www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(3), May-Aug 2025, 499-511. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i3.854

Comparative Evaluation of the Performance of Vocal Signals and EGG in the Classification of Vocal Pathologies

Virna V. Vela-Rincón[1], Dante Mújica-Vargas[1], Andrés Antonio Arenas Muñiz[1], Antonio Luna-Álvarez[1]
[1] Departamento de Ciencias Computacionales, Tecnológico Nacional de México/CENIDET, Interior Internado Palmira S/N, Cuernavaca, 62490, Morelos, México.
viryvela@cenidet.edu.mx, dante.mv@cenidet.tecnm.mx, d22ce043@cenidet.tecnm.mx, jesus.luna18ce@cenidet.edu.mx

Abstract. This study proposes a methodology for the classification of vocal pathologies by comparing voice signals with electroglottographic (EGG) signals. The segmentation of the voice signal into temporal components and its transformation into recurrence plots through intuitionistic fuzzy clustering provides input for a deep learning model to classify voices as healthy or pathological. The results obtained show that the Inception-v3 model, when using intuitionistic clustering, achieves superior accuracy — particularly with EGG signals — reaching a peak performance of 87.8%. Furthermore, the F1 score is 0.885 for EGG and 0.860 for speech, demonstrating better performance on EGG signals.

Keywords: Vocal pathologies, EGG signal, Recurrence Plot, Intuitionistic fuzzy clustering

Article Info
Received February 03, 2025
Accepted April 09, 2025

1 Introduction

Voice pathology refers to disorders affecting the quality, pitch, loudness, or function of the voice, often caused by structural abnormalities, neurological conditions, or psychological factors. The complexity of vocal signals, which depend on variables such as frequency, intensity and temporal characteristics, complicates the precise classification of pathologies (Abdulmajeed et al, 2022). Furthermore, traditional diagnostic methods that rely on auditory perception are susceptible to bias, owing to variability in the listener's experience and environmental conditions. In light of these challenges, there has been an increasing exploration of artificial intelligence (AI) techniques as a more objective and systematic approach to analyzing and classifying speech signals (Park et al., 2023; Ksibi et al., 2023). This has led to the development of novel methodologies and more specialized databases for training these models (Park et al., 2023; Ksibi et al., 2023; Harar et al., 2025).

A review of studies on vocal pathology detection reveals a diversity of approaches leveraging deep learning techniques, with each addressing unique challenges in the field. While several works (Islam et. al., 2022; Liu et al., 2023) focus on convolutional neural networks (CNNs) for feature extraction and classification, others integrate CNNs with recurrent neural networks (RNNs) to capture temporal dynamics (Ksibi et al., 2023), achieving higher accuracy. Additionally, feature extraction techniques have evolved to incorporate dynamic and static features from voice samples, allowing for more robust classification frameworks (Abdulmajeed et al, 2020; Omeroglu et al., 2022; Kumar et al., 2023). Support Vector Machines (SVM) and other machine learning algorithms continue to be widely utilized for their effectiveness in detecting specific pathologies like vocal nodules and polyps,

often achieving high accuracy rates (Mohammed et al, 2020). Other studies address class imbalance issues using fuzzy cluster oversampling or SMOTE, improving model robustness on datasets with underrepresented pathologies (Fan et al. 2021; Lee et al., 2023). Transfer learning has been shown to be effective in handling limited datasets by leveraging pre-trained models (Mittal & Sharma, 2023; Won & Kim, 2024), while novel features such as "pitch difference" and automated processes for specific disorders such as vocal cord polyps further refine detection methods (Changwei et al., 2020). Taken together, these studies demonstrate the potential of deep learning to advance voice pathology detection but also underscore the need for standardized datasets and validation protocols to ensure clinical applicability.

In addition, there has been a growing trend towards the integration of multimodal data in vocal pathology research, combining acoustic analysis with electroglottography (EGG) signals to better understand vocal fold function during phonation and improve diagnostic accuracy (Abdulmajeed et al, 2022). This approach allows researchers to correlate physiological and acoustic features, providing a deeper insight into the mechanics of voice production. There has also been an emphasis on standardised databases for training machine learning models to address variability in data quality and improve classification robustness. Notable datasets in the field of pathological voice classification include the Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorder Database includes over 1,400 samples of sustained vowels and passages but has limitations related to recording environments (Fang, 2019). The Saarbruecken Voice Database (SVD) offers high-quality recordings in German, making it suitable for various research applications (Woldert-Jokisz, 2007). The Arabic Voice Pathology Database (AVPD) features recordings from Arabic speakers collected under standardized conditions, addressing linguistic diversity (Mesallam, 2017). Additionally, the VOICED Database on PhysioNet contains clinically verified samples from 208 individuals, along with demographic and medical data (Goldberger et al., 2000). Together, these resources facilitate the study of various vocal pathologies, including cysts, vocal fold paralysis, and polyps, while enhancing the overall reliability of machine learning models in this field; but, are mostly accessible through specific permissions, licenses, or collaborations, and not all are entirely free.

