algorithms

MDPI

Article

Optimizing Speech Emotion Recognition with Deep Learning and Grey Wolf Optimization: A Multi-Dataset Approach

Suryakant Tyagi [1] and Sándor Szénási [2,3,*]

[1] Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, 1034 Budapest, Hungary; suryakant.tyagi@phd.uni-obuda.hu
[2] John von Neumann Faculty of Informatics, Óbuda University, 1034 Budapest, Hungary
[3] Faculty of Economics and Informatics, J. Selye University, 945 01 Komarno, Slovakia
* Correspondence: szenasi.sandor@nik.uni-obuda.hu or szenasis@ujs.sk

Abstract: Machine learning and speech emotion recognition are rapidly evolving fields, significantly impacting human-centered computing. Machine learning enables computers to learn from data and make predictions, while speech emotion recognition allows computers to identify and understand human emotions from speech. These technologies contribute to the creation of innovative human–computer interaction (HCI) applications. Deep learning algorithms, capable of learning high-level features directly from raw data, have given rise to new emotion recognition approaches employing models trained on advanced speech representations like spectrograms and time–frequency representations. This study introduces CNN and LSTM models with GWO optimization, aiming to determine optimal parameters for achieving enhanced accuracy within a specified parameter set. The proposed CNN and LSTM models with GWO optimization underwent performance testing on four diverse datasets—RAVDESS, SAVEE, TESS, and EMODB. The results indicated superior performance of the models compared to linear and kernelized SVM, with or without GWO optimizers.

Keywords: speech emotion recognition; neural network; deep learning; LSTM

Citation: Tyagi, S.; Szénási, S. Optimizing Speech Emotion Recognition with Deep Learning and Grey Wolf Optimization: A Multi-Dataset Approach. *Algorithms* **2024**, *17*, 90. https://doi.org/10.3390/a17030090

Academic Editors: Frank Werner and Roberto Montemanni

Received: 9 December 2023
Revised: 7 February 2024
Accepted: 15 February 2024
Published: 20 February 2024

1. Introduction

Speech emotion recognition (SER) is a research field aiming to develop systems that automatically recognize emotions from speech, with the potential to enhance user experiences in various applications like spoken dialogue systems, intelligent voice assistants, and computer games. However, the accuracy of current SER systems remains relatively low due to factors such as the intricate nature of human emotions, variability in human speech, and challenges in extracting reliable features from speech signals.

A typical SER system comprises two main components: feature extraction and classification. Feature extraction is tasked with capturing essential acoustic characteristics from the speech signal, including pitch and spectral features. Subsequently, the classification component assigns emotional labels to the speech signal based on extracted features.

In recent years, deep learning has emerged as a powerful tool for machine learning, finding success in domains like computer vision, speech recognition, and natural language processing. Deep learning is well suited for SER for several reasons. Firstly, deep learning models excel at capturing intricate relationships between features, crucial for accurately identifying emotions from speech. Secondly, deep learning models possess unique characteristics that enable them to efficiently leverage large speech datasets, a capability particularly advantageous in SER, where access to extensive annotated data facilitates model training and improves generalization to diverse speakers and emotional expressions. Additionally, while other models also claim the ability to generalize to new speakers and situations, deep learning models demonstrate a notable robustness and adaptability in handling variations, making them particularly effective in real-world SER applications.

Here are some other points to keep in mind before going forward with the research:

- The intricate nature of human emotions poses a challenge to developing accurate SER systems as emotions are complex and expressed in various ways. For instance, happiness may be conveyed through laughter, smiling, and a high-pitched voice, but also through tears, a frown, and a low-pitched voice.
- The variability in human speech further complicates SER systems, as individuals speak differently based on factors like age, gender, and accent. For example, a young woman from the United States may have a different speaking style than an older man from the United Kingdom.

In the dynamic landscape of speech emotion recognition (SER), marked by its rapid growth, this research aims to contribute to the evolution of computer interaction by leveraging deep learning advancements to advance real-time emotion recognition from speech. The primary objective involves extracting pivotal features from audio waveforms to construct a model that accurately predicts emotions. While Mel-frequency cepstral coefficients (MFCC) feature extraction was employed in this study, it is important to acknowledge that other feature extraction methods exist and could be explored in future research to potentially enhance performance. MFCC is widely used and has shown success in many SER applications due to its ability to capture spectral characteristics relevant to speech perception, it is essential to note that there is not a universally "best" technique. Different techniques may perform better in different contexts. Four datasets (RAVDESS, SAVEE, TESS, and EmoDB) were utilized to train and evaluate the model. The model underwent initial training with linear and kernelized support vector machines (SVMs), followed by comprehensive training with a convolutional neural network (CNN) model incorporating long short-term memory (LSTM). Further refinement of the datasets was achieved through the application of a Gray Wolf Optimizer, tailoring the parameters for optimal model fitting. It is important to clarify that, while the study contributed to refining the models and optimization techniques, the datasets themselves were not crafted by the authors. Despite the inherent challenges in doing so, the study underscores the potential of deep learning in enhancing the accuracy and efficiency of real-time emotion recognition systems.

2. Related Work

This section provides a concise overview of the speech emotion recognition (SER) research landscape, highlighting the importance of acoustic features [1,2]. Acoustic parameters play a crucial role in deciphering emotions, prompting studies to investigate emotion-specific profiles through these parameters. While the integration of diverse classifiers such as Bayesian, K-nearest neighbor (KNN), and decision trees has reshaped the research field, it's worth noting that models like Gaussian mixtures (GMMs) and hidden Markov models (HMMs) may lack insights into low-level feature distribution [3–14].

The introduction of deep learning in SER, particularly with end-to-end systems and deep neural networks (DNN), has led to significant accuracy improvements [15]. Explorations into feature fusion techniques integrating acoustic and lexical domains [16], breakthroughs like the RNN-ELM model addressing long-range context effects [17], and the preference for features like logarithmic Mel-frequency cepstral coefficients (logMel), MFCC, and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) over prosodic features [18] underscore the evolving landscape. The convolutional recurrent neural network for end-to-end emotion prediction from speech [19] and recent studies focusing on deep learning approaches utilizing spectrograms [20–27], including models with convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have demonstrated enhanced accuracy [21]. Ensemble models incorporating bagging and boosting techniques with support vector machines (SVMs) showcased significant accuracy gains [28].

