Machine Learning-Based Audio Classifier for Healthcare Monitoring

Author:

Andrés Martínez Almazán A01621042

Course:

Modeling Learning with Artificial Intelligence

Professor:

Dr. Omar Mendoza Montoya

Date:

June 10, 2025

1 Description of the Implemented Application and Collected Data

1.1 What does the application consist of?

The implemented application is a real-time classification system for environmental sounds based on audio amplitude, developed using supervised learning. This system uses data collected through the Phyphox application Staacks, Hütz, Heinke, and Stampfer (2019), which captures sound pressure levels (dB) and associated timestamps using a mobile device. The application employs a Support Vector Machine (SVM) model with RBF kernel, pre-trained with features extracted from 0.5-second audio windows, to classify sounds into six categories: silence/ambient, conversation, whisper, scream, music, and applause. The data is processed online through a remote connection with Phyphox, enabling continuous predictions and visualization of results with confidence probabilities.

1.2 What is it expected to do?

The application is expected to monitor the sound environment in real-time and automatically classify detected sound types with high precision, providing a useful tool for medical applications. Specifically, it should identify audio patterns that may be related to health conditions, such as detecting screams as distress indicators (e.g., calls for help, emergency whistles, or sounds of falling impacts), whispers as signs of vocal fatigue, or ambient noise levels affecting sensitive patients. Additionally, it is expected to offer a simple interface to visualize predictions and their confidence levels, facilitating its integration into alert or medical monitoring systems.

1.3 What examples of similar applications exist?

Similar applications exist in the medical and environmental monitoring fields. For example, systems like those developed by Philips with their patient monitoring technology use audio analysis to detect falls or critical events in hospitals Philips Healthcare (2020). Another application is Stanford University's "Sound Health" digital health platform, which explores the use of artificial intelligence to analyze vocal patterns in patients with neurological disorders Stanford University (2019). Additionally, devices like Apple wearables with fall detection employ audio algorithms to identify sounds associated with emergencies, showing a comparable approach to the one proposed Apple Inc. (2021).

1.4 What type of information will it monitor?

The application monitors audio amplitude in the form of sound pressure levels (dB) and associated timestamps, captured by a mobile device's microphone through Phyphox. This information is processed in 0.5-second windows to extract features such as mean, standard deviation, percentiles, derivatives, and zero-crossing rate. The system focuses on detecting temporal and dynamic patterns that differentiate the six defined sound types, with particular emphasis on variations that may be indicative of health states, such as amplitude peaks (screams) or soft patterns (whispers).

2 Description of Features Extracted from the Data

The features extracted from the data come from audio amplitude signals captured through Phyphox, processed in 0.5-second sliding windows to ensure consistency with real-time classification. This procedure generated a set of 31 variables per window, carefully designed to reflect both basic statistics and distinctive dynamic patterns among the six sound categories. Among these features are descriptive statistics, such as mean, standard deviation, skewness, kurtosis, median, minimum, maximum, range (peak-to-peak), variance, and percentiles (10, 25, 75, 90), which provide a complete view of the signal's distribution and dispersion.

Additionally, energy measures were calculated, including sum of squares, average power (mean of squares), and sum of absolute values, which evaluate sound intensity. To capture abrupt changes in amplitude, temporal variables were derived such as velocity (first derivative) and acceleration (second derivative), represented by their averages and standard deviations. Another key metric is the zero-crossing rate (ZCR), which measures the proportion of crossings through the mean level, being useful for identifying rhythmic or irregular patterns.

The analysis is complemented with statistical moments, such as the third and fourth moments, which evaluate asymmetry and distribution shape, respectively, along with the median absolute deviation (MAD), a robust measure of dispersion. Other features include an approximate signal-to-noise ratio (SNR), calculated as the quotient between the maximum and standard deviation (adjusted to avoid division by zero), and the dynamic range, defined as the difference between the 90th and 10th percentiles, reflecting extreme variability. The spectral centroid (a temporal approximation based on weighting absolute values by their position) and spectral flux (mean of absolute differences between consecutive samples) were also incorporated, indicating temporal "centroidicity" and rate of change, respectively. Finally, the root mean square (RMS) was added as the average signal magnitude, reinforcing energy measures.

All these features were extracted using signal processing methods, ensuring the identification of distinctive patterns among sound categories (silence/ambient, conversation, whisper,

scream, music, applause). The process included signal interpolation at a 20 Hz sampling rate to standardize the windows, resulting in 6408 training samples for the SVM model.

