Computer Speech and Language 97 (2026) 101898

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Feature extraction from speech signals using empirical mode decomposition for depression detection: A comparative study with machine learning models

Xavier Sánchez Corrales [a],*, Jordi Solé-Casals [a,b]

[a] Data and Signal Processing Research Group, University of Vic–Central University of Catalonia, Vic, Spain
[b] Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

ARTICLE INFO

Keywords:
Depression
IMF (Intrinsic Mode Functions)
Voice
EMD (Empirical Mode Decomposition)
Gaussian kernel
Machine learning

ABSTRACT

Depression is a prevalent mental disorder that affects quality of life, and its early detection through voice analysis could improve diagnosis. This study investigates the effectiveness of Intrinsic Mode Functions in differentiating depression through voice signals. We used data from the Distress Analysis Interview Corpus. Empirical Mode Decomposition was applied to extract IMFs, and statistical characteristics and similarities were analysed using a Gaussian kernel between depression and healthy groups according to sex. The results revealed significant differences in the mean of the first IMFs in men, but not in women, while there were no differences in other statistics. Gaussian kernel analysis showed variations in the probability density function in the first Intrinsic Mode Functions (up to IMF 7), with differences according to sex. Six Machine Learning models were experimentally tested, trained, and adjusted. The best accuracy results in women were obtained with Gradient Boosting (94.1%), while in men they were obtained with Gradient Boosting (88.0%). Intrinsic Mode Functions proved to be useful for detecting depression, suggesting their potential in developing non-invasive tools for early detection of this.

1. Introduction

Depression is a mental pathology that affects millions of people worldwide and, traditionally, its diagnosis is based on subjective methods such as clinical interviews and questionnaires. Questionnaires with items about which the patient, paradoxically, is not qualified to offer an objective response. Currently, there is a growing demand for more objective and quantifiable tools that do not depend on the patient's subjectivity and provide the clinician with an objective assessment. In this sense, voice analysis presents itself as a promising avenue (Corrales et al., 2025; Koops et al., 2023; France et al., 2000; Low et al., 2009) to detect mental pathologies such as depression.

Studies based on more classical methods analyse acoustic and prosodic features as potential biomarkers of depression (Menne et al., 2024). Other studies examine the voice through statistical differentiation, on MFCCs (Mel Frequency Cepstral Coefficients) (France et al., 2000; Donaghy et al., 2024), on energy (Sharma et al., 2017) or on PSDs (Power Spectral Density) (France et al., 2000; Liu et al., 2020; Akkaralaertsest and Yingthawornsuk, 2019) of IMFs (Intrinsic Mode Functions). Many of the studies based on the decomposition of the voice signal into IMFs agree that only the first IMFs present discriminating characteristics in signal analysis. Usually, in previous studies, all IMFs are analysed, or the selection of IMFs is made based on a statistical analysis of

* Corresponding author.
E-mail addresses: xavier.sanchez.corrales@uvic.cat (X. Sánchez Corrales), jordi.sole@uvic.cat (J. Solé-Casals).

https://doi.org/10.1016/j.csl.2025.101898
Received 26 June 2025; Received in revised form 15 September 2025; Accepted 14 October 2025

Available online 5 November 2025

these. We also find studies that incorporate into Machine Learning (ML) models as features, the results of data analysis by Gaussian models (Gaussian Mixture Models, GMM) and MFCCs (Sharma et al., 2017; Cummins et al., 2015). Or, on the other hand, they pass the processing of raw IMF data through a Gaussian model to an ML model such as SVM or KNN (Alghowinem et al., 2013).

Although there are increasingly more studies on the competence of models in voice signal analysis (Mao et al., 2023), there is still no consensus on what type of model is more effective in this field, and above all, there are no concrete references to why some models work better than others considering premises about voice data (complexity based on non-linearity and non-stationarity of the data). In fact, many studies continue to base their results on stationary models such as the Fourier Transform (FT) analysis, which is ineffective in non-linear and non-stationary data. This is coupled with the difficulty the researcher faces in obtaining sufficient clean and reliable data (from correctly identified patients with minimal noise (Donaghy et al., 2024; Solé-Casals et al., 2013)). We consider it essential to continue working from the data towards the models, prioritising a deep understanding of the data.

The voice signal analysis we present in this study was conducted using voice recordings of 31 women with depression and 54 without it, and 25 men with depression and 76 without it, from the Distress Analysis Interview Corpus (DAIC) (DAIC, 2022; Gratch et al., 2014), from the University of Southern California.

In the present study, we perform an analysis of voice signals in people with and without depression, separated by sex. Our main objective is to thoroughly understand the specific characteristics of the signals, which will allow us to identify the most relevant patterns. Based on these characteristics, we seek to select and optimise the models iteratively, using a recursive approach.

2. Materials and methods

The data downloaded from the Distress Analysis Interview Corpus (DAIC) has followed the protocol established by said University, according to anonymisation standards, in order to protect the identity of the participants. This database includes 189 folders of voice, text, and video data collection sessions. The DAIC database comes from users' responses to clinicians' questions about aspects of their lives and their mental health; therefore, it consists of spontaneous speech. For this study, only voice data and those related to scores on the PHQ-8 (Patient Health Questionnaire depression scale) (Kroenke et al., 2009).

Files were selected based on sex, and from the two groups, they were divided again according to the cut-off score on the PHQ-8, where a score of 10 or higher is considered positive for depression. In total, the data selection gives us 4 groups: female-depression, female-healthy, male-depression, and male-healthy. This group selection is done to be able to separate the signal between sexes, as we know that there are different characteristics between them in the voice signal (Whiteside and Irving, 1997) and could affect the results if the voice of both sexes were mixed. The voice data within each group were combined by randomly concatenating all available voice segments for that group. Following concatenation, silences exceeding 0.5 s were removed from the four resulting files, after which the signals were segmented into 20-s samples. This task was carried out with the open-source program Audacity version 3.5.1 (Audacity, 2023).

