Predictive Analytics for Pipe Organ Registration

Submitted by
Pablo Dumenil
pablo.dumenil@ircam.fr
From 20.03.2024 to 30.03.2024

Supervised by
Peter van Kranenburg
p.vankranenburg@uu.nl
Philippe Eisling
p.eisling@ircam.fr

22 août 2024

Table des matières

Abstract

The sound of the pipe organ has echoed through cathedrals, churches, and concert halls for centuries, captivating audiences with its rich timbres and harmonic complexity. Central to the organ's versatility and expressive range is its registration configuration, which determines the combination of stops and pipes engaged to produce a particular sound. However, decoding the registration of an organ piece remains a challenging task, especially in musical traditions where registration details are not always explicitly specified. In French organ music, registration is often indicated in the title of the organ pieces, while in German or Italian organ music, such details are typically omitted, creating confusion for modern performers.

To address this challenge, we created a unique dataset specifically for this study by collecting tracks from French Baroque composers using the OrganRoxx database and classifying them into registration classes through a heuristic approach with regular expressions. This dataset fills a significant gap, as no labeled dataset for organ music had previously been created.

To further our goal, we utilize machine learning techniques to analyze the subtle timbral changes between different organ music registrations. Indeed, after collecting the data, we extract meaningful audio embeddings using contrastive learning methods. Specifically, we employ Contrastive Language-Audio Pretraining (CLAP) [15] and Bootstrap Your Own Latent for Audio (BYOL-A) [10] to generate these embeddings. Dimensional reduction techniques, such as t-distributed Stochastic Neighbor Embedding (t-SNE), are then applied to visualize possible outliers and the relationship between audio embeddings and registration configurations. This visualization helps lay the groundwork for developing prediction models, such as multi-layer perceptrons, to capture the nuanced timbral features associated with specific registration settings in a low data context.

We aim to advance the analysis and classification of organ registrations, ultimately providing valuable tools for organists and scholars to decode and replicate authentic registration settings and preserve the rich legacy of organ music.

French - Le son de l'orgue a résonné à travers les cathédrales, les églises et les salles de concert pendant des siècles, captivant les auditoires avec ses timbres riches et sa complexité harmonique. Au cœur de la polyvalence et de la gamme expressive de l'orgue se trouve sa configuration de registration, qui détermine la combinaison de jeux et de tuyaux engagés pour produire un son particulier. Cependant, décoder la registration d'une pièce d'orgue reste une tâche complexe, surtout dans les traditions musicales où les détails de la registration ne sont pas toujours explicitement spécifiés. Dans la musique d'orgue française, la registration est souvent indiquée dans le titre des pièces d'orgue, tandis que dans la musique d'orgue allemande ou italienne, ces détails sont généralement omis, créant ainsi de la confusion pour les interprètes modernes.

Pour relever ce défi, nous avons créé un ensemble de données unique spécifiquement pour cette étude en collectant des morceaux de compositeurs baroques français à l'aide de la base de données OrganRoxx et en les classant en classes de registration par une approche heuristique utilisant des expressions régulières. Cet ensemble de données comble un vide significatif, aucun ensemble de données étiqueté pour la musique d'orgue n'ayant été préalablement créé.

Pour poursuivre notre objectif, nous utilisons des techniques d'apprentissage automatique pour analyser les subtils changements timbraux entre différentes registrations de musique d'orgue. En effet, après avoir collecté les données, nous extrayons des embeddings audio significatifs en utilisant des méthodes d'apprentissage contrastif. Plus précisément, nous employons le Contrastive Language-Audio Pretraining (CLAP) [15] et le Bootstrap Your Own Latent for Audio (BYOL-A) [10] pour générer ces embeddings. Des techniques de réduction de dimensionnalité, telles que l'Embedding Stochastique de Voisinage (t-SNE), sont ensuite appliquées pour visualiser les éventuels points atypiques et la relation entre les embeddings audio et les configurations de registration. Cette visualisation aide à poser les bases pour le développement de modèles de prédiction, tels que les perceptrons multicouches, afin de capturer les caractéristiques timbrales nuancées associées à des réglages de registration spécifiques dans un contexte de données limité.

Nous visons à faire progresser l'analyse et la classification des registrations d'orgue, fournissant ainsi des outils précieux pour les organistes et les chercheurs afin de décoder et de reproduire des réglages de registration authentiques et de préserver le riche héritage de la musique d'orgue.

Acknoledgement

My deepest gratitude goes to Peter van Kranenburg, whose rigor and perseverance guided me towards the realization of this project. As my supervisor, he provided invaluable guidance throughout the work and introduced me to the beautiful world of organ music, sharing his passion for this microcosm with me. This project has been a truly enjoyable journey, and working with Peter has been an enriching experience.

I would also like to extend my thanks to David Genova for providing me with valuable technical tips that significantly contributed to the success of this work.

Finally, I am sincerely grateful to Philippe Esling, with whom I share so many values in music and beyond, for supervising this internship from both Paris and Tokyo. His passion and knowledge in applying neural networks creatively in music have been both inspiring and educational. Thank you, Philippe, for being an inspiration in many ways.

1 Introduction

1.1 Overview of the Pipe Organ

1.1.1 Historical and Cultural Significance

The pipe organ is one of the oldest and most complex musical instruments, with a rich history dating back over two millennia. During the Renaissance and Baroque eras, the organ saw significant advancements in design and complexity, with notable composers like Johann Sebastian Bach elevating the instrument's status through their works. It became a symbol of religious and civic pride, often housed in churches, cathedrals, and concert halls.

1.1.2 Role of registration in organ timbre

Registration in the pipe organ refers to the selection and combination of different sets of pipes to produce various timbres and dynamic levels. There are stops available near the performer that allow to select the pipes where the air will flow through. Each stop control a specific rank of pipes that corresponds to a particular sound quality, ranging from flutes and strings to reeds and diapasons.

By skillfully choosing and blending stops, the organist can imitate the sounds of an orchestra, produce unique tonal effects, and adapt the instrument's sound to suit different musical styles and acoustical environments.

Registration is crucial for expressive playing, as it allows the organist to shape the music's emotional and dynamic contours. Mastery of registration techniques is essential for interpreting organ repertoire authentically and creatively, making it a key aspect of organ performance and artistry.

FIGURE 1 – Illustration of the pipe organ's keyboards and the stops (Peter van Kranenburg performing).

1.2 Principal Registrations

The following are key principal registrations commonly used in French Baroque organ music, each offering distinctive timbral qualities :

— **Plein Jeu**
  — Used for loud, bright music with a powerful, harmonically rich sound. It combines principal stops with mixtures to create a full, resonant tone. Suitable for pieces that require a commanding presence.
  — **Example Piece :** Last verse from Kyrie of Couperin's *Messe des Paroisses*.
— **Grand Jeu**
  — A robust, reed-heavy registration ideal for dramatic passages. It features reeds and Cornets along with foundation stops, producing a bold and assertive sound. Often used for climactic moments in the music.
  — **Example Pieces :** *Grand Dialogue en ut* by Louis Marchand ; *Caprice sur les grands jeux* by Clérambault.
— **Tierce en Taille**
  — Ideal for delicate and expressive music focusing on the tenor voice. This registration prominently features the Tierce stop combined with flutes and foundation stops to create a soft and sweet tone.
  — **Example Piece :** *Tierce en Taille* from Couperin's *Messe des Couvents*.
— **Cromorne en Taille**
  — Focuses on the Cromorne reed stop, providing a nasal, reed-like sound suitable for solo melodies. This registration is effective in showcasing the unique timbre of the Cromorne in the tenor register.
  — **Example Piece :** *Cromorne en Taille* by Guilain.
— **Récit de Cromorne**
  — Focuses on the Cromorne reed stop, providing a nasal, reed-like sound suitable for solo melodies. This registration is effective in showcasing the unique timbre of the Cromorne in the soprano register.
  — **Example Piece :** *Récit de Cromorne* by Nicoleas De Grigny.
— **Basse de Cromorne**
  — A bass registration with a warm, mellow sound, often used for a rich, expressive tone usin the Cromorne.
  — **Example Piece :** *Suite 4 Ton : Deposuit potentes* by Guilain.
— **Basse de Trompette**
  — Centered on the Trompette stop, this registration is bright and resonant, often used for strong bass lines or prominent melodic passages. It combines the Trompette with various foundation stops for flexibility in intensity. Often within one piece combined with Trompette in the dessus (soprano).
  — **Example Piece :** *Basse et dessus de Trompette* from Clérambault's *Livre d'orgue*.
— **Flûtes**
  — Various flute stops that offer a range of soft, sweet tones, useful for melodic lines and gentle accompaniment.
  — **Example Piece :** *Suite du 2ème ton pour flûtes* by Clérambault.
— **Duo (sur les Tierces)**
  — Used for duets, this registration prominently features the Tierce stop in both hands, creating a rich, shimmering sound ideal for inter-voice dialogue. It allows for a clear distinction between the two hands. This does not refer to timbre, but to a musical form. Duo is almost always registered with Grand Jeu de Tierce in the bass. Upper voice is mostly cornet (separé) or a reed (cromorne).
  — **Example Piece :** *Duo* from Clérambault.
— **Fond d'orgue**
  — A foundational registration for softer, sustained passages. It combines Bourdons, Montres, and occasionally soft reeds to create a warm, gentle sound suitable for quieter, more introspective moments.
  — **Example Piece :** *Fond d'orgue* by Louis Marchand.

