applied sciences

MDPI

Article

Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques

Tae-Wan Kim and Keun-Chang Kwak *

Department of Electronics Engineering, Interdisciplinary Program in IT-Bio Convergence System, Chosun University, Gwangju 61452, Republic of Korea; gyd03002@naver.com
* Correspondence: kwak@chosun.ac.kr; Tel.: +82-062-230-6086

Abstract: This study aims to establish a greater reliability compared to conventional speech emotion recognition (SER) studies. This is achieved through preprocessing techniques that reduce uncertainty elements, models that combine the structural features of each model, and the application of various explanatory techniques. The ability to interpret can be made more accurate by reducing uncertain learning data, applying data in different environments, and applying techniques that explain the reasoning behind the results. We designed a generalized model using three different datasets, and each speech was converted into a spectrogram image through STFT preprocessing. The spectrogram was divided into the time domain with overlapping to match the input size of the model. Each divided section is expressed as a Gaussian distribution, and the quality of the data is investigated by the correlation coefficient between distributions. As a result, the scale of the data is reduced, and uncertainty is minimized. VGGish and YAMNet are the most representative pretrained deep learning networks frequently used in conjunction with speech processing. In dealing with speech signal processing, it is frequently advantageous to use these pretrained models synergistically rather than exclusively, resulting in the construction of ensemble deep networks. And finally, various explainable models (Grad CAM, LIME, occlusion sensitivity) are used in analyzing classified results. The model exhibits adaptability to voices in various environments, yielding a classification accuracy of 87%, surpassing that of individual models. Additionally, output results are confirmed by an explainable model to extract essential emotional areas, converted into audio files for auditory analysis using Grad CAM in the time domain. Through this study, we enhance the uncertainty of activation areas that are generated by Grad CAM. We achieve this by applying the interpretable ability from previous studies, along with effective preprocessing and fusion models. We can analyze it from a more diverse perspective through other explainable techniques.

Keywords: speech emotion recognition; explainable model; deep learning; YAMNet; VGGish; audible feature

Citation: Kim, T.-W.; Kwak, K.-C. Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques. *Appl. Sci.* **2024**, *14*, 1553. https://doi.org/10.3390/app14041553

Academic Editor: Douglas O'Shaughnessy

Received: 21 January 2024
Revised: 9 February 2024
Accepted: 12 February 2024
Published: 15 February 2024

1. Introduction

In modern society, speech is used in various fields, including human–computer interaction, emotion analysis, and voice-based services. Speech is a representative way for users to express information, emotions, and thoughts. Therefore, it has become important to accurately receive and analyze speech data acquired from computers, smartphones, smartwatches [1], and smart homes to understand the emotions and thoughts of users. The information obtained from speech data is used to improve the quality of products and services. It is analyzed in various fields, such as mental state recognition, emotion classification, emotional analysis, the Internet of Things, and speech recognition systems [2]. It is also used for decision-making and health-status monitoring by understanding the emotional states of individuals or groups [3]. Unnecessary elements, such as environmental noise, breakage, and errors in speech classification add uncertainty to the inferred results. This leads to distrust in the model in terms of the correctness of the area focused on

and the performed classification. Applying appropriate preprocessing can remove these unnecessary elements while increasing the quality of voice data [4].

Additionally, appropriate feature extraction methods are required for analyzing speech complexity and diversity [5]. A spectrogram shows the shape of the extracted features, which allows visual inspection of the frequency component changes over time. Moreover, it can be analyzed by viewing it as an image and then applying a computer vision technique [6]. In a spectrogram study, an explainable model can be applied to determine the relationship between the input and output data to identify which time–frequency domain influences the results of the classification model. When the results of a classification model are analyzed using an explainable model, the reliability of the former is increased, and its performance can be improved by additional research and analysis [7].

This study converted acquired speech signals into spectrogram images for analysis in the time–frequency domain. Unnecessary elements (such as noise, silence, and disconnection, etc.) that could lead to an incorrect decision-making process through the classification model were processed. Such elements are inevitably created in the process of acquiring speech. Instead of manual removal, an algorithm using Gaussian distribution and correlation coefficients was applied to exclude unnecessary divisions, which reduced the learning time and computational resource consumption. Three datasets were used for emotion classification to ensure that the classification model could be generalized. To improve classification performance, a visual geometry group-like audio classification model (VGGish) and yet another mobile network (YAMNet) model were combined to effectively extract emotion-related features from data acquired in diverse environments and lines. The combined model collected more information and extracted more features than the independent models. In the combined model, a late-fusion method was used to classify emotions by learning the same training data independently and then sharing information in the layer before classification. Additionally, by applying three explainable models to the classified results, the emotional feature section in the input time–frequency domain images was confirmed in various ways. The activation region extracted with gradient-weighted class activation mapping (Grad CAM) was analyzed in the time domain, and an algorithm was designed to convert it into a speech file. The converted voice section contained features related to each emotion in the recorded speech. These features can be directly checked by a user to increase the model's reliability and serve as data for further research.

The remainder of this paper is structured as follows: Section 2 presents the related research on spectrogram image analysis using deep learning and interpretable techniques. Section 3 describes the spectrogram transformation methods and Gaussian data selection (GDS) mechanism used in the data preprocessing in this study. Section 4 presents the features of VGGish and YAMNet and the structure and characteristics of the fusion model designed using these models. Section 5 describes the interpretable techniques (Grad CAM, LIME, and occlusion sensitivity) that were used for identifying the activation area. Section 6 explains the three datasets employed for designing a generalized model applicable to various environments and contexts. Section 7 discusses the analysis of the experimental setup and the results, and Section 8 concludes the study.

2. Related Work

2.1. Studies Applying Deep Learning to Spectrogram Images

This section presents studies conducted on spectrograms in which a one-dimensional (1D) signal was converted into two-dimensional (2D) data by applying deep learning to obtain more information and applications. Each study used preprocessing methods and network designs for speech research.

Xi [8] added specific values to insufficient areas to reduce confusion in training and classification when dividing speech data containing imbalanced time. This was achieved by using the IEMOCAP dataset. Padded data were trained and classified by a model that used a convolutional neural network (CNN) and long short-term memory (LSTM). The model was evaluated by weighted accuracy (WA), which assigns weights according to the number

of samples when data are imbalanced. It was also evaluated using unweighted accuracy (UA), which reflects equally regardless of the number of samples. The evaluation of the accuracy of the emotion classification model by WA and UA showed that WA was evaluated at 71.45%, and UA was evaluated at 64.22% when specific values were padded. When the length of speech data was imbalanced, WA reached 68.86%, and UA was at 57.45%, showing slight variations. Badshah [9] constructed a model using a CNN and a fully connected layer. The model's performance was compared with that of the fine-tuned, pretrained AlexNet using seven-emotion data contained in the Berlin dataset for effective speech emotion recognition (SER). AlexNet showed an imbalance in that the accuracy was high in certain classes but low in others; therefore, the mean accuracy was 56.19%. The proposed model was confirmed to be even more effective than AlexNet, with a mean accuracy of 56.38%. Zhang [10] used speech data containing MUOC classroom atmosphere information to check and observe the atmospheric and environmental characteristics of a classroom containing many people. Filter bank, Mel-frequency Cepstral Coefficients (MFCC), and spectrogram features were collected to improve performance and reduce loss. Each feature was processed using an independently trained hybrid neural network combining CNN and LSTM models. The information collected from each feature map, independently processed by the network, was classified into six types of classroom atmospheres. When all the collected features were combined and classified, the lowest loss rate was 24.64%. Raja [11] conducted a study to confirm an effective data type when recognizing emotions using the Berlin dataset speech data. Speech data in 1D audio type was converted to a 2D image type by MFCC and used for training support vector machine (SVM), 1D-CNN, and 2D-CNN models. Each model classified seven emotions for each gender: SVM, using 1D audio-type data had the lowest performance, whereas the 2D-CNN model using 2D image-type data from the MFCC showed the highest performance, achieving 92.5% accuracy in all classes. Thus, transforming speech data into 2D image-type data is effective when classifying speech data. Zheng [12] proposed training NAO robots to respond according to a recognized user's emotion using CNN and random forest (RF) models with the CASIA emotion dataset from China. The data that the robots collected in everyday environments were mixed with noise and had lower purity than the data used for training. Therefore, to improve the generalization performance compared to those of existing methods, features of the speech data were extracted using the CNN model and the extracted feature map was classified using the RF classifier for six emotions. When classified through CNN, a performance of 81.43% was achieved, and when using RF together with CNN, the performance was increased to 84.68%, confirming that using CNN and RF together was effective in the learning process for NAO robots. Heng [13] converted speech data to spectrograms to extract time–frequency features from the IEMOCAP dataset and divided each spectrogram into certain sizes. Emotion-related features were extracted using a CNN, and temporal features were learned and extracted using an LSTM. The extracted features were classified into six classes and evaluated using the WA and UA methods. WA achieved 64%, while UA achieved 56%.

Classification studies applying deep learning to spectrogram images show high performance when the amount of data is small, but most studies using a lot of other data show low accuracy. It can be confirmed that emotional features were not properly extracted in a lot of data.

2.2. Studies Applying Deep Learning and Explainable Methods to Spectrograms

This section presents various studies in which deep learning was applied to 2D spectrogram images, using explainable models. The model's decision-making process was verified by applying explainable techniques to increase its reliability, and activation areas were also used as data.

Yuanyuan [14] converted speech data in the IEMOCAP dataset into spectrogram images and extracted feature maps using AlexNet. The feature maps processed attention using the tanh and softmax activation functions to highlight the essential feature channels

in the attention layer. They identified that the model, applied by CNN and LSTM, achieved a WA of 68.8% and UA of 59.4%; however, the accuracy was effectively improved to a WA of 70.4% and a UA of 63.9% when attention was applied. The focus area of the model was visually confirmed using Grad CAM. Carofilis [15] proposed recognizing a speaker's place of local origin using the VCTK dataset. Speech data were converted into spectrogram images and trained using CNN models. The essential areas created using Grad CAM were combined with a spectrogram to create new feature vectors. These feature vectors were applied to various machine learning algorithms (MLAs) to classify a speaker's place of local origin. The model performances were evaluated using the unweighted average recall (UAR) and macro average accuracy (MAA) and compared with those of other methods, such as the speech wave input and other MLA methods. SVM was measured as 0.348 in MAA and 0.340 in UAR, GaussianNB was measured as 0.262 in MAA and 0.262 in UAR, and passive aggressive was measured as 0.351 in MAA and 0.356 in UAR. Bicer [16] conducted a study on acoustic scene classification, in which sound signals were analyzed to identify the environment and scene to manage and analyze the surroundings. This study used the dataset employed in the DCASE 2016 challenge. Spectrogram images converted into speech signals were trained using ResNet to classify three different environments and identify the areas that influenced the classification decision. The classification achieved an accuracy of 91.82% and the time–frequency area related to the classified class is highlighted using Grad CAM. This highlighted area is then converted into a sound signal for confirmation. Cesarelli [17] used PCG signal data from Kaggle to train phonocardiogram signals with various hyperparameter values and 2D-CNN models and to classify normal and abnormal sounds. That model was evaluated with an accuracy of 86%, and the decision of the model to classify normal and abnormal sounds in the time domain of the heartbeat was confirmed using Grad CAM for the results. Lee [18] conducted a study to predict the recovery of patients who underwent thyroid cancer surgery by converting their speech sounds before and after surgery into spectrograms. The training involved the use of EfficientNet and LSTM, with patients and vocal data provided by DIRAMS. The results showed that the trained model achieved an evaluation score of 0.822 by AUROC. Additionally, Grad CAM was used to confirm the area affected in the time-frequency domain corresponding to the classification.