In this context, the present research proposes a methodology for the classification of vocal pathologies using both the voice signal and electroglottography (EGG) signals, with the additional objective of comparing the performance of the methodologies. Each signal is segmented into windows of a specific size in order to preserve important patterns. The generation of recurrence graphs for each window captures recurring patterns of vocal fold vibration. An intuitionistic fuzzy clustering approach is then employed to classify these segments into two homogeneous groups based on similarities. The recurrence graphs thus obtained are then processed by a deep learning model to identify relevant patterns and features in the visual data.

The rest of this paper is organized as follows. Section 2 presents some theoretical concepts, Section 3 describes in detail the composition of the methodology, with special emphasis on the generation of recurrence plots and their mathematical basis. Section 4 presents the experimental results. The final section offers conclusions deriver from results and recommendations for future work.

2 Preliminaries

This section will present the fundamental concepts necessary for the development of this work. The section will commence with a discussion of recurrence plots, which are utilized for the analysis of

temporal patterns. The different types of set theory, including classical, fuzzy and intuitionistic fuzzy, will also be addressed, as these allow us to deal with uncertainty in different ways. Finally, a brief description of the deep learning models that will be used throughout the study will be introduced, in order to contextualize their use and application.

2.1 Recurrence plot

A recurrence plot (RP) is a technique for the analysis of nonlinear data that can be considered as a visualization of a square matrix, where each element represents the time at which a state of a dynamic system is repeated. When applied to speech signals, the recurrence plot illustrates the instants at which the signal exhibits similar patterns or periodicities in time, thereby enabling the visualization of the signal's structure and recurrent characteristics in the time domain (Marwan, 2007). The recurrence plot can be expressed mathematically as follows:

$$R_{ij} = \Theta\left(\mathcal{E}_i - \| \vec{x_i} - \vec{x_j} \|\right), \qquad \vec{x_i} \in \Re^m, \qquad \mathrm{i,j} = 1, \dots, \mathrm{n} \tag{1}$$

where n is the number of states or moments considered x_i, $\mathcal{E}_i$ is a threshold distance, $\|\cdot\|$ a norm (e.g. Euclidian norm) and $\Theta(\cdot)$ is a Heaviside function.

2.2 Set theory: *Classical, Fuzzy and Intuitionistic Sets*

Set theory, a foundational branch of mathematics, is the study of the properties and relationships between collections of objects. In its classical form, a set is defined as a well-defined collection, where the membership of an element is determined in a precise and binary way: an element either belongs to the set or it does not. However, the classical theory is not always suitable for modeling situations where the membership of an element is uncertain or fuzzy. To address this limitation, fuzzy sets and intuitionistic fuzzy sets have emerged as extensions of classical set theory. These extensions allow for the handling of imprecision and uncertainty in a more flexible manner, thereby expanding the scope of the theory. In classical set theory (Kunen, 2014), a set is defined by a precise and absolute membership rule. That is to say, the membership of an element x to a set A is expressed by a binary proposition: $x \in A$ o $x \notin A$. In the context of classical data clustering, the best-known algorithm is K-means (Ahmed, 2020), which attempts to divide a data set into K groups (clusters) by minimising the squared distance between the points and the cluster centres. Its objective function is:

$$\mathrm{J}(X; U, V) = \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_i - v_j\|^2 \tag{2}$$

where n is the number samples $\mathrm{X} = x_1, x_2, \dots, x_n$, k the number of clusters, x_i is the *ith* point. v_j is the centroid of the *jth* cluster and $\| x_i - v_j \|^2$ is the squared Euclidean distance between data point x_i and cluster centroid v_j. The objective of the K-Means algorithm is to find the centroids v_j that minimize this function. These centroids are updated with the average of the points assigned to each cluster. It has been demonstrated that the efficacy of this algorithm is contingent upon two factors:

first, the elements must be clearly distinguishable and secondly, the boundaries between the sets must be exact.

