Preprocessing Method for Class Noise Treatment in Speech Emotion Recognition

ISSN 1870-4069

Randy Brandon Gallegos-Rodríguez[1],
Bernardo Garcia Bulle-Bueno[2],
Ana María Magdalena Saldaña-Pérez[1]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

[2] Institute of Data, Systems and Society,
MIT Cambridge,
USA

{rgallegosr2023, amagdasaldana}@cic.ipn.mx,
bernard0@mit.edu

Abstract. Speech emotion recognition datasets can have class noise due to subjectivity during the labeling process or because of automatic labeling. Class noise has been neglected until recently in machine learning research. In this work, we study the effects of controlled class noise in 10 datasets from different languages. We find that low levels of class noise (5-10%) do not significantly affect the performance of classifiers, but higher levels of class noise severely impact performance. Support Vector Machines (SVMs) appear to be the best candidate for handling class noise across most datasets and noise levels compared to other traditional algorithms. We propose a preprocessing method that effectively corrects mislabeled samples at moderate and high noise levels, enhancing the model's performance as measured by balanced accuracy.

Keywords: SER, class noise, support vector machines, preprocessing, ensemble.

1 Introduction

Speech Emotion Recognition (SER) is a task of affective computing, the subfield of artificial intelligence dedicated to recognizing human emotions, sentiments, and feelings [1]. The SER problem can be seen as a classification problem, where the aim is to find a proper model (classifier) that maps samples from the input space $\mathbf{X}$ to one of the c discrete labels or classes in a set of labels C. For the SER problem, as its name implies, the input samples are audio recordings, and the labels are the set of considered emotions. Fig. (1) is shows the most common methodologies used in SER. More details about the current state of the literature for the SER problem will be discussed in the *State of the Art* section.

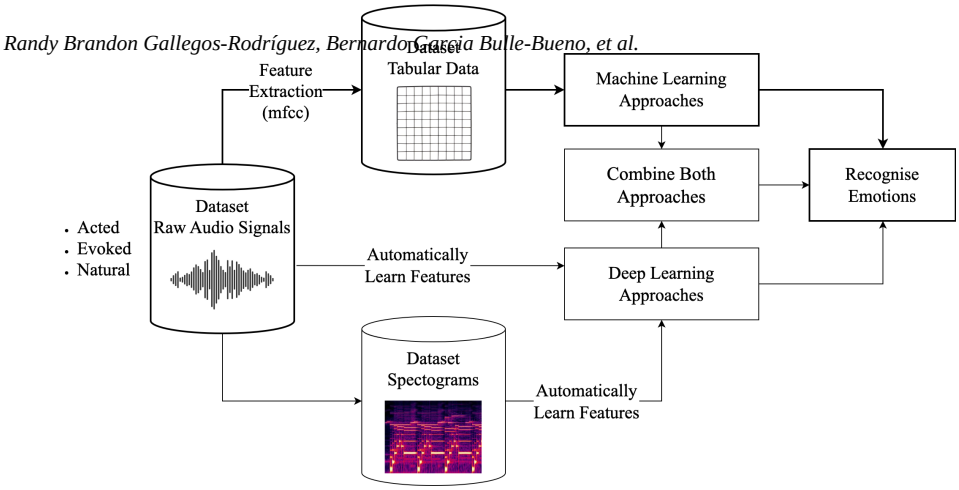

Fig. 1. Most common methodologies followed in the SER problem. Some strategies involve deep learning approaches. Thicker lines highlight the path followed in this work, consisting of extracting 13 Mel-frequency cepstral coefficients from audio recordings and implementing traditional machine learning algorithms. Image inspired by [2].

Since the SER problem can be treated as a classification problem, typical challenges in classification, such as noise, are also present in SER. Noise can be simply thought of as errors in the data; according to [8], it is also referred to as irregularities or corruptions in a dataset. For a more formal definition of noise, [9] briefly introduces the topic in the context of the Probably Approximately Correct (PAC) theory. PAC is a mathematical formalization of machine learning proposed by Valiant in [10]. For regression problems, the expected test error depends on variance, bias, and noise, as can be seen in Eq. 1:

$$Err(x_0) = \underbrace{[E\hat{f}(x_0) - f(x_0)]^2}_{Bias^2} + \underbrace{E[\hat{f}(x_0) - E\hat{f}(x_0)]^2}_{Variance} + \underbrace{\sigma^2}_{Noise}, \tag{1}$$

where x_0 is a sample from input space, $\hat{f}$ is the estimated function and f is the actual value according to the dataset. A complete derivation can be found in [11]. This result is studied in the bias-variance decomposition and is typically derived for regression due to its simplicity.