— **Récit de Cornet**
  — Featuring the Cornet stop, this registration is ideal for solo melodies, offering a bright and clear tone. It contrasts with softer accompaniments to highlight the solo line effectively.
  — **Example Piece :** *Récit de Cornet* by Couperin.
— **Recit de Nazard**
  — A registration providing a distinctive, slightly sharp timbre, used to add color and texture.
  — **Example Piece :***Suite du deuxième ton - récit de nazard* by Clérambault
— **Voix Humaine**
  — Mimics the human voice with a unique, vocal timbre, often used for expressive and lyrical passages.
  — **Example Piece :** *Récit de voix humaines* by Nivers

1.3 Challenges in Organ Registration

The task of predicting organ registration is particularly challenging due to the lack of explicit registration instructions in many historical organ compositions. In the case of Italian and German organ music, registration details are often absent from the music sheets, leaving the choice of stops to the performer. This omission requires organists to possess a deep understanding of the instrument's capabilities and the stylistic practices of the period in which the piece was composed. Consequently, registration can vary significantly between performances, offering a high degree of interpretative freedom and personal expression. Our classification task aims to address this variability by developing models that predict organ registration based on limited data, with the goal of enhancing both the performance and preservation of these historical works. The classification task of our study contributes to enhancing the performance and preservation of historical organ works by standardizing interpretative freedom, guiding informed performance, and creating digital records of registration practices. By predicting organ registration based on limited data, these models provide performers with historically informed suggestions that balance personal expression with authenticity. This not only aids in preserving traditional knowledge and enabling accurate recreations of historical performances but also supports creative innovation by allowing performers to use predictions as a starting point for new interpretations. Ultimately, these models ensure that historical organ music is performed and preserved in a way that honors its original intent while embracing modern insights and creativity.

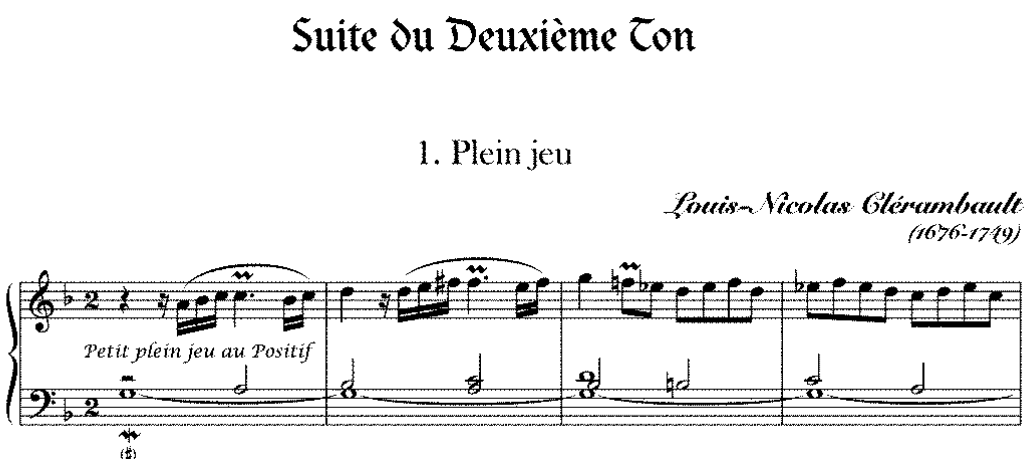


FIGURE 2 – French Organ music sheet with registration indication.

1.3.1 Registrations with Similar Timbre

Several registrations exhibit similar timbral characteristics, which may contribute to misclassification. For example, **Grand Jeu** and **Plein Jeu** are both full stops or registrations on the organ, producing similar tonal qualities with subtle differences in the blend of stops and overtones. **Basse de Trompette** and **Basse de Cromorne** are bass registrations with distinct yet subtle differences : Basse de Trompette

is brighter and more piercing, while Basse de Cromorne is warmer and mellower. Similarly, **Recit de Cromorne** and **Voix Humaine** are used for expressive purposes, with Recit de Cromorne being more mellow and soft compared to the more vocal quality of Voix Humaine. **Fond d'Orgue** and **Tierce En Taille** both provide a rich, full sound but differ in texture; Fond d'Orgue is more subdued, while Tierce En Taille has a brighter timbre. Lastly, **Cromorne En Taille** and **Tierce En Taille** are used in similar registers but have differing tonal characteristics, with Cromorne En Taille being softer and more reedy, and Tierce En Taille being sharper and more focused.

1.4 Technical Approach and Objectives

The complexity of organ registration, coupled with the interpretative freedom afforded to organists, presents a unique challenge in both performance and musicological research. To address these challenges, this study aims to apply machine learning techniques to analyze and classify organ registrations based on their timbral characteristics in organ music. A crucial component of this research is the creation and preprocessing of a novel dataset representing various organ registrations. This was a significant undertaking given the limited availability of labeled data in organ music, and it provides a valuable resource for the machine learning community interested in the rich and diverse universe of organ music.

A critical first step in our approach involves the detailed analysis of timbre features associated with different organ registrations, including harmonic content, envelope, and spectral characteristics, to quantify the unique sound qualities of each registration. One of the initial goals of this project was to determine whether organ registrations could be systematically classified using machine learning techniques. This foundational objective entailed evaluating whether the distinct timbral features of various organ registrations could be reliably distinguished from one another. The success of this initial phase was crucial in establishing the feasibility of further, more detailed analysis. However, our dataset faces the challenge of class imbalance, as most of the recordings are concentrated in a few registration classes such as Grand Jeu, Plein Jeu, Tierce En Taille, and Basse de Cromorne. This imbalance might necessitates the use of specialized techniques to address the overrepresentation of certain classes and the underrepresentation of others.

To support this work, we have integrated data from OrganRoxx Radio, a platform that specializes in broadcasting organ music from around the world. Their extensive collection has been invaluable for our study, providing a wide range of organ music recordings, which have been crucial in expanding our dataset. We would like to express our gratitude to OrganRoxx for sharing their data, which has significantly enriched our research.

By addressing these technical challenges, we aim to develop robust machine learning models capable of effectively analyzing and classifying organ registrations, ultimately enhancing our understanding of organ music and supporting the performance and study of this intricate art form.

1.4.1 Broader Applications

Beyond its direct applications in organ music analysis, our work has the potential to assist in the management of large audio collections of organ music, such as those curated by OrganRoxx. Additionally, the model could be used to support playlist generation, a feature that could enhance the listening experience for audiences on streaming platforms. We believe that our research not only contributes to the preservation and interpretation of organ music but also opens new avenues for the application of machine learning in the broader field of musicology.