As an extension of classical set theory, fuzzy set theory aims to model the uncertainty and imprecision inherent in many real-world phenomena. In a fuzzy set, the membership of an element x to a set A is not a binary proposition, but a numerical value $\mu_A(x)$ in the interval [0, 1], representing the membership degree of x to the set A. This function, called a membership function, makes it possible to describe vague or fuzzy concepts such as 'high temperature', 'light weight' or 'young person', where there is no exact boundary between the elements that belong to the set and those that do not. Fuzzy C-Means (FCM) is a fuzzy clustering algorithm that assigns each point a degree of membership to each cluster, as opposed to a binary assignment as in K-Means (Kahraman, 2016). The objective function of FCM is as follows:

$$J(X; U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m} d^2(x_i, v_j) \tag{3}$$

where c is the number of clusters, U is a membership matrix that contains the memberships degrees μ_{ij} and m is a fuzziness parameter, $d^2(x_i, v_j)$is the distance between the sample x_I and the cluster centroid v_j. The objective is to minimize this function by iteratively updating the centroids and membership values until convergence. The centroids v_j are recalculated as a weighted combination of all data points, where the weights are the fuzzy memberships μ_{ij}. This process continues until the centroids and membership values stabilize. The centroids, v_j, are recalculated as a weighted combination of all data points:

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \tag{4}$$

Meanwhile, the fuzzy memberships μ_{ij} are updated by a fuzzy membership function, which depends on the distance between each data point x_I and each centroid v_j.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - v_j \right\|^2}{\|x_i - v_k\|^2} \right)^{\frac{2}{m-1}}} \tag{5}$$

Intuitionistic fuzzy sets (IFS) (Xu & Wu, 2010) represent an extension of traditional fuzzy sets, incorporating not only the membership of an element to a set, but also the degree of non-membership, along with a third component termed "hesitant." An intuitionistic fuzzy set is formally characterised by a membership function, $\mu(x)$, a non-membership function, $\nu(x)$, and an indeterminacy function, $\pi(x)$, satisfying the relation $\mu(x) + \nu(x) \leq 1$. This approach facilitates enhanced flexibility in the representation of elements that are not exclusively partial members of a set, but also exhibit indeterminate characteristics with respect to their relationship with the set. This enables modelling scenarios where there is not only an uncertain degree of membership, but also an indeterminacy about whether or not an element should be incorporated into the set.

2.3 Deep Learning Models

Deep learning, as a subset of machine learning, utilizes artificial neural networks to model intricate patterns in data, emulating the cognitive functions of the human brain.This approach is characterized by its multi-layered architecture, which allows for the automatic learning of features from large datasets without the need for explicit programming for each specific task. The efficacy of deep learning has been particularly notable in various domains, including image recognition, natural language processing, and speech recognition. In these domains, deep learning has outperformed traditional methods by effectively capturing complex data representations (Lim & Zohren, 2021; Purwono et al., 2023).

Among the various architectures within the field of deep learning, SqueezeNet distinguishes itself by its capacity to attain accuracy comparable to that of more extensive models while concurrently diminishing the number of parameters. This efficiency renders it particularly well-suited for deployment in environments where resources are limited (Iandola, 2016). A similar approach is seen in GoogLeNet, which introduces inception modules to convolutional neural networks (CNNs). These modules utilize varying filter sizes within the same layer, offering a novel perspective on CNN design. This design enables the model to capture different aspects of the input data effectively, enhancing its performance in image classification tasks (Szegedy, 2015). InceptionV3, an evolution of GoogLeNet, further optimises performance through advanced techniques such as factorised convolutions and aggressive regularisation. These enhancements allow InceptionV3 to recognise complex patterns in images more efficiently than its predecessors (Szegedy, 2016). On the other hand, ResNet50 uses a unique strategy known as residual learning, which incorporates skip connections to facilitate the training of very deep networks - sometimes comprising hundreds of layers - without running into the vanishing gradient problem. This architecture has been widely adopted for image classification tasks due to its robustness and efficiency (He et al., 2016). These architectures exemplify the capabilities of deep learning models in addressing a range of challenges across diverse domains.

3 Proposed methodology

This research presents a methodology for classifying vocal pathologies, as shown in the general diagram in Figure 1. The methodology involves analyzing electroglottography (EGG) signals, which are normalize to a [0,1] range to ensure uniform magnitude. Each signal is then segmented into 350-sample windows, capturing local patterns and fine details. This segmentation further aids in data augmentation by generating multiple segments from the same signal, enriching the training dataset. Furthermore, a recurrence plot is generated for each signal window, thereby capturing recurrent patterns in vocal fold vibrations during phonation and facilitating the identification of salient features. These graphs are derived from an intuitionistic fuzzy clustering approach, which facilitates the categorization of signal segments into two homogeneous sets based on their similarities. After generating the recurrence plots, they are processed by deep learning.