2.1. Data preparation

Considering that the sampling rate of the original signal was 16 kHz and that processing at this frequency requires high computational consumption, after checking through spectrograms that the voice data in no case exceeded 10 kHz, the frequency was reduced to 10 kHz using the Librosa library version 0.10.1 of the Python programming language version 3.11.9 (version with which the rest of the study code has been worked). It internally implements an anti-aliasing filter that accounts for the Nyquist theorem. This ensures that the spectral content does not exceed the 5 kHz limit imposed by the new sampling rate. Librosa applies a resampling procedure with appropriate filtering that prevents aliasing and preserves the signal quality within the allowed frequency range. The original dataset comprises four groups of varying lengths: 31 depressed females, 51 healthy females, 25 depressed males, and 76 healthy males. These signals were concatenated, then silences longer than 0.5 s were removed from the concatenated vectors in each of the four groups, and afterwards the resulting four vectors were segmented into 20-s vectors. With a sampling rate of 10 kHz, each 20-s vector contains 200,000 data points. Table 1 shows the data structure for the four groups before and after EMD decomposition. In the first row, the first number in each group is the number of subjects, and the second number is the number of samples (data points) for each subject. The second row shows the structure after EMD decomposition: the first number is still the number of subjects, the second is the number of Intrinsic Mode Functions (IMFs) obtained from the decomposition, and the third is the number of samples per IMF.

Next, the extraction of Intrinsic Mode Functions (IMFs) was performed through Empirical Mode Decomposition (EMD), using the Python library pyemd version 1.0.0. The EMD procedure decomposes a complex signal through an iterative process that identifies local extrema (maxima and minima), constructs upper and lower envelopes by interpolation, computes their mean line, and subtracts it from the original signal to extract the first component. This process is repeated until the extracted component satisfies the criteria of an Intrinsic Mode Function (IMF), at which point the residual is used to obtain the next IMF. The result is a series of IMFs ordered from higher to lower frequency, together with a final residual, where each IMF represents a natural frequency band of the original signal. This enables an adaptive decomposition that preserves temporal characteristics and guarantees perfect reconstruction through the summation of all components (see Fig. 1).

This study implemented EMD in Python using the PyEMD library, applying it to normalised 20-s voice signals with a maximum of 16 IMFs (max_imf = 16). The algorithm's built-in stopping criterion, based on envelope comparison, together with the max_imf parameter, prevents oversifting and ensures that each IMF fulfils the conditions of local orthogonality and oscillation without overfitting noise.

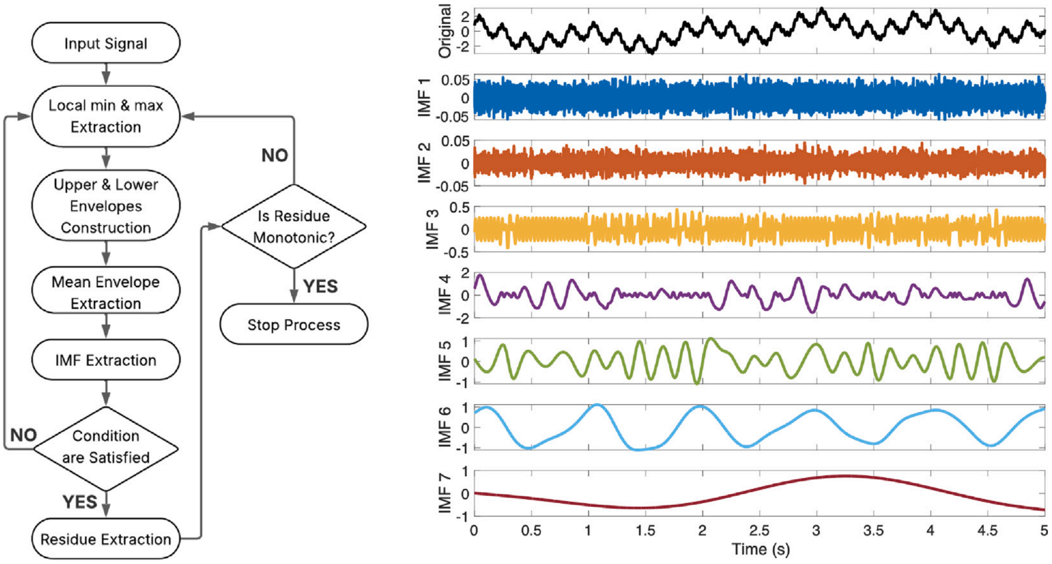

Fig. 1. Chartflow and graph of the Empirical Mode Decomposition of a voice signal. The signal is decomposed into Intrinsic Mode Functions (IMFs), ordered from highest to lowest frequency, together with a residual component (RES) that represents the remaining trend.

Table 1
First row shows the number of subjects and samples per group before EMD decomposition. Second row shows the number of subjects, IMFs, and samples per group after decomposition.

Female depression	Female healthy	Male depression	Male healthy
(42, 200 000)	(123, 200 000)	(61, 200 000)	(178, 200 000)
(42, 16, 200 000)	(123, 16, 200 000)	(61, 16, 200 000)	(178, 16, 200 000)

Between 10 and 20 decompositions were tested, with the effectiveness of the process being specified in 16 IMFs with a practically flat residue in the last one, without relevant information. For the input data in voice decomposition using EMD, both very long and excessively short voice samples cause issues during execution: the former due to high computational demands, and the latter due to information loss. After testing various voice segment lengths (5, 10, 15, 20, 25, and 30 s), 20-s samples were found to avoid these issues and also represent a clinically acceptable voice recording duration, roughly equivalent to 2 min of raw voice (after removing silences longer than 0.5 s).

Before presenting the data to the models that worked with balanced datasets, several balanced datasets were generated using random undersampling to prevent class bias. This involved identifying the indices of each class (healthy and depression), randomly selecting an equal number of samples from each, and combining them into a balanced dataset. The procedure is repeated each time the analysis is run, producing different balanced subsets on each execution and thus reducing bias from the original class distribution.

