Ingénierie des Systèmes d'Information
Vol. 30, No. 3, March, 2025, pp. 763-770
Journal homepage: http://iieta.org/journals/isi

An Audio Feature Extraction Approach Using Machine Learning for Spoken Audio File Indexing

Hortense N'dri Koua[1*], Marcellin Konan Brou[1], Maxime Seraphin Gnagne[2]

[1] Research and Innovation Unit in Mathematics and Digital Sciences, National Polytechnic Institute Félix Houphouët-Boigny (INP-HB), Yamoussoukro BP 1093, Côte d'Ivoire
[2] Department of Mathematics and Computer Science, University Félix Houphouët-Boigny, Abidjan 01 BP V34, Côte d'Ivoire

Corresponding Author Email: ndri.koua18@inphb.ci

https://doi.org/10.18280/isi.300320

Received: 7 January 2025
Revised: 20 February 2025
Accepted: 24 February 2025
Available online: 31 March 2025

Keywords:
audio descriptors, data mining, multilayer perceptrons

ABSTRACT

With the rise of voice and audio data processing, data mining now enables the automatic extraction of hidden knowledge from vast datasets, particularly in speech processing and spoken language dialogue. In light of the growing volume of voice data generated by applications such as e-learning and e-conferences, data warehouses hold a substantial amount of audio data requiring efficient indexing. However, traditional algorithms struggle to handle the specificities of such voice data. This work proposed a novel extraction approach leveraging heuristic algorithms based on original descriptors obtained through wavelet analysis, an audio feature engineering process, and an autoencoder for selecting relevant descriptors. We extracted 119 descriptors from 1,000 audio files in the GTZAN database. After the extraction phase, we used machine learning in the encoder's latent space to select 80 relevant descriptors that preserve the semantic meaning of the files. We evaluated the effectiveness of our approach using a multilayer perceptron. The neural network achieved an accuracy of 87.5% compared to 77.8% reported in the literature.

1. INTRODUCTION

With the rise of digital technologies and the continuous increase in data storage and processing capacities, the volume of audio and voice data has experienced exponential growth [1]. This evolution is primarily driven by the widespread adoption of digital platforms across various domains such as e-learning, e-conferences, automatic transcription systems, and intelligent voice assistants [2]. These systems generate massive amounts of audio data that require efficient indexing for optimal utilization. For instance, online courses and virtual conferences produce thousands of hours of audio recordings daily that need to be classified, searched, and reused.

However, audio signals exhibit complex variability due to several factors, such as variations in speech rhythm, pitch, and intensity, speaker-specific resonances, and background noise [3, 4]. In the face of this complexity, traditional methods for extracting audio descriptors, often based on heuristic algorithms for simple feature extraction (e.g., Mel-Frequency Cepstral Coefficients (MFCC) techniques), show significant limitations. They struggle to capture the rich spectral and temporal characteristics of audio signals, especially in noisy environments or for complex spoken files [5].

Recent advances in audio descriptor extraction techniques and machine learning have enabled new approaches capable of better handling this variability. For example, wavelet analysis can decompose audio signals into time-frequency coefficients, providing a multi-scale representation of the signal [6]. While commonly used models such as Support Vector Machines (SVM) with the One-vs-Rest (OVR) strategy have demonstrated robust performance for classifying audio files in multi-class environments [3], they generally underperform compared to ensemble learning methods. As indicated by Gnagne et al. [7], individual models struggle to surpass ensemble learning methods. Furthermore, traditional algorithms have difficulty managing the complex specificities of voice data, particularly when signals are subject to perturbations or nonlinear spectral variations.

A central question emerges: How can we extract and select optimal audio descriptors for efficient indexing of spoken audio files?

Based on the premise that descriptors distinctively characterize each audio file, we hypothesize that audio data are inherently nonlinear, requiring nonlinear models for analysis. From this hypothesis, we derive the following subsidiary questions:

RQ1: How can we extract and select the most relevant audio descriptors for indexing?

RQ2: What is the best kernel configuration for an SVM model?

RQ3: What classification approach for spoken audio files ensures better indexing?

Our solution involves developing a hybrid system that combines an innovative heuristic extraction method with a machine learning-based approach for selecting relevant descriptors, all evaluated using a neural network.

To address these concerns, we organize this article as follows: In the second section, we present a related work on

approaches for extracting and classifying audio files.

Section 3 presents our approach. In Section 4, we present and discuss our results, and finally, we conclude the article.

2. RELATED WORK

The indexing of spoken audio files relies on two essential components: the extraction and selection of audio descriptors, and supervised classification. Recent research focuses on optimizing the relevance of descriptors and selecting the most effective classification algorithms, based on rigorous mathematical foundations to refine these steps. The extraction and selection of audio descriptors are crucial steps in the analysis of vocal and audio data. These descriptors serve as numerical representations of the acoustic features of audio signals, including elements such as the spectrum, energy, as well as temporal and prosodic characteristics. Table 1 provides an overview of recent advances in this field.