Noise can reduce the performance of the model, increase the complexity of the model, and extend the training time [8]. In [12], two categories of noise are presented: attribute noise and class noise. More details about these types of noise and methods for their treatment will be provided in the *State of the Art* section.

1.1 Outline of the Work

The remainder of this paper is structured as follows: In the next section, State of the Art, we provide a detailed review of the current literature on

SER and noise in classification tasks. Following this, the Research Goal section underscores the significance of addressing class noise in SER and outlines the objectives of this study. In Criteria for Effective Noise Treatment, we present a theoretical framework that guides the experimental design, which is detailed in the Experimental Framework section. The remainder of the paper follows the conventional structure of a scientific paper, including the presentation of results, conclusions and future work.

2 State of Art

2.1 Speech Emotion Recognition

In [2], a systematic literature review is presented. They consider seven key questions related to SER; one of these questions is, "What type of speech datasets[3] are used for SER?" They proposed three main categories: acted datasets, evoked or elicited datasets, and natural datasets.

Acted Datasets Acted datasets represent the most common type of SER datasets [2]. One example of an acted dataset is the Mexican Emotional Speech Database (MESD) [4]. For this dataset (and for many other acted datasets), a variety of actors (differing in age, gender, or accent) are chosen to speak words or phrases while acting out an emotion.

Elicited Datasets Elicited datasets are created by eliciting emotions in participants. One example of an elicited dataset is EmoMatchSpanishDB [5]. For this dataset, the labeling process was done through a crowdsourcing method.

Natural Datasets Natural datasets can be created in different ways, such as from TV shows, interviews, and conversations between customer care agents and customers [2, 7]. One example of this type of dataset is the RECOLA dataset, which is a multimodal corpus, meaning it includes more than one modality. In this case, it includes audio and video data, as well as physiological data, namely electrocardiogram and electrodermal activity. The data were obtained through a video conference where participants were asked to complete tasks.

Feature Extraction SER task follows the typical pipeline of any classification problem. If it is opted to extract features from the audio signal, instead of working directly with the raw signal, extracting Mel Frequency Cepstral Coefficients (MFCCs) is the most widely set of features employed [7]. According to [21], MFCCs are derived by capturing the envelope of the short-time power

[3] In the SER field, it is common to refer to the set of labeled audios as a database. In this work, we adopt the term *dataset*, more commonly used in other fields of machine learning.

spectrum, representing the vocal tract shape. The process involves segmenting utterances and converting them into the frequency domain using the short-time discrete Fourier transform to obtain MFCCs. Subsequently, Mel filter banks are employed to calculate energies in several sub-bands, followed by computing the logarithm of respective sub-band energies. Finally, MFCCs are obtained by applying the inverse Fourier transform.

Classification algorithms used in SER According to [3], there is currently no consensus on a state-of-the-art algorithm for SER that performs optimally under all conditions. It is recommended to conduct preliminary research to choose the most appropriate classification algorithm. For this reason, different algorithms will be tested in this work to compare their performance under specific noise conditions.

2.2 Noise in Classification Tasks

Attribute Noise Attribute noise refers to errors or corruptions in instances of the input space X; in particular, for the SER problem, this type of noise involves corruptions in the audio data. Attribute noise in SER has been extensively studied. For example, [13] examines the effect of white noise, [14] explores the impact of reverberation and Gaussian noise, and [15] presents a survey of SER in natural environments, which often include background noise.

In [16], a strategy is proposed to handle attribute noise for general machine learning tasks (tabular data), achieving new state-of-the-art results. Their strategy aims to correct attribute noise rather than deleting samples with potential noise. This approach is proposed because, as discussed in [12], removing noisy samples can eliminate noise but may also discard valuable information in some cases. Similar approaches, where noisy samples in unstructured data are corrected rather than removed, can be found in [14]. In this study, white noise in audio is handled with speech enhancement, followed by feature extraction to convert the problem to tabular classification.

