sensors

MDPI

Article

A Wearable Assistant Device for the Hearing Impaired to Recognize Emergency Vehicle Sirens with Edge Computing

Chiun-Li Chin [1], Chia-Chun Lin [1], Jing-Wen Wang [1], Wei-Cheng Chin [1], Yu-Hsiang Chen [1], Sheng-Wen Chang [1], Pei-Chen Huang [1], Xin Zhu [2], Yu-Lun Hsu [3] and Shing-Hong Liu [4,*]

[1] Department of Medical Informatics, Chung Shan Medical University, Taichung 40201, Taiwan; ernestli@csmu.edu.tw (C.-L.C.); chiachunlin1223@gmail.com (C.-C.L.); s0958011@gm.csmu.edu.tw (J.-W.W.); s0958039@gm.csmu.edu.tw (W.-C.C.); s0958033@gm.csmu.edu.tw (Y.-H.C.); s0858048@gm.csmu.edu.tw (S.-W.C.); s0858009@gm.csmu.edu.tw (P.-C.H.)

[2] Division of Information Systems, School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Fukushima, Japan; zhuxin@u-aizu.ac.jp

[3] Bachelor's Program of Sports and Health Promotion, Fo Guang University, Yilan 26247, Taiwan; ylhsu@gm.fgu.edu.tw

[4] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung 41349, Taiwan

* Correspondence: shliu@cyut.edu.tw; Tel.: +886-4-2332-3000 (ext. 7811)

Abstract: Wearable assistant devices play an important role in daily life for people with disabilities. Those who have hearing impairments may face dangers while walking or driving on the road. The major danger is their inability to hear warning sounds from cars or ambulances. Thus, the aim of this study is to develop a wearable assistant device with edge computing, allowing the hearing impaired to recognize the warning sounds from vehicles on the road. An EfficientNet-based, fuzzy rank-based ensemble model was proposed to classify seven audio sounds, and it was embedded in an Arduino Nano 33 BLE Sense development board. The audio files were obtained from the CREMA-D dataset and the Large-Scale Audio dataset of emergency vehicle sirens on the road, with a total number of 8756 files. The seven audio sounds included four vocalizations and three sirens. The audio signal was converted into a spectrogram by using the short-time Fourier transform for feature extraction. When one of the three sirens was detected, the wearable assistant device presented alarms by vibrating and displaying messages on the OLED panel. The performances of the EfficientNet-based, fuzzy rank-based ensemble model in offline computing achieved an accuracy of 97.1%, precision of 97.79%, sensitivity of 96.8%, and specificity of 97.04%. In edge computing, the results comprised an accuracy of 95.2%, precision of 93.2%, sensitivity of 95.3%, and specificity of 95.1%. Thus, the proposed wearable assistant device has the potential benefit of helping the hearing impaired to avoid traffic accidents.

Keywords: edge computing; human vocalization; emergence vehicle siren; EfficientNet-based fuzzy rank-based ensemble model; hearing impairment

check for updates

Citation: Chin, C.-L.; Lin, C.-C.; Wang, J.-W.; Chin, W.-C.; Chen, Y.-H.; Chang, S.-W.; Huang, P.-C.; Zhu, X.; Hsu, Y.-L.; Liu, S.-H. A Wearable Assistant Device for the Hearing Impaired to Recognize Emergency Vehicle Sirens with Edge Computing. *Sensors* **2023**, *23*, 7454. http://doi.org/10.3390/s23177454

Academic Editor: Marc Parrilla

Received: 7 August 2023
Revised: 21 August 2023
Accepted: 24 August 2023
Published: 27 August 2023

1. Introduction

Hearing is an essential ability in daily life that aids in avoiding the risk of bodily injuries [1]. People with hearing impairments have reduced sensitivity to sound, leading to dangerous situations such as ignoring the warning sounds from vehicles. A study by Donmez and Gokkoca showed that elderly people are involved in 31.5% of traffic accidents in Turkey. Their hearing impairment was the main issue [2]. The reason for this is that elderly people who are afflicted with hearing impairments cannot perceive the warning sounds from vehicles. Tiwari and Ganveer proposed that 10.2% of those involved in traffic accidents are hearing impaired [3]. Therefore, the development of a wearable assistant device for recognizing the various warning sounds from ambulances and vehicles on the road could help the hearing impaired, reducing the risk of traffic accidents.

Recently, deep learning has been widely applied in voice recognition [4,5]. In 2021, Bonaventure et al. proposed the FSER architecture, which converts speech files into spectrograms and inputs them into a two-dimensional, convolutional neural network (2D CNN) for identification [6]. Its accuracy surpasses that of the one-dimensional convolutional neural network (1D CNN), as 2D CNN models can extract finer features from the spectrogram [7]. Kevin et al. aimed to build a more accurate sound classification model and proposed a two-stream neural network architecture that includes the EfficientNet-based model [8]. Lee et al. utilized preoperative and postoperative voice spectrograms as features to predict three-month postoperative vocal recovery [9]. This model could be widely applicable for transfer learning in sound classification. Lu et al. used the morphology of spectrograms as the input pattern to recognize speech using an EfficientNet model [10]. Padi et al. employed transfer learning to improve the accuracy of speech emotion recognition through spectrogram augmentation [7]. Additionally, Allamy and Koerich utilized a 1D CNN to classify music genres based on audio signals [11].