In a study that applied deep learning and explanatory techniques to the spectrogram, the decision-making process of the model was mainly confirmed using only Grad CAM. However, since the information that can be obtained with a heat map is limited, it is advantageous to use more diverse models and aspects.

3. Materials and Methods

This section describes the method for treating unnecessary data by preprocessing speech data. It also introduces a constructed network that can accurately classify and apply speech data across various contexts. Subsequently, explainable techniques and the method for their use to analyze the output of a model from various aspects are discussed. Figure 1 shows the common architecture of various proposed analyses that apply explainable models for emotion recognition.

In this study, speech data were converted using short-time Fourier transform (STFT), used as input for a CNN-based network, and then divided in time based on the network input size. Data selection was performed for each segmented region using a Gaussian distribution and correlation coefficients to treat the unnecessary data. This process resulted in a reduction of data scale, training time, and memory consumption. The network design independently extracted various features from the selected data, synthesizing them for training and classification. Three explainable models were applied to analyze the classified results. Based on these analyses, the decision-making process was examined, and the activation time–frequency area corresponding to emotions was identified. Subsequently, the activation area identified through Grad CAM was applied to the speech data to directly identify linguistic and phonetic characteristics.

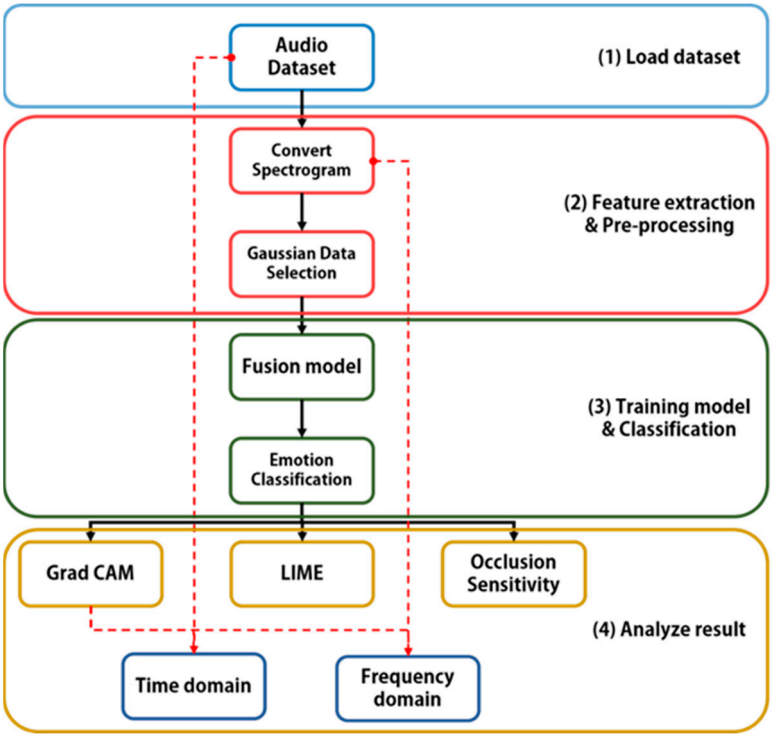

Figure 1. Diagram of speech emotion recognition and analysis using gradient-weighted class activation mapping (Grad CAM) and local interpretable model-agnostic explanations (LIME).

3.1. Convert Speech Data to Spectrogram Images

Speech contains a large amount of information in the time and frequency areas. Using raw speech data is inadequate for capturing all speech information. Therefore, this information can be converted to a spectrogram for study and analysis because it can confirm the change in the time–frequency area, allowing direct visualization. This study converted a speech signal to a spectrogram image using STFT, a frequently used method for conversion, using the parameters listed in Table 1. Figure 2 shows the progressing of STFT, a representative technology for converting speech signals into the time–frequency domain. It divides time into window sizes and applies a fast Fourier transform (FFT) to each section to obtain the frequency information. To obtain the desired frequency range, the filter parameters listed in Tables 2 and 3 were used when converting a spectrogram. The spectrogram was then divided according to the model input size, with an overlap of a certain size incorporated into the time domain.

Table 1. Summary of the related literature.

Author	Purpose	Dataset	Model	Performance
Xi	Speech emotion classification	IEMOCAP	CNN+LSTM	WA: 71.45, UA: 64.22
Badshah	Speech emotion recognition	Berlin	CNN, AlexNet	56.19%, 56.38%
Zhang	Classroom atmosphere classification	MUOC classroom atmosphere information	CNN+LSTM	Error rate: 24.64%
Raja	Speech emotion classification	Berlin dataset (EMO-DB)	SVM, 1D-CNN, 2D-CNN	92.5%
Zheng	Speech emotion classification	CASIA emotion dataset	CNN + random forest (RF)	75.57%
Heng	Speech emotion classification	IEMOCAP	CNN+LSTM	WA: 64, UA: 56
Yuanyuan	Speech emotion recognition	IEMOCAP	AlexNet, CNN+LSTM/Grad CAM	WA: 70.4, UA: 63.9

Table 1. *Cont.*

Author	Purpose	Dataset	Model	Performance
Carofilis	Speaker's place recognition	VCTK	CNN/Grad CAM	MAA: 0.351, UAR: 0.356
Bicer	Acoustic scene classification	DCASE 2016 challenge dataset	ResNet/Grad CAM	91.82%
Cesarelli	Heartbeat sound classification	Heart sound database in Kaggle	2D-CNN/Grad CAM	86%
Lee	Predict sound recovery	Patients and vocal data by DIRAMS	EfficientNet + LSTM/Grad CAM	AUROC: 0.822

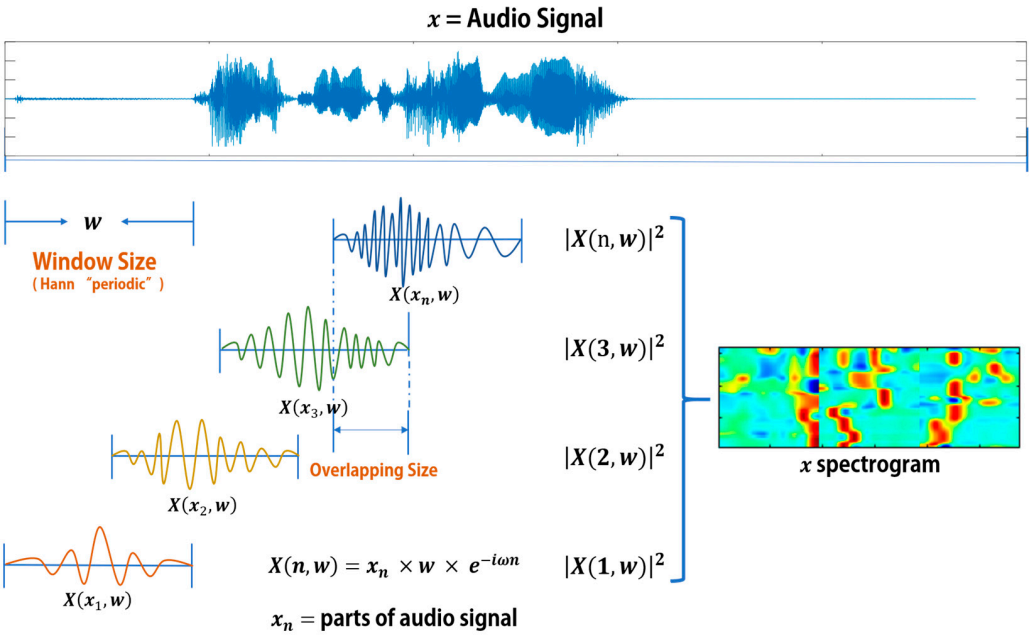

Figure 2. Speech-data-to-spectrogram conversion process.

Table 2. Short-time Fourier transform (STFT) parameters to convert log-Mel spectrogram.

	Window Size	Overlapping Size	Window Type	Range of Frequency
STFT parameter	1200	720	Hann "periodic"	One-way

Table 3. Audio filter parameters involving the fast Fourier transform (FFT).

	Number of Frequency Band	Sampling Rate	Frequency Scale	FFT Length
Audio filter	64 channels	48,000	Mel	1200

3.2. Gaussian Data Selection Mechanisms

When acquiring speech data, matching the length of the data is difficult, even if the same lines are used. The data length varies depending on the speaker's pronunciation speed, interruptions, and noise. The consumption of computing resources increases with the increase in the data scale, which leads to inclusion of unnecessary sections, unrelated to emotions. Various data preprocessing methods are being studied to select and exclude unnecessary sections.

Figure 3 illustrates the application of a Gaussian distribution and correlation coefficients to select unnecessary sections in the acquired data. The selection process excludes

heterogeneous data based on the correlation coefficients in the divided data section rather than simply removing sections with low or high data values by Algorithm 1. After calculating the mean and variance of the divided section, expressed as a Gaussian distribution, the correlation coefficients between all distributions are calculated, normalized, and selected based on the set threshold. Using the above method, sections having a low correlation with the data containing emotional features (i.e., sections that are heterogeneous from the Gaussian distribution of the emotional feature section) are excluded from learning. Accordingly, the data scale and computing resource consumption are reduced while maintaining classification accuracy.

Algorithm 1. Pseudocode of Gaussian data selection algorithm.

Gaussian Data Selection Algorithm

$$Audiodata = \{x_1, x_2, \ldots, x_N\}, N = number\ of\ data$$
$$S_1 = STFT(x_1), S_N = \{S_1, S_2, \ldots, S_N\}$$
$$S_i = \{S_{i1}, S_{i2}, \ldots, S_{iD}\}, D = number\ of\ divided\ for\ each\ data$$
$$\mu_{ij} = mean\left(S_{ij}\right), \sigma_{ij} = std\left(S_{ij}\right)$$
$$G_{ij} = Gaussian\left(S_{ij}; \mu_{ij}, \sigma_{ij}\right)$$
$$C_i = Sum(Correlation\ Coefficient(G_{i1}, G_{i2}, \ldots, G_{iD}), Dim = 1)$$
$$C_i = [C_1, \ldots, C_D], Score_i = normalize(C_i)$$
$$Selected\ data = \left\{S_{ij} \middle| Score_{ij} > T\right\}, T = Threshold$$

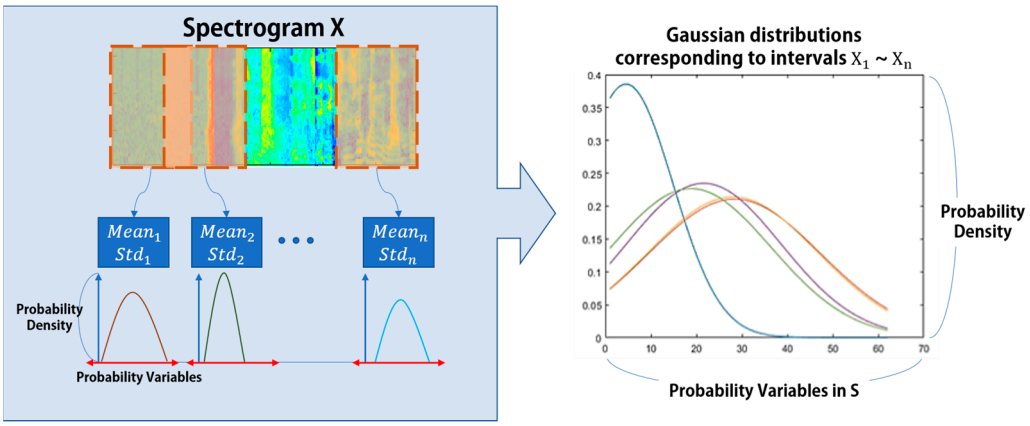

Figure 3. Applying Gaussian data selection to a speech spectrogram.