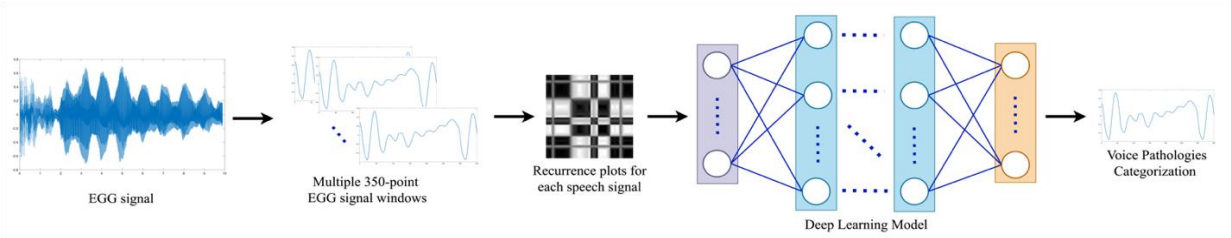


Fig. 1. Proposed methodology for speech processing.

The clustering process employed to generate the recurrence graphs is founded on the theory of intuitionistic fuzzy sets. Given a set of data $X = \{x_1, x_2, \ldots, x_n\}$ to be clustered into c groups. The formulation of a clustering algorithm necessitates the establishment of an objective function that seeks to minimize the distances between the centers of the groups and the data. This is due to the fact that the proximity of a data point to a center corresponds to an elevated degree of membership. The objective function that has been formulated is as follows:

$$J_m\left(X^{IFS}; U, V^{IFS}\right) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m}\, d^2(x_{ij}, v_{kj}) \tag{6}$$

where $U = (u_{ik})_{c\times n}$ is the intuitionistic fuzzy cluster partition of X^{IFS}, each u_{ik} is defined as the membership degree of the *ith* element (x_i) with respect to the *jth* cluster. $X^{IFS} = \{x_1, x_2, \ldots, x_n\}$ are N intuitionistic fuzzy elements, with each x_i is represented as intuitionistic fuzzy set of the form $x_i = \{\mu(x_i), \upsilon(x_i), \pi(x_i)\}$, where $\mu(x_i)$, $\upsilon(x_i)$and $\pi(x_i)$ stand for membership, nonmembership and hesitant degrees, respectively. $V^{IFS} = \left(v_1, v_2, \ldots, v_k\right)$is the prototypes-vector, with each component given by membership, non-membership and hesitant indexes, such as $v_k = \{\mu(v_k), \upsilon(v_k), \pi(v_k)\}$. $d^2(x_i, v_k)$ is a distance measure (e.g. Euclidean intuitionitic fuzzy distance) between v_k (cluster center) of each region and x_i (data points). m is a fuzziness parameter (e.g. $m = 2$), c the number of cluster, n is the number of elements. To minimize J_m, it is necessary to choose a membership matrix ($U = (u_{ik})_{c\times n}$) and v_k based on the following equation:

$$u_{ij} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d^2(x_i, v_k)}{d^2(x_i, v_l)}\right)^{\frac{2}{m-1}}} \tag{7}$$

$$\mu(v_k) = \frac{\sum_{i=1}^{n} u_{ik}^{m}\,\mu(x_i)}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{8}$$

$$v(v_k) = \frac{\sum_{i=1}^{n} u_{ik}^{m} v(x_i)}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{9}$$

$$\pi(v_k) = \frac{\sum_{i=1}^{n} u_{ik}^{m} \pi(x_i)}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{10}$$

Subsequently, these components must be integrated to generate prototypes for the purpose of computing:

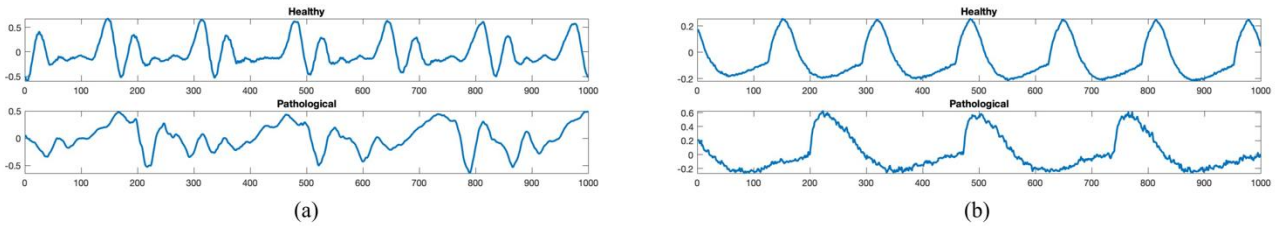
$$V_k^{IFS} = (\mu(v_k), v(v_k), \pi(v_k)) \tag{11}$$

4 Experimentation

This section presents the results of the experiments carried out. First, the dataset used for the evaluation are described. The experiment was designed in two phases: (1) to demonstrate, through the evaluation of several deep learning models, the contrast of the recurrence graphs obtained from classical, fuzzy and intuitive sets, (2) to compare the performance of the results of the classification of voice signals with the EGG signals to determine whether the integration of vocal cord vibration data improves the accuracy and reliability of the classification of voice signals.