3 Summary of Classifier Evaluation Results

The classifier evaluation was conducted using a processed dataset comprising 30 samples, each with 31 features extracted from audio signals collected through Phyphox. These data were scaled with a StandardScaler and evaluated using cross-validation with 5 folds (StratifiedK-Fold) to ensure balanced distribution of the six classes (silence/ambient, conversation, whisper, scream, music, applause). Below are the detailed results of the tested classifiers, organized into classical and unseen models, with tables summarizing the classification reports.

Table 1: Classification Report - Linear SVM

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.60	0.60	0.60	5
5.0 (Music)	0.67	0.40	0.50	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.80	30
Macro avg	0.81	0.80	0.79	30
Weighted avg	0.81	0.80	0.79	30

Table 2: Classification Report - SVM with RBF Kernel

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.60	0.60	0.60	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.86	0.83	0.84	30
Weighted avg	0.86	0.83	0.84	30

Table 3: Classification Report - LDA

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	0.80	0.89	5
2.0 (Conversation)	0.75	0.60	0.67	5
3.0 (Whisper)	0.83	1.00	0.91	5
4.0 (Scream)	0.71	1.00	0.83	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.84	0.83	0.83	30
Weighted avg	0.84	0.83	0.83	30

Table 4: Classification Report - K-NN

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.56	1.00	0.71	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.67	0.40	0.50	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.87	0.83	0.83	30
Weighted avg	0.87	0.83	0.83	30

Table 5: Classification Report - MLP

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.75	0.60	0.67	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.85	0.83	0.83	30
Weighted avg	0.85	0.83	0.83	30

Table 6: Classification Report - Gaussian Naive Bayes

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.80	0.80	0.80	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.80	0.80	0.80	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.93	30
Macro avg	0.93	0.93	0.93	30
Weighted avg	0.93	0.93	0.93	30

Table 7: Classification Report - Gradient Boosting

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.67	0.80	0.73	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	1.00	0.80	0.89	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.90	30
Macro avg	0.91	0.90	0.90	30
Weighted avg	0.91	0.90	0.90	30

Table 8: Classification Report - Extra Trees

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.80	0.80	0.80	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	1.00	1.00	1.00	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.93	30
Macro avg	0.93	0.93	0.93	30
Weighted avg	0.93	0.93	0.93	30

Table 9: Classification Report - Gaussian Process

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.50	0.60	0.55	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.67	0.60	0.63	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.80	30
Macro avg	0.82	0.80	0.81	30
Weighted avg	0.82	0.80	0.81	30

Table 10: Classification Report - Nearest Centroid

Class	Precision	Recall	F1-score	Support
1.0 (Silence/ambient)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Scream)	0.75	0.60	0.67	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.86	0.83	0.85	30
Weighted avg	0.86	0.83	0.84	30

3.1 Hyperparameter Optimization and Feature Selection

The hyperparameter optimization and feature selection process for SVM models with RBF kernel and Gradient Boosting was developed through the following procedure. Initially, hyperparameter performances were plotted through preliminary explorations to determine optimal values. For SVM with RBF kernel, the parameters C (regularization) and gamma (kernel width) were varied over a wide range, identifying that C = 20.0 and gamma = 0.15 offered the best performance according to the generated curves. For Gradient Boosting, n_estimators (number of trees), max_depth (maximum depth), and learning_rate were adjusted, highlighting that n_estimators = 50, max_depth = 2, and learning_rate = 0.05, along with k = 12 for feature selection, maximized accuracy in preliminary graphs.

Next, feature selection was performed. For SVM, SelectKBest was used with the f_classif criterion, selecting the 10 most relevant features based on their statistical score, corresponding to indices [4, 8, 10, 11, 12, 15, 18, 20, 22, 23], representing median, variance, 25th, 75th, and 90th percentiles, average power, velocity and acceleration standard deviations, and third and fourth moments. For Gradient Boosting, SelectKBest was also employed with k = 12, identifying indices [0, 4, 8, 10, 11, 12, 14, 15, 18, 20, 22, 23], which include mean, median,

variance, 25th, 75th, and 90th percentiles, total energy, average power, velocity and acceleration deviations, and third and fourth moments, according to nested cross-validation results.

Finally, definitive optimization was carried out through nested cross-validation with GridSearchCV in the inner loop. For SVM RBF, C = 20.0 and gamma = 0.15 were confirmed, achieving an accuracy of 0.9667 in the inner CV. For Gradient Boosting, n_estimators = 50, max_depth = 2, learning_rate = 0.05, and k = 12 for feature selection were validated, resulting in an overall average accuracy of 0.9333.