4. Late-Fusion Model Design with VGGish and YAMNet

4.1. Features of VGGish and YAMNet Models

Figure 4 shows the structure of the VGGish model developed by Google that is used for speech recognition and classification in this study. It is designed as a CNN by modifying an existing visual geometry group (VGG) network for image classification in speech data processing.

The model above, designed to extract features from 2D data rather than 1D data in the time domain, uses a convolutional layer that learns the spatial patterns and features of an input image through weights and batch normalization. An activation function layer, which allows learning highly complex patterns by applying nonlinear features, is used to output the spatial feature vectors of the audio images for analysis and classification. A global pooling layer is not applied to preserve the spatial feature information in the time–frequency domain of the input spectrogram. The number of channels in the fully connected layer is designed to be large enough to generate a feature map larger than the input data. The output information is synthesized and used for classification, utilizing the location characteristics of the time–frequency domain that can be obtained for each pooling layer.

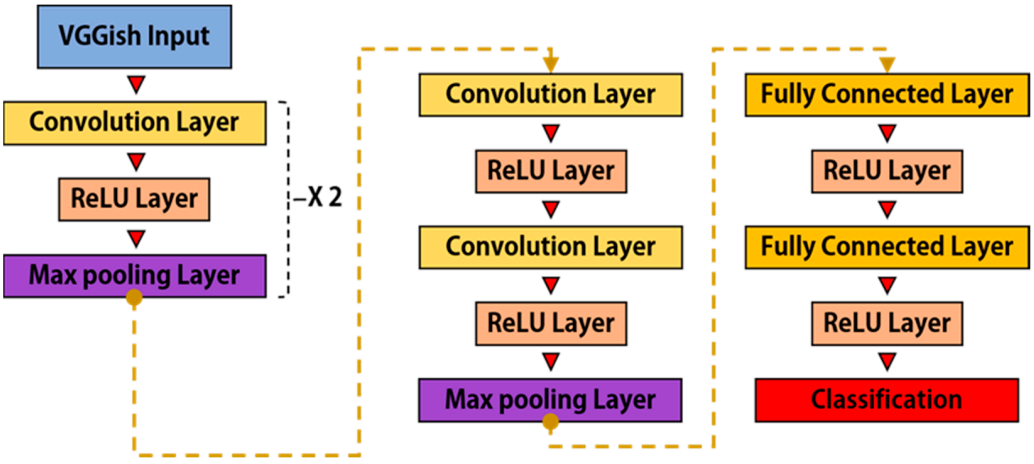

Figure 4. VGGish network structure for video game genre identification using spectrograms and headphones (VGGish), incorporating rectified linear unit (ReLU).

In VGGish, feature maps were extracted for each pooling layer, except for the first pooling layer. The size of the feature map extracted from the network is [24 × 16 × 128/12 × 8 × 256/6 × 4 × 512], respectively. Each feature map aggregates information that can be obtained from each layer by adjusting and fusing the number of channels.

Similar to VGGish, YAMNet, as shown in Figure 5, is an audio classification model developed by Google. It is a transfer learning model that learns from a large-scale dataset to perform 512 different speech recognition tasks, including animal sounds, cars, music, nature, and human conversations. This model is based on a CNN and consists of a convolution layer that can extract spatial patterns and features of multiple images. It also contains a pooling layer, which increases computational efficiency by reducing the spatial dimensions, as well as an activation function layer. Using a global pooling layer before the classification layer, the model focuses on the channel characteristics of the extracted feature map rather than the time–frequency positional characteristics of the feature map input from the previous layer. It is designed to identify the overall features of the input spectrogram image and does not rely on the positional features. Figure 6 shows the classification process of combining the feature maps obtained from each transfer learning model. Here, feature maps in VGGish were obtained by each pooling layer except for the first pooling layer.

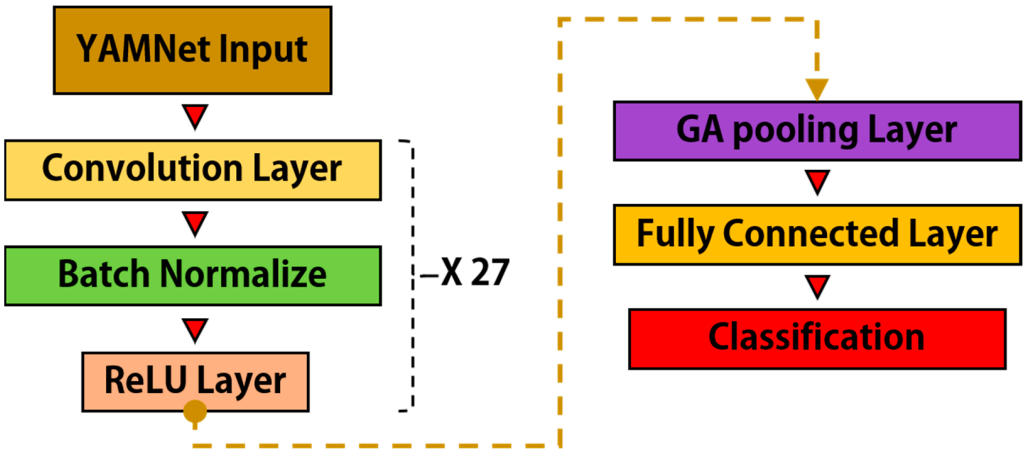

Figure 5. Diagram illustrating the network structure of YAMNet (yet another mobile network) and its integration with the genetic algorithm (GA).

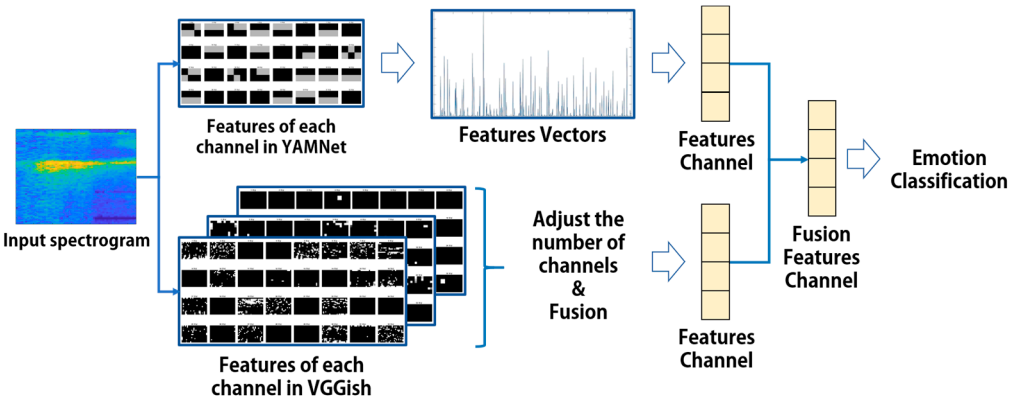

Figure 6. Feature extraction and combination in the fusion transfer learning model.

4.2. Introduction to the Late-Fusion Model

The fusion method combines the information extracted from different data sources or models to achieve a better performance than independent models. More cases can be learned by combining information from models with information from feature extraction methods, which are also used to integrate information from multiple sensor data.

This study selected a late-fusion method to input the same data into YAMNet and VGGish, learn in parallel, and appropriately combine the obtained information to improve the classification results. Figure 7 shows the area activated using Grad CAM in the layer before classification, when the same spectrogram image is input to each model. Owing to the structural differences between the models, VGGish achieves a high location resolution by extracting locational features in the time–frequency domain of the image. Conversely, YAMNet is designed to identify overall features of an image and does not rely on locational features to analyze information. Consequently, there is a difference in analyzing information.

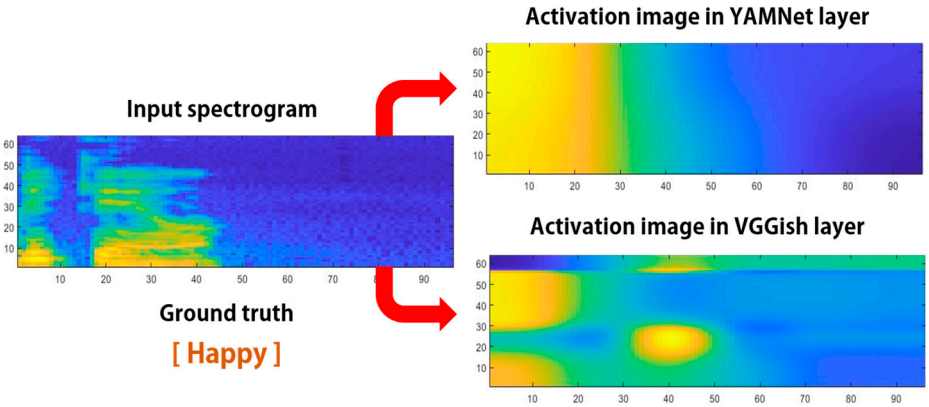

Figure 7. Feature extraction resolution comparison: YAMNet vs. VGGish.

Figure 8 shows the designed model that extracts patterns and features for emotion classification from preprocessed data and integrates information to classify the emotion classes. New feature information is generated by adding the information obtained from each layer of VGGish. Three types of data are synthesized, and a network that synthesizes the feature information and classifies the emotions in the final classification layer is designed. When combining features in a fusion model, the addition or multiplication method is mainly used. However, in the above network, to reduce the influence between the feature channels, a depth-combining method is used to combine the features extracted using the two models. This method not only maintains the information but also classifies it through the interaction of various information in the fully connected layer.

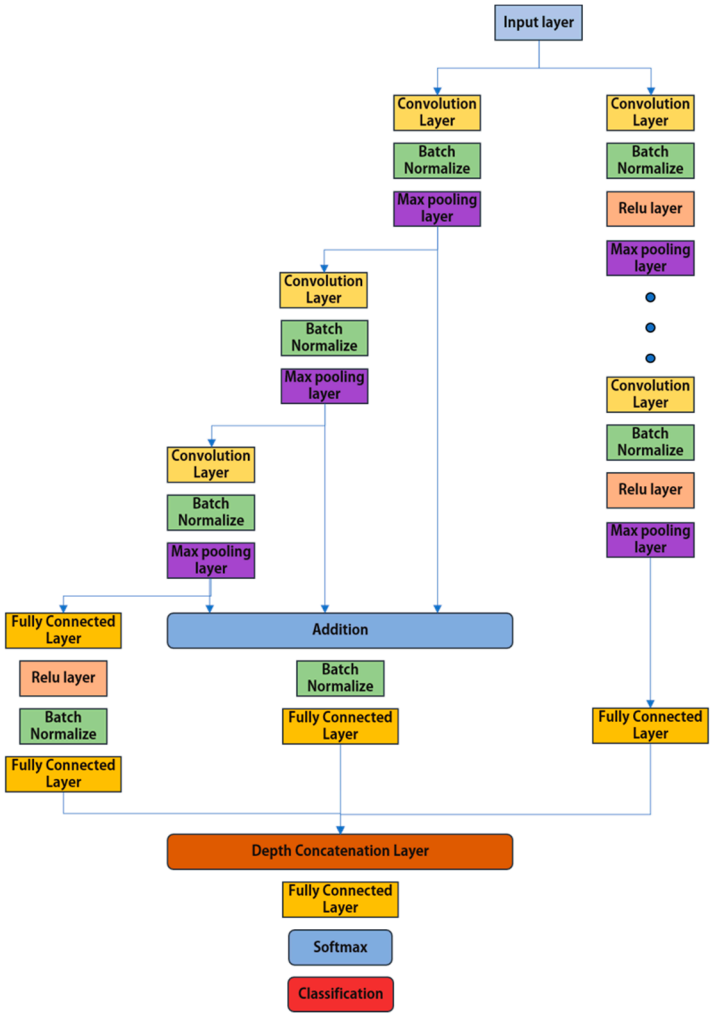

Figure 8. Late-fusion networks of YAMNet and VGGish.