Ensemble learning is a powerful technique that involves the amalgamation of predictions from multiple classifiers to create a single classifier, resulting in notably enhanced accuracy compared to any individual classifier [12,13]. Research has demonstrated that an effective ensemble consists of individual classifiers with similar accuracies, yet with distributed errors across different aspects [14,15]. Essentially, ensemble learning encompasses two necessary characteristics: the generation of distinct individual classifiers, and their subsequent fusion. Two common strategies for generating individual classifiers include the heterogeneous type, which employs various learning algorithms, and the homogeneous type, which uses the same learning algorithm but requires different settings. Thus, Tan et al. proposed ensemble learning to classify human activities, combining a gated recurrent unit (GRU), a CNN stacked on the GRU, and a deep neural network [16]. Xie et al. proposed three DNN-based ensemble methods that fused a series of classifiers whose inputs are representations of intermediate layers [17]. Erdal et al. proposed a voting-based ensemble to improve identification results in tuberculosis classification, traditionally using a single CNN model [12,18]. This fusion method uses a voting algorithm to determine the output. However, its disadvantage is that it simply votes on the model's output, only considering the number of predicted results and not the probability value of the predictions. Kavitha et al. proposed a weighted average-based ensemble to improve the accuracy of cell locations in cut electronic microscope images [19]. The disadvantage of this method is that when a large error occurs in the same prediction result, the weighted average result would be affected. Manna et al. proposed a fuzzy-based ensemble to improve the identification results of cervical cancer based on different CNN models. The output results of the CNN models, including InceptionV3, Xception, and DenseNet-169, were ensembled through fuzzy rank-based fusion [20]. The advantages of fuzzy rank-based ensembles include less computing time and memory consumption compared to fully connected layers.

The growth of the Internet of Everything (IoE) has led to great innovation in smart devices that connect to the internet, and in processing the necessary amounts of data. The innovation has aimed to resolve the problems of traditional cloud computing by decreasing burdensome bandwidth loads, increasing response speeds, and enhancing transmission security. To address these requirements, edge computing technologies have emerged as a promising solution [21,22]. Edge computing offers a more distributed and localized approach to data, allowing data to be processed in real time at the source. Hochst et al. proposed an edge artificial intelligence (AI) system to recognize bird species by their audio sounds, utilizing EfficientNet-B3 architecture based on an NVIDIA Jetson Nano board [23]. They demonstrated that the EfficientNet model could be efficiently implemented on an edge device. Rahman and Hossain developed an edge IoMT system using deep learning to detect various types of health-related COVID-19 symptoms based on a smartphone [24]. Nath et al. provided an overview of studies related to stress monitoring with edge computing, highlighting that computations performed using edge technology can reduce response time and are less vulnerable to external threats [25].

Based on the review of the above literature, the goal of this study is to develop a wearable assistant device with edge computing to help the hearing impaired to recognize the warning sounds from ambulances and vehicles on the road. An EfficientNet-based, fuzzy rank-based ensemble model was proposed to classify human vocalizations and warning sounds from vehicles. This model was embedded in an Arduino Nano 33 BLE Sense development board. The audio signals, including human vocalizations and the warning sounds of vehicles, were obtained from the CREMA-D dataset [26] and a Large-Scale Audio dataset of emergency vehicle sirens on the road [27], respectively. The categorization included seven types of audio sounds: neutral vocalization, anger vocalization, fear vocalization, happy vocalization, normal sound, car horn sound, siren sound, and ambulance siren sound. The spectrogram of the audio signal served as the feature. When one of the car horn, siren, or ambulance siren sounds was detected, the wearable assistant device presented alarms through a vibrator and displayed messages on the OLED panel. The results generated by edge computing were very close to those classified using offline computing. Moreover, we compared the performances between our proposed method and the iOS system, finding that our method outperformed the results of the iOS method. The contributions of this study included that the proposed EfficientNet-based, fuzzy rank-based ensemble model could be executed in the Arduino Nano 33 BLE Sense development board, and the performance of this model was better than the iOS system and the other deep learning models.

2. Materials and Methods

Figure 1 illustrates the architecture of the proposed method, including the data processing, training phase, and testing phase. The EfficientNet-based, fuzzy rank-based ensemble model is utilized to recognize human vocalizations and emergency vehicle sirens. During the training phase, this model is executed on a personal computer (PC). In contrast, during the testing phase, the proposed model is run on an Arduino Nano 33 BLE Sense development board. When the system detects specific sounds, such as car horns, sirens, or ambulance sirens, it alerts the user through the device's vibrator and displays messages on an OLED panel. The audio signal is converted into a spectrogram to serve as the input pattern. The EfficientNet-based, fuzzy rank-based ensemble model includes three individual EfficientNet-based models and one fuzzy rank-based model. The EfficientNet-based models estimate the weight of each class. Then, the fuzzy rank-based model determines the winner of the classes. Once trained, the model is implemented in the wearable assistant device, drawing the user's attention to warning sounds from vehicles.

2.1. Features

In this study, human vocalizations and emergency vehicle sirens on the road were classified. The human vocalizations, including neutral, anger, fear, and happy vocalizations, were sourced from the CREMA-D dataset, an emotional multimodal actor dataset [26]. A total of 5324 audio files were extracted from this dataset, and these files were augmented to create a total of 10,648 files. The emergency vehicle sirens on the road were acquired from the Large-Scale Audio dataset [27]. This dataset contains 3432 audio files featuring the car horn sound, siren sound, and ambulance siren sound. These files were augmented to a total of 6864 audio files using the RandAugment method [28]. The files were then divided into 80% for training, 10% for testing, and 10% for validation. The audio signals were segmented into 0.5 s intervals and transformed into spectrograms using the short-time Fourier transform. Each sample was represented by three spectrograms (1.5 s), which served as features for the classification of emergency vehicle sirens. Therefore, for the samples of emotional vocalizations, the numbers of training, testing, and validation samples were 25,555, 3195, and 3194, respectively. For the samples of emergency vehicle sirens, the numbers of training, testing, and validation samples were 16,474, 2058, and 2060, respectively.