2.2. Statistic analysis

In Fig. 2, the dominant frequencies of all the IMFs in the four groups were calculated after the IMF extractions to check whether visual differences between classes according to sex could be observed. The dominant frequencies of the IMFs are shown according to sex and class (depression, health). The first IMFs (1–7) show that in both women and men, the depression groups present higher values and a more uniform decline compared to the healthy groups, which maintain higher frequencies and greater dispersion in the first IMFs.

In order to compare the Intrinsic Mode Functions among themselves, calculations of mean, median, standard deviation, kurtosis, and skewness were performed on all IMFs in the 4 groups. To characterise the local and global dynamics of each IMF, each vector of 200,000 samples was segmented into 20 non-overlapping windows of 10,000 samples each (window_size $= 10{,}000$). Various window sizes were tested, ranging from 1000, 5000, 10,000, 20,000, to 50,000, with no significant differences observed within this range. Within each window, five descriptive statistics were calculated: skewness, kurtosis, median, standard deviation, and mean. Subsequently, for each IMF, these values were averaged across all windows, resulting in a single representative value per statistic and IMF. This approach enables the capture of both local variability and global trends within the signal, providing a robust and compact summary of the information contained in each IMF.

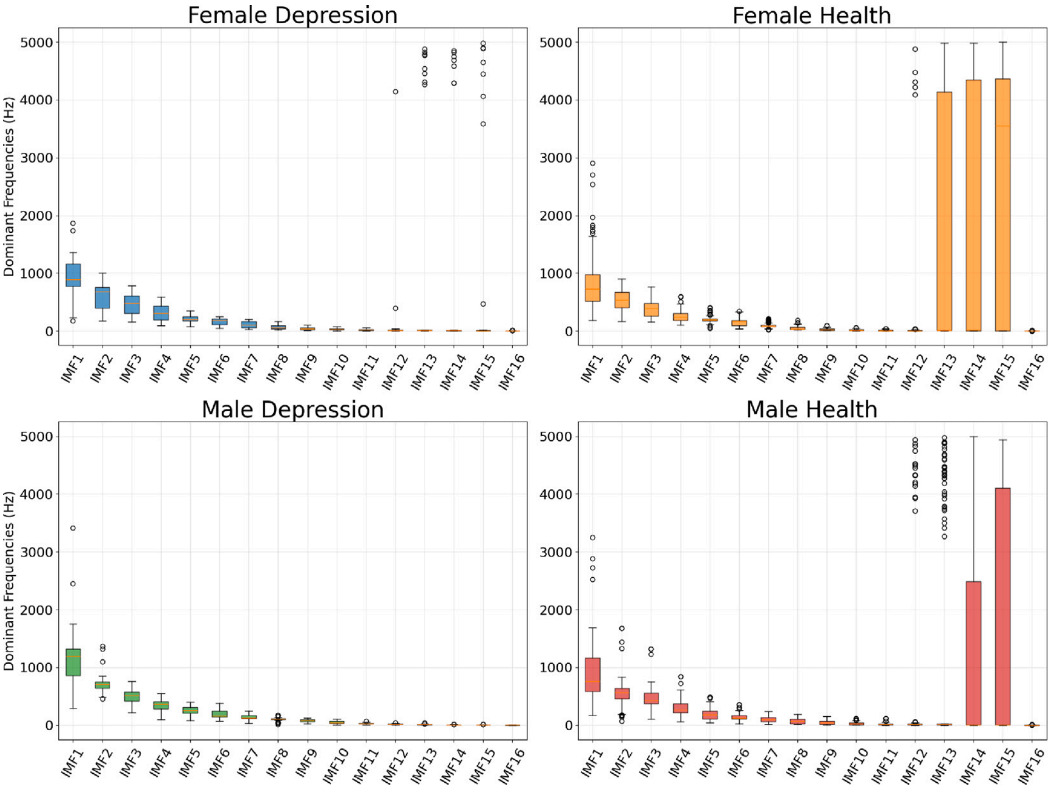

Fig. 2. Dominant frequencies in the IMFs according to sex and class (depression, health).

Subsequently, the results obtained in the indicated statistics were contrasted using the bootstrapping method, comparing female depression with female healthy, and male depression with male healthy. And then, the False Discovery Rate (FDR) correction was applied to the results of the bootstrapping method, in order to control the possibility of false positives among all significant results.

Taking into account that the comparison of statistics between classes by sex using Bootstrapping technique does not show differences, but the dominant frequencies of the IMFs do seem to show differences, a comparison of the similarity of the IMF distributions is performed by applying a Gaussian Radial Basis Function (RBF) kernel, which is a filtering technique that employs a normal (or Gaussian) distribution to find the similarity in the distributions. The Eq. (1) of the kernel that was applied to the IMFs is described below, where x and x' are feature vectors of two samples. Each point corresponds to a sample. For example, if you have two IMF1, a feature vector is extracted from each class (IMF1 depression and IMF1 healthy). The kernel measures their similarity: if the vectors are very close (small Euclidean distance), the kernel value is near 1; if they are very different, the value is near 0.

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{1}$$

The Gaussian bandwidth (σ) was optimised by implementing a recursive loop in which multiple bandwidth values were tested, and the one that provided the best discrimination between IMFs was selected.

2.3. Model comparisons

Subsequently, and taking into account all the information provided by the previous calculations, a test of 6 Machine Learning (ML) models is performed, namely: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest-Neighbour (KNN) and Gaussian Naive Bayes (GNB). For this, first, the IMFs with the most different distributions between classes for both sexes are selected. Then, the different statistics of the IMFs that showed the greatest differences are extracted, and finally, the calculations of these statistics are provided to the models. It is expected that the models will capture complex changes in the combination of the data that the Bootstrapping technique was not able to detect.