The extraction methods from previous works are generally either basic techniques or combined approaches. These methods have allowed the extraction of a large set of high-dimensional descriptors. A complementary approach of selecting a subset of features that retains the semantic content of the audio files has been introduced in some works, thereby improving the quality of the extracted descriptors. Most methods are evaluated using SVMs, and occasionally Random Forest. Although these approaches yield acceptable results, they do not leverage machine learning in the selection of the extracted descriptors.

Table 1. Summary of audio file extraction, selection, and classification methods from the state of the art

Ref.	Extraction and Selection Methods	Classification Methods	Contribution	Limit
Santiago et al. [2]	Audio descriptor engineering that focuses on efficient extraction and selection of relevant features	Generalized Linear Model (GLM), Random Forests (RF) and XGBoost	Optimization of functionalities for occupancy and activity detection in smart buildings	Specificity of functionalities for precise scenarios, limiting generalization
Besbes and Lachiri [3]	Extraction of spectral (MFCC, GFCC) and prosodic (Energy, formants, pich) features and combination of features for selection.	Multi-class SVM (OVR, OVO, Direct Acyclic Graph (DAG)	Improved speech recognition under stress	Method sensitive to noise and interlocutor variations
Kobayashi et al. [4]	Wavelet transform, rasterization and local feature extractions	Multi-class SVM. The unspecified binary transformation	Accurate classification of music genres based on innovative features	Not evaluated because evaluation procedure not specified
Jimenez et al. [5]	Sound engineering, Statistical measurement of the time series and the genetic algorithm	SVM à noyaux Kernel SVM with unspecified strategy	Proposal of a complete and automated method for the extraction and selection of audio descriptors	Performance dependent on the size and diversity of the datasets
Abdoune and Fezari [6]	Windowing process for extraction of temporal and frequency features and for Matching Pursuit (MP) selection, Fisher Discriminant Ratio (FDR) and Principal Component Analysis (PCA)	Gaussian Melange Model (GMM), Hidden Markov Model (HMM), Nearest Neighbors (KNN), SVM, Random Forest, TESPAR (Time Encoded Signal Processing and Recognition)	Detection and recognition of environmental sounds for distress situations	Lack of database standardization for performance validation
Besbes and Lachiri [8]	Using Mel frequency cepstral coefficients (MFCC) and multitaper MFCC	Gaussian kernel SVM (OVR) OAO, One Class SVM	Better consideration of variations due to stress in vocal signals	Limited reliability in the presence of high environmental noise
Barandas et al. [9]	Using the Time Series Feature Extraction Library (TSFEL)	Not specified	TSFEL offers more than 60 methods for feature extraction from time series, covering temporal, statistical and spectral domains.	Reliance on data quality and lack of integrated classification methods
Alam et al. [10]	Not taken into account	One-Class Support Vector Classifier (OCSVC)	Comprehensive review of available OCSVC algorithms	Unidentified in the context of audio file descriptors
Zaman et al. [11]	STFT (Short-Time Fourier Transform), Mel-spectrograms, MFCC	Random Forest, SVM	Comprehensive review of audio classification techniques using deep learning	Strong dependence on training data and significant computational costs
Bernard et al. [12]	Extraction of MFCC Estimation of pitch (pitch), Normalization with VTLN (Vocal Tract Length Normalization) coefficients Assessment of phoneme discriminability	Not specified	Open-source tool for speech feature extraction	Complexity of integration and limited applicability in an environment with noise
Hurbungs et al. [13]	Not specified	SVM (OVO, OVR, OVN)	Proposal of an innovative multi-class classification method to reduce interclassification errors	Method still limited to specific applications requiring more testing on varied real data

3. METHODOLOGY

3.1 Datasets used

To evaluate the proposed method, we used the GTZAN dataset [14, 15], a standard reference for audio classification tasks used by around a hundred works.

This dataset contains 1000 audio recordings evenly distributed across 10 music genres: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock [5]. Each recording lasts 30 seconds and is in WAV format, providing sufficient quality for extracting relevant audio features. The choice of the GTZAN dataset is based on several criteria:

• Data diversity: The 10 genres cover a wide range of sound characteristics, enabling robust analysis of various acoustic variations;

• Data size: With 1000 samples, the dataset is large enough to train and test supervised learning models;

• Popularity: This dataset is widely used in research, enabling direct performance comparisons with other existing methods [16].

In the case of our study, Table 2 shows the distribution of the 1000 files by genre. Each type of audio contains 100 files.