Class Noise Class noise refers to errors or corruptions in the attribute class, i.e., having samples mislabelled. In the context of SER, this means having audio samples in the dataset labelled with an inappropriate emotion. [12] found that class noise is generally more harmful than attribute noise for classification tasks. This may be due to the fact that a mislabelled sample (x_i, y_i) completely meaningless, which is the information received by the algorithm during training.

In [17], SVMs were used to detect suspicious labels and correct them through expert supervision. This method showed improvement but may be unsustainable at a large scale. Other strategies involve eliminating noisy samples and are called filters. Examples of these filters include *edited nearest neighbor*, which removes samples inconsistent with their k nearest neighbors; filters based on the voting of ensemble methods; or filters that eliminate misclassified samples

through cross-validation. For a more comprehensive review of filter methods, see [18].

According to [16], the two main strategies to handle both attribute and class noise are using robust algorithms (algorithms not sensitive to noise) or preprocessing techniques (such as filters).

3 Research Goal

3.1 Importance of Study Class Noise in SER Problem

In [5], it is mentioned that creating a SER dataset is expensive due to the need for actors. This limits the size of acted and elicited SER datasets. For this reason, deep learning approaches for SER have not yet reached their full potential [6]. These aspects highlight the need for new, scalable methods for creating SER datasets. For example, [19] exploits the current emergence of large-scale video available on social networks by implementing cross-modal techniques (audio and video in this case) and pseudo-labeling methods. In [6], data is augmented by using segments of labeled samples, but assigning the class of the utterance to the segment can create class noise. To address this problem, they use an iterative self-training method to label segments.

Another source of class noise in SER frequently discussed in the literature is the subjectivity of label annotators [20]. For all these reasons, class noise in SER problems has become an area of study. Current work addressing class noise in SER has improved performance by around 2% using BLSTM models [20]. Until recently, class noise in SER has been neglected, and there is still much work to be done. To our knowledge, there is no research on controlled class noise in SER across a significant number of datasets.

Objectives: This work has two primary objectives. The first is to study the robustness of several machine learning algorithms under controlled class noise conditions in the SER problem. We limit our study to well-known traditional algorithms, namely support vector machines (SVMs), random forest (RF), gradient boosting classifier (GB), AdaBoost, K-NN, and Ridge classifier. The second objective is to present a preprocessing method that satisfies the condition in Eq. (3) for specific noise levels. Finally, we compare the performance of the previously studied algorithms and our method alongside the most robust algorithm identified.

4 Criteria for Effective Noise Treatment

We propose the following formulation for the class noise problem, along with key conditions for a preprocessing method that aims to mitigate noise without discarding samples. For a single label classification dataset:

$$D = \{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{x}_i, y_i\}_{i=0}^{N},$$

we say it's noise free if for every $\mathbf{x}_i \in \mathbf{X}$ its label y_i is equal to the ground-truth label for that sample, we can denote this dataset as:

$$D_0 = \{\mathbf{X}, \mathbf{Y}_0\} = \{\mathbf{x}_i, y_i^0\}_{i=0}^N.$$

Similarly, we can extent the notion of class noise in a dataset D_r by measuring it as the fraction r of mislabelled samples, r can take values from 0 to 1; for example, a dataset with ten percent of mislabelled samples could be express as $D_{0.3}$. By definition, it is true that for a given dataset D_r, Eq. (2) is true:

$$\text{Acc}(D_r) = \text{Acc}(\mathbf{Y}_r, \mathbf{Y}_0) = \frac{1}{N}\sum_{i=0}^{N} \delta\big(y_i^r == y_i^0\big) = 1 - r. \tag{2}$$

Note, that $\text{Acc}(D_r)$ presented in Eq. (2) is equivalent to accuracy, one well-known metric to measure performance in classification tasks. In this context, we are usign this measure (fraction of correct labels in a dataset) as a internal measure of class noise in a given dataset. For a real scenario, we can't calculate this measure since $\mathbf{Y}_0$ is unkonwn.