To verify the importance of the differences in the probability distribution of the IMFs obtained using the Gaussian kernel, the models were tested both with all IMFs and only with those that showed significant differences in the Gaussian kernel. Similarly, the Machine Learning models were tested with balanced and unbalanced data in both sex groups. Several balanced datasets have been created through random undersampling, identifying the indices of each class (healthy and depression), randomly selecting the same number of samples from each class, and combining them to form a balanced dataset. This procedure is performed in each analysis, so the balanced subsets vary in every execution, helping to reduce the bias derived from the original class distribution.

Table 2
Scores of different Machine Learning models in balanced data. The data of the best models are presented in italics. The model names in the table refer to: RF = Random Forest, GB = Gradient Boosting, SVM = Support Vector Machine, LG = Logistic Regression, KNN = K-Nearest-Neighbours, GNB = Gaussian Naive Bayes.

Groups	Scores	RF	GB	SVM	LR	KNN	GNB
Female All IMFs	Accuracy	0.764	*0.823*	0.705	0.705	0.588	0.823
	Sensitivity	0.750	*0.875*	0.875	0.875	0.625	0.777
	Specificity	0.777	*0.777*	0.555	0.555	0.555	0.777
	ROC-AUC	0.890	*0.970*	0.750	0.740	0.670	0.880
Female Selected IMFs	Accuracy	0.823	*0.941*	0.764	0.823	0.588	0.764
	Sensitivity	0.875	*0.875*	1.000	1.000	0.750	0.875
	Specificity	0.777	*1.000*	0.555	0.666	0.444	0.666
	ROC-AUC	0.810	*0.930*	0.750	0.900	0.660	0.88
Male All IMFs	Accuracy	0.800	0.760	*0.840*	0.800	0.680	0.640
	Sensitivity	0.800	0.800	*0.900*	0.800	0.900	0.600
	Specificity	0.800	0.733	*0.800*	0.800	0.533	0.666
	ROC-AUC	0.910	0.920	*0.860*	0.810	0.810	0.710
Male Selected IMFs	Accuracy	0.800	*0.880*	0.840	0.720	0.800	0.760
	Sensitivity	0.800	*0.800*	0.900	0.700	1.000	0.900
	Specificity	0.800	*0.933*	0.800	0.733	0.666	0.666
	ROC-AUC	0.890	*0.980*	0.890	0.720	0.920	0.870

Different tests were carried out with IMF ranges in both sexes to evaluate the models, and the best results were obtained by selecting IMFs 1 to 7 inclusive in females and IMFs 2 to 6 inclusive in males. The Gaussian kernel was used as a weighting function to identify the IMFs that differ significantly between classes for each gender. Only these selected IMFs were then passed to the feature extraction stage and subsequently used in the classification models, shown in Fig. 5. With this procedure, we obtained 4 groups to be evaluated by the ML models: Female All IMFs, Female Selected IMFs, Male All IMFs, and Male Selected IMFs. Where in each of the groups, we look for a model with which we can predict the differences between depression and healthy, taking into account the IMFs.

Subsequently, a statistics extraction function was applied to all the IMFs, which included, in both groups according to sex: the mean, median, standard deviation, kurtosis, and skewness. These statistical values serve as features in the machine learning models, helping to distinguish between classes (depression and healthy) and allowing the patterns in the IMFs obtained from the signals in all groups to be summarised. They represent properties that capture the distribution and shape of each IMF, thereby enabling the models to detect key patterns. Two tests were then carried out: one with the statistics extracted from all the IMFs, and another only with those from the IMFs that showed differences in the Gaussian kernel test, with the aim of determining whether the differences between the IMF distributions constitute a relevant change in information for the ML models.

In model adjustments, it is important to consider that many machine learning algorithms, such as SVM and logistic regression, are sensitive to the scale of the data. Therefore, standardisation (using the z-score) was applied to both the training and test data of the model. The standardisation of the data was included in a pipeline within the model, which also incorporated a k-fold cross-validation method (parameterised as cv = 5, from the Sklearn library) and a hyperparameter search method (GridSearchCV from the same library), thereby ensuring that the results did not depend on a single data partition, which strengthens the robustness of the model analysis. The aforementioned machine learning models were then applied. Finally, both the confusion matrices for each model and the AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic) curves, as well as sensitivity (Recall) and specificity, were calculated and are presented in Tables 2 and 3.

3. Results

We present below the results obtained from the statistical comparison between the raw IMF data, the IMF distribution results, the comparison of IMFs in probability distribution density, and the comparative test of Machine Learning models, with all results separated by sex.

3.1. Statistic analysis

Analysing the voice dominant frequency graphs, significant differences are observed in the first IMFs (IMF1-IMF3) between depressive groups and healthy controls, where individuals with depression show notably higher dominant frequencies (1000–1300 Hz in women, 900–1300 Hz in men) with greater variability compared to healthy controls who exhibit lower and more stable frequencies.

An initial statistical analysis of the voice data between depression and healthy subjects for both sexes was carried out in order to determine whether there are statistical differences between the groups. Figs. 3 and 4 show the results of the bootstrapping analysis with False Discovery Rate (FDR) correction. In both cases, it can be observed that there are no differences between depression and healthy subjects according to the IMFs in the following statistics: mean, median, standard deviation, kurtosis, and skewness.