5. Explainable Artificial Intelligence Model

For artificial intelligence (AI) to assist user judgment and be used in activities such as production, confirming the basis and validity of the results is necessary. Figure 9 shows an explainable model (explainable AI) that uses data and labels in conjunction with machine and deep learning models to understand and interpret the factors and areas that affect the input data. This technology can improve the performance of the machine and deep learning modules by confirming incorrect classifications and predictions. It also improves reliability by making it easier for users to understand. A decision tree or linear regression model is simple. Therefore, the decision-making process can be confirmed using the model itself. Other complex models require additional processing, such as Grad CAM, activation, occlusion sensitivity, and LIME.

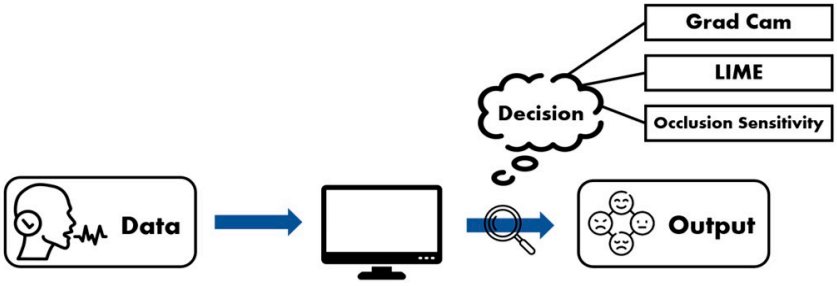

Figure 9. Explanation of explainable AI.

5.1. Grad CAM Technique

Grad CAM creates a focus area that uses the gradient of the classification class, which uses a backpropagation gradient to describe the class classification of a model [19]. An input image is passed through a CNN to generate a feature vector, and the class is classified using this feature vector. As the score of the classified class, the gradient collected for each channel in the previous layer is used as a weight to determine how much each channel influences the class score. The spatial area related to the class is obtained from the feature map, restored to the original size, and expressed visually. As can be seen in Figure 10, Grad CAM uses a spectrogram that is converted to examine the changes in both the time and frequency domains. Subsequently, the spectrogram passes through a CNN, and the results demonstrate which time and frequency section are essential for the classification class by restoring them to the size of the original.

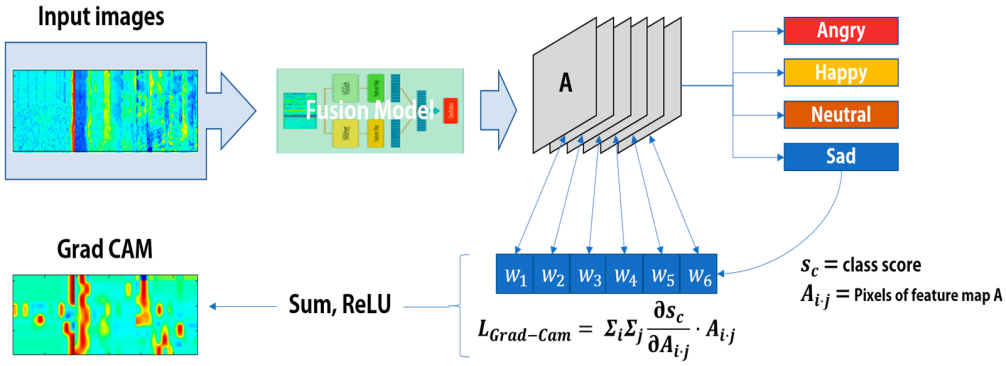

Figure 10. Creation process of Grad CAM.

5.2. LIME Technique

LIME is an explainable technique that determines how small areas, such as pixels of the input data, affect model classification [20]. As shown in Figure 11, the above method transforms a specific area of the input image by adding noise, inputs the transformed areas into the model, and measures the prediction results of the model. A vector is created from the measured feature values for each pixel or specific area, and training is performed using a simple model, such as linear regression, based on the prediction results of the images corresponding to the feature vector.

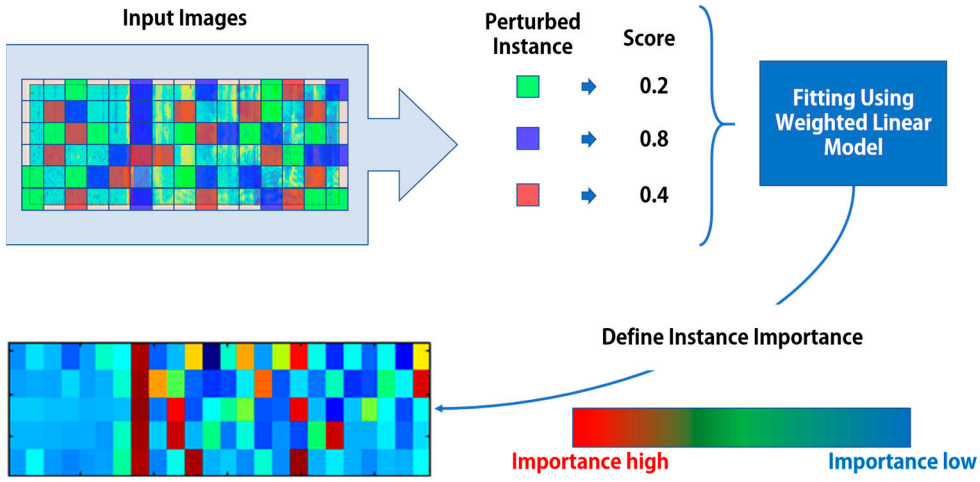

Figure 11. Creation process of LIME.

The trained model calculates the influence of each pixel or specific area of the original image in the image class classification and visualizes it as a heatmap. The predicted results

are measured by transforming various time–frequency pixel areas in the spectrogram image, and the model learns to determine which areas influenced the classification.

5.3. Occlusion Sensitivity Technique

Figure 12 shows how occlusion sensitivity can be achieved by blocking a specific area of the image inputted into the model. The impacted areas were determined by checking the results based on occlusion sensitivity. Continuously altering the position of the blocked area visually illustrates the extent of influence of each position on the model predictions. The above method examines the influence of the area by repeatedly covering the image and a specific part of the voice or text, filling the area with random or average values, and entering them into the model to check for changes in accuracy.

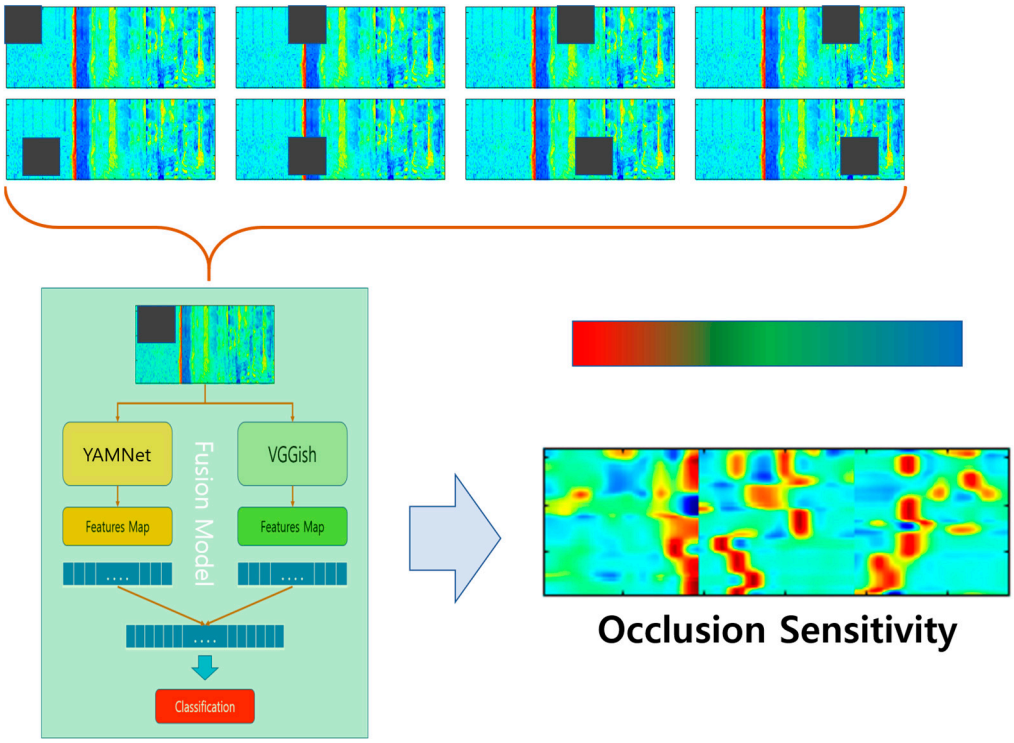

Figure 12. Occlusion sensitivity creation process.

6. Speech Emotion Datasets and Data Selection Method

To evaluate the performance of the proposed model, we use three datasets (CSU 2021, CSU2022, and AI-Hub), acquired from different environments. These datasets were constructed by Korean speech. The entire data from these datasets are combined and used for both training and validation.

6.1. CSU 2021 Speech Emotion Dataset

The above dataset was obtained from 100 people in 2021 at Chosun University, comprising the general public, theater actors, and AI characters. A total of 40 lines, with 10 lines related to each of the four distinct emotions from dramas or movies, were collected. As shown in Figure 13, the speeches were recorded by the participants in a quiet space to mitigate noise, using a Sony stereo lavalier microphone (ECM-LV1). This microphone is used to record ten voices for each of the four emotions from both theater actors and the general public, resulting in 40 voices. The AI character uses Typecast, an AI voice actor service, and the Prosody program to create ten basic voices and ten voices with adjusted speed and tone. All AI voices are produced by a single AI character in the same environment. Consequently, 80 voice data points, 20 for each emotion, were generated. The voices were

recorded through a microphone and configured, as shown in Figure 14. Data samples are as shown in Figure 15.

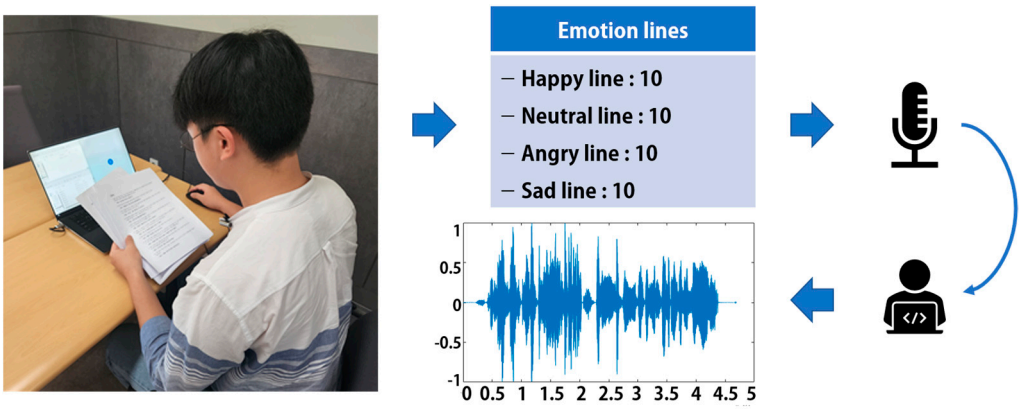

Figure 13. Data acquisition process.