A preprocessing method $P(D_r)$ without removal of potential noisy samples, generates a new dataset $D_{\hat{r}} = (\mathbf{X}, \hat{\mathbf{Y}}_r)$ with the same samples, but some of them are re-labeled. We propose that P is a candidate to be a good preprocessing method for class noise treatment if it fulfills condition presented in Eq. (3), for some noise level r over a significant amount of datasets:

$$\text{Acc}\Big(P(D_r)\Big) \geq 1 - r, \tag{3}$$

nonetheless, this condition is only appropiate for balanced datasets, i.e., datasets where class distribution is highly skewed amog classes [30]. A dataset is said to be imbalanced if its imbalance ratio (IR), see Eq. (4), is smaller than 1.5 [31]:

$$IR = \frac{card\big(\text{majority class}\big)}{card\big(\text{minority class}\big)}. \tag{4}$$

For a more general purpose, another metric should be used, such as balanced accuracy, see Eq. (5), which is the average of sensitivity by classes, allowing an equal representation of all classes. For an imbalanced dataset, there can be a preprocessing method which fulfills the condition of Eq. (3) by misrepresenting minority classes in the process of re-labeling. For this reason, in this work we examine both metrics:

$$\text{BAcc}(D_r) = \frac{1}{|C|}\sum_{c \in C} \frac{1}{|\text{class c}|}\sum_{i=1}^{N} \delta\left(y_i^r == y_i^0 \wedge y_i^0 == c\right). \tag{5}$$

In Eq. (5) is the set of classes in D_r. We say, that $P(D_r)$ is a candiadte to be a good preprocessing method and not a actual one, since the final goal of noise treatment in classsification can be thought as improvement in model performance after applying P to the training set and testing in complete unseen data (test set), and not only improving mislabelled samples inside training set.

Table 1. Datasets used in this work. The reported cardinalities and imbalance ratios refer only to the samples utilized for this study, and may differ from those reported in the original papers.

Dataset	Cardinality	Imbalance Ratio	Language	Type
Ravdess [22]	672	2.0	English	Acted
TESS [23]	1600	1.0	English	Acted
MESD [4]	574	1.0	Spanish	Acted
ShEMO [24]	2737	5.3	Persian	Semi-Natural
CaFE [25]	504	2.0	Canadian-French	Acted
EMO-DB [26]	339	2.0	German	Acted
CREMA-D [27]	4900	1.2	English	Acted
EMOVO [28]	336	1.0	Italian	Acted
EmoMatchSpanishDB [5]	1318	1.8	Spanish	Elicited
URDU [29]	400	1.0	Urdu	Natural

5 Experimental Framework

5.1 Algorithms' Robustness

The data for this study comprised 10 supervised SER datasets, as shown in Tab. 1. We only considered the emotions common to all datasets: anger, happiness, neutral, and sadness. All audio files were downsampled to a frequency of 16 kHz and converted to mono-channel signals. After standardizing the audio signals, 13 Mel-frequency cepstral coefficients (MFCCs) were extracted from each, transforming the unstructured data into tabular form. Various levels of class noise were introduced: 0%, 5%, 10%, 20%, 30%, 40%, and 50%. Noise was injected randomly, ensuring each class contained approximately the same amount. This resulted in a total of 60 datasets, with 10 datasets per noise level. This approach allowed us to control the noise levels in datasets assumed to be correctly labeled, enabling an analysis of the effects of varying noise levels on different algorithms.

To ensure a fair comparison between algorithms, we tested various hyperparameter instances for each classification scenario (SER dataset and noise level). A grid search was used to find the optimal set of hyperparameters. Despite using a substantial number of SER datasets, we performed 3×5-fold cross-validation. Within each 5-fold cross-validation, different noise injections, folds, and random states for algorithms were employed. Balanced accuracy was chosen as the evaluation metric, given that half of the datasets are imbalanced (see Tab. 1). The grid search space was selected based on prior knowledge of key hyperparameters for regularization in each algorithm.

5.2 Preprocessing Method Proposed

After evaluating the robustness of the algorithms, we found that SVM often performed the best, particularly with a radial basis function (RBF) kernel and a regularization parameter $C = 10$. The second-best algorithms were RF and

Table 2. Overall results of the preprocessing method's impact are summarized as follows. Datasets Improved show the number of cases (out of 10) with improved correct labels. The overall change represents the average change across all datasets for each noise level.