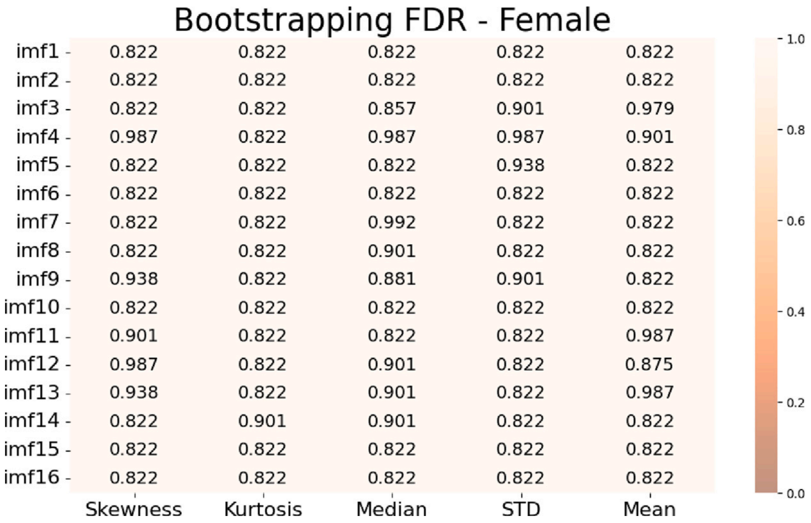

Fig. 3. Heatmap of p-values from the comparison of statistics between depression and health by IMFs in women. The Bootstrapping technique with FDR correction was applied.

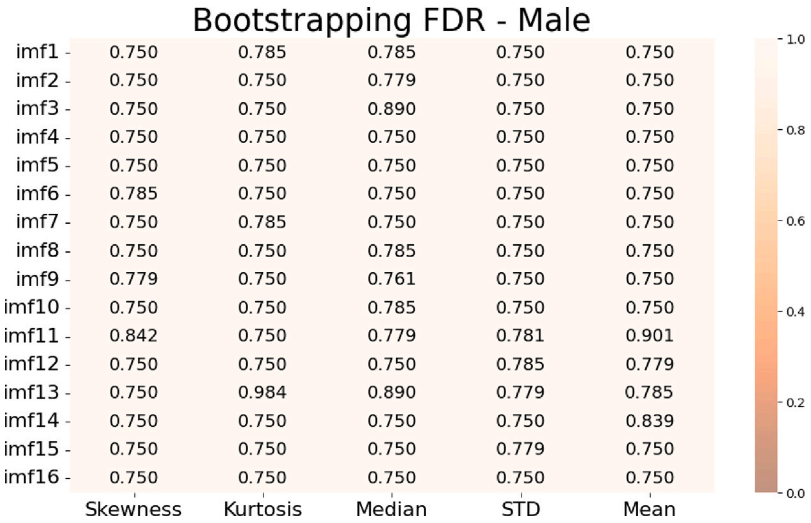

Fig. 4. Heatmap of p-values from the comparison of statistics between depression and health by IMFs in men. The Bootstrapping technique with FDR correction was applied.

The raw empirical decomposition data for each column (IMF) were processed using the bootstrapping method, with the parameterisation (n_iterations = 1000). Subsequently, FDR correction was applied to the obtained p-values. The results show no significant differences in any IMF for either sex.

The results of similarity in the distribution density function after applying the Gaussian kernel to the IMFs are presented in Fig. 5, separated by sex. This graph represents the relationship between the order of the Intrinsic Mode Functions (IMFs) and their degree of similarity in the probability density function between the depression and healthy groups. As shown in the legend, the bars in different colours indicate sex. The smaller the value represented on the y-axis (logarithmic scale), the greater the differences between depression and healthy subjects in the corresponding IMF.

These results show evident differences in the female group, particularly regarding the early IMFs, from 1 to 7, with a significant emphasis on IMF 3, and a considerable increase in similarity starting from IMF 7. In the male group, the differences are not as noticeable in the first IMF but are more pronounced in IMF 5, with similarity increasing from IMF 7.

In Table 2, we can see the results of the data processing by the models described in the Methods section. These results are presented in four groups: Female All IMFs: includes the statistical data for the female group (Results/Statistical Analysis) for all the IMFs. Female Selected IMFs: includes the models for the IMFs (1 to 7 inclusive) of the female group, selected by the Gaussian kernel as significantly different considering the sigma of the kernel for this sex. Male All IMFs: includes the statistical data for the male

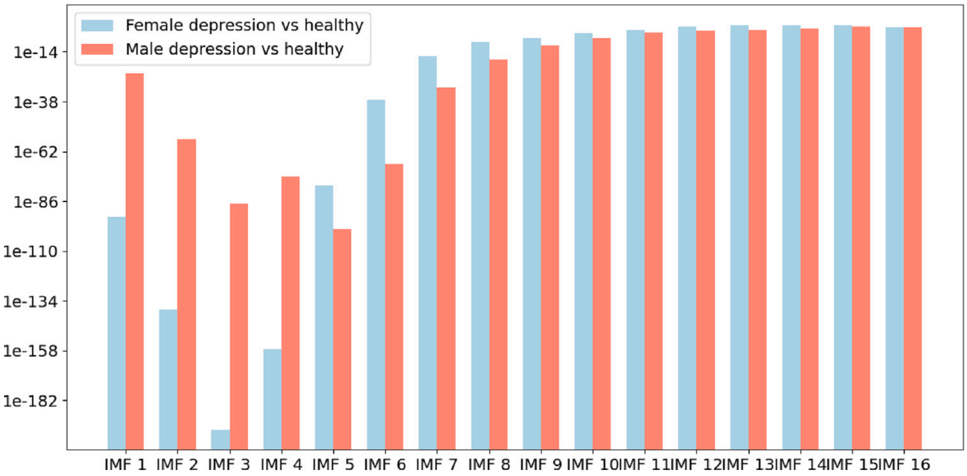

Fig. 5. Similarity of distribution density in the IMFs. A lower value indicates greater discrepancies in density between depression and health states (sigma = 0.5).

Table 3
Scores of different Machine Learning models in unbalanced data. The data of the best models are presented in italics. The model names in the table refer to: RF = Random Forest, GB = Gradient Boosting, SVM = Support Vector Machine, LG = Logistic Regression, KNN = K-Nearest-Neighbours, GNB = Gaussian Naive Bayes.