4 Online Application Results

The online application for real-time classification of environmental sounds was implemented using a previously trained SVM model with RBF kernel. Data were processed in 0.5-second windows with 50% overlap (0.25-second step), interpolated at a 20 Hz sampling rate, and 31 features were extracted per window, replicating the training method. These features were scaled and the 10 most relevant were selected using a pre-trained feature selector, before passing to the final SVM model with optimal hyperparameters (C = 20.0, gamma = 0.15). Below are the results observed during online testing, based on the provided output example.

4.1 Ambient Sound Classification

Ambient sound classification was perfect, with the model consistently identifying this category across multiple consecutive predictions (e.g., predictions with confidence levels between 0.80 and 0.99). This indicates that the model is highly effective at detecting silent environments or those with minimal background noise.

4.2 Music Classification

Music classification was accurate, as long as the volume was high, as the model was trained. In the output example, multiple correct predictions were observed with confidence levels ranging from 0.61 to 0.99. This demonstrates the model's ability to identify rhythmic and dynamic patterns associated with music at elevated volumes.

4.3 Applause Classification

Applause classification was very good, although not all predictions were identified as applause due to differences in timing between events. However, the model successfully identified most

applause, with confidence levels between 0.56 and 1.00. This suggests that the model is robust for detecting intermittent sound events, although temporal variability may generate some erroneous predictions.

4.4 Whisper Classification

Whisper classification was perfect as long as the whisper remained within a specific amplitude range. Multiple instances were correctly identified with confidence levels between 0.59 and 0.95. However, if the whisper was too loud, the model tended to confuse it with normal conversation.

4.5 Conversation Classification

Conversation classification was good, but required a specific and constant tone of voice. Conversations were identified in predictions with confidence levels between 0.49 and 0.78. However, if the tone was not consistent, the model tended to confuse conversation with a scream.

4.6 Scream Classification

Scream classification was the least accurate, requiring a very specific and constant tone of voice. Although some screams were identified, the model frequently confused them with music or applause, with variable confidence levels (0.45 to 0.89). This indicates that the model has difficulty differentiating screams from other high-amplitude sounds, suggesting the need to adjust the model or incorporate additional features, such as spectral analysis, to improve accuracy.

4.7 Overall Performance Analysis

4.7.1 Does it work the same for all team members?

The online application does not work identically for all team members due to variations in audio capture between devices and environments. Tests conducted indicated that microphone quality and environmental conditions (e.g., background noise, distance to microphone) influence prediction accuracy. For example, a member with a high-sensitivity device (such as a high-end smartphone) obtained better results for whispers and applause, while another with a lower-quality device showed inconsistencies in detecting screams and conversations. This suggests that hardware calibration and environment standardization are critical factors to ensure uniform performance among team members.

4.7.2 Is the online application performance as expected according to cross-validation results?

The online application performance is generally consistent with cross-validation results obtained for the SVM model with RBF kernel. In online tests, categories such as ambient sound, music, whisper, and applause showed success rates close to or above 80-100%, aligning with expectations based on cross-validation. However, scream classification was significantly less accurate, with frequent confusions with music and applause, suggesting performance below expectations. This could be due to variability in real-time recordings or the lack of specific spectral features in the current model, indicating a need for refinement to fully meet cross-validation projections.

5 Conclusion

This project successfully developed a real-time audio classification system for medical monitoring applications, demonstrating the feasibility of using amplitude-based audio analysis in healthcare environments. The optimized SVM model with RBF kernel achieved 96.67% validation accuracy, outperforming other evaluated classifiers after hyperparameter tuning.

The main limitation identified was the confusion between screams and music due to overlapping high-amplitude characteristics. Since the system relies exclusively on amplitude-based features, sounds with similar intensity but different frequency content are difficult to distinguish. Additionally, device variability and environmental conditions affect classification reliability.

Future work should incorporate spectral features such as MFCCs to improve differentiation between similar-amplitude sounds, and expand the training dataset across diverse devices and environments. This project provides a foundation for more sophisticated patient monitoring systems in healthcare settings.

References

Apple Inc. (2021). *Fall detection with audio algorithms in wearable devices.* Technical documentation. Retrieved 2025-06-10, from `https://developer.apple.com/documentation/healthkit`

Philips Healthcare. (2020). *Patient monitoring systems: Audio-based fall detection* (Technical report). Amsterdam, Netherlands: Philips Healthcare.

Staacks, S., Hütz, S., Heinke, H., & Stampfer, C. (2019). Advanced tools for smartphone-based experiments: Phyphox. *Physics Education*, *53*(4), 045009. doi: 10.1088/1361-6552/aac05e

Stanford University. (2019). *Sound health initiative: Analyzing vocal patterns in neurological disorders.* Research project. Retrieved 2025-06-10, from `https://med.stanford.edu/soundhealth`