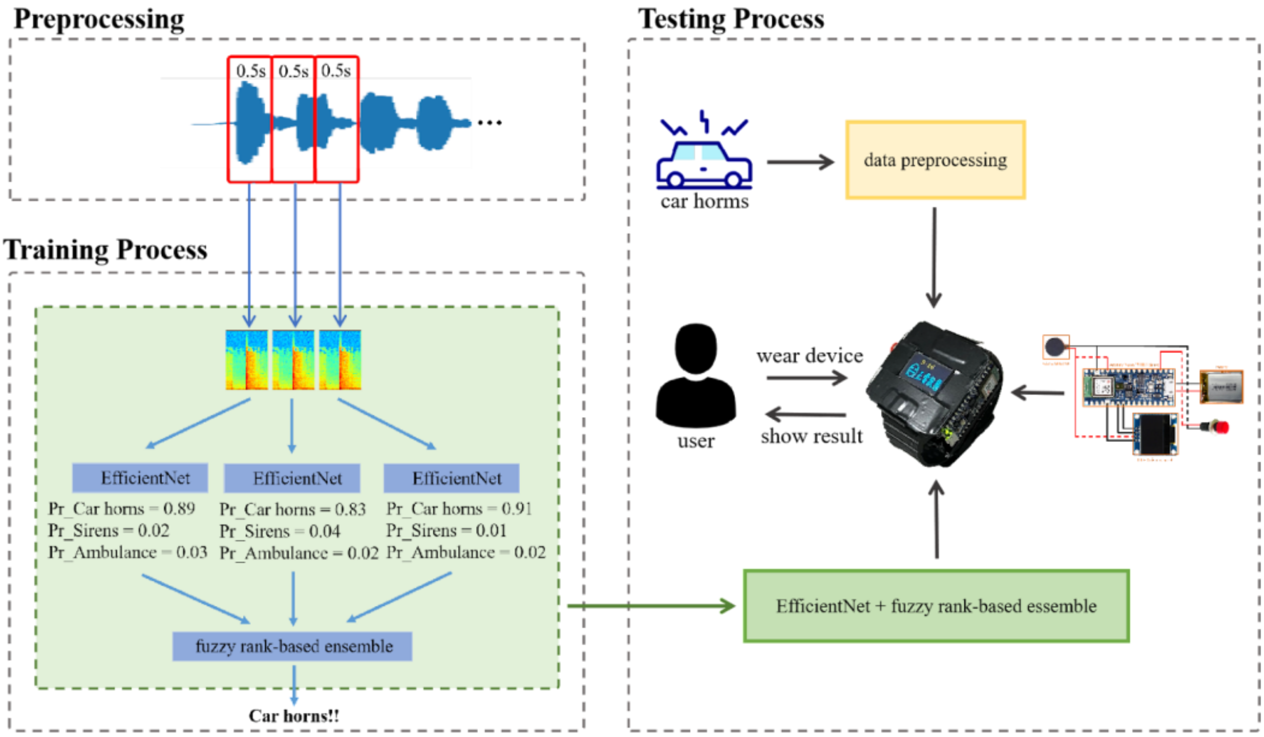

Figure 1. The architecture of proposed wearable assistant system allowing the hearing impaired to recognize warning sounds from vehicles.

2.2. EfficientNet-based Model

To recognize the human vocalizations and emergency vehicle sirens in real time with edge computing, the EfficientNet-based model was used to estimate the weights of all classes. The EfficientNet-based model is a deep learning model that has achieved top performance in image classification tasks and has demonstrated State-of-the-Art (SOTA) performance in the ImageNet image classification challenge [29]. The EfficientNet-based model builds on the base architectures of ResNet [30] and MobileNet [29,31] and leverages the compound scaling method [32] to achieve a balance between model size, computational efficiency, and accuracy. As a result, the EfficientNet-based model has become one of the most popular convolutional neural network models in current research. The structure of model is shown in Figure 2, which has one layer of Conv3 × 3, one layer of MBConv1, k3 × 3, six layers of MBConv6, k3 × 3, nine layers of MBConv6, k5 × 5, and one layer of full connection. The number of output layers is seven. The resolution and channel number of each layer are described in Table 1. Each row describes a stage i with $\hat{L}^{\mathrm{i}}$ layers, with input resolution $\{\hat{H}^{\mathrm{i}}, \hat{W}^{\mathrm{i}}\}$ and output channels $\hat{C}^{\mathrm{i}}$. The hyperparameters of the EfficientNet-based model, as shown in Table 2, were used for all experiments. The optimizer is the Adam, learning rate is 1×10^{-5}, batch size is 16, and the number of epochs is 1000. The sum of label smooth cross entropy loss function (L_{LSCE}) [33] and focal cross entropy loss function (L_{FCE}) [34] is defined as the total loss function (L_T) to validate the performance of EfficientNet-based model.

$$L_{LSCE} = \frac{-1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} P_{ji}\log\left(f\left(x_j\right)\right), \tag{1}$$

where N is the number of samples, M is the number of categories, $f(*)$ is the classifier, x_j is the sample. When x_j belongs to lth class, P_{ji} = $1 - \varepsilon$, and P_{ji} = $\varepsilon/M - 1$ for the other classes. ε is 0.2.

$$f(*) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise}' \end{cases} \tag{2}$$

where p is the probability of the target.

$$L_{FCE} = \frac{-1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M}\alpha_i\left(1 - f\left(x_j\right)\right)^{\gamma}\log\left(f\left(x_j\right)\right), \tag{3}$$

where α_i is the weight of loss function and γ is 2.