Table 2. Distribution of musical genres in the GTZAN dataset

Genres	**Pop**	**Blues**	**Reggae**	**Classical**	**Rock**
Labels	0	1	2	3	4
Genres	**Jazz**	**Metal**	**Hiphop**	**Disco**	**Country**
Labels	5	6	7	8	9

The ability of a model to adapt to different types of data is essential for evaluating its robustness and applicability in real-world scenarios. With this in mind, we chose to test our model on data from FMA (Free Music Archive), a vast database consisting of 106,574 music tracks categorized into a hierarchical taxonomy of 161 genres. FMA provides full, high-quality audio excerpts, with MP3-encoded files and various sizes of audio data:

fma_small: 8,000 tracks, 8 balanced genres (similar to GTZAN);

fma_medium: 25,000 tracks, 16 imbalanced genres;

fma_large: 106,574 tracks, 161 imbalanced genres.

We chose fma_small to maintain class balance, which helps avoid biases caused by imbalanced class distributions.

Table 3 shows the distribution of the 8000 files by genre. Each type of audio contains 100 files.

3.2 Audio descriptor extraction methods

Audio descriptors play a key role in the analysis and classification of spoken audio files. They represent numerical characteristics of the audio signal, allowing its distinctive properties to be captured. Our extraction approach combines the audio feature extraction method based on undecimated wavelet transform (UWT), used to decompose an audio signal into multiple sub-bands in order to analyze the different frequencies in detail [4], with techniques based on sound engineering, descriptive statistics of central tendency or variability, and time series [5]. Tables 4 and 5 summarize the list of descriptor classes used in our combination approach.

Table 3. Distribution of musical genres in the FMA dataset

Genres	**Pop**	**HipHop'**	**Rock**	**Electronic**
Labels	0	1	2	3
Genres	**Experimental**	**Instrumental**	**International**	**Folk**
Labels	4	5	6	7

Table 4. Classes of descriptors proposed by Kobayashi et al. [4]

Methods	**Type of Descriptors**	**Some Descriptors**
Descriptive statistics	Local sub-band statistics	Average of subband values: $\mu_{\propto}^{(j)}$
		Coefficient of variation: $v_{\propto}^{(j)}$
		Average coefficient of variation: $\mu_{\beta}^{(j)}$
		Coefficient of variation of coefficient of variation: $v_{\beta}^{(j)}$
	Correlations between sub-bands	Average of correlation coefficients between two sub-bands: $\mu_{\gamma}^{(je,j)}$
		Variance of correlation coefficients: $\sigma_{\gamma}^{(je,j)}$
	Global statistics	Average absolute value of the signal: A

Table 5. Classes of descriptors used by Jiménez et al. [5]

Methods	**Type of Descriptors**	**Some Descriptors**
Sound engineering	Acoustic descriptors	ACI (Acoustic Complexity Index), M (Acoustic Index based on the median of the amplitude envelope), NDSI (Normalized Difference Soundscape Index), Q (Quality Factor)
	Spectral descriptors	Mean frequency, Median frequency, Mode (dominant frequency), Quartiles (Q25, Q75), IQR (Interquartile Range), Roughness
Descriptive statistics	Statistics based on STFT time and frequency contours	Time P1 (The time initial percentile of the time contour), Time IPR (The time interpercentile range), Freq M (The frequency median), Freq P2 (The frequency terminal percentile)
	Statistical properties of a frequency spectrum	Mean: Mean frequency, Median: Median frequency, Kurtosis Kurtosis, a measure of peakedness
Time series	Statistical descriptors of time series	Entropy Spectral entropy, Stability Variance of the means for tiled (non-overlapping) windows
		Lumpiness Variance of the variances for tiled (non-overlapping) windows
		Crossing points the number of times a time series crosses the median line

3.3 Descriptor selection approach

In the process of using audio files for a particular task, two steps are essential. After the extraction phase, the selection of relevant descriptors is crucial to reduce the dimensionality of the feature space while preserving the file's semantics. Several approaches have been proposed, including heuristic methods such as Forward, Backward, and genetic algorithms [5]. We propose an approach based on machine learning, more specifically, an autoencoder. Generally, an autoencoder is modeled by a function $\mathcal{H}$ applied to a random $X \in \mathcal{R}^n$such that: $\|\mathcal{H}(x) - x\| \leq \varepsilon$ [14].

That is, the image of x by $\mathcal{H}$ is a reconstruction of x with an error ε. More specifically, $\mathcal{H} = \mathcal{D}o\mathcal{E}$ with $\mathcal{E}: \mathcal{R}^n \to \mathcal{R}^d$called the encoder and $\mathcal{D}: \mathcal{R}^d \to \mathcal{R}^n$the decoder ($d \leq n$). The encoder compresses x by reducing its dimension, while the decoder reconstructs x from the reduced code z. The space of reduced codes, $\mathcal{R}^d$, is called the latent space. It is this space that will provide us with the essential descriptors for our indexing through machine learning. The selection of the latent space dimension represents a good compromise between model complexity and generalization capability.

We opted for multiple trials followed by cross-validation, evaluated the model's performance using metrics such as reconstruction error, training loss, and validation loss, and ultimately selected the dimension that minimizes reconstruction error while avoiding overfitting.