Optimization algorithms for feature selection in SER, such as Cat Swarm Optimization (CSO), Grey Wolf Optimizer, and Enhanced Cat Swarm Optimization (ECSO), have shown promise in enhancing classification accuracy and reducing selected features [29,30]. The Whale-Imperialist Optimization algorithm (Whale-IpCA) introduced multiple support vector neural network (Multi-SVNN) classifiers for emotion identification [31], while feature

selection methods employing metaheuristic search algorithms like Cuckoo Search and Non-dominated Sorting Genetic Algorithm-II (NSGA-II) demonstrated effective emotion classification with reduced features [32]. Comprehensive speech emotion recognition systems leverage diverse machine learning algorithms, including recurrent neural networks (RNNs), SVMs, and multivariate linear regression (MLR) [33]. Innovative feature selection approaches, such as the combination of Golden Ratio Optimization (GRO) and Equilibrium Optimization (EO) algorithms, have been explored [34–36].

Advanced architectures like dual-channel Long Short-Term Memory (LSTM) compressed capsule networks have been proposed for improved emotion recognition [37]. Furthermore, clustering-based Genetic Algorithm (GA) optimization techniques have been utilized to enhance feature sets for SER [38]. Recent research has also investigated the effectiveness of weighted binary Cuckoo Search algorithms for feature selection and emotion recognition from speech [39]. Moreover, the application of wavelet transform in SER has been explored for its potential in capturing relevant emotional features [40]. Ensemble methods like Bagged Support Vector Machines (SVMs) have shown promise in achieving robust emotion recognition from speech signals [41]. Convolutional Neural Networks (CNNs) have been employed to extract salient features for SER, leveraging their capability to capture hierarchical representations [42].

Additionally, novel methods for feature selection in SER, such as a hybrid metaheuristic approach combining Golden Ratio and Equilibrium Optimization algorithms, have been proposed [43]. Recent studies have explored the application of modulation spectral features for emotion recognition using deep neural networks [44]. Transfer learning frameworks, like EmoNet, have been developed to leverage multi-corpus data for improved SER performance [45–47]. Feature pooling techniques for modulation spectrum features have also been investigated to enhance SER accuracy, particularly in real-world scenarios [48,49].

3. Dataset

This study uses four different datasets, namely the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS), Emotional Database (EmoDB), and Surrey Audio-Visual Expressed Emotion (SAVEE). Details of the dataset and their limitations can be seen in Tables 1 and 2 respectively.

Table 1. The datasets used and their details.

Dataset	Description
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a publicly available dataset designed for affective computing and emotion recognition research. It contains audio and video recordings of 24 actors performing eight different emotions in both speech and song. The dataset provides a diverse set of emotional expressions for study and is widely used in emotion recognition research.
TESS	The Toronto Emotional Speech Set (TESS) is a comprehensive and publicly available collection of 2803 audio recordings, featuring professional actors portraying seven distinct emotional categories. The dataset includes balanced representation from both male and female actors, making it valuable for developing and evaluating emotion recognition models.
SAVEE	The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset is designed for research into speech emotion recognition, featuring 480 audio recordings with a single male actor portraying seven emotional states. The dataset provides a standardized resource for studying emotional speech and has been widely used in affective computing research.
EmoDB	The Emotional Database (EmoDB) is utilized for studying emotional speech recognition and consists of recordings from ten professional German actors portraying different emotions. The dataset, developed by the Technical University of Berlin, is valuable for developing and evaluating algorithms for automatic emotion classification from speech signals.

Table 2. Observations and considerations about datasets used.

Observations	Dataset-Specific Observations
Limited Actor Diversity	RAVDESS: 24 actors may not fully represent diverse vocal characteristics. SAVEE: Only four male actors may limit diversity and variability. EmoDB: Ten actors may not fully capture wide-ranging vocal characteristics.
Imbalanced Emotional Classes	RAVDESS, SAVEE, EmoDB: Some emotions have fewer instances, impacting model performance.
Controlled Recording Conditions	RAVDESS, SAVEE, EmoDB: Recordings in controlled studios lack natural variability, affecting generalizability to real-world scenarios.
Limited Contextual Information	RAVDESS, SAVEE, EmoDB: A lack of contextual cues in datasets may limit the applicability to real-world scenarios influenced by various factors.
Limited Language Representation	RAVDESS: Primarily English, limiting cross-lingual applications. EmoDB: Primarily German, affecting cross-lingual usability.
Limited Emotional Variability	SAVEE: Four basic emotions may restrict generalizability. EmoDB: Seven discrete emotions may not cover the full spectrum of human emotional experiences.
TESS Advantages	Diverse emotional expressions, large number of actors, naturalistic recording conditions, high-quality recordings, detailed metadata, and multimodal data: the TESS dataset stands out for its richness, naturalness, and comprehensive features, contributing to its reliability and robustness in emotional speech analysis.

Despite limitations in some datasets, they remain valuable for emotional speech analysis research. The TESS dataset, in particular, excels due to its diverse emotional expressions, large actor pool, natural recording conditions, high-quality data, detailed metadata, and multimodal features, making it a robust resource in emotional speech analysis research. Researchers should be aware of dataset-specific considerations when interpreting results and generalizing findings beyond a dataset's scope.

4. Experimental Setup

The experimental setup entails the development and evaluation of a recurrent neural network (RNN) with long short-term memory (LSTM) architecture for emotion recognition using audio datasets. Inspired by [20] work on CNNs for environmental sound recognition, this study extends their approach to emotion recognition. The LSTM model, comprising four LSTM layers, a dropout layer, and a dense layer, aims to effectively identify speech sections with relevant information. Additionally, the study introduces the Gray Wolf Optimizer (GWO) requiring optimization. In terms of model architecture, the CNN-LSTM ensemble incorporates both convolutional and recurrent layers to capture spatial and temporal dependencies in the audio data. Convolutional layers extract relevant features from the raw audio waveforms, while LSTM layers process sequences of features over time. Meanwhile, the SVM serves as a traditional yet robust classifier for emotion recognition tasks, leveraging its ability to find the optimal hyperplane to separate different emotion classes in the feature space.

Furthermore, the methodology unfolds in three main stages: feature extraction, feature selection (using GWO), and classification. Feature extraction involves extracting meaningful representations from the raw audio waveforms, such as Mel-frequency cepstral coefficients (MFCCs) or logMel features, to capture acoustic characteristics related to different emotions. Feature selection using GWO aims to identify the most discriminative features for emotion recognition, enhancing the model's performance. Subsequently, the selected features are fed into the classification stage, where SVM and CNNs handle the task of classifying emotions based on the extracted features. Specifically, GWO-SVM, GWO-CNN, and GWO-LSTM approaches are explored, each involving initialization, evaluation, updating the GWO population, and a stopping criterion. Key components include solution representation and fitness function, crucial for feature selection optimization. This com-

prehensive approach leverages both traditional and deep learning techniques, along with optimization algorithms, to optimize emotion recognition performance.