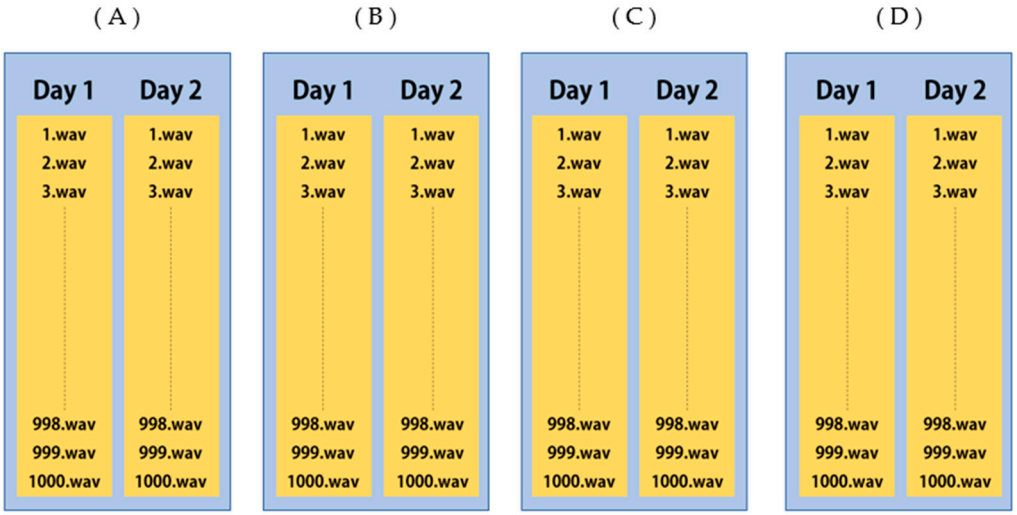

Figure 14. CSU 2021 speech emotion dataset—general public: (**A**) angry, (**B**) happy, (**C**) neutral, and (**D**) sad.

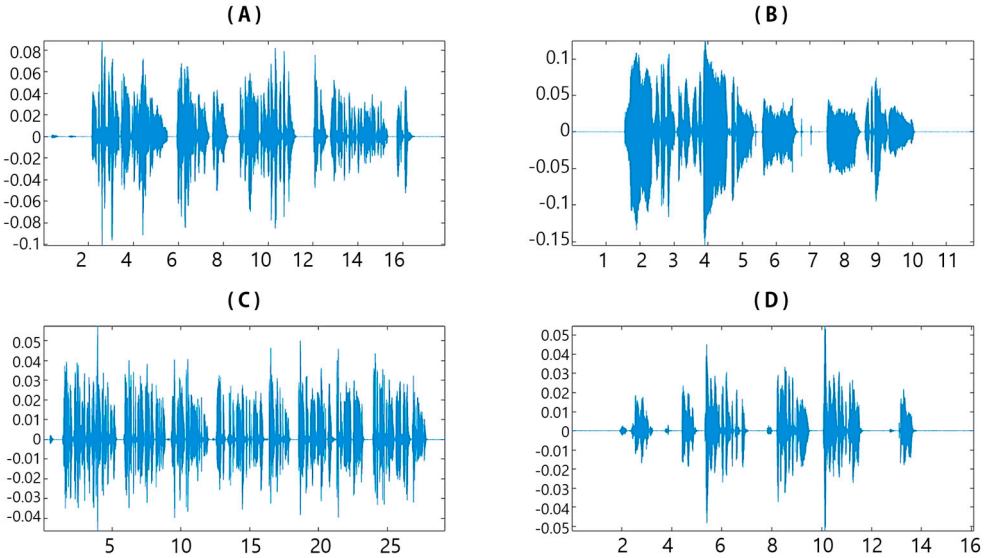

Figure 15. CSU 2021 data samples: (**A**) angry, (**B**) happy, (C) neutral, and (**D**) sad.

6.2. CSU 2022 Speech Emotion Dataset

The dataset above, created by Chosun University in 2022, comprises data from 200 members of the general public. It was acquired for classification into eight emotional states: happy, neutral, angry, sad, chagrin, disgust, fear, and surprise. For each emotion, ten situations and ten short lines suitable for it were selected to prompt participants to act with the intended emotions. A Sony stereo lavalier microphone (ECM-LV1) was used as the acquisition equipment to record signals at 48 kHz. They were recorded alone in a quiet space to minimize noise as much as possible. The data were organized in the WAV format, with each participant contributing ten files per emotion, a total of 80 files per person. The overall dataset, consisting of 16,000 files, was configured as shown in Figure 16. Data samples are as shown in Figure 17.

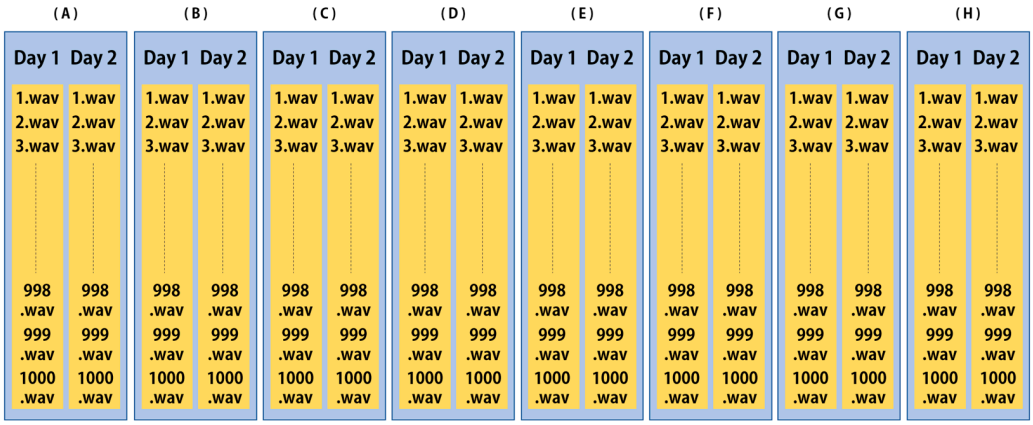

Figure 16. CSU 2022 speech emotion dataset—general public: (**A**) happy, (**B**) neutral, (**C**) angry, (**D**) sad, (**E**) chagrin, (**F**) disgust, (**G**) fear, and (**H**) surprised.

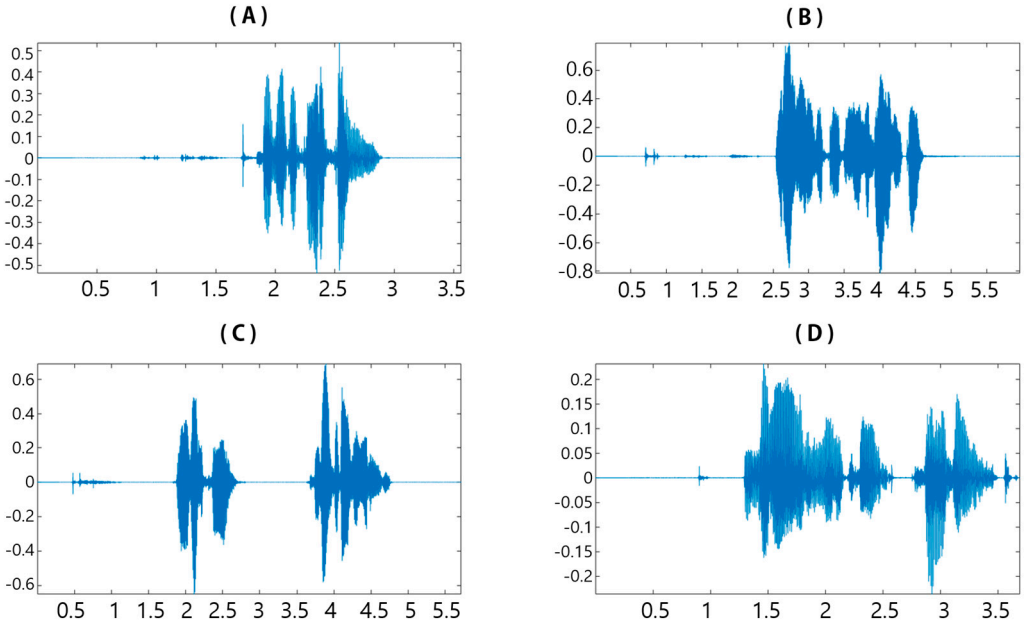

Figure 17. CSU 2022 data samples: (**A**) angry, (**B**) happy, (**C**) neutral, and (**D**) sad.

6.3. AI-Hub Dataset for Emotion Classification

The above data are from the AI-Hub's public data [21]. AI-Hub is an AI-integrated platform that anyone can utilize and participate in by supporting the AI infrastructure (AI data, AI SW API, computing resources) necessary for developing AI technology, products, and services. Among the existing public data, FER 2013 shows a large difference in the

amount of data for each emotion, and the facial emotion data are acquired in low quality. The SFEW data were high-quality data but small in size, with approximately 2000 cases. In the case of CMU-MOSI, the speech domain was limited to movie reviews and the emotion classification was limited to positive and negative; thus, AI-Hub's public data were compiled. The dataset was gathered from among 100 aspiring actors and acting experts, acquiring seven emotional classes: happy, surprised, neutral, fearful, disgusted, angry, and sad. A total of 10,351 videos and voices were included by performing speech and acting, 100 times per emotion. The speech files were configured as shown in Figure 18. Data samples are as shown in Figure 19.

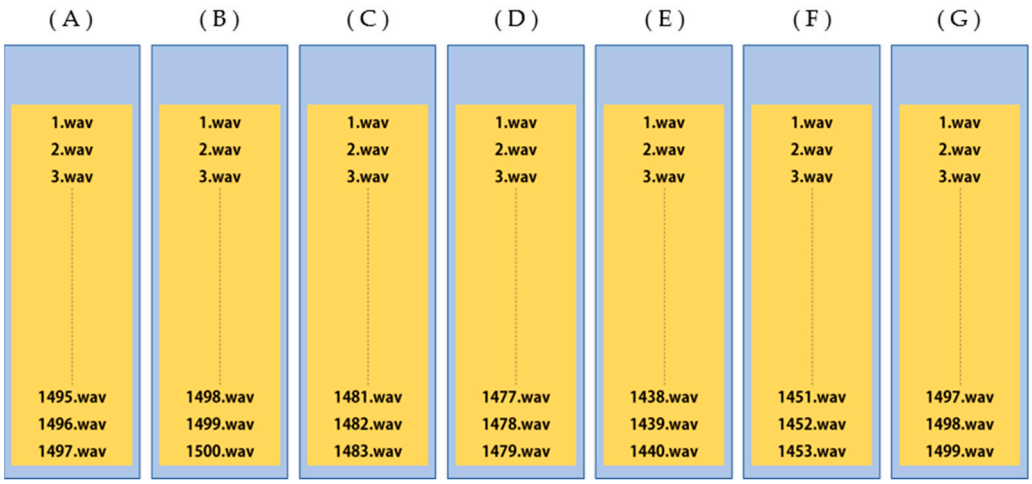

Figure 18. AI-Hub dataset for emotion classification: (**A**) happy, (**B**) neutral, (**C**) angry, (**D**) sad, (**E**) disgust, (**F**) fear, and (**G**) surprised.