We will evaluate the relevance of the selected descriptors using a dense neural network. The process of the overall approach unfolds in three essential phases. The first step involves extracting potential relevant descriptors using heuristic methods. Next, with an autoencoder, we select a subset of essential features through machine learning in the latent space of the encoder. Finally, we evaluate the relevance of the extracted descriptors using a neural network. The process of the overall approach is illustrated in Figure 1.

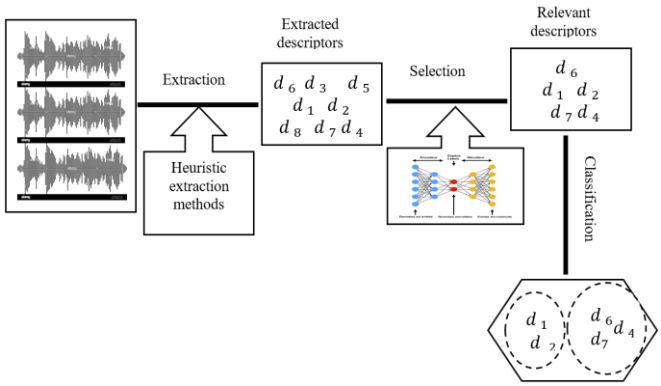

Figure 1. Process of extracting and selecting audio descriptors

3.4 Evaluation of descriptors selection

Supervised classification is a crucial step in evaluating the relevance of the selected descriptors. In the context of our study, several models will be experimented with.

3.4.1 Classification using SVM and OVR strategy

Support vector Machine (SVM) are particularly well-suited for tasks requiring cleat separation between multiple classes.

Since audio files are complex data, they are generally not linearly separable [17].

We will therefore use kernel SVM adapted to this type of data. Several kernel SVM models are described in the literature.

They solve optimization problems under the following constraint.

$$\begin{cases} \text{Minimise } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{under the contraints} \\ y_i(w.\phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \forall i \in \{1,2,\dots,n\} \end{cases} \quad (1)$$

The decision function is translated as follows:

$$f(x) = \sum_{i=1}^{n} \alpha_i\, y_i\, \mathrm{K}(x_i, x) + b \quad (2)$$

The dual problem is defined by:

$$\begin{cases} max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\, \mathrm{K}(x_i, x_j) \\ c \geq \alpha_i \geq 0, \qquad i = 1,\dots,n \\ \sum_{i=1}^{n}\alpha_i y_i = 0 \end{cases} \quad (3)$$

where:

- C: the regularization parameter that controls the trade-off between a large margin and minimizing errors;
- c: Regularization parameter;
- ξ_i: relaxation variable;
- ϕ: a kernel function that projects the data into a higher-dimensional space to handle nonlinear relationships;
- K: new kernel function using kernel trick;
- The α_i are the Lagrange multipliers from the dual problem; for better numerical stability, b is obtained from the average over the set I of vectors for which $0 < \alpha_i < c$.

Our evaluation was based on the commonly used kernels presented below:

- The Gaussian kernel: $\mathrm{K}(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$;
- The polynomial kernel: $\mathrm{K}(x_i, x_j) = (< x_i, x_j > + b)^d$;
- The linear kernel: $\mathrm{K}(x_i, x_j) = < x_i, x_j >$.

The initiative is due to the fact that for some studies [8, 17], the SVM model with a gaussian kernel is the most effective, while Jimenez et al. [5] argue that linear kernel SVM outperform other kernel-based models.

To solve our multi-class problem with SVM, we use the One-vs-Rest (OVR) strategy to transform it into several binary classification, as it is robust for complex and noisy data, such as spoken audio files.

It is mathematically formulated as follows:

$$y_i = \begin{cases} 1, si\ classe\ i \\ 0, \quad sinon \end{cases} \quad (4)$$

For a new observation, the class is determined as:

$$classe\ predite = argmax\ f_i(x) \quad (5)$$

where, f_i is the decision function for class i.

3.4.2 Ensemble learning model

We formalize the bagging technique as proposed by Gnagne et al. [7].

Let $Z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the initial sample where x_i is a code snippet, $y_i \in \{0, 1\}$ and $Z^b = \{(x_1^b, y_1^b), \ldots, (x_n^b, y_n^b)\}$ be the bootstrap samples of n observations with b=1, …, B.

The bagging ensemble model is defined as follows:

$$\hat{f}(Z) = VoteMajoritaire_{i=1}^{B}\{\hat{f}_b(Z^b)\} \quad (6)$$

where, $\hat{f}_b(Z^b)$ represents the predictions of the weak model with $\hat{f}_b$ trained using the bootstrap sample b.

We selected 100 Random Forests as the base estimator. This choice was derived from a comparative study of performance metrics (accuracy, precision, recall, and F1-score) for our ensemble model using different latent space dimensions (25, 50, 75, 100, 125, 150, 175, 200), as described in the table. We observed that the accuracy remained stable starting from 100 Random Forests. Table 6 shows performance measurement based on latent dimensions.