The GWO optimizer was introduced by Seyedali Mirjalili and Seyed Mohammad Mirjalili in 2014 [50]. GWO mimics wolf hunting strategies to solve optimization problems. It maintains a population of alpha, beta, delta, and omega wolves, updating their positions iteratively based on fitness and social interactions. GWO efficiently explores and exploits the search space through equations simulating hunting behavior. The methodology involves three steps: feature extraction, feature selection, and classification. Figure 1 displays the architecture of the conventional RNN-LSTM model for feature extraction and classification.

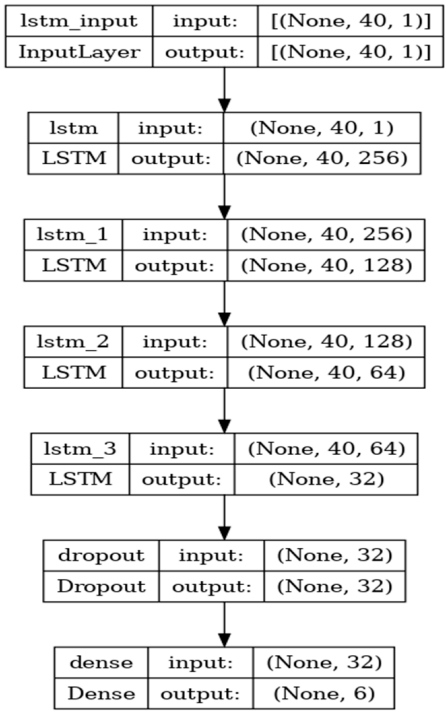


Figure 1. RNN-LSTM architecture.

GWO adapts as a feature selection method to identify crucial emotional features. Support vector machines (SVM) and convolutional neural networks (CNNs) handle the classification task. GWO-SVM, GWO-CNN, and GWO-LSTM approaches involve initializing, evaluating, and updating the GWO population and the stopping criterion. Key components include solution representation and fitness function, crucial for feature selection optimization.

This streamlined research methodology leverages GWO to optimize emotion feature selection, demonstrating its adaptability with SVM and CNNs to provide accurate classification. Figure 2 gives and abstract overview of proposed research.

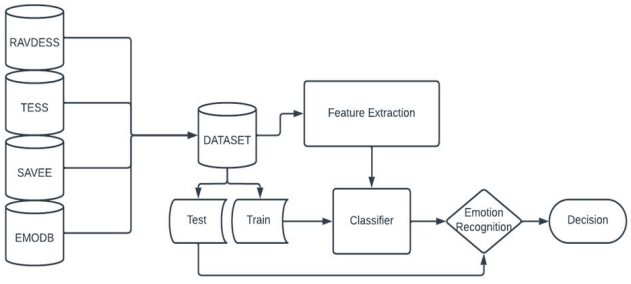


Figure 2. Methodology.

4.1. Social Hierarchy

In the social structure of gray wolves, a cohesive cultural dominance system is observed, comprising four distinct categories: alpha, beta, delta, and omega. The alpha, occupying the highest position, serves as the dominant authority and demonstrates superior intelligence in pack management. Assisting the alpha are the beta wolves, who play decision-making roles at the second level of the social order. The omega wolves make up the majority of the pack and are positioned at the lowest tier. Although not explicitly mentioned in the hierarchy levels, the delta wolves follow the beta wolves and function as leaders among the omega-level members. In the context of GWO solutions, they are classified based on the grey wolf social order, with the alpha wolf considered the most fit, followed by the beta and delta wolves. Figure 3 provides a visual representation of the social hierarchy within a wolf pack.

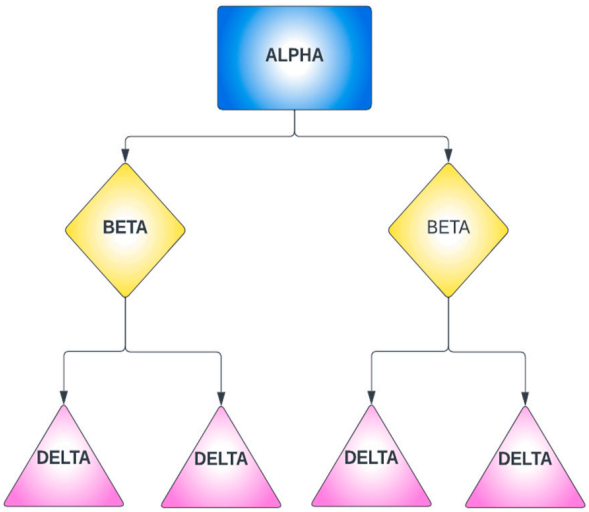


Figure 3. Social Hierarchy of the wolves within a pack.

4.2. Hunting Strategy

The hunting tactics utilized by a wolf pack entail a strategic process with three key stages: pursuit, encirclement, and assault. In the encirclement phase, a wolf exhibits the skill to tactically adjust its position, skillfully surrounding the prey within a defined area. This encircling behavior of gray wolves can be expressed mathematically as follows:

$$\mu = |v \times X'(i) - X(i)|, \tag{1}$$

$$X(i+1) - X'(i) - \eta \times \mu, \tag{2}$$

where η and v represent two controlling factors, X' represents the prey's location, X denotes the wolf's current location, and *i* indicates the present iteration. The values of the controllers η and v) are determined using the following calculations:

$$\eta = 2 \times \iota \times r_1 - \iota \tag{3}$$

$$v = 2 \times r_2 \tag{4}$$

where ι linearly decreases from 2 to 0 over the iterations, while the values r_1 and r_2 are random values within the range of [0, 1]. The range of the controller η is bounded by the interval $[-2\iota, 2\iota]$, which is determined by the value of ι.

4.3. Prey Search (Exploration)

The arrangement of alpha, beta, and delta wolves significantly influences the search behavior exhibited by other members of the pack as they pursue their prey. Throughout the

prey search, the wolves disperse from one another and subsequently converge to encircle and launch an attack on their target. The wolves' dispersal is symbolized by the variable η, which assumes random values greater than 1 or less than −1, guiding the movement of search agents away from the prey's location. This process is designed to ensure thorough exploration and enhance the GWO capacity for a comprehensive global search to attain the most optimal solution. In the GWO algorithm, the variable v governs the exploration phase, comprising random values between 0 and 2. This enables the random emphasis or de-emphasis of the prey's significance, contingent on whether v exceeds or is less than 1, respectively. Unlike η, v does not undergo linear diminishment. The utilization of random values in GWO serves to emphasize both exploitation and exploration, extending to the final iteration. This feature proves crucial in scenarios where search agents may risk being confined to local optima, particularly in the later stages of the search process.