4.1 Database

The Saarbrücken Voice Database (SVD) is a rich collection of voice recordings established at Saarland University in Germany, designed to support research in speech processing and linguistics. It includes recordings from over 2,000 speakers, representing a diverse array of dialects, accents, and speech pathologies. Each recording session is conducted under controlled conditions to ensure high audio quality and is accompanied by electroglottography (EGG) signals that provide insights into vocal fold vibration during speech. The Figure 2 presents an example of the signals. The database prioritizes speaker anonymity by assigning unique identification numbers, allowing researchers to access specific recordings without compromising privacy. Its user-friendly web interface enables detailed searches based on various criteria such as age, gender, and speech characteristics, while offering multiple download formats for flexibility in research applications (Woldert-Jokisz, 2007). The performance of the proposal is evaluated through the consideration of two categories: normal voice and pathological voice. A balanced subset of 660 samples is utilized for each category. The training of deep learning models is conducted with 70% of the data, while the remaining 30% is allocated for validation purposes.



(a) (b)

Fig. 2. Healthy and pathological samples in time domain. (a) Voice Signal (b) EGG Signal

4.2 Metrics

The evaluation of vocal pathology classification models will be conducted using the confusion matrix (Figure 3) and the several parameters derived from it. The following performance metrics are presented to evaluate the effectiveness of the models used, which are derived from the results of the confusion matrix (Sankaran & Kumar, 2024).

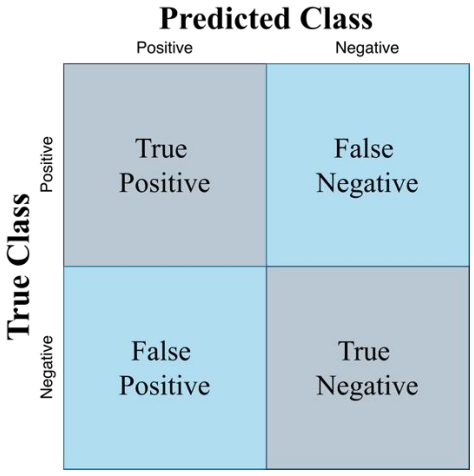


Fig. 3. Confusion matrix

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - Score = \frac{Precision \cdot Recall}{Precision + Recall} \tag{15}$$

$$Specificity = \frac{TN}{TN + FP} \tag{16}$$

$$Error = \frac{FP + FN}{TP + TN + FP + FN} \tag{17}$$

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

where TP are true positives, TN are true negatives, FP are false positives and FN are false negatives. *Accuracy* is measured as the proportion of correct predictions out of all predictions made. *Precision* measures the proportion of predicted positives that are truly positive, while *recall*, or sensitivity, reflects the ability to correctly identify positive instances. The *F1-score* is the harmonic mean of precision and recall, useful in imbalanced classes. Additionally, *specificity* measures the ability to correctly identify negative instances, complementing recall. The *error rate*, defined as the proportion of incorrect predictions, is another fundamental metric, though its utility is limited in imbalanced classes. The *false positive rate* (FPR) assesses the proportion of negative instances misclassified as positive, and a low FPR indicates a model's ability to avoid incorrectly classifying negatives. These metrics provide a comprehensive evaluation of model performance, particularly in cases with class imbalance or nuanced error analysis needs. In addition, The Kappa coefficient (Vieira et al., 2010) and Matthews correlation coefficient (MCC) (Chicco & Jurman, 2020) provide deeper insights into model performance. Kappa measures agreement, with values close to 1 indicating strong concordance. The MCC, ranging from -1 to 1, considers all confusion matrix elements and is especially useful in imbalanced datasets.

$$Kappa = \frac{T \times (TP + TN) - [(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)]}{T^2 - [(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)]} \tag{19}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN))}} \tag{20}$$

where T is the total number of observations, i.e., the number of elements in the confusion matrix. These metrics provide a more detailed understanding of how the model is performing in terms of its errors and its ability to distinguish between classes. The combined use of these metrics allows for a more accurate and complete evaluation of the model's performance on classification tasks.