Table 6. Performance measurement based on latent dimensions

Latent Dimension	Accuracy	Precision	Recall	F1-score
25	0.61	0.62	0.63	0.62
50	0.62	0.62	0.64	0.63
75	0.62	0.63	0.64	0.63
100	0.63	0.64	0.65	0.63
125	0.63	0.63	0.64	0.63
150	0.63	0.64	0.65	0.64
200	0.63	0.63	0.65	0.63

3.4.3 Multilayer perceptron

The multilayer perceptron (MLP) is a type of artificial neural network based on fully connected (or dense) layers. The deep learning approach is favored over traditional machine learning methods, such as SVM, due to its superior performance in sound classification [18]. Although computational cost and data size are critical factors in choosing the classification method, particularly for neural networks that require large datasets and significant computational cost to achieve good performance, these two aspects are set aside in the context of our study. Indeed, our future research will focus on large-scale environments, where these challenges will be more effectively addressed.

The architecture can be represented in Figures 2 and 3.

In a simplified way, for any input data x, the perceptron assigns a weight vector W_1 to the neurons in the first layer, which will be activated by the activation function σ. The result of this activation will give an output $\hat{x}$, which will represent the input to the second layer. The second will undergo the same process with the weight vector W_2 and the activation φ to produce the output y. in this simplified case, the bias b_i is not represented.

The graphic representation above corresponds to the following model [19]:

$$\hat{y} = \varphi_{out}\left(\sum_i W_i^{(2)} h_i^{(2)} + b^{(2)}\right)$$

$$\forall i, h_i^{(1)} = \varphi\left(\sum_j W_{ij}^{(0)} x_j + b_i^{(0)}\right)$$

$$\forall i, h_i^{(2)} = \varphi\left(\sum_j W_{ij}^{(1)} h_j^{(1)} + b_i^{(1)}\right)$$

With φ an activation function and $W_{ij}^{(l)}$ and $b_i^{(l)}$ respectively the weight and bias of the *ith* neuron of layer (l).

Rigor in the selection and optimization of hyperparameters is essential for the deployment of a high-performance model. We therefore used the grid search technique to determine the optimal hyperparameters.

The hyperparameters used are defined in Table 7.

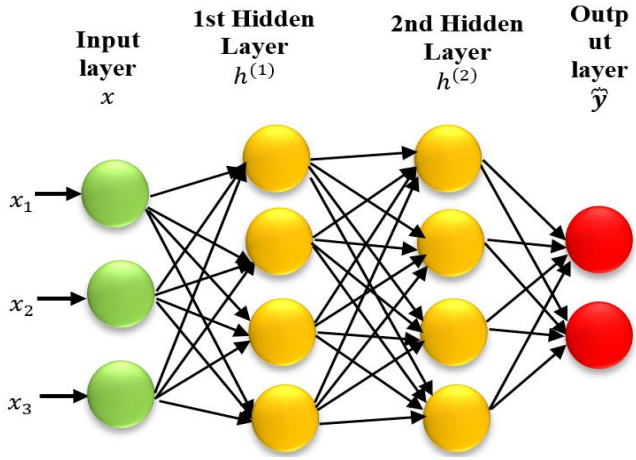

Figure 2. Structure of a two-layer perceptron

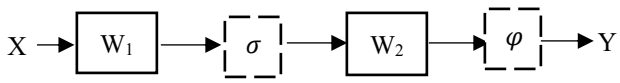

Figure 3. Simplified notation of the two-layer perceptron [7]

Table 7. The hyperparameters used

Hyperparameters	Values
Input layer	(100; 64)
Hidden layers	3 layers (64; 32) et (32; 16)
Output layer	(16; 10)
Learning rate	0.001
Activation function	Relu
Loss function	CrossEntropyLoss()
Optimization function	Adam()

4. RESULTS AND DISCUSSION

4.1 Results

Our approach, which we call KouaExtract, enabled the extraction of 119 descriptors compared to the 300 extracted by Jimenez et al. [5]. Our strategy involved designing heuristic algorithms that utilized all descriptor classes from the studies [4, 5], returning potentially relevant descriptors. To refine the obtained set of descriptors, we have, through automatic learning using an autoencoder, experimentally varied the dimension of the latent space.

Our goal was to identify a dimension that would

significantly reduce reconstruction error. The relevance of our selection approach is demonstrated by the convergence of the loss function curve, as shown in Figure 4.

It is worth nothing that the convergence of the loss function towards zero indicates a low reconstruction error, which reflets a good reproduction of the original data.

Our KouaExtract approach demonstrates improved audio file recognition with a performance of 87.5%, indicating effective extraction and optimal selection of descriptors. Table 8 provide a summary of these comparison.