The mathematical model for encircling the prey involves a linear deduction of the value of ι. The fluctuation range of η is reduced to a similar degree by the parameter ι. In situations where η consists of random values between 1 and −1, the location of the search agent can be updated to a position near the prey's location, facilitating a more focused exploitation of the prey. In situations where η consists of random values between 1 and −1, the location of the search agent can be updated to a position near the prey's location, facilitating a more focused exploitation of the prey.

5. Results and Discussion

In this study, Python serves as the programming language for implementing our methodologies. Our central objective revolves around the integration of the Grey Wolf Optimizer (GWO) to optimize emotion features, thereby enhancing emotion recognition performance. To validate the efficacy of our approach, we conduct experiments across four distinct emotion datasets: Emotion Database, RAVDESS, TESS, and SAVEE. Our evaluation metrics encompass classification accuracy, precision, recall, and F1-score, providing comprehensive insights into the effectiveness of the proposed methodology.

It can be seen from Table 3 that the GWO optimizer, when used with classifiers like SVM, LSTM, and CNN, surpasses the conventional classification methods on all evaluation metrics. In EmoDB, the traditional Kernalized SVM shows a classification accuracy of only 55%, which is the worst among the six models. The use of GWO improved accuracy to 85%. The difference in accuracy is an improvement of 30% for EmoDB and RAVDESS, whereas the precision doubled for EmoDB and RAVDESS and improved 7× for SAVEE, which is a huge jump from that of traditional SVM classifiers. While the use of CNNs is not as accurate as SVMs, the GWO technique managed to improve all the measurement parameters of CNN as well.

Figure 4a,b display the confusion matrix representing a set of emotions comprising six categories: anger, happiness, neutral, fear, disgust, and sadness for RAVDESS. The proposed model achieved an overall accuracy of 53% when evaluated using this particular emotion set and an accuracy of 60% when evaluated with the GWO technique. Upon analyzing the confusion matrix (Figure 4a,b), it becomes apparent that the accuracy rates for the fear and neutral classes are relatively high compared to those of the other classes. In contrast, the angry class exhibits a lower classification accuracy. Additionally, there is a significant misclassification of neutral as sad and disgust as happy. Despite the underperformance of the angry class, the proposed model demonstrated effective distinction among the other emotions within this emotion set. Figure 4b displays the classification after using GWO, which, as can be seen, lowers the misclassifications by a small degree.

Figure 4c,d display the confusion matrix representing a set of emotions comprising six categories: anger, happiness, neutral, fear, disgust, and sadness for SAVEE. The proposed model achieved an accuracy of 57% when evaluated using this particular emotion set and an accuracy of 68% when evaluated with the Gray Wolf Optimization technique. Upon analyzing the confusion matrix (Figure 4c,d), it becomes apparent that the accuracy rates for the angry class are relatively high compared to the other classes, whereas the fear class

exhibits the lowest classification accuracy. Despite the underperformance of all the classes, the proposed model demonstrated effective distinction among the other emotions within this emotion set. Figure 4d displays classification after using GWO, which, as can be seen, reduces the incidence of misclassifications by a small degree.

Figure 4e,f display the confusion matrix representing a set of emotions comprising six categories: anger, happiness, neutral, fear, disgust, and sadness for EmoDB. The proposed model achieved an overall accuracy of 75% when evaluated using this particular emotion set and an accuracy of 78% when evaluated with the Gray Wolf Optimization technique. Upon analyzing the confusion matrix, it becomes apparent that the accuracy rates for the disgust class are relatively high compared to the other classes, whereas the fear class exhibits a lower classification accuracy. The fear class is rarely classified accurately and is the lowest correctly identified class. Figure 4f shows the classification after using GWO, which, as can be seen, lowers the misclassifications by a small degree.

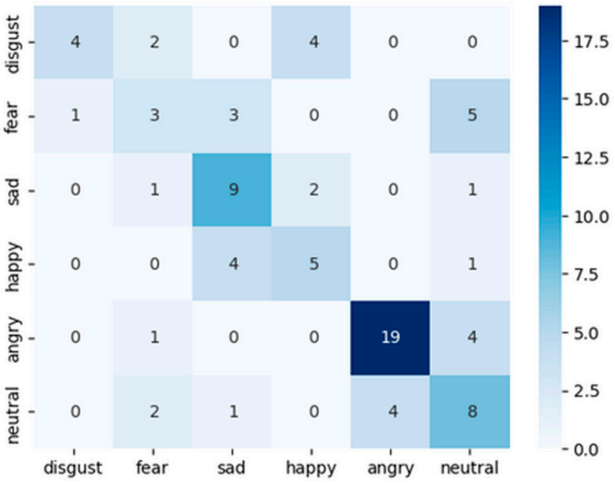
Table 3. Classification Performance of 6 classifiers on the four datasets.

		Without GWO Optimizer				GWO Optimizer			
Datasets	**Measurements**	**SVM**	**K-SVM**	**CNN**	**LSTM**	**SVM**	**K-SVM**	**CNN**	**LSTM**
EmoDB	Accuracy	0.74	0.55	0.75	0.76	0.85	0.85	0.78	0.83
	Precision	0.71	0.42	0.74	0.73	0.82	0.87	0.77	0.82
	Recall	0.70	0.45	0.69	0.71	0.82	0.84	0.77	0.87
	F1-Score	0.70	0.39	0.71	0.71	0.82	0.84	0.77	0.84
RAVDESS	Accuracy	0.74	0.55	0.53	0.73	0.85	0.87	0.60	0.73
	Precision	0.71	0.42	0.53	0.73	0.84	0.88	0.59	0.69
	Recall	0.70	0.45	0.52	0.82	0.83	0.85	0.56	0.71
	F1-Score	0.70	0.39	0.52	0.76	0.83	0.86	0.59	0.71
SAVEE	Accuracy	0.54	.35	0.57	0.62	0.76	0.75	0.68	0.71
	Precision	0.53	0.10	0.59	0.53	0.74	0.72	0.66	0.59
	Recall	0.51	0.25	0.51	0.57	0.75	0.68	0.64	0.72
	F1-Score	0.52	0.15	0.53	0.59	0.73	0.68	0.64	0.71
TESS	Accuracy	0.98	0.91	0.99	0.97	0.99	0.99	0.99	0.99
	Precision	0.98	0.93	0.99	0.96	0.99	0.99	0.99	0.99
	Recall	0.98	0.91	0.99	0.95	0.99	0.99	0.99	0.99
	F1-Score	0.98	0.91	0.99	0.96	0.99	0.99	0.99	0.99