4.3 Implementation Details

The hyperparameter tuning was performed through a manual procedure based on validation, aiming to improve the model's accuracy and generalization while aligning with the specific characteristics of the dataset. The learning rate was fine-tuned within the range of 0.0001 to 0.001 to ensure stable convergence. The dropout rate was set between 0.3 and 0.5 to mitigate overfitting, while L2 regularization was tested with coefficients ranging from 0.0001 to 0.001 to enhance generalization. The Adam optimizer was employed for its efficiency in updating parameters. Batch sizes of 32, 64, and 128 were assessed, and the number of epochs varied between 10 and 50. Cross-validation was used to evaluate the different hyperparameter combinations, enabling the selection of the most appropriate values for each model. The primary focus of this research was model accuracy; therefore, processing times were not a consideration. However, the broader goal is to optimize computational efficiency and refine the model for potential deployment in clinical or real-time environments.

4.4 Performance Results

The contrast between the recurrence plots generated by the classical, fuzzy and intuitionistic fuzzy clustering techniques is shown in Figure 4. It presents examples of the plots generated by each clustering technique for both a pathological and a healthy signal. In these plots, the differences between each type of clustering can be seen, revealing that fuzzy intuitionistic clustering manages to condense the information more completely than the other methods. This ability to condense information makes it possible to identify relevant patterns more clearly, which is useful for the classification of speech signals. This behaviour is observed in both speech and EEG signals, highlighting the usefulness of the approach in finding structure in the data and improving the analysis of signals in both pathological and healthy contexts.

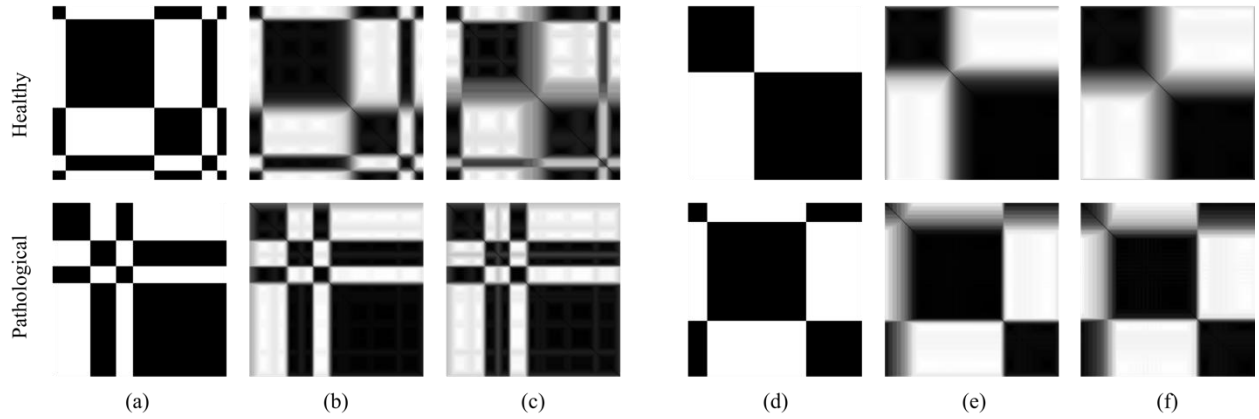


Fig. 4 Sample recurrence plot obtained. Voice Signal Methods: (a) Classical, (b) Fuzzy, (c) Intuitionistic. EGG Signal Methods: (d) Classical (e) Fuzzy (f) Intuitionistic.

In order to validate these results and to assess the performance of the classification models, a more detailed and quantitative analysis is performed. Table 1 shows the comparative accuracy results of different classification models applied to speech and EGG signals. The results reveal that intuitionistic methods consistently outperform classical and fuzzy approaches for both categories of signals, with Inception-v3 standing out by achieving a maximum accuracy of 0.878 on the EGG signal using the intuitionistic method. While ResNet-50 shows inferior performance when classifying speech signals, its performance improves significantly with EGG signals, reaching an accuracy of 0.720 with the classical approach. Furthermore, the EGG signal generally shows better results compared to the voice signal, suggesting that it contains more distinctive features that facilitate classification. This analysis highlights the effectiveness of intuitionistic techniques in improving the accuracy of classification models in specific contexts, and suggests that the choice of model and clustering technique is crucial for optimising performance in classification tasks.

Table 1. Accuracy results of different models using recurrence plots obtained by clusterings techniques.