$$L_T = L_{LSCE} + L_{FCE} \tag{4}$$

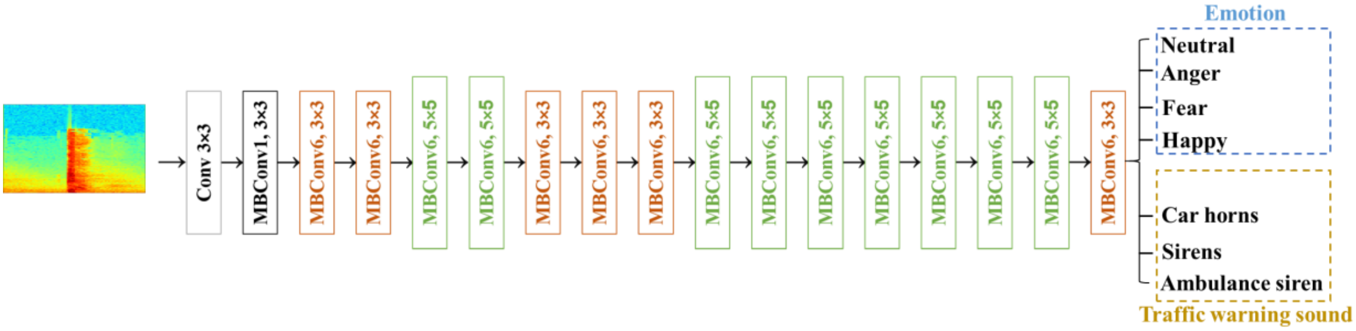

Figure 2. The architecture of EfficientNet-based model.

Table 1. EfficientNet-based model, the stage i with $\hat{L}_i$ layers, with input resolution $\left\langle \widehat{H}_I, \widehat{W}_i \right\rangle$ and output channels $\hat{C}_i$.

stage i	Operator $\widehat{\mathbf{F}}_i$.	Resolution $\widehat{\mathbf{H}}_I \times \widehat{\mathbf{W}}_i$	#Channels $\widehat{\mathbf{C}}_i$	#Layers $\hat{\mathbf{L}}_i$
1	Conv3 × 3	224 × 192	32	1
2	MBConv1, k3 × 3	112 × 96	16	1
3	MBConv6, k3 × 3	112 × 96	24	2
4	MBConv6, k5 × 5	56 × 48	40	2
5	MBConv6, k3 × 3	28 × 24	80	3
6	MBConv6, k5 × 5	28 × 24	112	3
7	MBConv6, k5 × 5	14 × 12	192	4
8	MBConv6, k3 × 3	7 × 6	320	1
9	Conv1 × 1 & Flatten & FC	7 × 6	1280	1

Table 2. Hyperparameters for training.

Hyperparameter	Selected Value
Loss function	$L_{Total} = L_{LSCE} + L_{FL}$
Optimizer	Adam
Learning rate	1×10^{-5}
Batch size	16
Epoch	1000

2.3. Fuzzy Rank-Based Model

The fuzzy rank-based model takes the results from the three EfficientNet-based models and calculates two fuzzy ranks by applying the exponential function and hyperbolic tangent function transformations, R_1 and R_2 [20].

$$R_{1ik} = 1 - \exp(-\frac{(P_{ik} - 1)^2}{2}),\ k = 1,\ 2,\ 3 \tag{5}$$

$$R_{2ik} = 1 - tanh(\frac{(P_{ik} - 1)^2}{2}),\ k = 1,\ 2,\ 3 \tag{6}$$

where P_{ik} is the estimating weight of ith category on kth EfficientNet-based model. RS is the fused rank score.

$$RS_{ik} = R_{1ik} \times R_{2ik} \tag{7}$$

The confidence score of a particular category, CS, is the sum of RS_{sk}.

$$CS_i = \sum_{k=1}^{3} RS_{ik} \tag{8}$$

The category with the minimum confidence score is considered the winner.

2.4. Wearable Assistant Device

This system operates on the Arduino Nano 33 BLE Sense development board [35], which is linked to various hardware modules such as an OLED panel, vibrator, GPS positioning module, microphone, and relay module, as depicted in Figure 3. The relay module governs the power of the OLED panel to conserve energy, while the microphone captures the audio sound. When the wearable assistant device detects a car horn, siren, or ambulance siren, it activates the vibration module and displays a message on the OLED panel. Figure 4 shows a photo of the wearable assistant device being worn on the wrist. Its size is 3.9 cm × 3.1 cm × 2.5 cm.

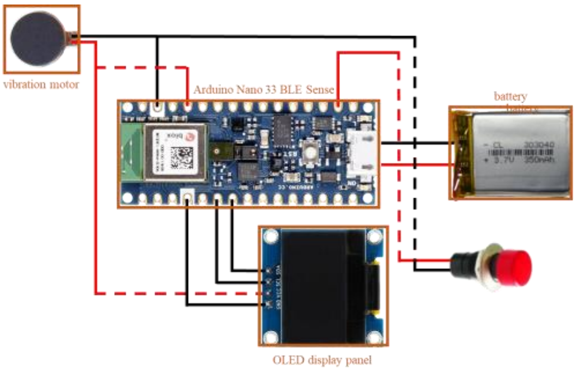

Figure 3. The Arduino Nano 33 BLE Sense development board.