2 State-of-the-art

2.1 Problem of Machine Learning and Basic Definitions

Machine learning aims to build models that generalize well from training data to unseen data. The generalization capability of a model is its ability to perform well on new inputs after being trained on a finite dataset. A fundamental challenge in machine learning is the trade-off between bias and variance, encapsulated in the bias-variance trade-off theory [2]. A high-bias model is too simple and may underfit the data, while a high-variance model is too complex and may overfit, capturing noise instead of the underlying data distribution.

Formally, given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the input features and $y_i \in \mathbb{R}$ are the corresponding outputs, a machine learning model seeks to approximate a function $f : \mathbb{R}^d \to \mathbb{R}$ that minimizes the expected loss :

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(f(x), y)],$$

where $L(\cdot, \cdot)$ is a loss function such as Mean Squared Error (MSE) for regression or cross-entropy for classification.

2.2 Optimization and Gradient Descent

Optimization is the process of finding the best parameters for a model that minimize the loss function. In machine learning, most models are trained using gradient-based optimization techniques. The most widely used method is *gradient descent*, which iteratively updates the model's parameters in the opposite direction of the gradient of the loss function with respect to the parameters [12].

Mathematically, given a model with parameters θ, the update rule for gradient descent is :

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta L(\theta),$$

where η is the learning rate, and $\nabla_\theta L(\theta)$ is the gradient of the loss function with respect to the parameters.

There are several variants of gradient descent, including :

— *Stochastic Gradient Descent (SGD)* : Updates the parameters using a single or a few training examples, leading to faster convergence but noisier updates.
— *Mini-Batch Gradient Descent* : A compromise between batch gradient descent and SGD, where updates are made based on small batches of data.
— *Momentum* : Accelerates convergence by adding a fraction of the previous update to the current one.
— *Adam (Adaptive Moment Estimation)* : Combines the benefits of both momentum and RMSProp, using adaptive learning rates and momentum to improve convergence [5].

2.3 Neural Networks : Neurons and Layers

Neural networks are the backbone of modern machine learning, especially deep learning. A neural network consists of layers of interconnected units called neurons, which mimic the behavior of biological neurons. The architecture of a neural network includes :

— *Input Layer* : The layer that receives the input data.
— *Hidden Layers* : Intermediate layers that perform computations, transforming the input into a form the output layer can use.
— *Output Layer* : The final layer that produces the prediction.

Mathematically, a neuron in a layer l computes :

$$a_j^{(l)} = \sigma\left(\sum_i w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right),$$

where $w_{ji}^{(l)}$ is the weight connecting neuron i in layer $l-1$ to neuron j in layer l, $b_j^{(l)}$ is the bias term, and $\sigma(\cdot)$ is an activation function such as the ReLU (Rectified Linear Unit) or sigmoid function.

2.4 Backpropagation

Backpropagation is the algorithm used to train neural networks by computing gradients of the loss function with respect to each weight through the chain rule of calculus [12]. This method efficiently propagates the error backwards through the network, enabling the optimization algorithm to update the weights.

Formally, the gradient of the loss function L with respect to a weight $w_{ji}^{(l)}$ in a network is calculated as :

$$\frac{\partial L}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)},$$

where $\delta_j^{(l)}$ is the error signal for neuron j in layer l, propagated from the subsequent layer.

The error signal $\delta_j^{(l)}$ is computed as :

$$\delta_j^{(l)} = \left(\sum_k \delta_k^{(l+1)} w_{kj}^{(l+1)} \right) \sigma'(z_j^{(l)}),$$

where $z_j^{(l)}$ is the input to the activation function at neuron j in layer l, and $\sigma'(\cdot)$ is the derivative of the activation function.

Backpropagation allows for efficient computation of these gradients, making it feasible to train deep networks with many layers.

Recent advancements in Music Information Retrieval (MIR) have significantly shaped the field, particularly in musical instrument recognition, timbre analysis, and contrastive learning techniques. In this context, musical instrument recognition within MIR has seen substantial progress with the adoption of deep learning methodologies. State-of-the-art models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures, have demonstrated impressive performance in classifying instruments from audio recordings. These models benefit from large-scale datasets and advanced preprocessing techniques, such as Mel-spectrograms and wavelet transforms, which enhance their ability to learn discriminative features [8, 3]. However, specific challenges persist, particularly in the context of organ music, where the intricate and overlapping harmonic structures can complicate the recognition process. Research has focused on developing specialized architectures and training paradigms to address these issues, leveraging techniques like data augmentation and transfer learning to improve model robustness and generalization [1, 6].

In the domain of timbre recognition, it plays a vital role in musical analysis by identifying the unique characteristics of sounds that distinguish different instruments or voices. Techniques for timbre recognition have evolved significantly with the advent of machine learning and deep learning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been extensively used to model the spectral and temporal features of audio signals, respectively. CNNs typically operate on spectrograms, transforming audio into a 2D representation where temporal and frequency features are captured. However, CNNs can struggle with long-term dependencies in audio due to their limited receptive fields and reliance on fixed window sizes. RNNs, on the other hand, are designed to handle sequential data and capture temporal dependencies, but they can suffer from issues related to vanishing and exploding gradients, making them less effective for long-term time dependencies [11, 7].

2.5 Regularization Techniques

In the quest to improve the generalization capabilities of neural networks and mitigate overfitting, several regularization techniques have been developed. Two of the most impactful are *dropout* and *batch normalization*.

2.5.1 Dropout

Dropout is a technique designed to prevent overfitting in neural networks by randomly "dropping out" units (neurons) during the training phase. This means that, at each training iteration, a certain percentage of neurons are ignored, i.e., their contribution to the activation of subsequent layers is temporarily set to

zero [14]. The remaining neurons must compensate by adjusting their weights, which helps the network to learn more robust and generalized features. Dropout is a regularization technique used to prevent overfitting by randomly deactivating a fraction of neurons during training. The following figure illustrates the concept of dropout in a neural network layer.

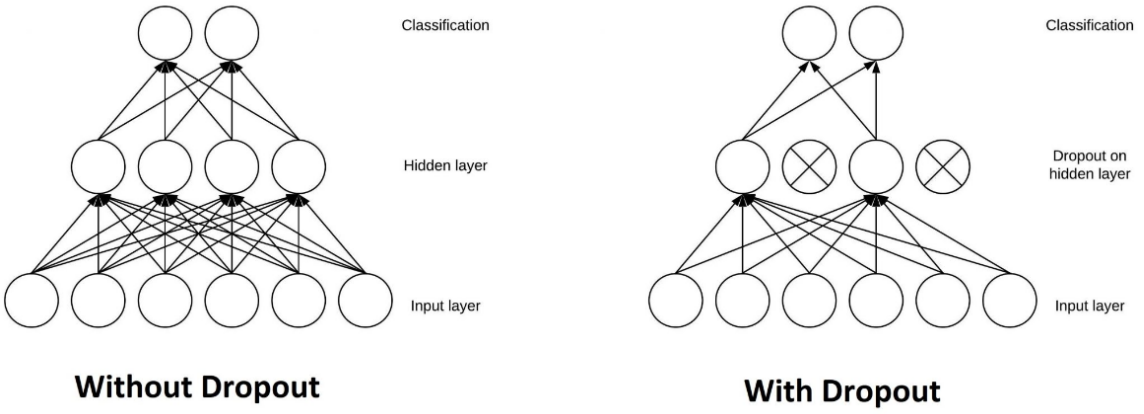


FIGURE 3 – Illustration of dropout in a neural network.

Formally, for a layer with activations $a^{(l)}$, dropout creates a mask $m^{(l)}$ of the same size, where each element $m_i^{(l)}$ is a Bernoulli random variable with a probability p of being 1 (and $1-p$ of being 0). The activations during training are then :

$$\tilde{a}^{(l)} = m^{(l)} \odot a^{(l)},$$

where $\odot$ denotes the element-wise multiplication. During inference, to maintain the same expected output, the activations are scaled by the dropout probability p.

Dropout forces the network to not rely too heavily on any single neuron, encouraging the emergence of redundant representations and thus reducing the likelihood of overfitting.

