral therapy. *Curr. HIV/AIDS Rep.* **2019**, *16*, 82–95. [CrossRef] [PubMed]
7. Arseniou, S.; Arvaniti, A.; Samakouri, M. HIV infection and depression. *Psychiatry Clin. Neurosci.* **2014**, *68*, 96–109. [CrossRef]
8. Wang, T.; Fu, H.; Kaminga, A.C.; Li, Z.; Guo, G.; Chen, L.; Li, Q. Prevalence of depression or depressive symptoms among people living with HIV/AIDS in China: A systematic review and meta-analysis. *BMC Psychiatry* **2018**, *18*, 1–14. [CrossRef] [PubMed]
9. Padilla, M.; Frazier, E.L.; Carree, T.; Shouse, R.L.; Fagan, J. Mental health, substance use and HIV risk behaviors among HIV-positive adults who experienced homelessness in the United States–Medical Monitoring Project, 2009–2015. *AIDS Care* **2020**, *32*, 594–599. [CrossRef]
10. Miller, T.R.; Halkitis, P.N.; Durvasula, R. A biopsychosocial approach to managing HIV-related pain and associated substance abuse in older adults: A review. *Ageing Int.* **2019**, *44*, 74–116. [CrossRef]
11. Treisman, G.; Angelino, A. Interrelation between psychiatric disorders and the prevention and treatment of HIV infection. *Clin. Infect. Dis.* **2007**, *45*, S313–S317. [CrossRef] [PubMed]
12. Yousuf, A.; Arifin, S.R.M.; Musa, M.L.M.R. Depression and HIV disease progression: A mini-review. *Clin. Pract. Epidemiol. Ment. Health CP EMH* **2019**, *15*, 153. [CrossRef] [PubMed]
13. Wojna, V.; Nath, A. Challenges to the diagnosis and management of HIV dementia. *AIDS Read.* **2006**, *16*, 615–616, 621.
14. Babiloni, C.; Vecchio, F.; Buffo, P.; Onorati, P.; Muratori, C.; Ferracuti, S.; Roma, P.; Battuello, M.; Donato, N.; Pellegrini, P.; et al. Cortical sources of resting-state EEG rhythms are abnormal in naïve HIV subjects. *Clin. Neurophysiol.* **2012**, *123*, 2163–2171. [CrossRef] [PubMed]
15. Cook, J.A.; Grey, D.; Burke, J.; Cohen, M.H.; Gurtman, A.C.; Richardson, J.L.; Wilson, T.E.; Young, M.A.; Hessol, N.A. Depressive symptoms and AIDS-related mortality among a multisite cohort of HIV-positive women. *Am. J. Public Health* **2004**, *94*, 1133–1140. [CrossRef]

16. Keiser, O.; Spoerri, A.; Brinkhof, M.W.; Hasse, B.; Gayet-Ageron, A.; Tissot, F.; Christen, A.; Battegay, M.; Schmid, P.; Bernasconi, E.; et al. Suicide in HIV-infected individuals and the general population in Switzerland, 1988–2008. *Am. J. Psychiatry* **2010**, *167*, 143–150. [CrossRef]
17. Catalan, J.; Harding, R.; Sibley, E.; Clucas, C.; Croome, N.; Sherr, L. HIV infection and mental health: Suicidal behaviour–systematic review. *Psychol. Health Med.* **2011**, *16*, 588–611. [CrossRef]
18. Shim, E.-J.; Lee, S.H.; Kim, N.J.; Kim, E.S.; Bang, J.H.; Sohn, B.K.; Park, H.Y.; Son, K.-L.; Hwang, H.; Lee, K.-M.; et al. Suicide risk in persons with HIV/AIDS in South Korea: A partial test of the interpersonal theory of suicide. *Int. J. Behav. Med.* **2019**, *26*, 38–49. [CrossRef] [PubMed]
19. Brown, L.A.; Majeed, I.; Mu, W.; McCann, J.; Durborow, S.; Chen, S.; Blank, M.B. Suicide risk among persons living with HIV. *AIDS Care* **2021**, *33*, 616–622. [CrossRef]
20. Kumar, S.; Nilsen, W.; Pavel, M.; Srivastava, M. Mobile health: Revolutionizing healthcare through transdisciplinary research. *Computer* **2012**, *46*, 28–35. [CrossRef]
21. Swendeman, D.; Ramanathan, N.; Baetscher, L.; Medich, M.; Scheffler, A.; Comulada, W.S.; Estrin, D. Smartphone self-monitoring to support self-management among people living with HIV: Perceived benefits and theory of change from a mixed-methods, randomized pilot study. *J. Acquir. Immune Defic. Syndr.* **2015**, *69*, S80. [CrossRef] [PubMed]
22. van Luenen, S.; Garnefski, N.; Spinhoven, P.; Kraaij, V. Guided internet-based intervention for people with HIV and depressive symptoms: A randomised controlled trial in the Netherlands. *Lancet HIV* **2018**, *5*, e488–e497. [CrossRef]