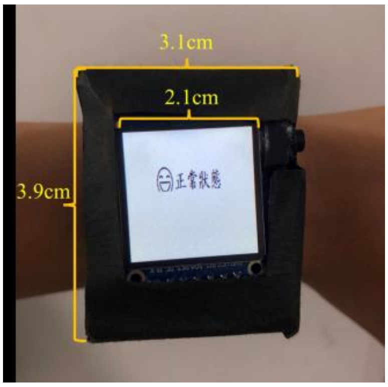

Figure 4. The photo of the wearable assistant device being worn on the wrist, the size of which is 3.9 cm × 3.1 cm × 2.5 cm.

2.5. Performances

The results of the tests are summarized in a confusion matrix, outlined in Table 3. This matrix illustrates the relationship between the actual and estimated classes in the test set, with each row corresponding to the actual classes and each column representing the estimated classes. According to the proposed method, a sample is classified as a true positive (TP) if the activity is correctly recognized, a false positive (FP) if the activity is incorrectly recognized, a true negative (TN) if the activity classification is correctly rejected, and a false negative (FN) if the activity classification is incorrectly rejected. To evaluate the performance of the proposed model, metrics such as accuracy, sensitivity, specificity, and precision are utilized, as described in Equations (9)–(12).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

Table 3. Confusion matrix.

		Estimated Class	
		Positive	Negative
Actual Class	Positive	*TP* (True Positive)	*FN* (False Negative)
	Negative	*FP* (False Positive)	*TN* (True Negative)

3. Results

The proposed model underwent training, validation, and testing processes on a computer equipped with an 8-core CPU (Intel Xeon W-3223), 64 GB of RAM, a GPU (RTX 3090) with 24 GB of graphics memory, and 10,496 CUDA cores. PyTorch was the framework used for the implementation.

Table 4 presents the confusion matrix for the EfficientNet-based, fuzzy rank-based ensemble model in offline computing. The results clearly show that the study achieved an accuracy of 97.05%, a precision of 97.79%, a sensitivity of 96.8%, and a specificity of 97.04%.

Table 4. The confusion matrix of EfficientNet-based, fuzzy rank-based ensemble model in offline computing.

	Neutral	**Anger**	**Fear**	**Happy**	**Car Horns**	**Sirens**	**Ambulance Siren**
Neutral	2590	10	5	4	0	0	0
Anger	10	2588	14	5	5	0	0
Fear	6	11	2594	3	0	2	0
Happy	8	6	2	2598	0	0	0
Car Horns	0	0	0	0	1094	17	25
Sirens	0	0	0	0	13	1105	35
Ambulance Siren	0	0	1	0	26	21	1096

In order to evaluate the robustness of the EfficientNet-based, fuzzy rank-based ensemble model in a noisy environment, we collected audio samples of road noise and superimposed the spectrogram of road noise over the testing samples. In this experiment, the number of testing samples is the same as the number used with the EfficientNet-based, fuzzy rank-based ensemble model in offline computing. Table 5 shows the confusion matrix of the EfficientNet-based, fuzzy rank-based ensemble model in offline computing under the noisy environment. It is evident that this study achieved an accuracy of 96.84%, precision of 96.17%, sensitivity of 96.13%, and specificity of 96.90%. We found that our proposed model exhibited high robustness. The accuracy, precision, sensitivity, and specificity only dropped 0.21%, 1.62%, 0.67%, and 0.14%, respectively.

Table 5. The confusion matrix of the EfficientNet-based, fuzzy rank-based ensemble model in offline computing under the noisy environment. The testing samples were superimposed over the road noise.

	Neutral	**Anger**	**Fear**	**Happy**	**Car Horns**	**Sirens**	**Ambulance Siren**
Neutral	2538	20	10	12	0	0	0
Anger	35	2549	33	20	0	0	0
Fear	24	27	2568	7	0	0	0
Happy	17	19	3	2571	0	0	0
Car Horns	0	0	0	0	1079	66	46
Sirens	0	0	0	0	18	1067	28
Ambulance Siren	0	0	0	0	39	12	1082

The iOS system, developed by Apple Inc., Cupertino, CA, USA, is one of the most prevalent mobile operating systems, holding a substantial market share. Thanks to its vast user base, the iOS system has become highly popular. Starting with version 14, iOS has integrated sound recognition functionality, providing options to recognize various sounds such as alarms, car horns, and shouting, among others. To compare our proposed method with the iOS system, we utilized the sound recognition feature in iOS to classify emergency vehicle sirens [36]. The training, validation, and testing samples were the same as those used in the offline computing experiment with the EfficientNet-based, fuzzy rank-based ensemble model. Table 6 displays the confusion matrix for the iOS system in offline computing, showing an accuracy of 70.82%, precision of 76.67%, recall of 76.22%, and specificity of 71.02%. Comparing the results in Tables 4 and 6, our proposed model outperformed the iOS system. The accuracy, precision, sensitivity, and specificity of our model exceeded those of the iOS system by 26.23%, 21.57%, 20.58%, and 26.02%, respectively.

Table 6. The confusion matrix of iOS system in the offline computing. All samples are the same as in Table 4.