2.5.2 Batch Normalization

Batch normalization addresses the issue of internal covariate shift, which refers to the changes in the distribution of network activations during training, causing training to slow down or even diverge. Proposed by Ioffe and Szegedy [4], batch normalization normalizes the input to each layer so that it has a mean of zero and a standard deviation of one. This normalization is performed for each mini-batch, hence the name.

The process can be described mathematically as follows : given a mini-batch of activations $\{a_i^{(l)}\}_{i=1}^m$ from layer l, batch normalization computes :

$$\mu^{(l)} = \frac{1}{m}\sum_{i=1}^{m} a_i^{(l)}, \quad \sigma^{(l)2} = \frac{1}{m}\sum_{i=1}^{m}\left(a_i^{(l)} - \mu^{(l)}\right)^2,$$

$$\hat{a}_i^{(l)} = \frac{a_i^{(l)} - \mu^{(l)}}{\sqrt{\sigma^{(l)2} + \epsilon}},$$

where $\mu^{(l)}$ and $\sigma^{(l)}$ are the mean and variance of the activations in the mini-batch, and ϵ is a small constant added for numerical stability.

To retain the expressiveness of the network, batch normalization also introduces two learnable parameters, $\gamma^{(l)}$ and $\beta^{(l)}$, which scale and shift the normalized output :

$$a_i^{(l)\text{normalized}} = \gamma^{(l)}\hat{a}_i^{(l)} + \beta^{(l)}.$$

Batch normalization has several benefits, including improved training speed, reduced sensitivity to hyperparameters such as learning rates, and often acting as a regularizer, thereby reducing the need for other regularization techniques like dropout.

These techniques, dropout and batch normalization, have become standard tools in the training of deep neural networks, contributing significantly to the success of modern machine learning models in various tasks.

2.6 Embedding Extraction

Embeddings extraction plays a crucial role in machine learning, especially in tasks involving complex data types like audio. By transforming raw data into a structured form that captures essential patterns and features, embeddings serve as the foundation for model learning. In this section, we delve into various contrastive learning techniques, emphasizing their role in generating effective embeddings. Specifically, we focus on two advanced frameworks : CLAP [15] and BYOL-A [10]. Both methods are employed to extract meaningful embeddings from audio data, enabling models to learn and generalize from these representations effectively. This process of embedding extraction is essential for leveraging the full potential of the data in downstream tasks.

2.6.1 CLAP (Contrastive Language-Audio Pretraining)

CLAP builds on contrastive learning by focusing on aligning audio and language embeddings in a shared space. This is particularly valuable for multi-modal tasks where understanding the relationship between audio and text is crucial. CLAP utilizes a contrastive loss that brings corresponding audio and text embeddings closer together while pushing apart non-corresponding pairs [15].

a. Audio Input Preprocessing

Before feeding audio into the CLAP framework, the signal undergoes preprocessing :

— Resampling and Conversion : The audio signal is resampled to 48 kHz and converted to mono to ensure consistency.
— Mel-Spectrogram Transformation : The resampled audio is transformed into a Mel-spectrogram using the Short-Time Fourier Transform (STFT) and mapping the frequencies onto the Mel scale, providing a more perceptually relevant representation.

b. Audio Encoder

The Mel-spectrogram is then processed by the audio encoder, which extracts a high-dimensional feature vector :

— Encoder Architectures :
  — PANN (Pretrained Audio Neural Network) : Utilizes convolutional neural networks (CNNs) to capture local patterns in the spectrogram.
  — HTSAT (Hierarchical Transformer with Self-Attention and Transformer layers) : Employs Transformer layers to capture global contextual information and hierarchical features.
— Feature Extraction : The encoder outputs a high-dimensional feature vector :

$$f_{\text{audio}}(X_a)$$

where X_a is the input audio signal.

c. Projection Layer

The feature vector $f_{\text{audio}}(X_a)$ is projected into a shared embedding space using a Multi-Layer Perceptron (MLP) :

— MLP (Multi-Layer Perceptron) : The 2-layer MLP with ReLU activation projects the feature vector to an embedding :

$$E_a = \text{MLP}_{\text{audio}}(f_{\text{audio}}(X_a))$$

Downstream Applications : The embedding E_a is applicable to various tasks, such as retrieval, classification, or multi-modal learning.

CLAP has significant implications for multimodal tasks, such as audio-visual scene understanding and cross-modal retrieval. By learning a shared embedding space, CLAP can facilitate better integration of different types of data, enhancing performance on tasks that require understanding of both text and audio. It is designed to capture meaningful relationships between different segments within audio samples by leveraging contrastive losses.The approach used in CLAP trains neural networks to distinguish between positive pairs (segments from the same audio) and negative pairs (segments from different audios), effectively encoding complex temporal dependencies and preserving subtle audio features. Unlike traditional approaches where RNNs and CNNs process raw inputs such as Mel-spectrograms or waveform data, CLAP generates high-dimensional embeddings, such as [1,512] vectors. These embeddings offer a compact and rich representation of audio data, capturing detailed temporal and spectral information more effectively than the raw inputs typically used by CNNs and RNNs. This compact representation facilitates the training of state-of-the-art classifiers, enabling them to handle complex audio data with greater precision and interpretability [15, 10].

2.6.2 BYOL (Bootstrap Your Own Latent)

Bootstrap Your Own Latent (BYOL) introduces a novel approach to self-supervised learning by eliminating the need for negative samples, which are commonly used in contrastive learning to differentiate between positive and negative pairs of data. Instead, BYOL focuses on learning high-quality representations by comparing different augmented views of the same input, using two distinct networks : an *online network* and a *target network*.

The online network, denoted as $\text{Enc}_{\text{online}}$, consists of an encoder, a projector, and a predictor. The target network, $\text{Enc}_{\text{target}}$, mirrors the architecture of the online network but does not include the predictor. The key difference between these networks lies in the updating mechanism : while the online network's parameters are updated via gradient descent, the target network's parameters are updated as an exponential moving average of the online network's parameters. This slow update strategy for the target network ensures stable targets for learning and prevents the networks from collapsing into trivial solutions, such as outputting the same representation for all inputs.

BYOL's learning process begins by applying two different sets of augmentations to an input sample x_i, resulting in two distinct views v_i and v_i'. The online network processes v_i, producing an embedding $z_{\text{online}} = \text{Enc}_{\text{online}}(v_i)$, while the target network processes v_i', yielding an embedding $z_{\text{target}} = \text{Enc}_{\text{target}}(v_i')$. The online network's predictor then attempts to predict the target network's embedding, resulting in a predicted embedding $q_{\text{online}}(z_{\text{online}})$.

The BYOL loss function, which drives the learning process, is defined as the mean squared error (MSE) between the predicted online network's embedding and the target network's embedding :

$$L_{\text{BYOL}} = \frac{1}{N} \sum_{i=1}^{N} \left\| q_{\text{online}}(z_{\text{online}}(x_i)) - z_{\text{target}}(x_i) \right\|^2 ,$$

where N is the number of samples in the batch. This loss measures how well the online network's predicted embeddings align with those from the target network, ensuring that the online network progressively learns to generate embeddings that accurately capture the essence of the augmented views.

One of the key advantages of BYOL is that it avoids the complexities and potential pitfalls associated with negative pairs, such as the need for large batch sizes or sophisticated sampling strategies. By focusing solely on positive pairs (i.e., different views of the same input), BYOL simplifies the training process and improves robustness. This method has demonstrated state-of-the-art performance in image representation learning, achieving top accuracy benchmarks without relying on negative pairs, thereby proving the effectiveness of its self-supervised learning approach.

The online and target networks work in tandem to iteratively refine the embeddings, bootstrapping each other to higher levels of representational quality. By the end of training, only the online network's encoder is retained, and its learned embeddings are used for downstream tasks such as classification, clustering, or transfer learning.

Researchers have explored CLAP's application across various domains within MIR, showcasing its efficacy in tasks ranging from music genre classification to content-based music retrieval. By generating embeddings that encapsulate the nuanced features of audio signals, CLAP enhances the robustness and accuracy of MIR systems. Moreover, advancements in CLAP have paved the way for innovative approaches in deep learning architectures, facilitating the development of more sophisticated models capable of handling complex audio data. While BYOL-A (Bootstrap Your Own Latent for Audio) [10] has also shown promise in self-supervised audio representation learning, CLAP's focus on contrastive language-audio pretraining provides a more nuanced approach to capturing temporal dependencies, making it particularly suitable for complex MIR tasks.