Level of Noise	0	5	10	20	30	40	50	(%)
Datasets Improved	0	5	6	7	9	9	10	
Overall Change	-4.6	-2.2	-0.1	4.0	10.0	13.2	17.8	(%)

k-NN, which showed similar performance across all noise levels. However, k-NN can be implemented using KD-trees, making it significantly faster than RF. Therefore, we decided to use both algorithms in an ensemble approach to detect noisy samples and re-label them. The process is as follows:

Given a noisy dataset D_r with an unknown noise level r and unknown ground-truth labels $\mathbf{Y}0$ (which are not available in real scenarios), two models, $hsvm$ and h_{knn}, are trained on $(\mathbf{X}, \mathbf{Y}_r)$. These models are then used to estimate the emotion for each sample across D_r, re-labeling them as indicated in Eq. (6):

$$P(\mathbf{x}_i) = \begin{cases} h_{svm}(x_i) & h_{svm}(x_i) = h_{knn}(x_i) \\ y_i^r & \text{otherwise} \end{cases}. \tag{6}$$

To test whether the proposed method is a valid class noise corrector, we applied it to the 10 datasets under all noise conditions. We assessed whether the similarity, based on 2 and 5, between $\hat{\mathbf{Y}}_r$ and $\mathbf{Y}_0$ improved compared to $\mathbf{Y}_r$ and $\mathbf{Y}_0$.

To evaluate whether the proposed method results in actual improvements in model performance, rather than merely correcting a fraction of mislabeled data, we tested the method within the same cross-validation framework used to assess algorithm robustness. Specifically, for each dataset D_r, 5-fold cross-validation was employed. In each iteration, the preprocessing method P was applied to the four training folds. An SVC was then trained on the re-labeled training folds and used to estimate emotions in the testing fold. Performance was measured using the unknown $\mathbf{Y}_0$. This process was repeated three times for each dataset and noise level, each time with different random injections and folds.

6 Results and Discussion

6.1 Algorithms Robustness

Results for the best hyperparameter configurations tested for each algorithm are shown in Fig. (5), along with the results of the preprocessing method. We found that the robustness of some algorithms depends significantly on the chosen hyperparameters. Fig. (2) displays the performance for two instances of k-NN and SVMs.

As shown in Fig. (2), k-NN's performance does not vary significantly with the number of neighbors, k. Although a greater number of neighbors is required as

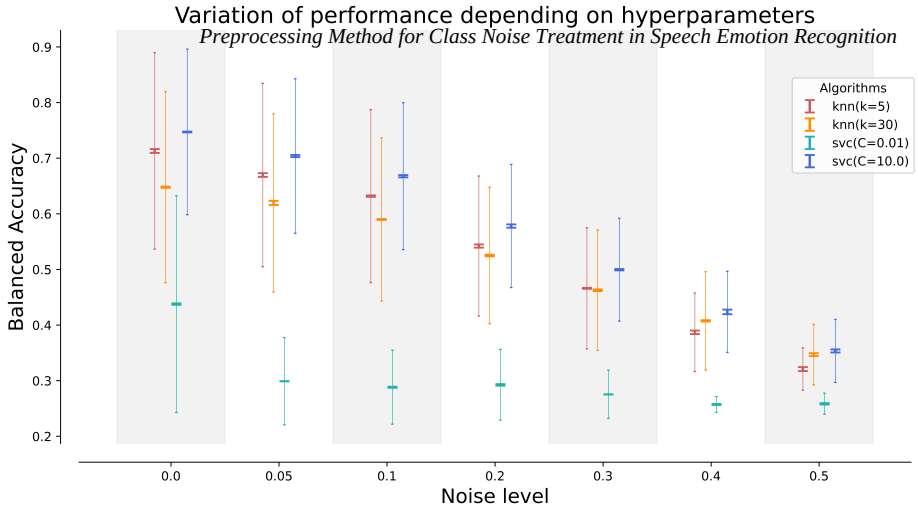

Fig. 2. Performances for each algorithm and different evaluations at one hyperparameter are displayed with double error bars. Thicker, smaller error bars represent the standard deviation of the mean performance across all datasets for each 5-fold cross-validation. Larger error bars represent the standard deviation across all datasets.

noise increases, the performance changes are less pronounced compared to SVMs with respect to the regularization parameter C. For each model, we selected the best hyperparameter configuration across all datasets and noise levels. Generally, SVM was the best algorithm for all noise levels, particularly with balanced class weights, an RBF kernel, and $C = 10$. For some datasets, k-NN and random forest performed best at low noise levels. Our results are publicly available on GitHub for more detailed analysis.

6.2 Impact of Preprocessing Method at Different Levels of Noise

To measure the effects of the preprocessing method under noisy conditions, we estimate the changes in the percentage of correct labels, as defined in Eq. (2). Fig. (3) shows the changes in extreme cases: no noise and 50% noise injection.