	Neutral	**Anger**	**Fear**	**Happy**	**Car Horns**	**Sirens**	**Ambulance Siren**
Neutral	1927	215	192	794	0	0	0
Anger	143	1589	776	118	0	0	0
Fear	230	743	1521	101	0	0	0
Happy	314	68	127	1597	0	0	0
Car Horns	0	0	0	0	1065	23	16
Sirens	0	0	0	0	26	1045	43
Ambulance Siren	0	0	1	0	47	77	1097

In this study, we embedded the trained EfficientNet-based fuzzy rank-based ensemble model into the Arduino Nano 33 BLE Sense development board. The testing samples were played through the speaker of a PC. The board was recording the sounds and recognized the class of each sound. If the sound belonged to a car horn, siren, or ambulance siren, it would start the vibrator and send a message to the OLED. Then, we counted the numbers of all categories. The testing samples were the same as those used in the experiment on the EfficientNet-based, fuzzy rank-based ensemble model in offline computing. Table 7 shows the confusion matrix of the EfficientNet-based, fuzzy rank-based ensemble model in edge computing. The performances of edge computing achieved an accuracy of 95.22%, precision of 93.19%, sensitivity of 95.27%, and specificity of 95.09%. According to the results shown in Tables 4 and 7, we find that the performances of edge computing are close to those of offline computing. Our proposed model exhibited the better performance. The accuracy, precision, sensitivity, and specificity only dropped 1.83%, 4.6%, 1.53%, and 1.95%, respectively. The testing video shows that the motorcycle horn and ambulance siren trigger the wearable assistant device, displaying the siren’s category [37].

Table 7. The confusion matrix of the EfficientNet-based fuzzy rank-based ensemble model in edge computing. The testing samples were the same as in the experiment involving offline computing.

	Neutral	Anger	Fear	Happy	Car Horns	Sirens	Ambulance Siren
Neutral	1272	14	12	11	8	12	10
Anger	12	1274	13	23	10	8	12
Fear	23	8	1268	12	5	10	11
Happy	9	10	4	1267	7	3	5
Car Horns	15	20	10	8	1084	14	6
Sirens	7	2	18	9	13	1079	7
Ambulance Siren	10	12	9	14	9	5	1095

4. Discussion

In ensemble learning, the voting and weight average methods are popular methods [38,39]. When a predicted class achieves the maximum voting number, the class is considered the winner. The main disadvantage of the voting method that it ignores the probabilities of true and false classes. It only focuses on the numbers of true classes. Although the weight average method could create a balance of true and false classes, this method would determine the wrong winner when the number of false classes is larger than the number of true classes. In this study, we proposed the fuzzy rank-based model to determine the winner amongst the classes. The model uses the two nonlinear functions, exponential function and hyperbolic tangent function, respectively, to estimate the ranks. The two functions are antisymmetric functions. Thus, they can inhibit the incorrect decisions caused by large amounts of false classes. Figure 5 shows the performances of the non-using ensemble (the EfficientNet-based model), voting and weight average methods (the EfficientNet-based voting ensemble model and the EfficientNet-based weight ensemble model), and the EfficientNet-based, fuzzy rank-based ensemble model. The results are: the accuracies of 83.31%, 87.23%, and 91.25%; the precisions of 82.87%, 86.82%, and 91.73%; the sensitivities of 83.01%, 87.11%, and 90.96%; and the specificities of 83.33%, 86.74%, and 91.17%. The performances of the EfficientNet-based, fuzzy rank-based ensemble model are the best, with an accuracy of 97.05%, precision of 97.79%, sensitivity of 96.8%, and specificity of 97.04%.

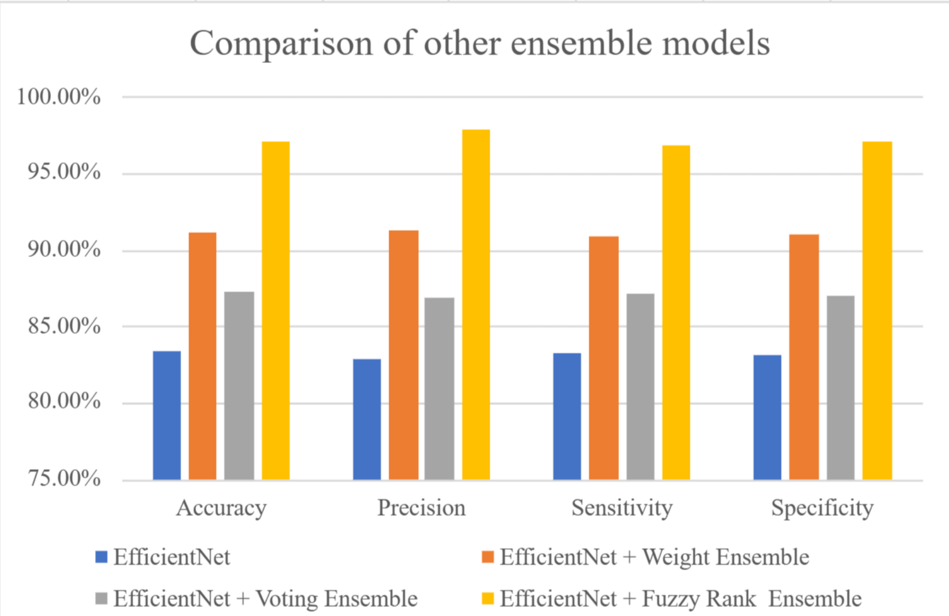

Figure 5. The comparison of other ensemble models. The EfficientNet-based, fuzzy rank-based ensemble model has an accuracy of 97.05%, precision of 97.79%, sensitivity of 96.8%, and specificity of 97.04%.

When compared to the iOS system, our method demonstrated significantly higher accuracy, outperforming it by a remarkable margin of 26.23%. This performance could be considered the main contribution of this study. The advantage of the EfficientNet-based, fuzzy rank-based ensemble model is that it not only recognized the warning sounds, but also identified the emotions within human vocalizations. The iOS system lacks the ability to recognize emotion. This reason indicates that our method is better than the iOS system.