Organ music presents a unique challenge in timbre recognition due to the intricate and overlapping harmonic structures that characterize this instrument, as well as the subtle timbre variations introduced by changing registrations. The organ's complex sound production mechanisms, involving multiple ranks of pipes and a wide range of stops, produce a rich and varied timbral palette. This complexity can make it difficult to distinguish between subtle variations in timbre, particularly when different stops are used simultaneously.

While significant strides have been made in MIR, the complexity of organ music, with its subtle timbral variations and intricate harmonic structures, make it hard to classify it. Advanced techniques such as CLAP and BYOL-A offer promising avenues for improvement, but ongoing research and innovation are essential to fully capture the richness of organ timbre and enhance the accuracy of musical instrument recognition systems.

2.7 Data visualization

In the analysis of complex datasets, visualizing high-dimensional data is crucial for understanding underlying patterns, clusters, and relationships between data points. However, directly interpreting high-dimensional data can be challenging. To address this, dimensionality reduction techniques such as T-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) are employed. These techniques project high-dimensional data into a lower-dimensional space, making it easier to visualize and interpret. This section explores the principles and applications of t-SNE and PCA in the context of data visualization.

While PCA focuses on capturing the global structure of the data by maximizing variance, t-SNE is more effective for visualizing complex, non-linear relationships by preserving local pairwise similarities. PCA can serve as a useful preprocessing step before applying t-SNE, reducing the dimensionality to a manageable level and removing noise, which helps improve the efficiency and effectiveness of the t-SNE algorithm. See in the annex for more details on mathematical formulation of PCA and t-SNE.

2.7.1 Clustering for isolating outtliers

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [13] is a popular density-based clustering algorithm that groups together closely packed points and identifies points that lie alone in low-density regions as outliers. It does not require the number of clusters to be specified in advance and can find clusters of arbitrary shapes. The two main parameters of DBSCAN are :

— **Epsilon (ϵ)** : The maximum distance between two points to be considered as neighbors.
— **MinPts** : The minimum number of points required to form a dense region (a cluster).

- Mathematical Formulation

1. Epsilon Neighborhood

For a point p in the dataset, the ϵ-neighborhood of p is defined as :

$$N_\epsilon(p) = \{q \in \mathbb{R}^d \mid \|p - q\| \leq \epsilon\}, \tag{1}$$

where $\|p - q\|$ denotes the Euclidean distance between points p and q.

2. Core Point Definition

A point p is considered a core point if :

$$|N_\epsilon(p)| \geq \text{MinPts}, \tag{2}$$

where $|N_\epsilon(p)|$ is the number of points within the ϵ-neighborhood of p.

3. Cluster Formation

- Initialization : Start with an arbitrary point. If it is a core point, create a new cluster. - Expansion : Add all points within its ϵ-neighborhood to the cluster. Recursively add the ϵ-neighborhood points of the added points. - Termination : Repeat the process until all points are processed. Points that cannot be reached from any core points are classified as noise.

4. Density Reachability

— **Directly Reachable** : A point p is directly reachable from a core point o if p is in the ϵ-neighborhood of o and o is a core point.

— **Density Reachable** : A point p is density reachable from a core point o if there is a chain of core points $o_1, o_2, \ldots, o_n$ such that p is directly reachable from o_1, o_1 is directly reachable from o_2, and so on.

When working with MIR datasets, where the data can be highly complex and noise-prone, clustering algorithms like DBSCAN are essential for identifying and isolating outliers. DBSCAN's ability to form clusters based on the density of data points rather than predefined shapes makes it particularly effective in MIR tasks, where the structure of the data may be non-linear or irregular. This capability is crucial for ensuring that outlier data points, which might represent noise or anomalies, do not negatively impact the performance of machine learning models.

3 Methodology

Organ registration, the process of selecting and combining stops on an organ, plays a crucial role in shaping the timbre and overall character of the music. These subtle timbre variations are complex to differentiate, even for the human ear. Our methodology is designed to tackle this problem using machine learning techniques to process audio recordings, with a particular focus on comparing the effectiveness of embeddings generated by two contrastive learning models : CLAP and BYOL-A.

We begin by curating a dataset of organ music from notable French Baroque composers, ensuring accurate labeling and preprocessing. Next, we extract meaningful features from the audio data using both the CLAP and BYOL-A models. These models produce high-dimensional embeddings that encapsulate the essential characteristics of each piece. By employing both models, we aim to evaluate which approach yields embeddings that better support the classification of organ registration.

To visualize and further process these embeddings, we apply t-SNE for dimensionality reduction and DBSCAN for clustering, which help in identifying and isolating outliers that may represent misclassifications or anomalies. Following this, we develop and train a custom multi-layer perceptron (MLP) model, optimized for this task, using the embeddings from both CLAP and BYOL-A.

The experiment then involves comparing the performance of the MLP model when trained and fine-tuned with embeddings from CLAP versus those from BYOL-A. This comparative analysis allows us to assess which contrastive learning model provides more effective representations for the classification of organ registration. Our methodology thus offers a comprehensive framework that combines state-of-the-art feature extraction, dimensionality reduction, and machine learning techniques, while also rigorously evaluating the performance of different embedding strategies in this unique domain.

3.1 Dataset

In our pursuit to classify organ registrations within French Baroque music, we meticulously curated a dataset that encapsulates the richness and diversity of this genre. The initial phase involved identifying prominent composers from the French organ school, for which we consulted reputable sources, notably the comprehensive list available on Wikipedia [1]. This list served as a foundational reference, ensuring that our selection was both authoritative and representative of the period.

Subsequently, we turned to the OrganRoxx database, a vast repository renowned for its extensive collection of organ music recordings. Leveraging this resource, we extracted all tracks attributed to the previously identified French Baroque composers. This extraction provided a substantial corpus of organ works, each potentially reflecting distinct registration practices characteristic of the era.

To categorize these tracks based on their registration, we adopted a heuristic approach centered around the utilization of regular expressions. For each predefined registration class, we crafted a set of regular expressions designed to detect specific patterns within track titles. When a track's title matched one of these patterns, it was accordingly assigned to the pertinent registration class. This method, while systematic and efficient, inherently relied on the accuracy and descriptiveness of the track titles.

Post-classification, we implemented a stringent filtering process to refine our dataset further. Only tracks that were unequivocally assigned to a single registration class were retained. This criterion was pivotal in eliminating ambiguities and ensuring the clarity of our dataset labels.

It is imperative to acknowledge the potential pitfalls associated with our reliance on regular expressions for classification. Track titles, despite often being indicative of content, can occasionally be ambiguous or lack the granularity required for precise classification. Such nuances may lead to inadvertent mislabeling, introducing anomalies within the dataset.

3.2 Data processing

The first step is to ensure that all files are normalized and sampled at a consistent rate adapted to the pretrained models. This uniformity is important for subsequent processing steps. After the audio data is converted to tensor format, we pass it through both the pre-trained CLAP and BYOL-A models

1. https://en.wikipedia.org/wiki/French_organ_school

to generate embeddings. These embeddings are high-dimensional representations of the audio data that capture its essential features.

We start by utilizing the CLAP model, which analyzes audio recordings of organ music and generates numerical representations of size [1,512] that encapsulate the essential features of the audio content [15]. In parallel, we extract embeddings using the BYOL-A model, which also provides rich numerical representations of size [1,2048] by focusing on self-supervised learning from augmented views of the same audio data.

For data visualization, we apply the t-SNE algorithm to reduce the dimensionality of these high-dimensional embeddings while preserving the structure of the data as much as possible. We experiment with different perplexity values to find the optimal configuration that best represents the data in a lower-dimensional space, allowing us to compare how well the embeddings from CLAP and BYOL-A capture the underlying patterns in the audio data.

This comparative approach enables us to evaluate the effectiveness of the CLAP and BYOL-A models in generating embeddings that are most suitable for subsequent classification tasks in our study.