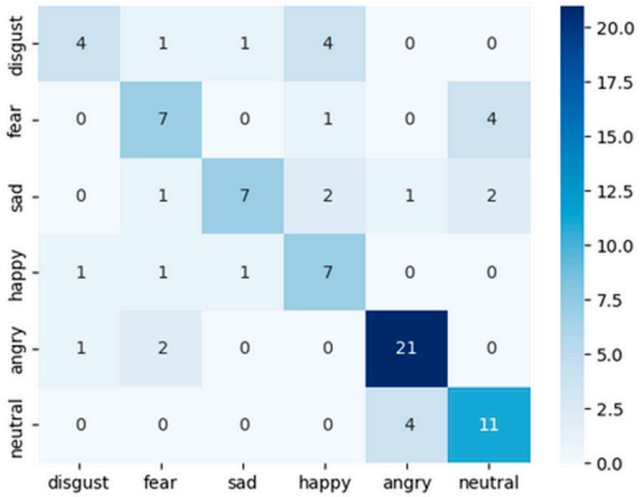
(**a**) Confusion Matrix for RAVDESS CNN

(**b**) Confusion Matrix for RAVDESS GWO-CNN

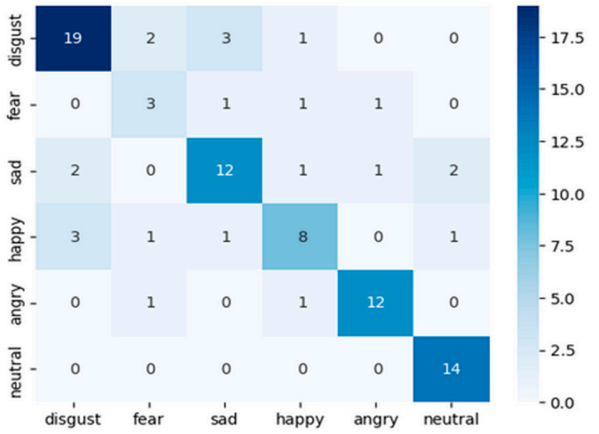
Figure 4. *Cont.*



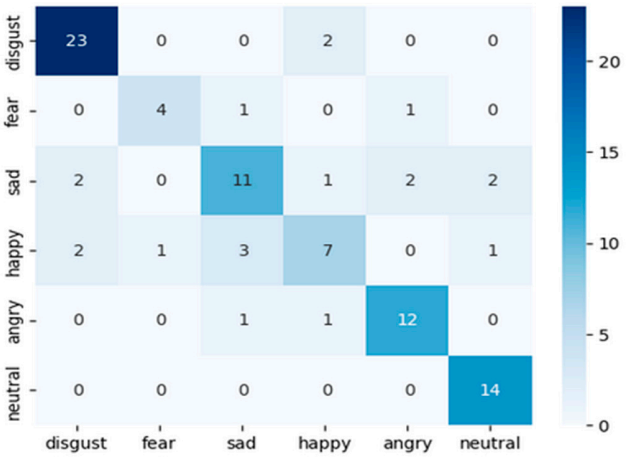
(**c**) Confusion Matrix for SAVEE CNN



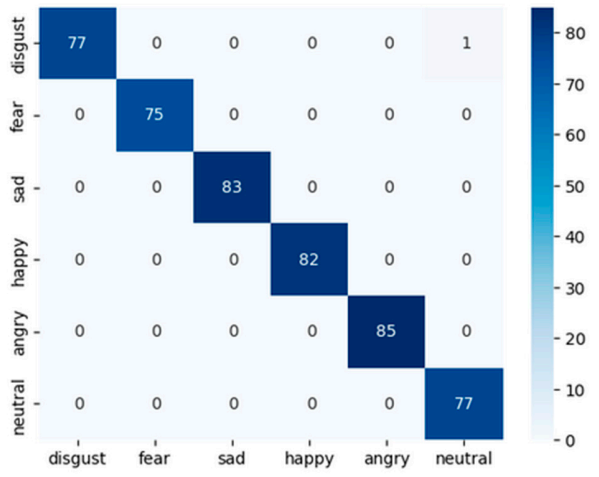
(**d**) Confusion Matrix for SAVEE GWO-CNN



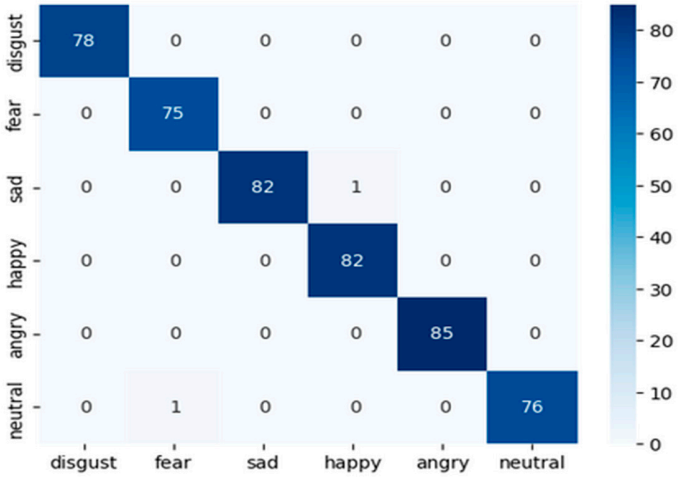
(**e**) Confusion Matrix for EmoDB-CNN



(**f**) Confusion Matrix for EmoDB GWO-CNN



(**g**) Confusion Matrix for TESS CNN



(**h**) Confusion Matrix for TESS GWO-CNN

Figure 4. The confusion matrix of GWO-CNN and CNN model was tested on four datasets—RAVDESS, SAVEE, TESS, and EMODB.

Figure 4g,h display the confusion matrix, representing a set of emotions comprising six categories: anger, happiness, neutral, fear, disgust, and sadness for TESS. The proposed model achieved an overall accuracy of 99% when evaluated using this particular emotion set and an accuracy of 100% when evaluated with the Gray Wolf Optimization technique. The TESS data are highly recommendable for speech emotion recognition as all the emotions are accurately classified by our model as well. The TESS dataset does not need Gray Wolf Optimization since it is already good enough to be used for accurate emotion classification via traditional methods.