Groups	Scores	RF	GB	SVM	LR	KNN	GNB
Female All IMFs	Accuracy	0.529	0.235	0.529	0.470	0.470	0.529
	Sensitivity	0.625	0.375	0.000	0.625	0.500	0.000
	Specificity	0.444	0.111	1.000	0.333	0.444	1.000
	ROC-AUC	0.640	0.310	0.500	0.480	0.470	0.500
Female Selected IMFs	Accuracy	0.393	0.515	0.515	0.424	0.696	0.636
	Sensitivity	0.800	0.200	0.500	0.300	0.400	0.400
	Specificity	0.217	0.652	0.521	0.478	0.826	0.739
	ROC-AUC	0.480	0.550	0.540	0.390	0.460	0.570
Male All IMFs	Accuracy	0.600	0.680	0.600	0.760	0.680	0.640
	Sensitivity	0.000	0.300	0.000	0.600	0.400	0.100
	Specificity	1.000	0.933	1.000	0.866	0.866	1.000
	ROC-AUC	0.920	0.760	0.500	0.720	0.860	0.550
Male Selected IMFs	Accuracy	0.729	0.645	0.729	0.645	0.708	0.729
	Sensitivity	0.000	0.461	0.000	0.846	0.000	0.000
	Specificity	1.000	0.714	1.000	0.571	0.971	1.000
	ROC-AUC	0.660	0.680	0.500	0.700	0.640	0.500

group (Results/Statistical Analysis) for all the IMFs. Male Selected IMFs: corresponds to the models for the IMFs (2 to 6 inclusive) of the male group, selected from the Gaussian kernel results as significantly different.

As we can see in Table 2 and in Fig. 6 of the ROC curve, regarding the balanced data, all groups showed better performance in the ROC curve with the Gradient Boosting (GB) model.

In the first group, Female All IMFs, the Gradient Boosting model achieved the best overall performance, with the highest accuracy of 0.823, a sensitivity of 0.875, a specificity of 0.777, and an ROC-AUC of 0.970 when using all the IMFs. Although it has a high discriminative power between classes, the model has difficulty identifying false positives.

Female Selected IMFs with balanced metrics, we observe that GB again demonstrates exceptional performance with the selected IMFs, achieving an accuracy of 0.941 and an ROC-AUC of 0.930, indicating an excellent balance between precision and sensitivity (0.875), with a maximum specificity of 1.000, producing highly reliable and balanced results.

In our third group, corresponding to Male All IMFs with balanced metrics, the Support Vector Machine (SVM) achieves an accuracy of 0.840 and an ROC-AUC of 0.860, compared to 0.930 for GB, suggesting a better balance between true and false positive rates. This model also shows a high sensitivity (0.900).

Finally, in the group Male Selected IMFs with balanced data, Gradient Boosting shows better scores in this case, with an accuracy of 0.880 and an ROC-AUC of 0.980, a sensitivity of 0.800 and a specificity of 0.933. These results indicate slightly lower identification of positives but a better overall balance compared to the group using all IMFs.

In Table 3 and in Fig. 7, we can observe the results for the unbalanced data. In general, all groups show worse metrics than in the balanced groups. Considering these metrics, it is difficult to regard any model as truly effective in class evaluation.

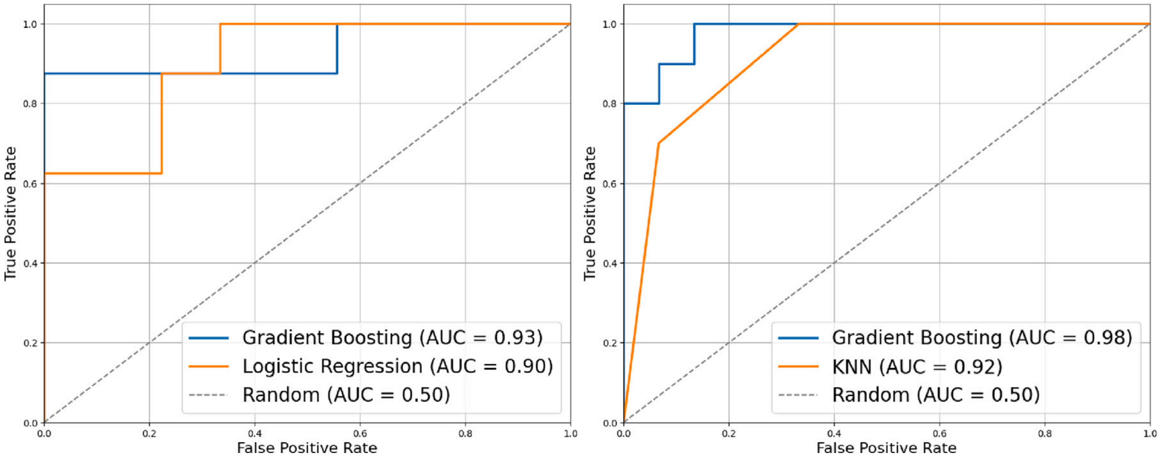

Fig. 6. Results of ROC curves on balanced data for selected IMFs: left panel for women and right panel for men.

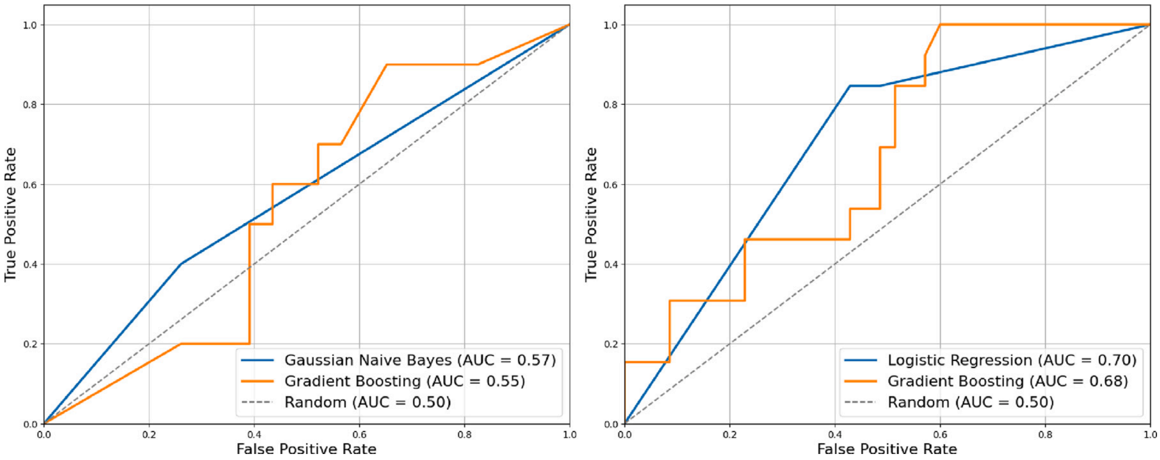

Fig. 7. Results of ROC curves on unbalanced data for selected IMFs: left panel for women and right panel for men.