With t-SNE, we aim to identify irregularities and isolate them. Once outlier clusters are observed, we use the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to isolate these outliers. DBSCAN is a clustering method that identifies clusters based on the density of data points and is particularly useful for identifying clusters of varying shapes and sizes.

We visualize the clustering results by creating scatter plots where different clusters are represented in different colors. These plots help us understand the distribution of the data, observe irregularities, and assess the effectiveness of the clustering process. Based on the clustering results, we identify and exclude specific points classified as outliers by DBSCAN, as these can skew results and affect the performance of subsequent analyses.

In addition to using DBSCAN for outlier clusters, we incorporate a subjective evaluation step to address isolated outliers, that DBSCAN might not group into clusters. We listen to all tracks with extreme values in either component 1 or component 2 of the t-SNE embedding and manually check the ground truth labels. This subjective listening process allows to ensure that these lonely outliers were correctly labeled and won't unduly influence subsequent classification.

3.3 Model and Training

We design a multi-layer perceptron (MLP) model tailored for both CLAP and BYOL-A embeddings classification. This model includes multiple layers equipped with ReLU activations and dropout for regularization, ensuring that the architecture is well-suited to handle the extracted features.

3.3.1 Model Architecture and Optimization

Figure 4 provides a schematic overview of our model architecture. The MLP model consists of the following key components :

- **Input Layer :** Receives the input feature vectors derived from the audio data.
- **Hidden Layer 1 :** A fully connected layer with 488 neurons, followed by Batch Normalization and a ReLU activation. A dropout layer with a rate of 0.30 is applied to prevent overfitting.
- **Hidden Layer 2 :** This layer has 334 neurons and similarly follows with BatchNorm, ReLU activation, and dropout.
- **Hidden Layer 3 :** Contains 179 neurons with the same subsequent operations, ensuring consistency across the model.
- **Output Layer :** The final fully connected layer maps the features to the output classes.

The model's architecture has been optimized using the Optuna framework to fine-tune hyperparameters. The use of Batch Normalization and Dropout further enhances model generalization, reducing the risk of overfitting.

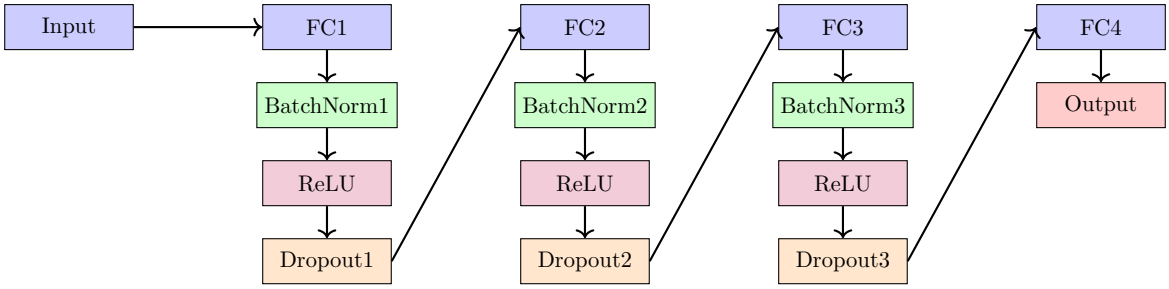


FIGURE 4 – Schematic representation of the MLP model architecture.

3.3.2 Training the Model

The training process involves the following steps :

— The model is trained using the Adam optimizer, which is chosen for its adaptive learning rate capabilities, crucial for handling the complex feature space of audio data.
— A learning rate of 0.000723 is initially set and is adjusted using a StepLR scheduler to ensure optimal convergence during training.
— Training is conducted over 15 epochs, with both training and validation datasets used to monitor performance.
— The model is trained on two different sets of embeddings : BYOL-A and CLAP. This dual training approach allows us to compare the effectiveness of these embedding techniques. By evaluating the model's performance on both, we can determine which embedding method better captures the audio features relevant to our task.

3.4 Evaluation and Metrics

To assess the model's performance, we employ a comprehensive evaluation strategy :

— **Average Loss and Accuracy Monitoring through 20 runs on both embeddings :** We track the loss and accuracy on both the trainings and validation sets across all epochs and all runs.
— **Confusion Matrix :** The confusion matrix is analyzed to identify specific classes where the model may struggle, guiding further improvements.
— **Classification Report :** This report provides detailed metrics, including precision, recall, and F1-score for each class.

In conclusion, this methodology combines a neural network architecture with thorough performance evaluation, ensuring reliable classification of audio data.

3.5 Detection of Mislabeled Instances and Analysis of Misclassified Data

To ensure the robustness and accuracy of the model, we conduct a detailed analysis of misclassified data and review the dataset for potential mislabeled instances.

3.5.1 Subjective Listening and Dataset Review

To identify potential mislabeled data, we analyze t-SNE figures to observe data points with extreme values that are significantly distant from their respective clusters. These outliers are then subjected to subjective listening, where we listen to the audio recordings and compare them to their stated labels. By understanding why these points are so far from their clusters, we can spot and address any discrepancies. If mislabeling is detected and the label is obviously wrong, the dataset is updated by correcting or removing these tracks, ensuring that the dataset is accurate and that the model's performance accurately reflects its true capabilities. It's important to note that while this procedure enhances the accuracy of the dataset by correcting obvious mislabeling, it could also introduce a positive bias to the classification performance. Specifically, if a track originally belongs to class A but is incorrectly labeled as class B in the

ground-truth and happens to fall within the cluster of class B in the t-SNE embedding, this mislabeling may go undetected. As a result, the model could appear to perform better than it actually does, since such mislabeled data would reinforce the incorrect label rather than reveal a discrepancy. However, it is reasonable to consider such cases as rare, given the assumption that most tracks are labeled correctly and that extreme outliers are more likely to indicate mislabeling.

3.5.2 Identifying Misclassified Data

During both training and evaluation, we track misclassified samples. After each epoch, indices of misclassified training samples are recorded, providing insights into which samples the model struggles with and highlighting potential issues with either the dataset or model performance. Similarly, misclassified validation samples are noted to assess model performance in a controlled setting.

These indices are analyzed and visualized to identify any patterns or common characteristics among misclassified samples.

3.5.3 Model Retraining and Evaluation

After correcting the ground-truth labels, the model is retrained with the corrected data. The model's performance is then re-evaluated using the same metrics as before, including loss, accuracy, and confusion matrix. The misclassification analysis is repeated to confirm that the adjustments have resolved previous issues and to verify improvements in model performance.

This iterative approach enhances dataset quality and model accuracy by addressing data labeling issues and refining model performance.

4 Results

4.1 Dataset Overview

One of the primary contributions of this work is the creation of a curated dataset specifically designed for the classification of organ registrations within organ music. This dataset represents a significant advancement in the field of music information retrieval and audio classification since it was never created in this domain.

4.1.1 Curated Dataset

The dataset includes organ music recordings from prominent French Baroque composers, meticulously sourced and labeled. Key aspects of the dataset include :

— **Composition and Size :** The dataset consists of **1916** tracks from various composers (such as François Couperin, Louis-Nicolas Clérambault, Jean Adam Guilain etc...) covering a diverse range of organ registrations and styles typical of the French baroque organ music. Each track is labeled with the specific organ registration used, ensuring a high level of detail and accuracy.

— **Labeling and Classification :** Tracks were classified into predefined registration classes using a combination of heuristic methods and expert knowledge. This process involved extracting and categorizing data based on patterns found in track titles and descriptions.

— **Data Quality and Processing :** The dataset was rigorously reviewed for accuracy, with misclassified and mislabeled instances being identified and corrected through subjective listening and manual validation. This ensures that the dataset not only represents the music accurately but is also reliable for training and evaluating machine learning models (see section 4.2, 4.3 and 4.4).

4.1.2 Visualization and Analysis

To illustrate the dataset's diversity and quality, we present several visualizations :

— **Distribution of Registrations :** A plot showing the distribution of different organ registrations within the dataset highlights the variety and balance of the collected samples.