In Figure 5a, the horizontal axis represents the number of epochs, and the vertical axis represents the accuracy and the validation accuracy through training. The accuracy peaks at nearly 90% once, while the validation accuracy rises with it to around 70% and then flattens. The model was set to train for 100 epochs with an Early Stopping Callback. This stopped the training at 61 epochs, denoting that training the model will no longer improve the results. As visible in Figure 5b, the horizontal axis represents the number of epochs, and the vertical axis represents the loss and the validation loss through training. The loss is relatively high at the beginning of training but decreases gradually. After a while, the validation loss starts to saturate and does not fall below the 0.11 mark. As visible in Figure 5c, the accuracy is relatively low at the beginning of training, but it increases gradually. The accuracy peaks at nearly 99% once, while the validation accuracy rises slower than that. The model was set to train for 100 epochs with an Early Stopping Callback. This stopped the training at 67 epochs, denoting that training the model will no longer improve the results. The testing accuracy is only 53%. As visible in Figure 5d, the x-axis represents the number of epochs, and the y-axis represents the loss and the validation loss through training. It is evident that the validation loss does not decrease and instead fluctuates between 0.14 and 0.12 at all times. This is why the testing accuracy and the validation accuracy are very low.

In Figure 5e, the horizontal axis represents the number of epochs, and the vertical axis represents the accuracy and the validation accuracy through training. The training accuracy climbs rapidly, whereas the callback stops the network after 56 epochs because further training will not yield useful results and will only result in overtraining. This is because the validation loss does not decrease. The model cannot accurately distinguish the emotions in the testing phase, and the validation loss always stays between 0.11 and 0.09.

Figure 5g shows the accuracy graph for the TESS dataset. There is again an exception to note, namely that the TESS dataset being balanced and great for SER produces excellent results. The accuracy climbs to greater than 95% within the first five epochs and then saturates around the 99% mark, ranging from 99.1% to 99.73% on training and validation. This is shown in the loss graph (Figure 5h) as well, which shows the loss nearing 0.01 as the epochs progress. The TESS dataset exhibits the highest values in all the parameters measured, with an average value of 99% in F1-score, precision, recall, and accuracy.

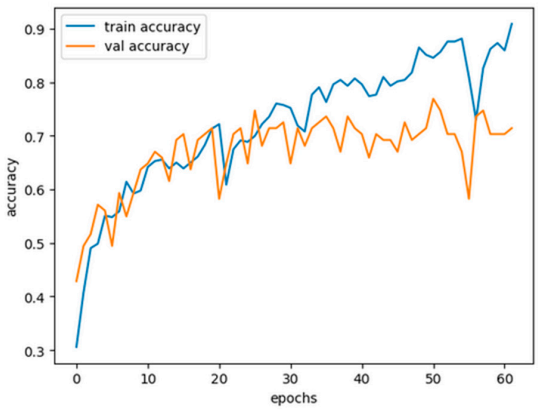
Table 4 shows a comparison of the proposed model with the existing studies conducted for the respective datasets. It can be seen that our proposed method outperforms the existing methods by good margins. Since TESS is a dataset that often displays an accuracy greater than 99%, there are not many comparisons for it. That is why RAVDESS, SAVEE, and EmoDB were given more emphasis during research and comparison. The average accuracy for the RAVDESS dataset lies between 60–75%, whereas our model achieved an accuracy of 87% with GWO optimization. The CNN and LSTM, however, underperformed for both RAVDESS and SAVEE datasets without the GWO optimizer. The GWO-SVM beat the traditional DNN frameworks and existing SVM models.

Table 4. Comparison with existing studies.

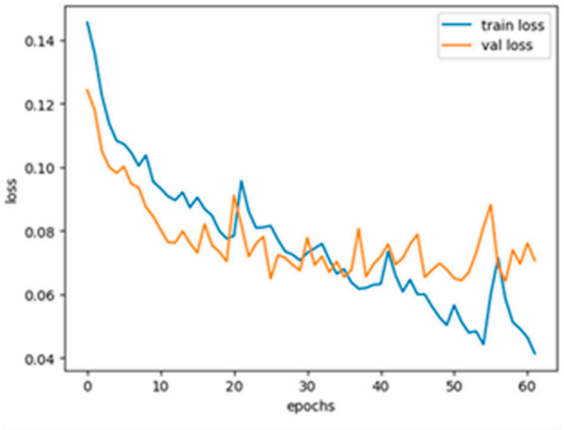
Reference	Dataset	Classifier Used	Accuracy
Bhavan et al. [44]	RAVDESS	Bagged ensemble of SVMs	75.69%
Zeng et al. [42]	RAVDESS	DNNs	64.52%
Shegokar and Sircar [43]	RAVDESS	SVMs	60.1%

Table 4. *Cont.*

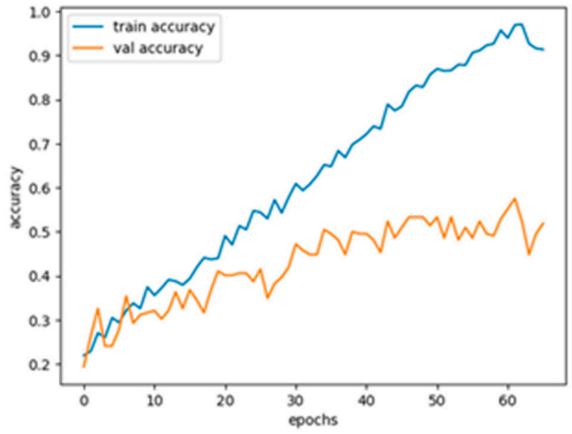
Reference	Dataset	Classifier Used	Accuracy
This Work (Proposed Method)	SAVEE	GWO-SVM	75%
		GWO-CNN	65.47%
This Work (Proposed Method)	EmoDB	GWO-SVM	85%
		GWO-CNN	78%
This Work (Proposed Method)	TESS	GWO-SVM	99.97%
		GWO-CNN	99.93%