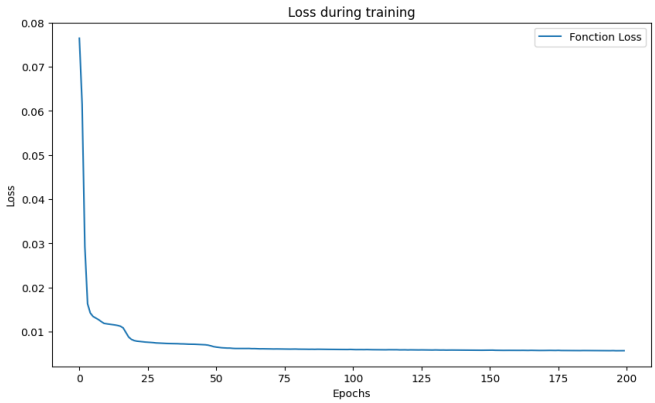

Figure 4. Loss function curve

Table 8. Performance comparison

Models	Number of Descriptors Extracted	Number of Descriptors Selected	Accuracy
Jimenez et al. [5]	300	57	75.3% (2.5%)
Kobayashi et al. [4]	239	0	81.3%
KouaExtract	119	100	**87.5%**

A visualization of the accuracies is illustrated in Figure 5.

To evaluate the performance of our novel method, Kouaextract, compared to existing methods, we conducted a rigorous statistical analysis. This approach is essential to distinguish significant results from random fluctuations inherent in the data. Our study focused on a set of 20 trials, a number exceeding the threshold of 10 recommended to ensure the reliability of statistical tests. As the data were paired and of multiclass nature, and not necessarily following a normal distribution, we opted for the non-parametric Wilcoxon signed-rank test.

The hypotheses of our test were as follows:

Null hypothesis (H0): There is no significant difference between the performances (accuracies) of the two models.

Alternative hypothesis (H1): There is a significant difference between the performances of the two models.

The Wilcoxon test revealed a p-value of 0.001, which is lower than the significance level α of 0.05. Therefore, we rejected the null hypothesis (H0). This conclusion supports the existence of a significant difference between the performances of Kouaextract and the Marvin model.

Our results confirm the superiority of Kouaextract in terms of performance. All of our calculations were performed using the wilcoxon function from the scipy.stats library in Python.

To perform a comparative study of the classification model accuracies based on the same dataset, we conducted two experiments on the "Features 3-sec.csv" dataset from GTZAN, which contains 10,000 audio files with the same number of descriptors as Jimenez et al. [5]. We then reduced the number of files to 1,000 as described by the authors. Tables 9 and 10 provide a summary of these experiments.

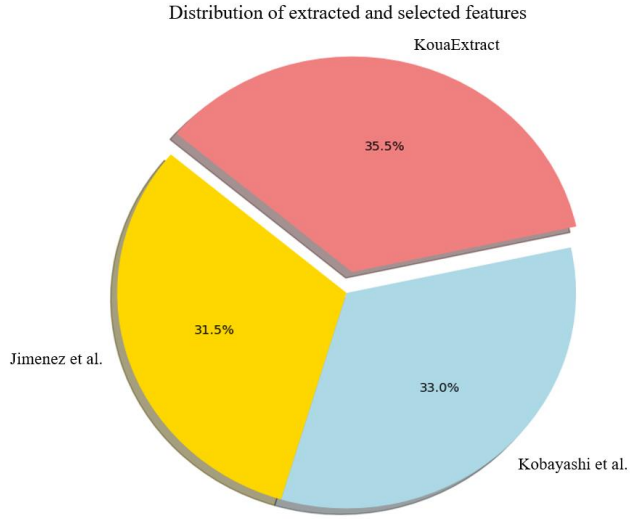

Figure 5. Distribution of extracted and selected descriptors

Table 9. Accuracy of models applied to 10,000 3-second audio files

Models	Accuracy
SVM with Gaussian kernel	84.8%
SVM with polynomial kernel	87.1%
SVM with linear kernel [5]	7.5% (75.3% [2.2])
Bagging with Random Forest	86.2%
Multilayer Perceptron	**92.8%**

Table 10. Accuracy of models applied to 1,000 3-second audio files

Models	Accuracy
SVM with Gaussian kernel	60%
SVM with polynomial kernel	60%
SVM with linear kernel [5]	55.5%
Bagging with Random Forest	63%
Multilayer Perceptron	**75%**

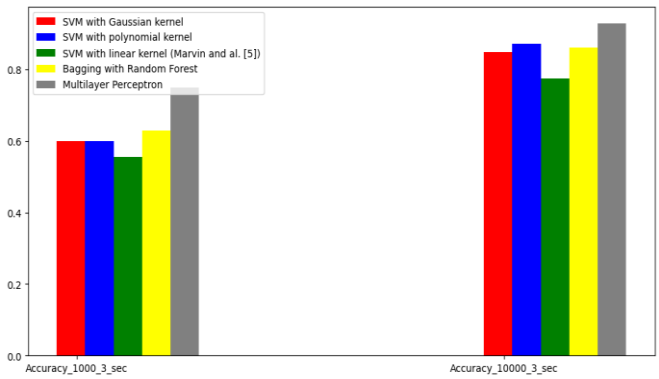

Figure 6. Distribution of model accuracies

Figure 6 illustrates the distribution of accuracies summarized in Tables 9 and 10.