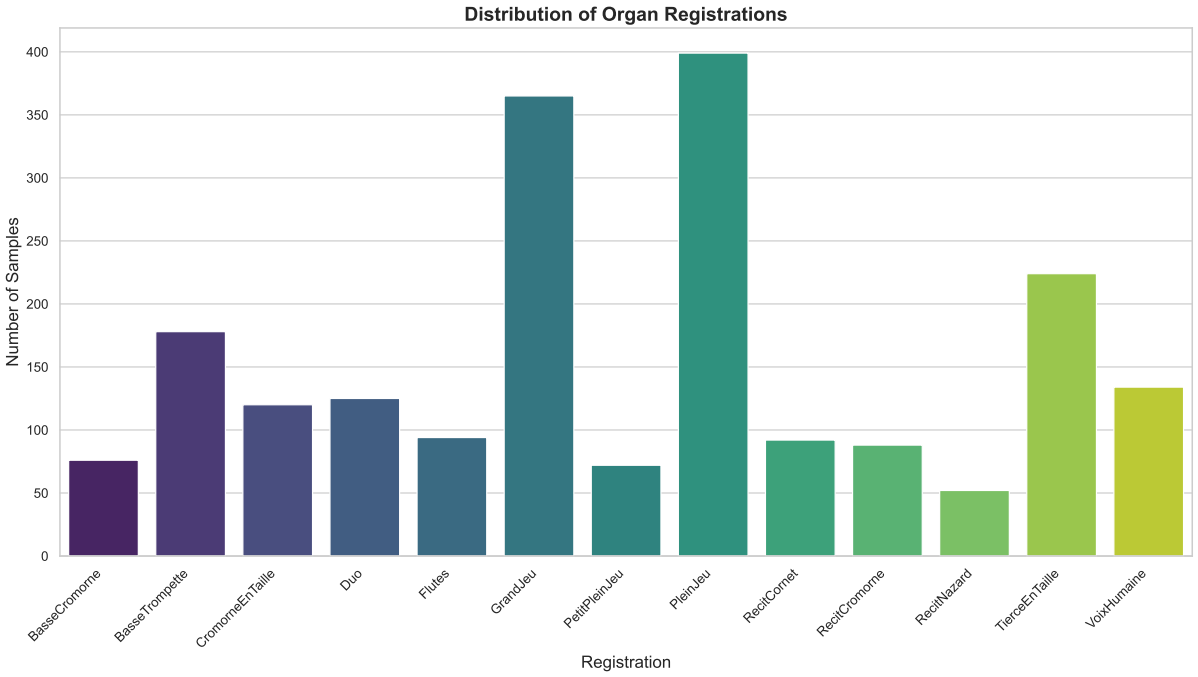


FIGURE 5 – Distribution of instances through all classes.

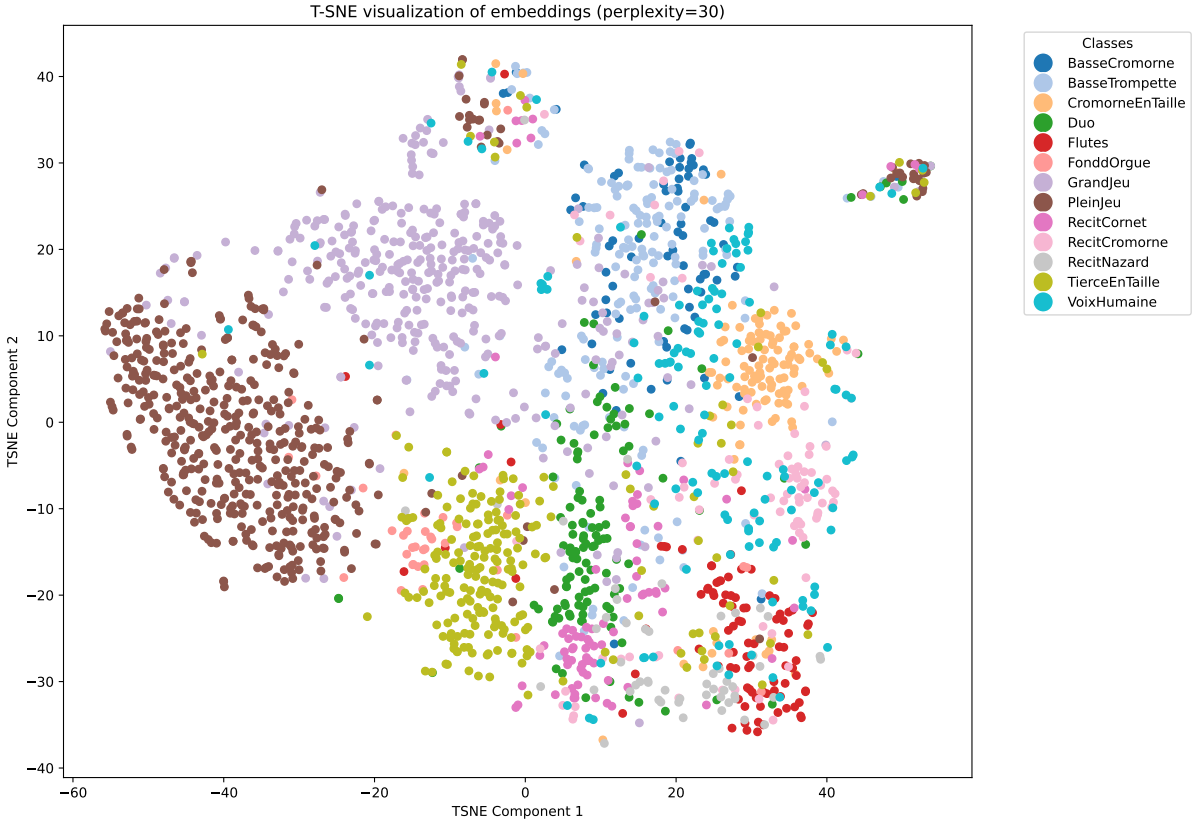


FIGURE 6 – t-SNE Visualization of CLAP embeddings.

We merge 'Plein Jeu' and 'Petit Plein Jeu' into a single class, 'Plein Jeu,' because 'Petit Plein Jeu' is essentially a variant or subset of 'Plein Jeu'. In many organ contexts, 'Petit Plein Jeu' represents a more delicate or subdued version of the 'Plein Jeu' sound, providing a softer and less intense expression while retaining the core character of 'Plein Jeu'. By combining these two registrations, we acknowledge that 'Petit Plein Jeu' is a specific instance of the broader 'Plein Jeu' category, reflecting its role as a nuanced adaptation rather than a distinct registration in its own right. Furthermore, the Grand Plein Jeu of a small organ can have roughly a similar sound as the Petit Plein Jeu of a big organ.

4.2 T-SNE Visualization of CLAP Embeddings

The presence of clusters in the t-SNE visualization of audio embeddings (Figure 6) suggests that the CLAP embeddings effectively capture meaningful information about the audio samples. This is indicative of a well-structured embedding space where similar audio samples are mapped close to each other, facilitating easy differentiation between classes of registration. The observed cluster separation and density provide insights into the consistency and distinctiveness of audio characteristics within each group. Moreover, outliers are present, it signify anomalies or rare instances in the dataset. Interpreting the clusters in conjunction with domain knowledge can reveal underlying patterns related to specific sounds, instruments, genres, or other audio features.

In our analysis of the t-SNE clusters, we observe the following :

— **Overlap of Classes :** The classes *Basse de Trompette* and *Basse de Cromorne* exhibit heavy overlap, which is not unexpected given their similar characteristics. *Voix Humaine* shows partial clustering and overlap with *Recit de Cromorne*, which is also consistent with their musical similarities.

— **Problematic Classes :** The classes related to flues, such as *Fond d'Orgue*, *Flûtes*, and *Recit de Nasard*, remain problematic and are not well-separated.

— **Clear Clusters :** On the other hand, there are distinct clusters observed for *Plein Jeu*, *Grand Jeu*, *Tierce en Taille*, and *Cromorne en Taille*.

We observe a heterogeneous cluster in the top right of Figure 7, which contains all classes. To understand the nature of this cluster, we isolate it using the DBSCAN algorithm and review all tracks within it.