Models	Voice signal			EGG signal		
	Classical	Fuzzy	Intuitionistic	Classical	Fuzzy	Intuitionistic
SqueezeNet	0.630	0.756	***0.781***	0.619	0.649	*0.706*
GoogLeNet	0.650	0.739	***0.787***	0.615	0.688	*0.766*
ResNet-50	0.594	0.783	*0.825*	0.720	0.814	***0.852***
Inception-v3	0.728	0.828	*0.845*	0.744	0.851	***0.878***

To complete the analysis, Table 2 presents various performance metrics of the Inception-v3 model using the intuitionistic clustering approach on the speech and EGG signals. The results show that the EGG signal is the most suitable choice for classification, with a higher F1 score (0.885) compared to the speech signal (0.860), reflecting a better balance between precision and recall. Moreover, the recall is significantly higher for the EGG signal (0.937 vs. 0.876), indicating a higher effectiveness in detecting positive cases. Although the specificity is slightly higher for the speech signal (0.857) than for the EGG signal (0.812), the overall performance in terms of precision and ability to detect positives makes the EGG signal preferable in applications where correct identification of vocal features is crucial. The low error values (0.155 for speech and 0.122 for EGG) and the high results for the Matthews Compensation Coefficient (MCC) and Kappa confirm the effectiveness of the model in classifying both signals, highlighting the performance of the Inception-v3 model with intuitionistic clustering in this context.

Table 2. Inception-v3 performance comparison with speech and EGG signals.

Metrics	Voice Signal	EGG Signal
F1_score	0.860	**0.885**
Specificity	**0.857**	0.812
Precision	**0.866**	0.834
Error	0.155	**0.122**
Recall	0.876	**0.937**
FPR	**0.143**	0.188
MCC	0.705	**0.816**
Kappa	0.690	**0.885**

Table 3 presents a comparison with some recent methods documented in the literature, selected within a similar setting to ensure a fair comparison. The results indicate that the proposed method demonstrates competitive performance, outperforming the majority of the compared approaches in terms of accuracy. Notably, it surpasses methods such as Islam et al. (2022), which, despite employing convolutional neural networks (CNNs) and raw speech signals, do not attain the same level of accuracy as the proposed method. In a similar vein, Won & Kim's (2024) approach attained commendable outcomes with EGG signals and transfer learning, underscoring its efficacy in terms of accuracy and reduced computational demands. Although Kumar et al.'s (2023) approach achieves superior accuracy, it employs a multimodal strategy that necessitates the additional processing of specific features, thereby augmenting its complexity. In contrast, the proposed method utilises raw signals and employs a transformation to recurrence graphs using intuitionistic fuzzy clustering, a process which contributes to reducing computational complexity.

Table 3. Comparison with some current methods in the literature.

Authors	Samples[a]	Input Data Type	Method	Accuracy	
				Voice	EGG
Islam et. al., 2022	H:150 P:65	Raw temporal data	2D CNN	0.803	0.721
Omeroglu et al., 2022	H:687 P:1354	Spectogram	Pretrained AlexNet + SVM	0.691	0.693
Won & Kim, 2024	H:869 P:520	MelSpectogram	Few-shot transfer learning (Pretrained Res-Net-18)	0.737	0.826
Kumar et al., 2023	H:303 P:303	EGG-Multimodal	ERB Spectrum + Gammatone, ensemble learner classifier	-	0.931
Proposal	H: 660 P: 660	Recurrence plot	Intuitionistic Fuzzy Recurrence Plot (Inception-v3)	*0.845*	*0.878*

[a] H:Healthy and P:Pathological

5 Conclusions

This research presents a methodology to classify vocal pathologies by comparing both voice and electroglottography signals using clustering and deep learning techniques. For each signal, recurrence plots were generated capturing the relevant patterns using three clustering approaches: classical, fuzzy and intuitionistic. The intuitionistic approach showed a superior ability to condense the information more completely, which facilitated the identification of discriminative patterns in the signals. The obtained recurrence plots were processed by deep learning models to classify the signals and to compare the performance of the voice and EGG signals. The results obtained show that the Inception-v3 model, using intuitionistic clustering, achieves excellent accuracy, especially on the EGG signals, with a maximum performance of 87.8%. Furthermore, the F1 score is 0.885 for EGG against 0.860 for voice, and the recall rates are 0.937 for EGG and 0.876 for voice. Although the specificity is slightly higher for voice signals (0.857) than for EGG (0.812), the error values are low, 0.122 for EGG and 0.155 for voice, supporting the effectiveness of the Inception-v3 model with intuitionistic clustering in the classification of voice pathologies. In addition, when evaluated in a fair manner against established literature-based methods, the proposed approach exhibits competitive performance, surpassing the majority of the compared approaches in terms of accuracy. In contrast to more intricate methods, the proposed approach maintains both high accuracy and efficiency by employing raw signals and reducing computational complexity through recurrence graphs with intuitionistic fuzzy clustering. Future work will explore multimodal approaches that combine both speech and EGG signals to improve classification accuracy and reliability.