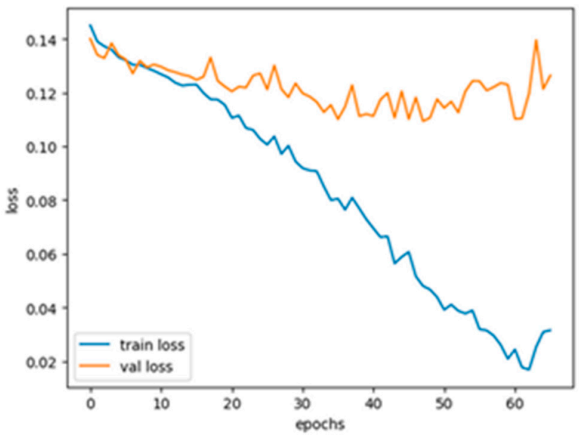
(**a**) Accuracy vs. Epochs for EmoDB



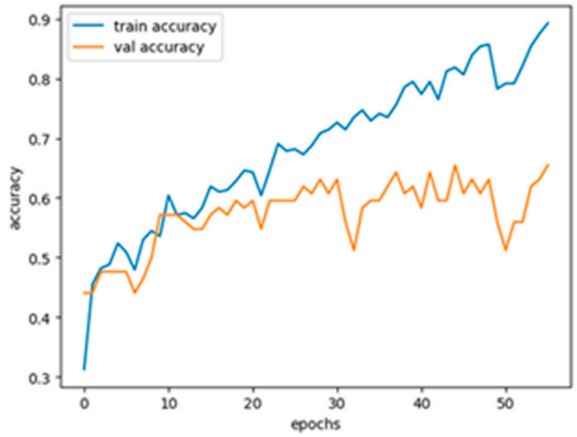
(**b**) Loss vs. Epochs for EmoDB



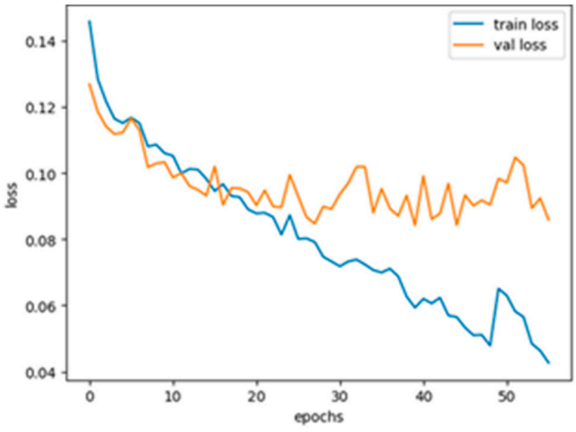
(**c**) Accuracy vs. Epochs for RAVDESS



(**d**) Loss vs. Epochs for RAVDESS

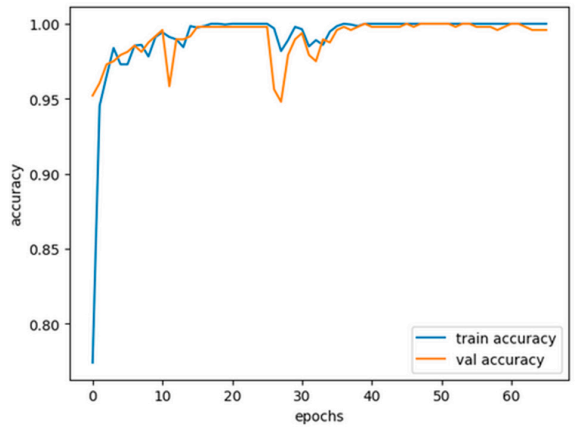


(**e**) Accuracy vs. Epochs for SAVEE

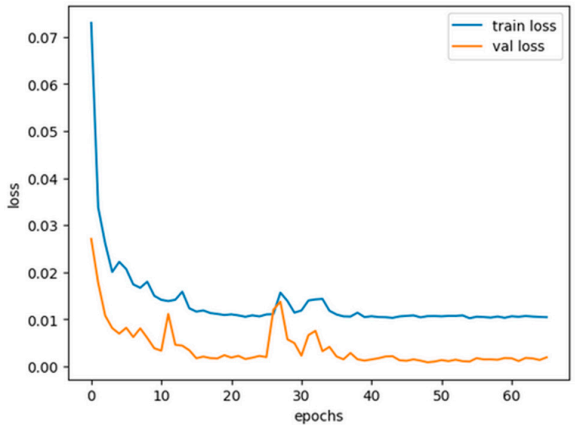


(**f**) Loss vs. Epochs for SAVEE

Figure 5. *Cont.*



(**g**) Accuracy vs. Epochs for TESS



(**h**) Loss vs. Epochs for TESS

Figure 5. The accuracy and loss by epochs on four different datasets—RAVDESS, SAVEE, TESS, and EMODB, they should be listed as: (5 (**a**–**h**)).

6. Conclusions and Future Work

In this research study, a deep learning model was introduced for speech emotion recognition. The effectiveness of the proposed model was assessed using four datasets: EmoDB, RAVDESS, TESS, and SAVEE. The experimental findings demonstrated that the proposed model achieved better results with GWO, irrespective of the model used on the four datasets. The evaluations revealed that the results obtained were comparable to the accuracy of other convolutional neural networks (CNNs) utilizing spectrogram features. Consequently, it can be concluded that the proposed approach is highly suitable for emotion recognition.

It is recommended to evaluate the proposed approach further using additional emotion datasets. Additionally, the proposed model can be enhanced by utilizing large datasets to improve the recognition of all emotions. Another solution to increase the accuracy of the model could be to combine all four datasets present and extract common emotions from them. This will result in a more balanced dataset that will show promising results based on overall SER instead of individual datasets, with constraints being used for the same objective.

In the future, human assistance will no longer be necessary for speech recognition tasks as they transition into automated processes. Voice recognition is likely to become a seamless and automatic function. It would not be surprising if we are eventually all carrying earpieces like C-3PO that listen to our conversations. Ongoing research and development in deep learning are expanding the possibilities of speech recognition. Exciting advancements are being made, including the utilization of neural networks to analyze sound patterns, resulting in improved artificial intelligence algorithms. Speech recognition, a subset of deep learning, has been a subject of study for over five decades. Presently, speech recognition systems exhibit higher accuracy levels than ever before. The future holds even more promise for speech recognition as deep learning techniques enable training models on larger datasets and utilize increased computing power, enhancing the algorithms' data processing capabilities.

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

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. The datasets RAVDESS, TESS, SAVEE, and EmoDB might be available on Kaggle (https://www.kaggle.com) (5 November 2023).

Acknowledgments: The authors would like to thank the "Doctoral School of Applied Informatics and Applied Mathematics" and the "High Performance Computing Research Group" of Óbuda University for their valuable support. The authors would like to thank NVIDIA Corporation for providing graphics hardware for the experiments.