In general, we found better metrics for the male sex than for the female sex. The Logistic Regression model appears to have a better overall fit for males compared to the others, although still far from the performance of models with balanced data. For females, the models face significant difficulties in discriminating between classes.

4. Discussion

The analysis of dominant frequencies on the IMFs decomposed by EMD indicates possible changes in phonation in depression compared to healthy states, particularly in females. This justifies the use of such techniques in identifying the most relevant IMFs for depression assessment through voice, since direct evaluation or the use of classical statistics on such a complex signal proves ineffective.

The results of the initial statistical analysis using bootstrapping show the difficulty of identifying differences between depression and healthy subjects in signal samples (IMFs). This reinforces our idea that voice signal data have a structure that should be considered globally, bringing us closer to the shape of the data distribution rather than its statistical analysis for pattern extraction.

Furthermore, in the comparison of the similarity of the IMFs according to the depression and healthy groups in the density distribution analysis, the data distributions in the lower IMFs (1–7) differ between the states of depression and healthy. The characteristics of the signal represented by the lower IMFs seem to manifest differently in depression depending on the sex. This may indicate that the underlying patterns of depression could have some gender differences. In principle, this comparison of distributions suggests that we will find more differences in the first IMFs, but also that the way to evaluate these differences by sex may vary, as the data distribution appears to be different in each case. In this regard, we also note that male voices generally exhibit greater spectral variability, which may require the inclusion of a larger number of IMFs to capture relevant signal information. This observation is consistent with previous findings in the literature indicating sex-based differences in the spectral and prosodic characteristics of speech.

Regarding the comparison of machine learning models, IMF selection improves model performance in all cases, but the effect is not as pronounced as with balanced data. This suggests that, while IMF selection is important, it alone cannot fully compensate for data imbalance. In the comparison of models with unbalanced data, the confusion matrices struggle to differentiate between classes and to correctly identify positive cases. Given the clinical interest in minimising false negatives, it is important to balance the data beforehand to avoid Type II errors (evaluating depression as healthy).

We then found that the selection of IMFs effectively contributes to the models, likely because the remaining IMFs add more noise than differentiable features between the depression and healthy groups, as we have seen in the density distribution similarity, where the first IMFs(1 to 7) presented more differences.

Regarding the models, in the case of selected IMFs versus all, in general, all models improve significantly when using selected IMFs instead of all IMFs, especially Gradient Boosting models, which show notable increases in performance. Feature selection seems to help eliminate noise and enhance predictive capability. Gradient Boosting outperforms SVM in the male group, suggesting a more uniform data structure with the presence of simpler or more easily separable patterns. In the case of the female group, GB also appears to perform better with the selection of IMFs.

5. Conclusions

Like most researchers, we believe that voice signal data, despite their low vector dimensionality (1D), present a complexity that must be taken into account when designing an analysis and classification method. The non-linearity and non-stationarity of this type of signal make it particularly challenging when identifying specific patterns, as in our study. In this research, we have been able to demonstrate the complexity of voice data and the difficulties associated with its analysis. Nevertheless, considering this situation, it is possible to create prediction models with a good fit.

The differences in voice characteristics found in the spectral analysis and their subsequent confirmation in the separation between depressive and healthy individuals as significant for ML models indicate that voice changes in depression are differentiable and assessable with this methodology. These differences in high frequencies suggest increased vocal cord tension characteristic of anxiety and stress states associated with depression, loss of fine phonatory control evidenced by the greater variability, and changes in laryngeal configuration due to emotional constriction and respiratory alterations, being more pronounced in women, possibly due to greater vocal emotional expressiveness. These findings in the first IMFs, which capture the rapid oscillations of the vocal signal, reflect the psychophysiological changes that depression produces in the neuromuscular control of the phonatory apparatus and could serve as objective biomarkers for early detection, treatment monitoring, and complementary assessment of emotional state in depression.

The analysis of the density distribution of IMF data appears to be a key element in differentiating between depression and healthy states. Based on this finding, the use of statistical measures on IMFs can be combined with models capable of handling complex data structures (e.g., stratified or kernel-based approaches), providing a more robust framework for classification.

Future studies should combine voice-based features with psychological and clinical assessments to validate the predictive value of the models in real-world settings. Expanding the dataset with more diverse and balanced samples, including different age groups and languages, could improve the generalisability of the findings. It would also be valuable to explore deep learning approaches, multimodal data (e.g., text and facial cues), and longitudinal voice recordings to monitor changes over time. Finally, testing these models in clinical environments would help evaluate their usability as practical screening tools for depression.

CRediT authorship contribution statement

Xavier Sánchez Corrales: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Jordi Solé-Casals:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Conceptualization.

Data availability

The `VOMEDIAL` algorithm implementation is publicly available at https://github.com/Xavisanco/VOMEDIAL-Voice-Mental-Disorders-Algorithm—Depression under a restricted academic use licence. The code is provided exclusively for academic and research purposes. Commercial use or redistribution requires explicit written permission from the authors. Researchers interested in using the algorithm should refer to the repository's `LICENSE` file for complete terms and conditions.