Table 8 presents a comparative analysis of our proposed method against other studies that utilized the CREMA-D dataset [26]. Previous studies [40–43] only classified four sounds, while our study classified seven. As shown, the proposed EfficientNet-based, fuzzy rank-based ensemble model achieved an accuracy of 97.05%, which ranks the best result within these literatures. For the Large-Scale Audio dataset of emergency vehicle sirens on the road [16], a previous study showed that the best result for the classification of four sounds was an accuracy of 97%. This result was very close to that of our study. However, we emphasized that our approach classified seven sounds, broadening the scope of sound recognition.

Table 8. Comparative results of various methods using the CREMA-D dataset.

Ref.	Classification Method	F_1-Score (%)	Accuracy (%)
[40]	ResNet18	NA	57.42%
[41]	CNN-LSTM	79.23%	78.52%
[42]	Metric Learning-Based Multimodal	NA	65.01%
[43]	Triplet Loss-based modal	NA	58.72%
NA	EfficientNet-based fuzzy rank-based ensemble model	NA	97.05%.

In this study, we used two datasets, the CREMA-D dataset and the Large-Scale Audio dataset to evaluate the performance of the EfficientNet-based, fuzzy rank-based ensemble model. The number of each category is very different. The samples' numbers of neutral vocalizations, anger vocalizations, fear vocalizations, happy vocalizations, car horn sounds, siren sounds, and ambulance sirens sound were 7921, 7834, 7621, 8568, 7128, 7484, and 5979, respectively. Moreover, the target output used one-hot encoding. If the model uses general cross entropy loss, the model may have poor training results due to a sample imbalance. The label smooth cross entropy loss function and focal loss function could overcome these problems [33,34]. Thus, we used the sum of two loss functions to validate the performance of the proposed model to avoid overfitting and local minimum problems in the training phase.

5. Conclusions

We proposed a wearable assistant device to help the hearing impaired recognize the warning sounds from vehicles on the road. By employing edge computing in the Arduino Nano 33 BLE Sense development board, we executed the EfficientNet-based, fuzzy rank-based ensemble model. To evaluate the performance of this model, we used the CREMA-D dataset and the Large-Scale Audio dataset for training and testing. In offline computing, the accuracy reached 97.05%, while in edge computing, it also achieved 95.22% accuracy. Because the audio signals were obtained from the open datasets, we did not explore the performance of the wearable assistant device under different types of audio signals and background noises. These problems will directly affect the response time of the user when faced with an approaching vehicle. In the future, we will explore the sensitivity and robustness of the wearable assistant device and increase the response time of the user.

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

Funding: This research was supported in part by the National Science and Technology Council, Taiwan, under grants NSTC 111-2221-E-324-003-MY3.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://paperswithcode.com/dataset/crema-d, and https://github.com/tabarkarajab/Large-Scale-Audio-dataset (accessed on 6 January 2023).

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

References

1. Cornelius, C.; Marois, Z.; Sober, J.; Peterson, R.; Mare, S.; Kotz, D. Vocal resonance as a passive biometric. In *Computer Science Technical Report*; Dartmouth Digital Commons: Hanover, NH, USA, 2014; pp. 747–761.
2. Donmez, L.; Gokkoca, Z. Accident profile of older people in Antalya city center, Turkey. *Arch. Gerontol. Geriatr.* **2003**, *37*, 99–108. [CrossRef] [PubMed]
3. Tiwari, R.R.; Ganveer, G.B. A study on human risk factors in non-fatal road traffic accidents at Nagpur. *Indian J. Public Health* **2008**, *52*, 197–199. [PubMed]
4. Lee, W.; Seong, J.J.; Ozlu, B.; Shim, B.S.; Marakhimov, A.; Lee, S. Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review. *Sensors* **2021**, *21*, 1399. [CrossRef]
5. 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]
6. Dossou, B.F.P.; Gbenou, Y.K.S. FSER: Deep Convolutional Neural Networks for Speech Emotion Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021) Workshops, Montreal, QC, Canada, 11–17 October 2021; pp. 3533–3538.
7. Padi, S.; Sadjadi, S.O.; Sriram, R.D.; Manocha, D. Improved Speech Emotion Recognition Using Transfer Learning and Spectrogram Augmentation. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI 2021), Montréal, QC, Canada, 18–22 October 2021; pp. 645–652.
8. Gunawan, K.W.; Hidayat, A.A.; Cenggoro, T.W.; Pardamean, B. Repurposing transfer learning strategy of computer vision for owl sound classification. *Procedia Comput. Sci.* **2023**, *216*, 424–430. [CrossRef]
9. 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]
10. Lu, Q.; Li, Y.; Qin, Z.; Liu, X.; Xie, Y. Speech Recognition using EfficientNet. In Proceedings of the 2020 5th International Conference on Multimedia Systems and Signal Processing (ICMSSP 2020), Chengdu, China, 28–30 May 2020; pp. 64–68.
11. Allamy, S.; Koerich, A.L. 1D CNN Architectures for Music Genre Classification. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI 2021), Orlando, FL, USA, 5–7 December 2021; pp. 1–7.
12. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
13. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
14. Opitz, D.W.; Shavlik, J.W. Actively Searching for an Effective Neural Network Ensemble. *Connect. Sci.* **1996**, *8*, 337–353. [CrossRef]
15. Hashem, S. Optimal linear combinations of neural networks. *Neural Netw.* **1997**, *10*, 599–614. [CrossRef]
16. Tan, T.H.; Wu, J.Y.; Liu, S.H.; Gochoo, M. Human Activity Recognition Using an Ensemble Learning Algorithm with Smartphone Sensor Data. *Electronics* **2022**, *11*, 322. [CrossRef]
17. Xie, J.; Xu, B.; Chuang, Z. Horizontal and vertical ensemble with deep representation for classification. *arXiv* **2013**, arXiv:1306.2759.
18. Tasci, E.; Uluturk, C.; Ugur, A. A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Comput. Appli.* **2021**, *33*, 15541–15555. [CrossRef] [PubMed]
19. Shaga Devan, K.; Kestler, H.A.; Read, C.; Walther, P. Weighted average ensemble-based semantic segmentation in biological electron microscopy images. *Histochem. Cell Biol.* **2022**, *158*, 447–462. [CrossRef] [PubMed]
20. Manna, A.; Kundu, R.; Kaplun, D.; Sinitca, A.; Sarkar, R. A fuzzy rank-based ensemble of CNN models for classification of cervical cytology. *Sci. Rep.* **2021**, *11*, 14538. [CrossRef] [PubMed]
21. Cao, K.; Liu, Y.; Meng, G.; Sun, Q. An Overview on Edge Computing Research. *IEEE Access* **2020**, *8*, 85714–85728. [CrossRef]
22. Varghese, B.; Wang, N.; Barbhuiya, S.; Kilpatrick, P.; Nikolopoulos, D.S. Challenges and Opportunities in Edge Computing. In Proceedings of the 2016 IEEE International Conference on Smart Cloud (SmartCloud 2016), New York, NY, USA, 18–20 November 2016; pp. 20–26.