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

References

1. Banse, R.; Scherer, K.R. Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psychol.* **1996**, *70*, 614–634. [CrossRef]
2. Mustafa, M.B.; Yusoof, A.M.; Don, Z.M.; Malekzadeh, M. Speech emotion recognition research: An analysis of research focus. *Int. J. Speech Technol.* **2018**, *21*, 137–156. [CrossRef]
3. Schuller, B.; Rigoll, G.; Lang, M. Hidden markov model-based speech emotion recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, China, 6–10 April 2003; Volume 2, p. II-1.
4. Hu, H.; Xu, M.-X.; Wu, W. GMM supervector based SVM with spectral features for speech emotion recognition. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. IV-413–IV-416.
5. Lee, C.; Mower, E.; Busso, C.; Lee, S.; Narayanan, S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* **2009**, *4*, 320–323.
6. Kim, Y.; Mower, E. Provost, Emotion classification via utterance level dynamics: A pattern-based approach to characterizing affective expressions. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013.
7. Eyben, F.; Wollmer, M.; Schuller, B. Openear—Introducing the munich open-source emotion and affect recognition toolkit. In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction (ACII), Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–6.
8. Mower, E.; Mataric, M.J.; Narayanan, S. A framework for automatic human emotion classification using emotion profiles. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 1057–1070. [CrossRef]
9. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the INTERSPEECH 2014, Singapore, 7–10 September 2014; pp. 223–227.
10. Jin, Q.; Li, C.; Chen, S.; Wu, H. Speech emotion recognition with acoustic and lexical features. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 4749–4753.
11. Lee, J.; Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In Proceedings of the INTERSPEECH 2015, Dresden, Germany, 6–10 September 2015; pp. 223–227.
12. Neumann, M.; Vu, N.T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1263–1267.
13. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Zafeiriou, S.; Schuller, B. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
14. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and recurrent neural networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Republic of Korea, 13–16 December 2016; pp. 1–4.
15. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
16. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
17. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion recognition from variable-length speech segments using deep learning on spectrograms. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 3683–3687.
18. Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, P. Speech emotion recognition using spectrogram phoneme embedding. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 3688–3692.
19. Guo, L.; Wang, L.; Dang, J.; Zhang, L.; Guan, H. A feature fusion method based on extreme learning machine for speech emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2666–2670.

20. Dai, W.; Dai, C.; Qu, S.; Li, J.; Das, S. Very deep convolutional neural networks for raw waveforms. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 421–425.
21. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the INTERSPEECH 2005, Libon, Portugal, 4–8 September 2005; pp. 1517–1520.
22. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
23. Shao, S.; Saleem, A.; Salim, H.; Pratik, S.; Sonia, S.; Abdessamad, M. AI-based Arabic Language and Speech Tutor. In Proceedings of the 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 5–8 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.
24. Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; Tarokh, V. Speech emotion recognition with dual-sequence LSTM architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6474–6478.
25. Chernykh, V.; Sterling, G.; Prihodko, P. Emotion recognition from speech with recurrent neural networks. *arXiv* **2017**, arXiv:1701.08071.
26. Sathiyabhama, B.; Kumar, S.U.; Jayanthi, J.; Sathiya, T.; Ilavarasi, A.K.; Yuvarajan, V.; Gopikrishna, K. A novel feature selection framework based on grey wolf optimizer for mammogram image analysis. *Neural Comput. Appl.* **2021**, *33*, 14583–14602. [CrossRef]
27. Sreedharan, N.P.N.; Ganesan, B.; Raveendran, R.; Sarala, P.; Dennis, B.; Boothalingam, R.R. Grey wolf optimisation-based feature selection and classification for facial emotion recognition. *IET Biom.* **2018**, *7*, 490–499. [CrossRef]
28. Dey, A.; Chattopadhyay, S.; Singh, P.K.; Ahmadian, A.; Ferrara, M.; Sarkar, R. A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. *IEEE Access* **2020**, *8*, 200953–200970. [CrossRef]
29. Shetty, S.; Hegde, S. Automatic classification of carnatic music instruments using MFCC and LPC. In *Data Management, Analytics and Innovation*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 463–474.
30. Saldanha, J.C.; Suvarna, M. Perceptual linear prediction feature as an indicator of dysphonia. In *Advances in Control Instrumentation Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 51–64.
31. Mannepalli, K.; Sastry, P.N.; Suman, M. Emotion recognition in speech signals using optimization based multi-SVNN classifier. *J. King Saud Univ.-Comput. Inf. Sci.* **2018**, *34*, 384–397. [CrossRef]
32. Yildirim, S.; Kaya, Y.; Kılıç, F. A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. *Appl. Acoust.* **2021**, *173*, 107721. [CrossRef]
33. Kerkeni, L.; Serrestou, Y.; Mbarki, M.; Raoof, K.; Mahjoub, M.A.; Cleder, C. Automatic speech emotion recognition using machine learning. In *Social Media and Machine Learning*; IntechOpen: London, UK, 2019.
34. Shen, P.; Changjun, Z.; Chen, X. Automatic speech emotion recognition using support vector machine. In Proceedings of the 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, Harbin, China, 12–14 August 2011; IEEE: Piscataway, NJ, USA, 2011; Volume 2, pp. 621–625.
35. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
36. Gomathy, M. Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. *Int. J. Speech Technol.* **2021**, *24*, 155–163. [CrossRef]
37. Daneshfar, F.; Kabudian, S.J. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimed. Tools Appl.* **2020**, *79*, 1261–1289. [CrossRef]
38. Shahin, I.; Hindawi, N.; Nassif, A.B.; Alhudhaif, A.; Polat, K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Syst. Appl.* **2022**, *188*, 116080. [CrossRef]
39. Kanwal, S.; Asghar, S. Speech emotion recognition using clustering based GA- optimized feature set. *IEEE Access* **2021**, *9*, 125830–125842. [CrossRef]
40. Zhang, Z. Speech feature selection and emotion recognition based on weighted binary cuckoo search. *Alex. Eng. J.* **2021**, *60*, 1499–1507. [CrossRef]
41. Wolpert, D.H. The lack of a priori distinctions between learning algorithms. *Neural Comput.* **1996**, *8*, 1341–1390. [CrossRef]
42. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. *Multimed. Tools Appl.* **2019**, *78*, 3705–3722. [CrossRef]
43. Shegokar, P.; Sircar, P. Continuous wavelet transform based speech emotion recognition. In Proceedings of the 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), Surfers Paradise, QLD, Australia, 19–21 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–8.
44. Bhavan, A.; Chauhan, P.; Shah, R.R.; Hitkul. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [CrossRef]
45. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
46. Özseven, T. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326. [CrossRef]

47. Singh, P.; Sahidullah; Saha, G. Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Commun.* **2023**, *146*, 53–69. [CrossRef]
48. Gerczuk, M.; Amiriparian, S.; Ottl, S.; Schuller, B.W. EmoNet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1472–1487. [CrossRef]
49. Avila, A.R.; Akhtar, Z.; Santos, J.F.; Oshaughnessy, D.; Falk, T.H. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Trans. Affect. Comput.* **2021**, *12*, 177–188. [CrossRef]
50. Seyedali, M.; Mohammad, I.S.; Andrew, L. Grey Wolf Optimizer. *Adv. Eng. Softw.* **2014**, *69*, 46–61.