It can be seen that the preprocessing method improves the labeling of samples with 50% noise across all datasets but mislabels samples when applied in noise-free conditions. A summary of the results is provided in Tab. (2).

From Tab. (2), it is evident that the proposed method shows improvements for conditions with at least 20% noise injection. The enhancements in correct labels are more significant under high noise conditions compared to the decreases observed in low noise conditions.

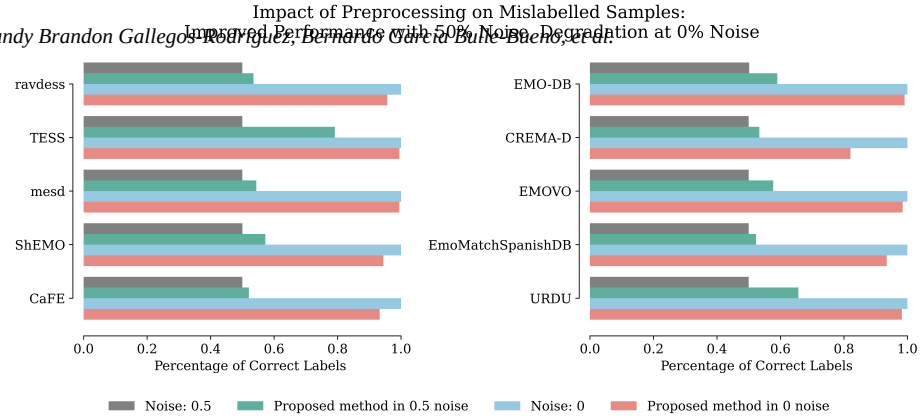

Fig. 3. The percentage of correct labels shown is the average from repeating the process three times for each dataset and noise level.

To assess whether these improvements compromise minority classes, we compared the percentage changes between $\text{BAcc}(\mathbf{Y}_r, \mathbf{Y}_0)$ and $\text{BAcc}(\hat{\mathbf{Y}}_r, \mathbf{Y}_0)$. The results, shown in Fig. (4), are consistent with those presented in Tab. (2).

According to Fig. (4), the proposed method may not be suitable for low noise levels (0-10%) due to a decrease in similarity to the ground-truth labels. For moderate noise levels (20-30%), the method appears to be effective, with most datasets showing improvement. Notably, the ShEMO dataset, which is the most imbalanced with a ratio of 5.3, also shows improvement at these noise levels, suggesting that the method may be resilient to imbalanced data. Finally, for higher noise levels (40-50%), the proposed method delivers better and more consistent results across all datasets.

6.3 Improvement in Model's Performance on Cross Validation Settings

After confirming that the proposed method is a viable preprocessing option for certain noise levels according to our definition, we need to ensure that these changes result in actual improvements in model performance. We followed the procedure described in the experimental framework. Results are shown in Fig. (5), along with the most robust configurations for each algorithm studied. Fig. (5) indicates that the proposed method performs worse under noise-free conditions, falling below even SVMs and most other algorithms. At 5% noise, the method remains nearly resilient and shows performance close to that of SVMs. At 10% noise, the proposed method achieves the best results compared to all other algorithms. For higher noise levels, while the performance of the preprocessing technique begins to degrade, it remains significantly better than that of all other algorithms.