23. Höchst, J.; Bellafkir, H.; Lampe, P.; Vogelbacher, M.; Mühling, M.; Schneider, D.; Lindner, K.; Röcner, S.; Schabo, D.G.; Farwig, N.; et al. Bird@ Edge: Bird Species Recognition at the Edge. In Proceedings of the International Conference on Networked Systems (NETYS 2022), Networked Systems, Virtual, 17–19 May 2022; pp. 69–86.
24. Rahman, M.A.; Hossain, M.S. An Internet-of-Medical-Things-Enabled Edge Computing Framework for Tackling COVID-19. *IEEE Internet Things J.* **2021**, *8*, 15847–15854. [CrossRef]
25. Nath, R.K.; Thapliyal, H.; Caban-Holt, A.; Mohanty, S.P. Machine Learning Based Solutions for Real-Time Stress Monitoring. *IEEE Consum. Electron. Mag.* **2020**, *9*, 34–41. [CrossRef]
26. CREMA-D Dataset. Available online: https://paperswithcode.com/dataset/crema-d (accessed on 2 May 2020).
27. Asif, M.; Usaid, M.; Rashid, M.; Rajab, T.; Hussain, S.; Wasi, S. Large-scale audio dataset for emergency vehicle sirens and road noises. *Sci. Data* **2022**, *9*, 599. [CrossRef]
28. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020) Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 702–703.
29. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (PMLR 2019), Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
30. Liu, S.H.; Li, R.X.; Wang, J.J.; Chen, W.; Su, C.H. Classification of Photoplethysmographic Signal Quality with Deep Convolution Neural Networks for Accurate Measurement of Cardiac Stroke Volume. *Appl. Sci.* **2020**, *10*, 4612. [CrossRef]
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.30.
32. Bello, I.; Fedus, W.; Du, X.; Cubuk, E.D.; Srinivas, A.; Lin, T.-Y.; Shlens, J.; Zoph, B. Revisiting ResNets: Improved Training and Scaling Strategies. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online, 6–14 December 2021.
33. Chen, Y.; Xu, K.; Zhou, P.; Ban, X.; He, D. Improved cross entropy loss for noisy labels in vision leaf disease classification. *IET Image Process.* **2022**, *16*, 1511–1519. [CrossRef]
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. Kurniawan, A. Arduino Nano 33 BLE Sense Board Development. In *IoT Projects with Arduino Nano 33 BLE Sense*; Apress: Berkeley, CA, USA, 2021; pp. 21–74. ISBN 978-1-4842-6457-7.
36. Classifying Live Audio Input with a Built-in Sound Classifier. Available online: https://developer.apple.com/documentation/soundanalysis/classifying_live_audio_input_with_a_built-in_sound_classifier (accessed on 3 January 2023).
37. Available online: https://drive.google.com/file/d/1NO1FpZ4LxTDhiH1B0ZW7RQoiTlb0gQlF/view?usp=drive_link (accessed on 6 January 2023).
38. Leon, F.; Floria, S.A.; Bădică, C. Evaluating the Effect of Voting Methods on Ensemble-Based Classification. In Proceedings of the 2017 IEEE International Conference on INnovations in Intelligent Systems and Applications (INISTA 2017), Gdynia, Poland, 3–5 July 2017; pp. 1–6.
39. Huang, F.; Xie, G.; Xiao, R. Research on Ensemble Learning. In Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence (AICI 2009), Shanghai, China, 7–8 November 2009; Volume 3, pp. 249–252.
40. Zielonka, M.; Piastowski, A.; Czyżewski, A.; Nadachowski, P.; Operlejn, M.; Kaczor, K. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics* **2022**, *11*, 3831. [CrossRef]
41. Hans, A.S.A.; Rao, S. A CNN-LSTM based deep neural networks for facial emotion detection in videos. *Int. J. Adv. Signal Image Sci.* **2021**, *7*, 11–20. [CrossRef]
42. Beard, R.; Das, R.; Ng, R.W.; Gopalakrishnan, P.K.; Eerens, L.; Swietojanski, P.; Miksik, O. Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition. In Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), Brussels, Belgium, 31 October–1 November 2018; pp. 251–259.
43. Kumar, P.; Jain, S.; Raman, B.; Roy, P.P.; Iwamura, M. End-to-end Triplet Loss based Emotion Embedding System for Speech Emotion Recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR 2021), Milan, Italy, 10–15 January 2021; pp. 8766–8773.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.