This comparative study indicates three essential results. Firstly, it does not confirm the assertions in the literature

concerning the best SVM model with kernels, then it shows that neural networks are better classifiers compared to commonly used SVM models and finally shows the impact of data size in performance extraction and selection approaches. To validate the generalization capability of our approach, we evaluated our model on a separate database, namely FMA. We first selected 100 relevant descriptors from the 140 available and then conducted a second experiment with all 140 descriptors. Table 11 presents the results of these experiments.

Table 11. Generalization of Kouaextract

Dataset	Number of Descriptors Extracted	Number of Descriptors Selected	Accuracy
FMA_small	140	100	58%
FMA_small	140	0	45%
GTZAN	119	100	87.5%

4.2 Discussion

The performance of the KouaExtract model, compared to the studies [4, 5], highlights the importance of a robust descriptor selection approach. KouaExtract optimally reduces the dimensionality of the descriptor space while preserving essential information. This reduction significantly decreases the computational cost, and unlike a genetic algorithm, once the auto-encoder is trained, it adapts more effectively to complex data and converges quickly. The relevance of the selection indicates that simple descriptor extraction may contain non-significant information. This assertion is corroborated by the results obtained with the FMA database: 58% with 100 selected descriptors versus 45% with 140 extracted descriptors without a selection process. Research into relevant descriptor selection approaches should be further explored after the extraction phase.

We achieved an accuracy of 58% on the FMA dataset, demonstrating a certain degree of generalization of our approach, although it is lower than the model's performance on the training set, which reached 87.5%. However, it is important to note that the similarity in format between FMA and GTZAN poses a challenge to the generalization principle, potentially limiting the model's adaptability to new and diverse audio data.

The comparative study of SVM kernel models reveals that the polynomial SVM approach yields better results, contrary to previous studies. This suggests that the performance of an SVM model depends on factors such as hyperparameters. Therefore, in a given context, simulating different models is necessary to determine the most suitable one. However, the superior performance observed with neural networks, even when compared to ensemble learning methods, underscores their suitability for evaluation tasks.

We also observe that dataset size significantly impacts the performance of classification models. This reflects the principles of machine learning: the larger the training dataset, the better the model's generalization capabilities.

While deep learning methods are powerful and often superior for audio recognition, they require considerable resources in terms of computing power and large datasets for training. This can be problematic for real-time applications. However, the increasing diversity of files and the ever-growing size of databases, such as the Million Song Dataset (MSD) [20], require a compromise between computational cost and achieving optimal results. In the context of our study, the computation time does not exceed 70 seconds. It is also worth noting that the use of GPUs for processing large datasets provides a solution to temporal complexity. The GTZAN files were extracted from commercial music tracks. Although they are relatively clean in terms of sound quality, some background noise or minor imperfections may be present. However, they were not explicitly designed to include background noise or interferences. In certain genres, such as jazz, hip-hop, or metal, ambient sounds (like crowd noises or non-musical instruments) may occasionally be heard, but they are very subtle. Notably, GTZAN does not contain specific environmental background noises. Consequently, noise considerations were not included in this study. Future work will explore the use of noisy datasets to evaluate model robustness.

5. CONCLUSIONS AND PERSPECTIVES

Effective application of tasks to audio files requires an extraction and selection phase for their optimal descriptors. Existing approaches do not leverage the potential of machine learning models to select relevant features. This study proposed an innovative approach for the extraction and selection process of spoken audio files. We introduced heuristic descriptor extraction methods based on various descriptor classes from the literature. Our approach successfully identified potentially relevant descriptors. After the extraction phase, we reduced the dimensionality of our feature space while preserving the semantics of the audio file using an autoencoder.

We evaluated the effectiveness of our approach by training a multilayer perceptron on the selected dataset. The results demonstrated that neural networks generalized better on the test dataset compared to SVM and Random Forest models. Furthermore, we showed that the size of the descriptor set had little influence on the accuracy of the model used. These contributions confirm the relevance of the proposed approach for vocal data mining and open up new perspectives for applications in various contexts.

The KouaExtract method may not be well-suited for extremely noisy environments, as performance can vary significantly with high noise levels. Our future research will focus on developing preprocessing techniques to reduce noise before feature extraction. Additionally, we plan to adapt the model using audio data from diverse contexts (e.g., different accents or sound environments) to enhance generalization and robustness.