For this study, we used data from the Distress Analysis Interview Corpus (DAIC) (Akkaralaertsest and Yingthawornsuk, 2019) provided by the University of Southern California. This dataset includes carefully collected and anonymised voice recordings to ensure the privacy and confidentiality of the participants.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the University of Southern California for providing voice data and questionnaire information, without which this research would not have been possible. This work was carried out as part of the doctoral programme in Bioinformatics at the University of Vic—Central University of Catalonia

References

Akkaralaertsest, T., Yingthawornsuk, T., 2019. Classification of Depressed Speech Samples with Spectral Energy Ratios as Depression Indicator. http://dx.doi.org/10.1109/iSAI-NLP48611.2019.9045167.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., Parker, G., 2013. A comparative study of different classifiers for detecting depression from spontaneous speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, New York, pp. 8022–8026, URL https://www.webofscience.com/wos/woscc/full-record/WOS:000329611508037. ISSN: 1520-6149 Num Pages: 5 Series Title: International Conference on Acoustics Speech and Signal Processing ICASSP Web of Science ID: WOS:000329611508037.

2023. Audacity. URL https://sourceforge.net/projects/audacity/.

Corrales, X.S., Solé-Casals, J., García, E.A., Vidal, D.P., 2025. Analyzing male depression using empirical mode decomposition. In: Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS. SciTePress, INSTICC, pp. 886–892. http://dx.doi.org/10.5220/0013157600003911.

Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J., 2015. Analysis of acoustic space variability in speech affected by depression. Speech Commun. 75, 27–49. http://dx.doi.org/10.1016/j.specom.2015.09.003.

2022. DAIC-WOZ database. URL https://dcapswoz.ict.usc.edu/.

Donaghy, P., Ennis, E., Mulvenna, M., Bond, R., Kennedy, N., McTear, M., O'Connell, H., Blaylock, N., Brueckner, R., 2024. A review of studies using machine learning to detect voice biomarkers for depression. J. Technol. Behav. Sci. 9, 892–915. http://dx.doi.org/10.1007/s41347-024-00454-2, Narrative review with a systematic search of the efficacy of voice biomarkers for depression :contentReference[oaicite:0]index=0.

France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 47 (7), 829–837. http://dx.doi.org/10.1109/10.846676, URL https://ieeexplore.ieee.org/document/846676. Conference Name: IEEE Transactions on Biomedical Engineering.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., Morency, L.-P., 2014. The distress analysis interview corpus of human and computer interviews. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), LREC 2014 - Ninth International Conference on Language Resources and Evaluation. European Language Resources Assoc-Elra, Paris, pp. 3123–3128, URL https://www.webofscience.com/wos/woscc/full-record/WOS:000355611004122. Num Pages: 6 Web of Science ID: WOS:000355611004122.

Koops, S., Brederoo, S.G., De Boer, J.N., Nadema, F.G., Voppel, A.E., Sommer, I.E., 2023. Speech as a biomarker for depression. CNS Neurol. Disord. - Drug Targets 22 (2), 152–160. http://dx.doi.org/10.2174/1871527320666211213125847, URL https://www.eurekaselect.com/198825/article.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B.W., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. J. Affect. Disord. 114 (1–3), 163–173. http://dx.doi.org/10.1016/j.jad.2008.06.026.

Liu, Z., Xu, Y., Ding, Z., Chen, Q., 2020. Time-frequency Analysis Based on Hilbert–Huang Transform for Depression Recognition in Speech. pp. 1072–1076. http://dx.doi.org/10.1109/BIBM49941.2020.9313587.

Low, L.-S., Maddage, N., Lech, M., Allen, N., 2009. Mel Frequency Cepstral Feature and Gaussian Mixtures for Modeling Clinical Depression in Adolescents. pp. 346–350. http://dx.doi.org/10.1109/COGINF.2009.5250714.

Mao, K., Wu, Y., Chen, J., 2023. A systematic review on automated clinical depression diagnosis. Npj Ment. Heal. Res. 2 (1), 9. http://dx.doi.org/10.1038/s44184-023-00040-z, Open Access systematic review examining speech, text and facial modalities for automated diagnosis of depression using PRISMA framework :contentReference[oaicite:0]index=0.

Menne, F., Dörr, F., Schräder, J., Tröger, J., Habel, U., König, A., Wagels, L., 2024. The voice of depression: Speech features as biomarkers for major depressive disorder. BMC Psychiatry 24 (1), 794. http://dx.doi.org/10.1186/s12888-024-06253-6, Open Access study identifying discriminating speech features and constructing an SVM model (AUC=0.93) for distinguishing patients with MDD from healthy controls :contentReference[oaicite:0]index=0.

Sharma, R., Prasanna, S.R.M., Bhukya, R.K., Das, R.K., 2017. Analysis of the intrinsic mode functions for speaker information. Speech Commun. 91, 1–16. http://dx.doi.org/10.1016/j.specom.2017.04.006, URL https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=DOISource&SrcApp=WOS&KeyAID=10.1016/j.specom.2017.04.006&DestApp=DOI&SrcAppSID=EUW1ED0BBA5o9FxXhj6r07bu4HEuL&SrcJTitle=SPEECH+COMMUNICATION&DestDOIRegistrantName=Elsevier. Num Pages: 16 Place: Amsterdam Publisher: Elsevier Web of Science ID: WOS:000403858100001.

Solé-Casals, J., Gallego-Jutglà, E., Martí-Puig, P., Travieso, C., Alonso, J., 2013. Speech enhancement: A multivariate empirical mode decomposition approach. Lect. Not. Comput. Sci. (Including Subser. Lect. Not. Artif. Intell. Lect. Not. Bioinform. 7911 LNAI, 192–199. http://dx.doi.org/10.1007/978-3-642-38847-7_25, ISBN: 9783642388460.

Whiteside, S., Irving, C., 1997. Speakers' sex differences in voice onset time: Some preliminary findings. Percept. Mot. Skills 85 (2), 459–463. http://dx.doi.org/10.2466/pms.1997.85.2.459.