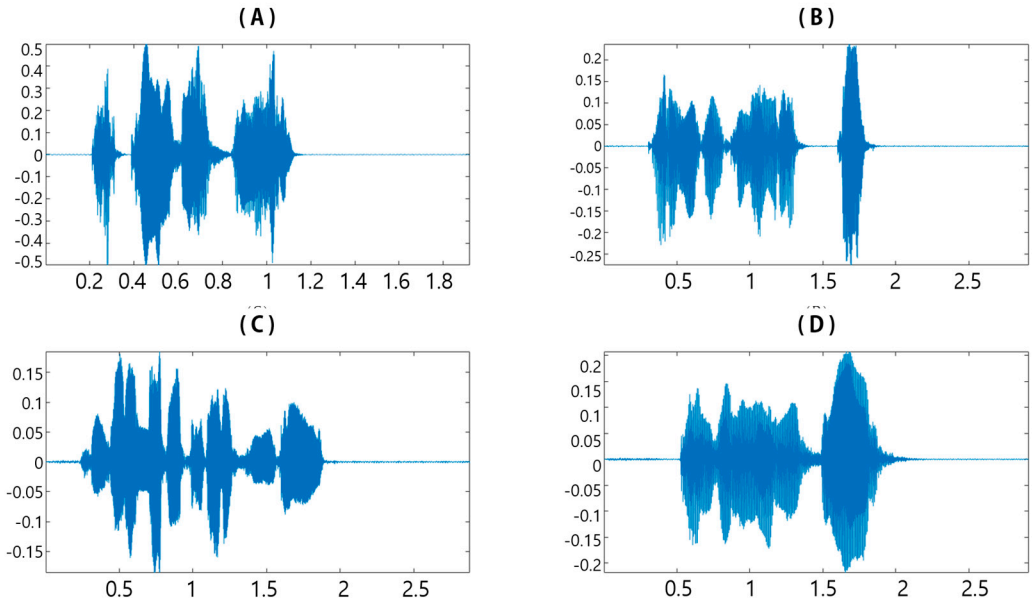

Figure 19. AI-Hub emotion classification dataset samples: (**A**) angry, (**B**) happy, (**C**) neutral, and (**D**) sad.

6.4. Gaussian Data Selection

When speech is divided into time domains according to the model input size, unnecessary sections unrelated to the features to be classified and predicted are generated. To effectively select and exclude this section, GDS preprocessing using a Gaussian distribution and correlation coefficients was performed. Figure 20 shows the number of sequence

segmentations for each file in the training data before applying the GDS and the number of sequence segmentations for each file in the training data after the application. The data before application are divided, consisting of 165,240 data points. After the application, unnecessary sections are reduced to 140,885, a total of 24,385 data points are excluded, and the data are reduced by 15%. These results confirm that the learning time and the time consumed computing resources can be saved.

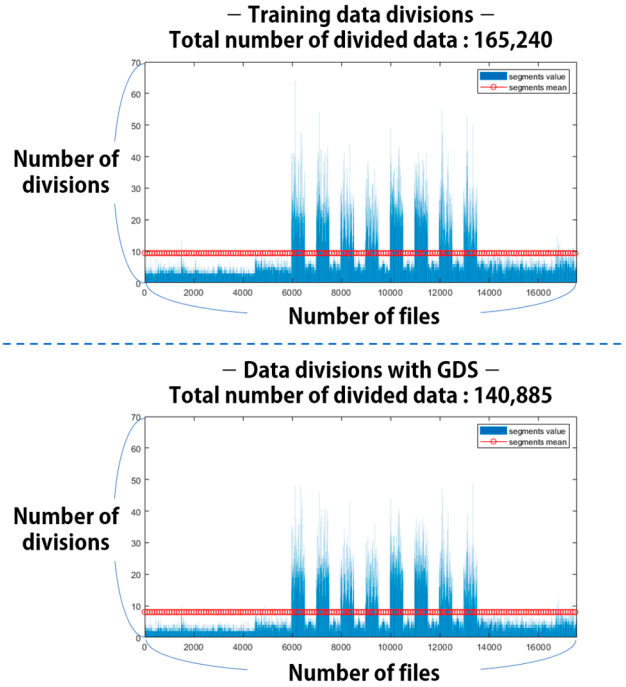

Figure 20. Change in the number of sequence segmentations after Gaussian data selection (GDS).

7. Experimental Findings and Results

7.1. Training Environment and Parameters

The hardware used for the experiments was NVIDIA GeForce GTX TITAN X, 32G RAM, and Intel(R) Xeon(R) CPU E5-1650 v3. The parameters of the audio filter and STFT for conversion to a log-Mel spectrogram were adjusted considering the hardware performance and experimental time. Table 4 lists the hyperparameters used to learn the designed network after data preprocessing. The optimization function uses adaptive moment estimation (Adam), which is frequently implemented in deep learning.

Table 4. Training hyper-parameter.

	Solver	Training Rate	Epoch	Batch Size
Training hyper-parameter	Adam	0.01	20	128

7.2. Accuracy Analysis per Class

Classification was performed on the sequence segmentation sections using the above model. The accuracy values of the sections corresponding to the original data were synthesized, and the average value was used to define the class. Learning and validation data were conducted by mixing data from the entire dataset and dividing it into 85:15 ratios. Figure 21 shows the accuracy of the developed fusion model, demonstrating high accuracy in classifying the four core emotions. As shown in Figure 22, the GDS is applied to remove data that do not affect learning. Its accuracy was analyzed using test data in which the same data selection was applied. The total number of data points was not affected, and as the number of sequence segmentation sections corresponding to each data point decreased by 15%, the learning time decreased, while the accuracy remained similar.

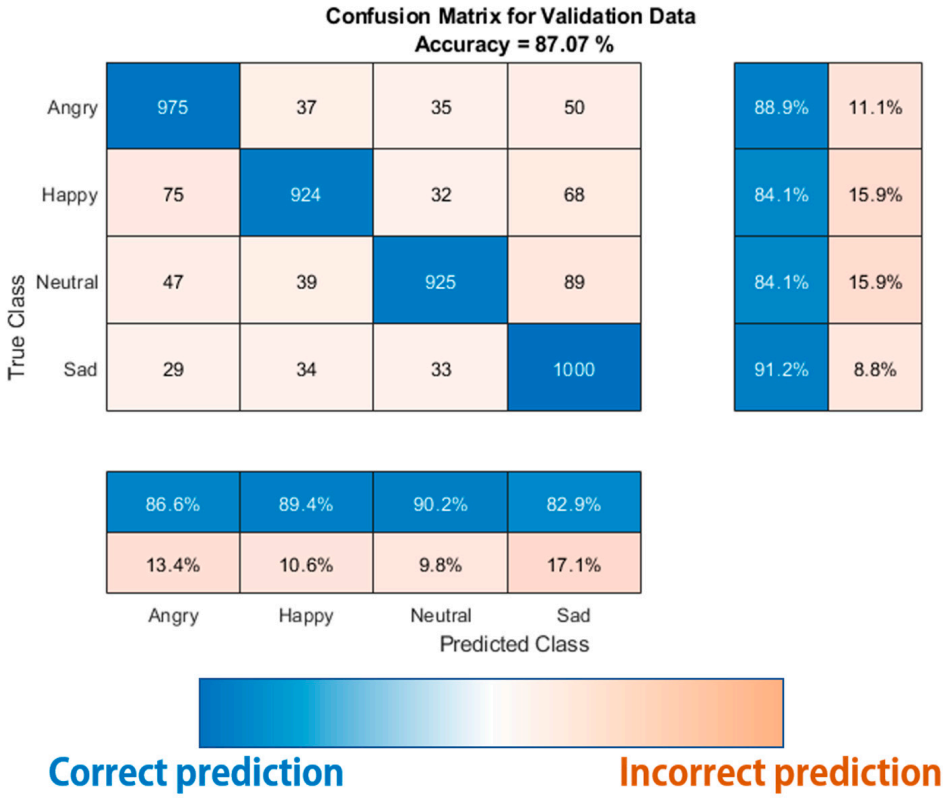

Figure 21. Custom YAMNet–VGGish late-fusion network—classification accuracy.

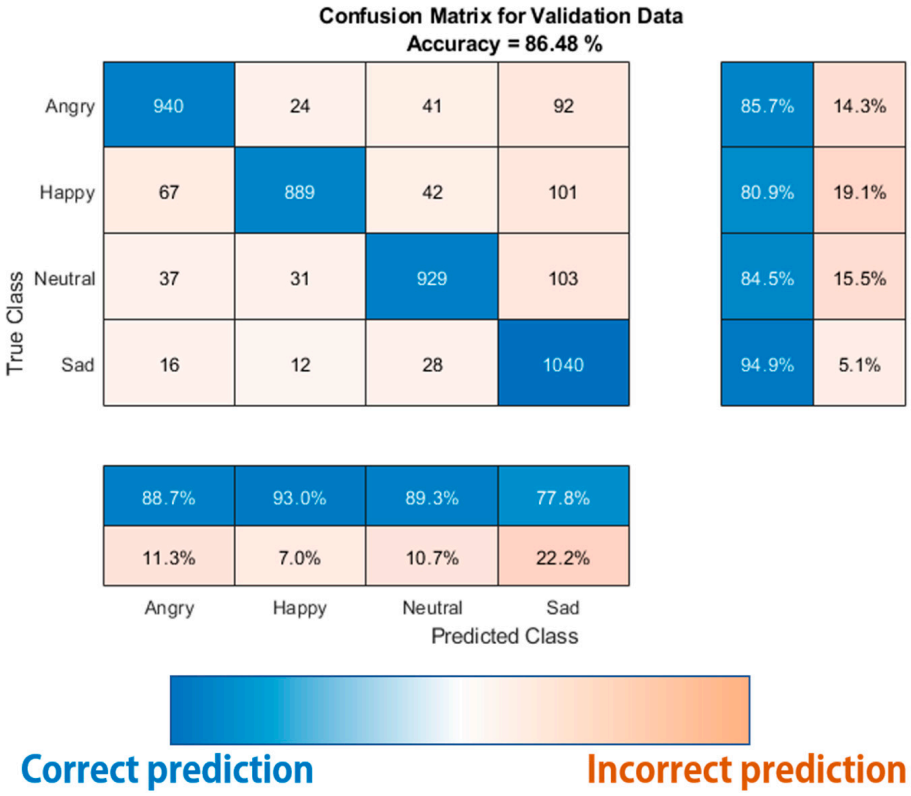

Figure 22. Classification accuracy with Gaussian data selection.

Table 5 shows the evaluation of the classification performance before and after applying the GDS. The results indicate a similar classification accuracy with and without GDS. However, the training time demonstrates a reduction from approximately 90 min to 70 min, resulting in a 22% decrease in time and memory consumption. Following the classification, various explainable models were applied to the network to analyze the focused areas. This analysis aimed to identify the sections used for correct classification and those leading to incorrect classifications. Such insights allow model users to understand the model behavior and facilitate efforts to improve its performance in the future.

Table 5. Comparison of performance before and after Gaussian data selection (GDS).