REFERENCES

[1] Hemmerling, D., Skalski, A., Gajda, J. (2016). Voice data mining for laryngeal pathology assessment. Computers in Biology and Medicine, 69: 270-276. https://doi.org/10.1016/j.compbiomed.2015.10.001

[2] Santiago, G., Jiménez, M., Aguilar, J., Montoya, E. (2021). Audio feature engineering for occupancy and activity estimation in smart buildings. Electronics, 10(21): 2599. https://doi.org/10.3390/electronics10212599

[3] Besbes, S., Lachiri, Z. (2016). Multi-class SVM for stressed speech recognition. In 2016 2nd international conference on advanced technologies for signal and

image processing (ATSIP), Monastir, Tunisia, pp. 782-787. https://doi.org/10.1109/ATSIP.2016.7523188

[4] Kobayashi, T., Kubota, A., Suzuki, Y. (2018). Audio feature extraction based on sub-band signal correlations for music genre classification. In 2018 IEEE international symposium on multimedia (ISM), Taichung, Taiwan, pp. 180-181. https://doi.org/10.1109/ISM.2018.00-15

[5] Jimenez, M., Aguilar, J., Monsalve-Pulido, J., Montoya, E. (2021). An automatic approach of audio feature engineering for the extraction, analysis and selection of descriptors. International Journal of Multimedia Information Retrieval, 10: 33-42. https://doi.org/10.1007/s13735-020-00202-1

[6] Abdoune, L., Fezari, M. (2017). Feature extraction for everyday life sounds. In 5th International Conference on Control & Signal Processing (CSP-2017), Proceeding of Engineering and Technology-PET, pp. 186-191.

[7] Gnagne, M.S., Dosso, M., Diarra, M., Oumtanaga, S. (2024). An approach to detect structural development defects in object-oriented programs. Open Journal of Applied Sciences, 14(2): 494-510. https://doi.org/10.4236/ojapps.2024.142036

[8] Besbes, S., Lachiri, Z. (2019). Stressed speech recognition using MMFCC and kernel-based classification. In 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, pp. 377-382. https://doi.org/10.1109/SSD.2019.8893273

[9] Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T., Gamboa, H. (2020). TSFEL: Time series feature extraction library. SoftwareX, 11: 100456. https://doi.org/10.1016/j.softx.2020.100456

[10] Alam, S., Sonbhadra, S.K., Agarwal, S., Nagabhushan, P. (2020). One-class support vector classifiers: A survey. Knowledge-Based Systems, 196: 105754. https://doi.org/10.1016/j.knosys.2020.105754

[11] Zaman, K., Sah, M., Direkoglu, C., Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access, 11: 106620-106649. https://doi.org/10.1109/ACCESS.2023.3318015

[12] Bernard, M., Poli, M., Karadayi, J., Dupoux, E. (2023). Shennong: A Python toolbox for audio speech features extraction. Behavior Research Methods, 55(8): 4489-4501. https://doi.org/10.3758/s13428-022-02029-6

[13] Hurbungs, V., Bassoo, V., Fowdur, T.P. (2024). A novel One-vs-Next approach for multiclass classification. In 2024 IEEE Symposium on Computers and Communications (ISCC), Paris, France, pp. 1-6. https://doi.org/10.1109/ISCC61673.2024.10733675

[14] Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X. (2016). FMA: A dataset for music analysis. arXiv preprint arXiv:1612.01840. https://doi.org/10.48550/arXiv.1612.01840

[15] RCP211-Artificial Intelligence Certificate-Cnam. Introduction aux modèles génératifs. Consulté le: 28 décembre 2024. https://cedric.cnam.fr/vertigo/cours/RCP211/introduction-modeles-generatifs.html#auto-encodeurs.

[16] McFee, B., Bertin-Mahieux, T., Ellis, D.P., Lanckriet, G.R. (2012). The million song dataset challenge. In Proceedings of the 21st International Conference on World Wide Web, Lyon France, pp. 909-916. https://doi.org/10.1145/2187980.2188222

[17] Besbes, S., Lachiri, Z. (2017). Classification of speech under stress based on cepstral features and one-class SVM. In 2017 International Conference on Control, Automation and Diagnosis (ICCAD), Hammamet, Tunisia, pp. 213-218. https://doi.org/10.1109/CADIAG.2017.8075659

[18] Shah, M., Pujara, N., Mangaroliya, K., Gohil, L., Vyas, T., Degadwala, S. (2022). Music genre classification using deep learning. In 2022 6th international conference on computing methodologies and communication (ICCMC), Erode, India, pp. 974-978. https://doi.org/10.1109/ICCMC53470.2022.9753953

[19] Abdoune, L., Fezari, M., Dib, A. (2024). Indoor sound classification with One-vs-Rest: State of the art and experimentation. International Journal of Computational Methods and Experimental Measurements, 12(3): 269-279. https://doi.org/10.18280/ijcmem.120307

[20] Tzanetakis, G., Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5): 293-302. https://doi.org/10.1109/TSA.2002.800560