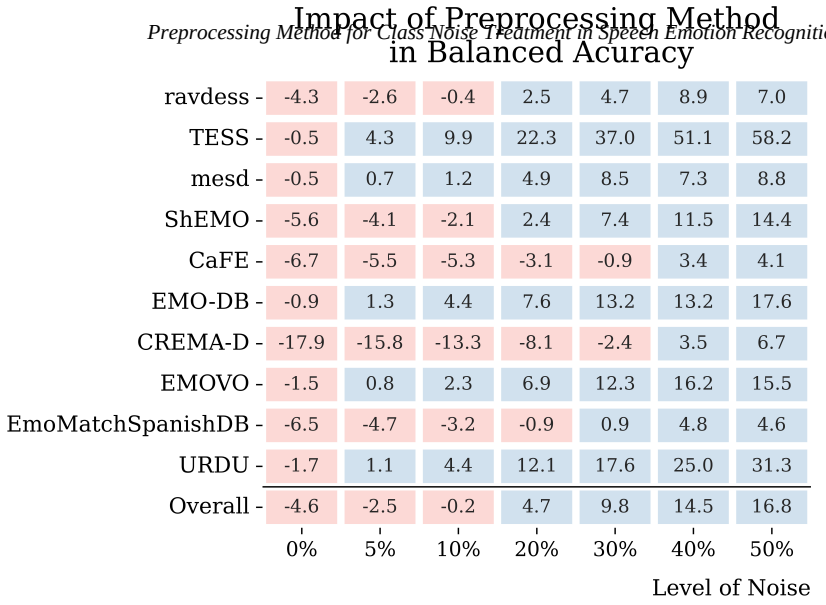

Fig. 4. Percentage changes in balanced accuracy serve as an internal measure for dataset D_r, estimating the similarity between $\mathbf{Y}_r$ and $\mathbf{Y}_0$ before applying the preprocessing method and $\hat{\mathbf{Y}}_r$ and $\mathbf{Y}_0$ after applying the method.

7 Conclusions and Future Work

In this work, we conducted a systematic study on the effects of noise in the Speech Emotion Recognition (SER) problem. Our primary contributions are twofold. First, we established that among the algorithms studied, Support Vector Machines (SVM) consistently deliver superior performance across both low and high noise levels. Second, we proposed a preprocessing technique that effectively outperforms traditional machine learning algorithms in high noise conditions.

The proposed method has shown promise in enhancing data quality for SER datasets. It significantly improves performance in noisy environments, outperforming all other algorithms studied, showing its potential as a valuable tool for preprocessing in real-world applications.

Future work will focus on enhancing the generalizability of this method for cross-corpora problems, where models are trained and tested on datasets with different distributions, such as actors or speakers from various regions or languages. We aim to investigate how well this preprocessing technique can support a broader range of algorithms, including neural networks, which were not covered in this study. By extending the scope of our research, we hope to improve the robustness and adaptability of SER systems across diverse and challenging conditions.

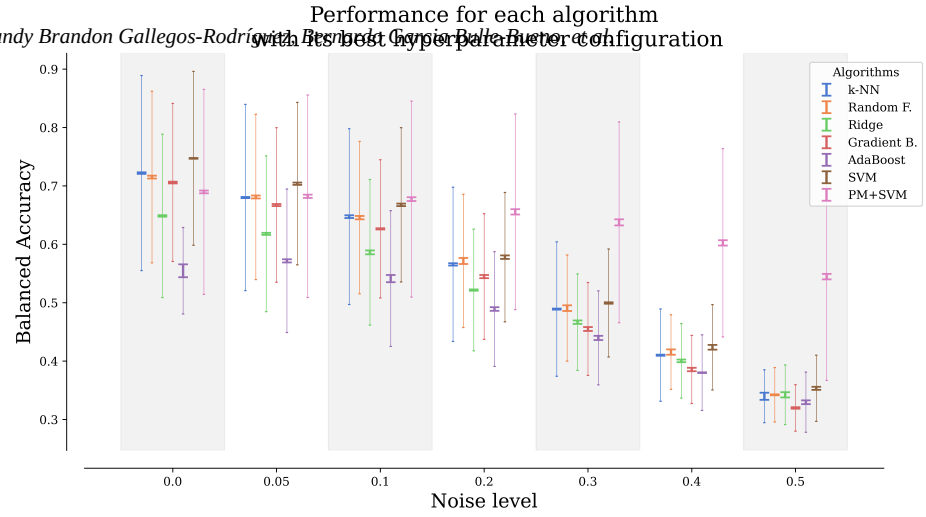

Fig. 5. Performance results for each algorithm are displayed with double error bars. Thicker, smaller error bars represent the standard deviation of the mean performance over all datasets for each 5-fold cross-validation. Larger error bars represent the standard deviation across all datasets. The proposed method is labeled as 'PM'.

References

1. Afzal, S., Khan, H. A., Piran, M. J., Lee, J. W.: A comprehensive survey on affective computing; challenges, trends, applications, and future directions. IEEE Access (2024)
2. Singh, Y. B., Goel, S.: A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, 245–263 (2022)
3. Hashem, A., Arif, M., Alghamdi, M.: Speech emotion recognition approaches: A systematic review. Speech Communication, 102974 (2023)
4. Duville, M. M., Alonso-Valerdi, L. M., Ibarra-Zarate, D. I.: The Mexican Emotional Speech Database (MESD): elaboration and assessment based on machine learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), pp. 1644–1647, IEEE (2021)
5. Garcia-Cuesta, E., Salvador, A. B., Pãez, D. G.: EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database. Multimedia Tools and Applications, 83(5), 13093–13112 (2024)
6. Mao, S., Ching, P. C., Lee, T.: Enhancing segment-based speech emotion recognition by iterative self-learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 123–134 (2021)
7. Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., Ambikairajah, E.: A comprehensive review of speech emotion recognition systems. IEEE access, 9, 47795–47814 (2021)
8. Hasan, R., Chu, C.: Noise in Datasets: What Are the Impacts on Classification Performance?[Noise in Datasets: What Are the Impacts on Classification Performance? In: Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (2021)