	Accuracy	Recall	F1 Score	Training Time
Before GDS Results	86.54%	0.8681	0.8653	89.87 (min)
After GDS Results	85.82%	0.8614	0.8583	70.62 (min)

Figure 23 compares the performances before and after applying the GDS to the designed fusion model and existing transfer learning models, VGGish and YAMNet, in terms of accuracy, recall, and F1 score. Recall is an evaluation indicator representing the ratio of what the model predicts as positive to actual positives. The F1 score is the harmonic average of the model precision and recall and is an indicator of the evaluation by considering the two. YAMNet shows the lowest performance, and VGGish outperforms YAMNet. The designed model shows improved performance in terms of all evaluation indicators compared to the existing transfer learning model. There is only a small difference before and after applying the GDS.

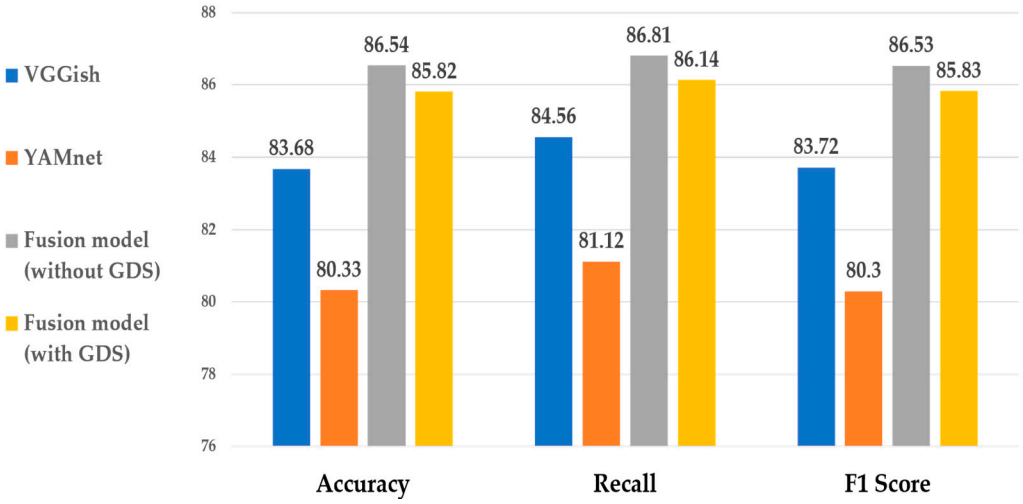

Figure 23. Performance comparison with existing models (accuracy, recall, and F1 score).

7.3. Focus Area Analysis with Explainable Model

Various preprocessing and training were performed to train the time–frequency characteristics related to emotion in the input data. Explainable techniques were subsequently used to confirm the features' components used for classifying the emotions in the model. Figure 24 shows the characteristics of the frequency domain. The activation for each emotion can be determined based on the statistics of the entire dataset. In the case of anger and happiness, relatively high-frequency regions are commonly activated, whereas in the case of neutrals and sadness, lower-frequency regions are activated.

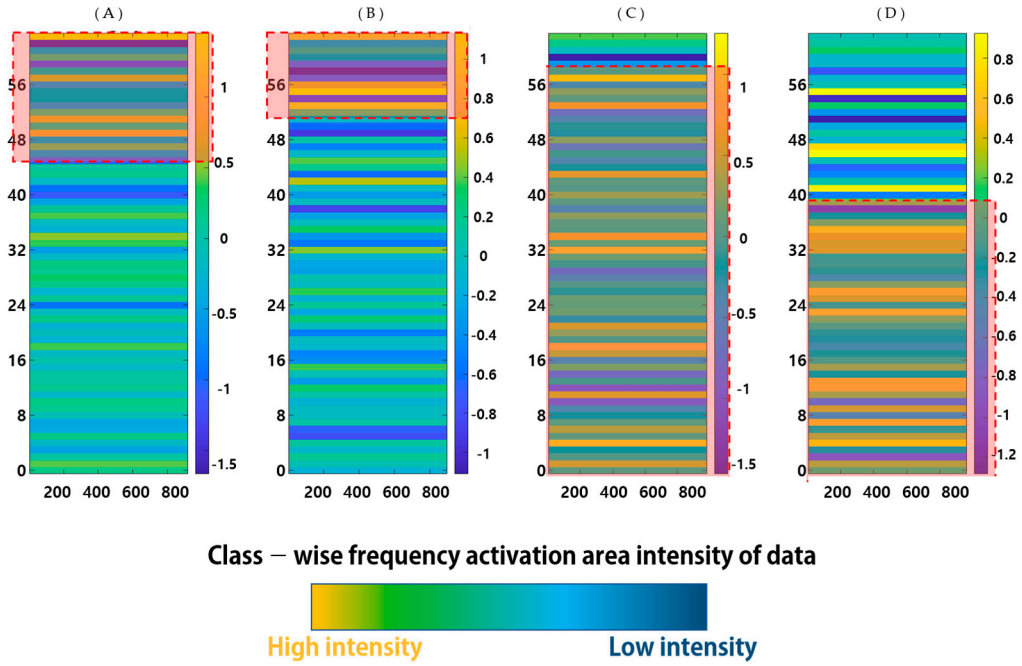

Figure 24. Emotional activation frequency region: (**A**) angry, (**B**) happy, (**C**) neutral, and (**D**) sad.

Figure 25 shows the activated area when the speech obtained using lines related to anger is converted into a spectrogram. The focused area is analyzed using the trained model and Grad CAM. The above method can be seen in the red area, where the time–frequency region influences the decision-making of the model. It is confirmed that the model focuses on areas with large values and combines areas with various values to classify emotions.

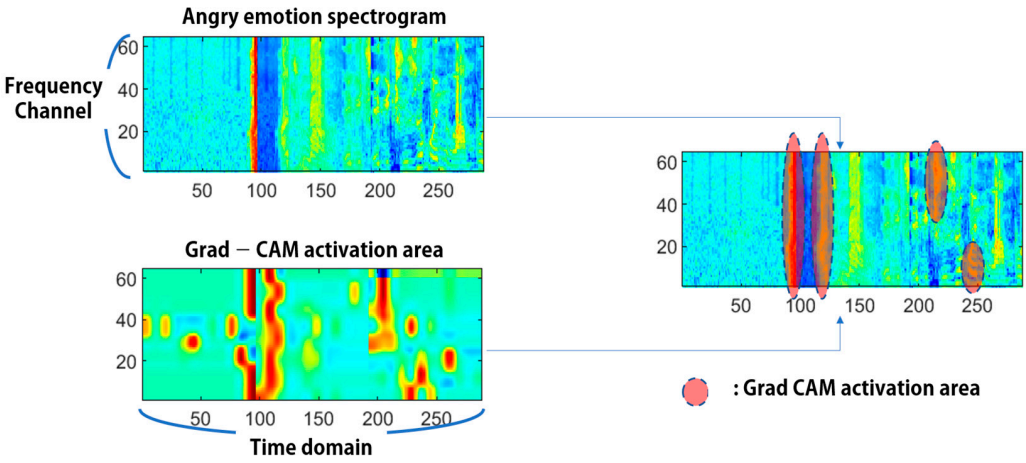

Figure 25. Anger emotion spectrogram and Grad CAM focus area.

Figures 26–29 show images that verify the model's decision-making process by applying Grad CAM, LIME, and occlusion sensitivity to the speech spectrograms. In images in which the explainable technique is applied, areas of the input spectrogram that influence the model's decision-making can be identified. Grad CAM uses gradient information through class scores in the model. Therefore, it has a high level of interpretation regarding which areas are important. In the case of LIME and occlusion sensitivity, by checking the extent to which a certain area influences the model classification, the areas that affect classification differently from Grad CAM can be determined.

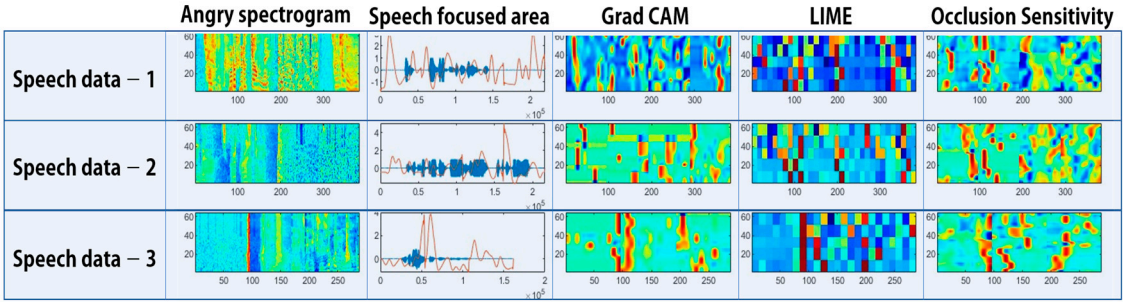

Figure 26. Anger emotion spectrogram (raw signal, Grad CAM, LIME, and occlusion sensitivity).

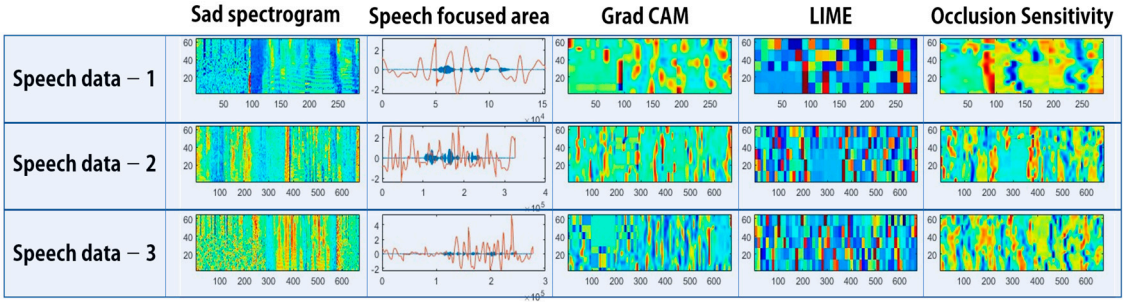

Figure 27. Sad emotion spectrogram (raw signal, Grad CAM, LIME, and occlusion sensitivity).

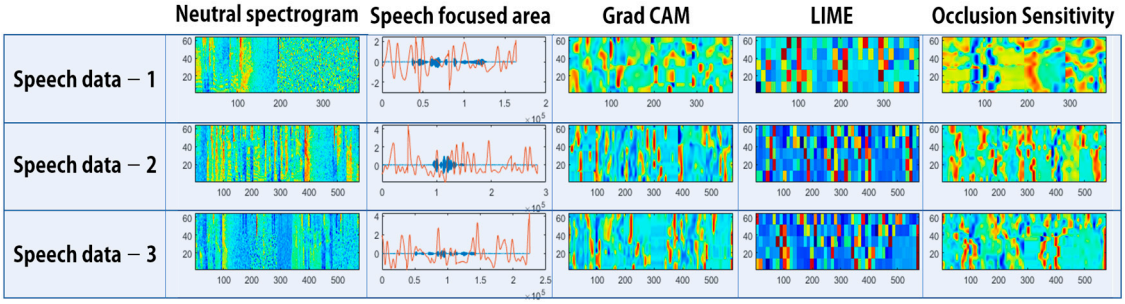

Figure 28. Neutral emotion spectrogram (raw signal, Grad CAM, LIME, and occlusion sensitivity).