References

Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2022). A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. *Journal of Intelligent Systems, 31*(1), 855–875.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295.

Changwei, Z., Lili, Z., Xiaojun, Z., Yuanbo, W., Di, W., & Zhi, T. (2020, October). Classification of normal and pathological voices using convolutional neural network. In *2020 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD)* (pp. 325–329). IEEE.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6.

Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C., & Wang, C. T. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice, 33*(5), 634–641.

Fan, Z., Wu, Y., Zhou, C., Zhang, X., & Tao, Z. (2021). Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. *Applied Sciences, 11*(8), 3450.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220. https://physionet.org/content/voiced/1.0.0/

Harar, P., Alonso-Hernández, J. B., Mekyska, J., Galaz, Z., & Burget, R. (2025). Voice pathology detection using machine learning algorithms based on different voice databases. *Results in Engineering*. https://doi.org/10.1016/j.rineng.2025.100123

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Iandola, F. N. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint*, arXiv:1602.07360.

Islam, R., Abdel-Raheem, E., & Tarique, M. (2022). Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. *Computer Methods and Programs in Biomedicine Update, 2*, 100074.

Kahraman, C., Öztayşi, B., & Çevik Onar, S. (2016). A comprehensive literature review of 50 years of fuzzy set theory. *International Journal of Computational Intelligence Systems, 9*(sup1), 3–24.

Ksibi, A., Hakami, N. A., Alturki, N., Asiri, M. M., Zakariah, M., & Ayadi, M. (2023). Voice pathology detection using a two-level classifier based on combined CNN–RNN architecture. *Sustainability, 15*(4), 3204. https://doi.org/10.3390/su15043204

Kumar, D., Satija, U., & Kumar, P. (2023, February). Analysis and classification of electroglottography signals for the detection of speech disorders. In *2023 National Conference on Communications (NCC)* (pp. 1–6). IEEE.

Kunen, K. (2014). *Set theory: An introduction to independence proofs*. Elsevier.

Lee, J. N., & Lee, J. Y. (2023). An efficient SMOTE-based deep learning model for voice pathology detection. *Applied Sciences, 13*(6), 3571.

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

Liu, G. S., Hodges, J. M., Yu, J., Sung, C. K., Erickson-DiRenzo, E., & Doyle, P. C. (2023). End-to-end deep learning classification of vocal pathology using stacked vowels. *Laryngoscope Investigative Otolaryngology, 8*(5), 1312–1318. https://doi.org/10.1002/lio2.1144

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*(5–6), 237–329.

Mesallam, T. A., Farahat, M., Malki, K. H., Alsulaiman, M., Ali, Z., Al-Nasheri, A., & Muhammad, G. (2017). Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. *Journal of Healthcare Engineering, 2017*(1), 8783751.

Mittal, V., & Sharma, R. K. (2021). Deep learning approach for voice pathology detection and classification. *International Journal of Healthcare Information Systems and Informatics, 16*(4), 1–30.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., ... & Al-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences, 10*(11), 3723. https://doi.org/10.3390/app10113723

Omeroglu, A. N., Mohammed, H. M., & Oral, E. A. (2022). Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Engineering Science and Technology, an International Journal, 36*, 101148.

Park, D., Kim, H. K., & Gwangju Institute of Science and Technology. (2023). Adversarial continual learning to transfer self-supervised speech representations for voice pathology detection. *IEEE Signal Processing Letters*. https://doi.org/10.1109/LSP.2023.3298532

Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & Haq, Q. M. U. (2023). Understanding of convolutional neural network (CNN): A review. *International Journal of Robotics and Control Systems, 2*(4), 739–748. https://doi.org/10.31763/ijrcs.v2i4.888

Sankaran, A., & Kumar, L. S. (2024). Advances in automated voice pathology detection: A comprehensive review of speech signal analysis techniques. *IEEE Access*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010, July). Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems* (pp. 1–8). IEEE.

Won, J. H., & Kim, D. H. (2024). Metric-based few-shot transfer learning approach for voice pathology detection. *IEEE Access*.

Woldert-Jokisz, B. (2007). Saarbruecken voice database. *Institute of Phonetics, University of Saarland*, Saarbrücken, Germany. Tech. Rep.