23. Li, Y.; Guo, Y.; Hong, Y.A.; Zhu, M.; Zeng, C.; Qiao, J.; Xu, Z.; Zhang, H.; Zeng, Y.; Cai, W. Mechanisms and effects of a WeChat-based intervention on suicide among people living with HIV and depression: Path model analysis of a randomized controlled trial. *J. Med. Internet Res.* **2019**, *21*, e14729. [CrossRef]
24. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A Database of German Emotional Speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
25. Gangamohan, P.; Kadiri, S.R.; Yegnanarayana, B. Analysis of emotional speech—A review. In *Toward Robotic Socially Believable Behaving Systems-Volume I*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 205–238.
26. Williamson, J.R.; Young, D.; Nierenberg, A.A.; Niemi, J.; Helfer, B.S.; Quatieri, T.F. Tracking depression severity from audio and video based on speech articulatory coordination. *Comput. Speech Lang.* **2019**, *55*, 40–56. [CrossRef]
27. Wang, S.; Ling, X.; Zhang, F.; Tong, J. *Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network, Proceedings of the 2010 International Conference on Measuring Technology and Mechatronics Automation, Changsha, China, 3–14 March 2010*; IEEE: New York, NY, USA, 2010; pp. 437–440.
28. Shah, A.F.; Krishnan, V.V.; Sukumar, A.R.; Jayakumar, A.; Anto, P.B. *Speaker Independent Automatic Emotion Recognition from Speech: A Comparison of MFCCs and Discrete Wavelet Transforms, Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, India, 27–28 October 2009*; IEEE: New York, NY, USA, 2009; pp. 528–531.
29. Mishra, H.K.; Sekhar, C.C. Variational. In *Gaussian Mixture Models for Speech Emotion Recognition. Proceedings of the 2009 Seventh International Conference on Advances in Pattern Recognition, Kolkata, India, 4–6 February 2009*; IEEE: New York, NY, USA, 2009; pp. 183–186.
30. Wu, C.-H.; Liang, W.-B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans. Affect. Comput.* **2010**, *2*, 10–21.
31. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
32. Ekman, P.; Friesen, W.V.; Ellsworth, P. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 11.
33. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons: Sussex, UK, 1999; Volume 98, p. 16.
34. 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. [CrossRef]
35. Liu, Z.-T.; Li, K.; Li, D.-Y.; Chen, L.-F.; Tan, G.-Z. *Emotional Feature Selection of Speaker-Independent Speech Based on Correlation Analysis and Fisher, Proceedings of the 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 28–30 July 2015*; IEEE: New York, NY, USA, 2015; pp. 3780–3784.
36. EI Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *4*, 572–587. [CrossRef]
37. Taiwan Centers for Disease Control. AIDS Statistics. 2021. Available online: https://www.cdc.gov.tw/En/Category/MPage/kt6yIoEGURtMQubQ3nQ7pA (accessed on 2 September 2021).
38. Leung, C.; Wing, Y.; Kwong, P.; Shum, A.L.K. Validation of the Chinese-Cantonese version of the Hospital Anxiety and Depression Scale and comparison with the Hamilton Rating Scale of Depression. *Acta Psychiatr. Scand.* **1999**, *100*, 456–461. [CrossRef]
39. Schuller, B.; Steidl, S.; Batliner, A. The Interspeech 2009 Emotion Challenge. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
40. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D. The HTK book. *Camb. Univ. Eng. Dep.* **2002**, *3*, 12.
41. Cao, W.-H.; Xu, J.-P.; Liu, Z.-T. *Speaker-Independent Speech Emotion Recognition Based on Random Forest Feature SELECTION ALGOrithm, Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017*; IEEE: New York, NY, USA, 2017; pp. 10995–10998.