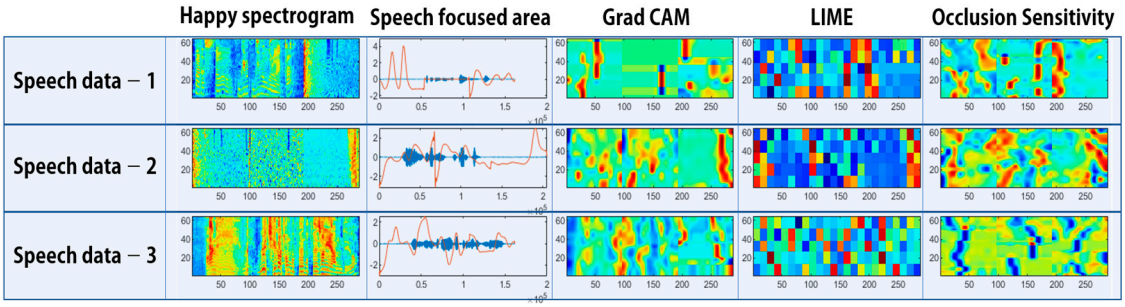

Figure 29. Happy emotion spectrogram (raw signal, Grad CAM, LIME, and occlusion sensitivity).

7.4. Applying the Focus Areas of the Explainable Models to Audio

Figure 30 shows that the influence of the frequency, obtained from Grad CAM, is added to the time domain to examine time features rather than frequency features in the input image. Many frequency areas are considered to be concentrated in specific periods, indicating that words or intonation during these times play an important role in classification. When the parameters used for spectrogram conversion are reversed to restore the size of the original speech signal, it becomes difficult to perceive it audibly due to noise. Therefore, convolution with a filter value of one is applied to a rapidly vibrating

part and postprocessed to stabilize it, ensuring it sounds natural when heard. Speech from areas with positive activation values are then extracted to determine which words, phrases, and intonations influence the classification decision.

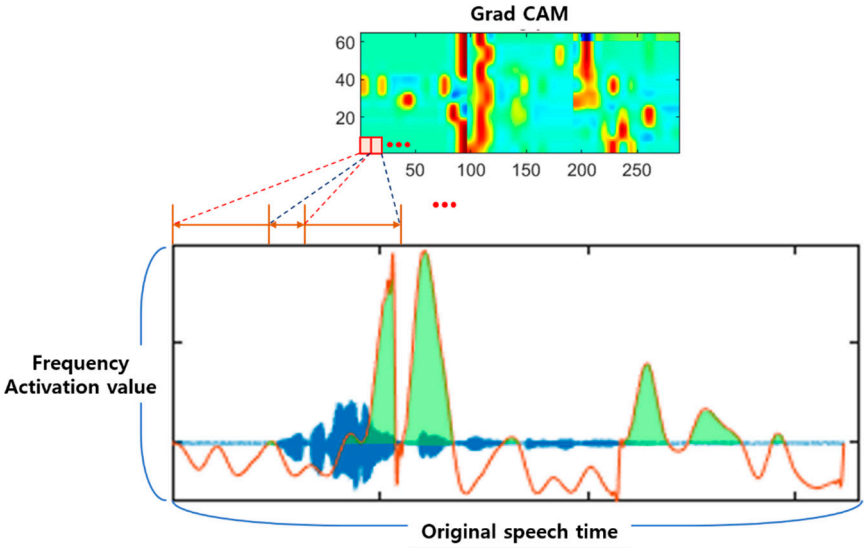

Figure 30. Restoration of audio signal using Grad CAM with a focus on specific areas in the image.

Figure 31 shows, among the happy emotion lines, the line "Call for sure, I love skiing so much." The line is analyzed by restoring the audio resolution to the area that the model focused on when classifying the emotion. When analyzing the visually expressed activation area by applying it to speech, the model focuses on the sections "for sure" and "so much". Therefore, it is confirmed that the focus is on the emotional characteristics that appear in each emotion's intonation or word elements during speech. In the case of a neutral emotion, we identified the focus on uniform distribution throughout the dialogue.

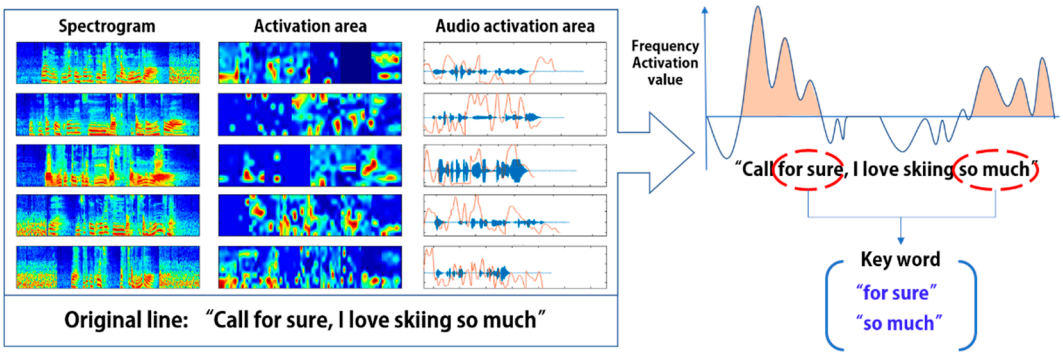

Figure 31. Applying the happy emotion Grad CAM focus area to speech data.

8. Conclusions

In this study, we conducted a study to learn the emotion classification model through various Korean-based data and to increase the reliability of the classification results. In order to enhance the accuracy of model decision-making, uncertain factors must be minimized during the learning process, and generalization capabilities must be achieved by extracting a lot of information from input data. To do this, we created refined data through a Gaussian data selection algorithm that selects the data needed as a correlation coefficient by representing data from three datasets converted into spectrograms with STFTs as Gaussian distributions for each segmented interval. Through this algorithm, unnecessary elements were reduced to make learning clearer, and the required learning time and computing resources could be reduced. In order to learn accurate features from the

refined data, YAMNet and VGGish were combined, resulting in the generation of structural features for each model. This confirmed that combining each model in the proposed manner has higher performance than using it individually. The trained model confirmed the decision-making process in many aspects by applying various explanatory techniques. By analyzing the two-dimensional activation area in both time and frequency domains, respectively, the frequency domain activated for each emotion was identified, and the model's concentration area was converted into an audio file to directly check the process. In this study, the activation area of the model was converted into an audio file so that it could be confirmed. However, there is a limitation wherein people may not understand the output if the model does not focus on word features. In future studies, it is expected that this research will contribute to the classification of voice emotions by assessing whether high performance is maintained in data from other languages and by further investigating whether the activation area can be correctly identified.

Author Contributions: Conceptualization, T.-W.K. and K.-C.K.; methodology, T.-W.K. and K.-C.K.; software, T.-W.K. and K.-C.K.; validation, T.-W.K. and K.-C.K.; formal analysis, T.-W.K. and K.-C.K.; investigation, T.-W.K.; resources, K.-C.K.; data curation, K.-C.K.; writing—original draft preparation, T.-W.K.; writing—review and editing, K.-C.K.; visualization, T.-W.K. and K.-C.K.; supervision, K.-C.K.; project administration, K.-C.K.; funding acquisition, K.-C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by a research fund from Chosun University, 2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

SER	Speech emotion recognition	IEMOCAP	Interactive emotional dyadic motion capture database
STFT	Short-time Fourier transform	MUOC	Massive open online course
Grad CAM	Gradient-weighted class activation mapping	EMO	Emotional speech database
LIME	Local interpretable model-agnostic explanations	CASIA	Chinese Academy of Sciences Institute of Automation
VGGish	Visual geometry group-like audio classification model	VCTK	Voice cloning toolkit
YAMNet	Yet another mobile network	DCASE	Detection and classification of acoustic scenes and events
GDS	Gaussian data selection	DIRAMS	Dongnam Institute of Radiological & Medical Sciences
CNN	Convolutional neural network	AI	Artificial intelligence
LSTM	Long short-term memory	CSU	Chosun University
UA	Unweighted accuracy	RF	Random forest
WA	Weighted accuracy	MLAs	Machine learning algorithms
MFCC	Mel-frequency cepstral coefficients	UAR	Unweighted average recall
SVM	Support vector machine	MAA	Macro average accuracy

References

1. Mekruksavanich, S.; Jitpattanakul, A. Sensor-based Complex Human Activity Recognition from Smartwatch Data Using Hybrid Deep Learning Network. In Proceedings of the 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Republic of Korea, 27–30 June 2021; pp. 1–4.
2. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech recognition using deep neural networks: A systematic review. *IEEE Access* **2019**, *7*, 19143–19165. [CrossRef]
3. Latif, S.; Qadir, J.; Qayyum, A.; Usama, M.; Younis, S. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Rev. Biomed. Eng.* **2020**, *14*, 342–356. [CrossRef] [PubMed]
4. Cho, J.; Kim, B. Performance analysis of speech recognition model based on neuromorphic architecture of speech data preprocessing technique. *J. Inst. Internet Broadcast Commun.* **2022**, *22*, 69–74.

5. Lee, S.; Park, H. Deep-learning-based Gender Recognition Using Various Voice Features. In Proceedings of the Symposium of the Korean Institute of Communications and Information Sciences, Seoul, Republic of Korea, 17–19 November 2021; pp. 18–19.
6. Fonseca, A.H.; Santana, G.M.; Bosque Ortiz, G.M.; Bampi, S.; Dietrich, M.O. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. *Elife* **2021**, *10*, e59161. [CrossRef] [PubMed]
7. Lee, Y.; Lim, S.; Kwak, I.Y. CNN-based acoustic scene classification system. *Electronics* **2021**, *10*, 371. [CrossRef]
8. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion recognition from variable-length speech segments using deep learning on spectrograms. *Proc. Interspeech* **2018**, *2018*, 3683–3687.
9. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.
10. Zhang, S.; Li, C. Research on feature fusion speech emotion recognition technology for smart teaching. *Mob. Inf. Syst.* **2022**, *2022*, 7785929. [CrossRef]
11. Subramanian, R.R.; Sireesha, Y.; Reddy, Y.S.P.K.; Bindamrutha, T.; Harika, M.; Sudharsan, R.R. Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Virtual Conference, 8–9 October 2021; pp. 1–6.
12. Zheng, L.; Li, Q.; Ban, H.; Liu, S. Speech Emotion Recognition Based on Convolution Neural Network Combined with Random Forest. In Proceedings of the 2018 Chinese Control and Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 4143–4147.
13. Li, H.; Zhang, X.; Wang, M.J. Research on speech Emotion Recognition Based on Deep Neural Network. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; pp. 795–799.
14. Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention-based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1771–1775.
15. Carofilis, A.; Alegre, E.; Fidalgo, E.; Fernández-Robles, L. Improvement of accent classification models through grad-transfer from spectrograms and gradient-weighted class activation mapping. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2859–2871. [CrossRef]
16. Bicer, H.N.; Götz, P.; Tuna, C.; Habets, E.A. Explainable Acoustic Scene Classification: Making Decisions Audible. In Proceedings of the 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Bamberg, Germany, 5–8 September 2022; pp. 1–5.
17. Cesarelli, M.; Di Giammarco, M.; Iadarola, G.; Martinelli, F.; Mercaldo, F.; Santone, A. Deep Learning for Heartbeat Phonocardiogram Signals Explainable Classification. In Proceedings of the 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 7–9 November 2022; pp. 75–78.
18. Lee, J.H.; Lee, C.Y.; Eom, J.S.; Pak, M.; Jeong, H.S.; Son, H.Y. Predictions for three-month postoperative vocal recovery after thyroid surgery from spectrograms with deep neural network. *Sensors* **2022**, *22*, 6387. [CrossRef] [PubMed]
19. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 618–626.
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
21. Available online: http://www.aihub.or.kr/aihubdata/data/view.do?currMenu=120&topMenu=100&dataSetSn=259&aihubDataSe=extrldata (accessed on 13 October 2023).