9. Mohri, M., et al.: Foundations of Machine Learning. Cambridge, Massachusetts, The Mit Press (2018)
10. Valiant, L. G.: A theory of the learnable. Communications of the ACM, 27(11), 1134–1142 (1984)
11. Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction, Vol. 2, pp. 1–758, New York: Springer (2009)
12. Zhu, X., Wu, X.: Class noise vs. attribute noise: A quantitative study. Artificial intelligence review, 22, 177–210 (2009)
13. Huang, C., Chen, G., Yu, H., Bao, Y, Zhao, L.: Speech emotion recognition under white noise. Archives of Acoustics, 38(4), pp. 457–463 (2017)
14. Heracleous, P., Yasuda, K., Sugaya, F., Yoneyama, A., Hashimoto, M.: Speech emotion recognition in noisy and reverberant environments. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 262-266, IEEE (2017)
15. Fahad, M. S., Ranjan, A., Yadav, J., Deepak, A.: A survey of speech emotion recognition in natural environment. Digital signal processing, 110, 102951 (2021)
16. Sáez, J. A., Corchado, E.: ANCES: A novel method to repair attribute noise in classification problems. Pattern Recognition, 121, 108198 (2022)
17. Ekambaram, R., Fefilatyev, S., Shreve, M., Kramer, K., Hall, L. O., Goldgof, D. B., Kasturi, R.: Active cleaning of label noise. Pattern Recognition, 51, 463–480 (2016)
18. Chen, Q., Jiang, G., Cao, F., Men, C., Wang, W.: A general elevating framework for label noise filters. Pattern Recognition, 147, 110072 (2024)
19. Zhang, S., Chen, M., Chen, J., Li, Y. F., Wu, Y., Li, M., Zhu, C.: Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. Knowledge-Based Systems, 229, 107340 (2021)
20. Fujioka, T., Homma, T., Nagamatsu, K.: Meta-learning for speech emotion recognition considering ambiguity of emotion labels. In: INTERSPEECH (2020)
21. Gupta, D., Bansal, P., Choudhary, K.: The state of the art of feature extraction techniques in speech recognition. Speech and Language Processing for Human-Machine Communications: Proceedings of CSI 2015, 195-207 (2018)
22. Livingstone, S.R., Russo, F.A.: The Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):e0196391 (2018)
23. Dupuis, K., Pichora-Fuller, M.: Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set. Canadian Acoustics, 39(3):182-3. Available from: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2471 (2011)
24. Mohamad Nezami, O., Jamshid Lou, P., Karami, M.: ShEMO: A large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53, 1–16 (2019)
25. Gournay, P., Lahaie, O., Lefebvre, R.: A Canadian French emotional speech dataset. In: Proceedings of the 9th ACM multimedia systems conference, pp. 399–402 (2018)
26. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B.: A database of German emotional speech. In: Interspeech, Vol. 5, pp. 1517–1520 (2005)
27. Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377–390 (2014)

28. Costantini, G., Iaderola, I., Paoloni, A., Todisco, M.: EMOVO corpus: An Italian emotional speech database. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), pp. 3501–3504, European Language Resources Association (ELRA) (2014)
29. Latif, S., Qayyum, A., Usman, M., Qadir, J.: Cross lingual speech emotion recognition: Urdu vs. Western languages. In: 2018 International conference on frontiers of information technology (FIT), pp. 88–93, IEEE (2018)
30. Alcal-Fdez, J., Fernndez, A., Luengo, J., Derrac, J., Garca, S., Snchez, L., Herrera, F.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing, 17(2-3), 255–287 (2011)
31. Villuendas-Rey, Y., Yáñez-Márquez, C., Camacho-Nieto, O.: Ant-based feature and instance selection for multiclass imbalanced data. IEEE Access (2024)