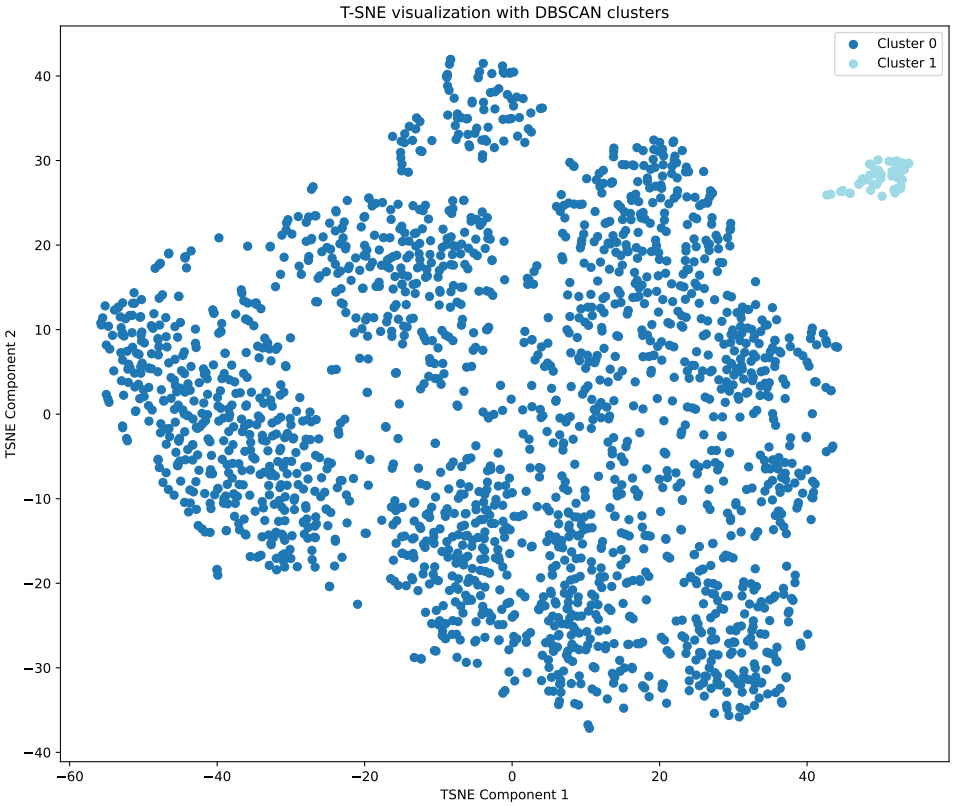


FIGURE 7 – Cluster isolated with DBSCAN.

Upon inspection, we discover that the tracks within this cluster predominantly contain voices (choir singing). Interestingly, some instances of voices not in the are properly classified. For example, in the track *J.A. Guilain Suite pour le Magnificat du Second Ton/5*, voices are present at the beginning for the first thirty seconds, after which the *Flûtes* registration is used for over two minutes. This track is not part of the cluster. The reason for this discrepancy is that CLAP, the representation learning model used for extracting embeddings, takes the average over the entire track [15], which may dilute the distinctive features of the voices, leading to its exclusion from the cluster. Additionally, we choose not to classify clusters containing isolated choir singing, as this falls outside the scope of our study.

Despite the observation from t-SNE visualizations, which show the choir cluster distinctly separated in the non-linear representation space, excluding this cluster does not lead to improvements in model training.

4.3 T-SNE Visualization of Byol-A Embeddings

The t-SNE visualization of BYOL-A embeddings (Figure 8) reveals an even more pronounced separation between clusters compared to CLAP embeddings, indicating that BYOL-A embeddings capture the nuances of the audio data with greater precision. The reduced overlap between clusters suggests that the model is better at distinguishing between different classes of registration, enhancing the clarity and distinctiveness of the embedding space. However, despite this improvement, certain classes, such as *Basse de Trompette* and *Basse de Cromorne*, appear close to each other, which is due to the inherent similarities in their timbral characteristics. Additionally, the presence of outliers in clusters that do not align with their respective categories highlights potential mislabeling in the dataset. Identifying and addressing these outliers is crucial for refining the dataset and improving the overall model performance, as they may represent mislabeled instances or rare variations that could otherwise distort the learning process.

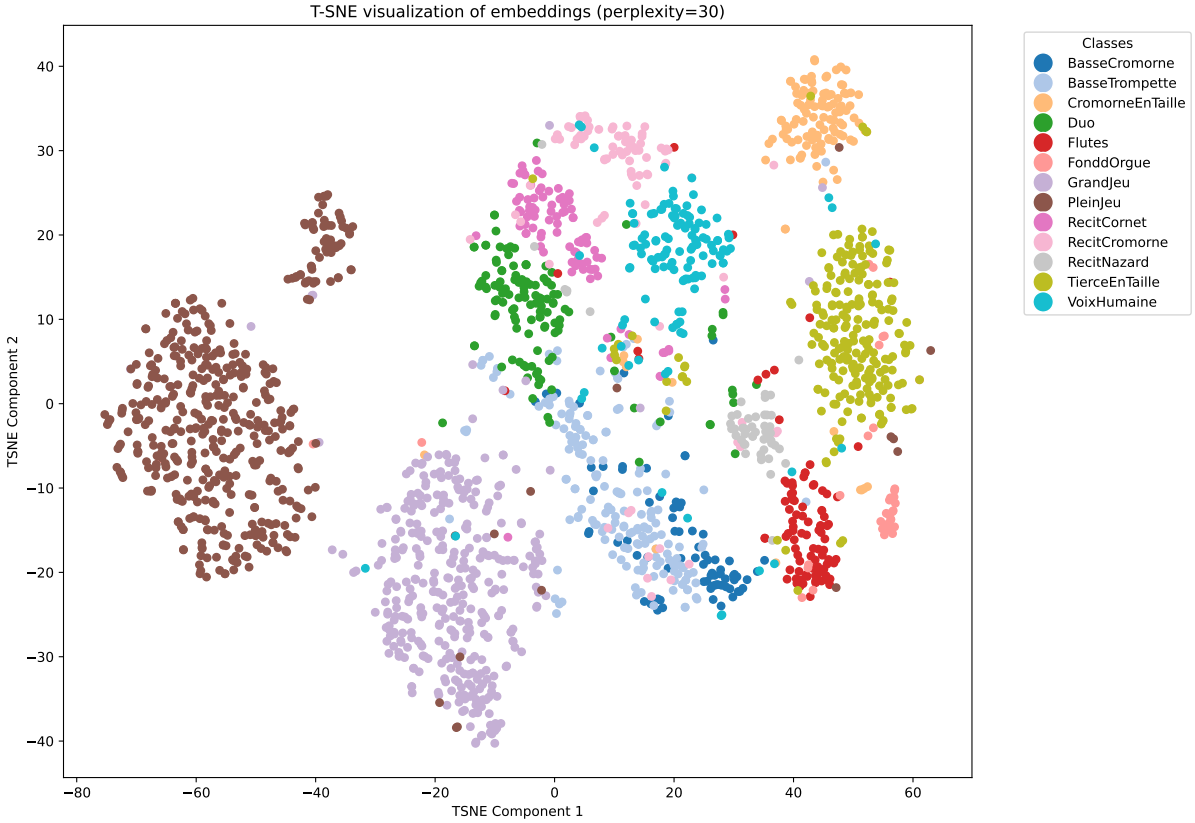


FIGURE 8 – t-SNE visualization of BYOL-A Embeddings

4.4 Data Cleaning

We listened to all tracks with extreme values in either component 1 or component 2 of the t-SNE embedding and manually checked the ground truth labels. Through this process, we identified 34 mislabeled tracks, which were likely mislabeled due to the heuristic collection of the dataset that may have induced some labeling mistakes. Additionally, we identified 139 tracks where the registration had too much variation over time (multiple classes in one track, E.g, a Basse de Trompette and Recit de Cornet in one track), making it impossible to associate them with a single class. Also sometimes wrong registration were chosen by the organist performing (E.g. Basse de Trompette with reeds in accompaniment). These tracks have been set aside and are reserved for the creative experiment described in section 5.

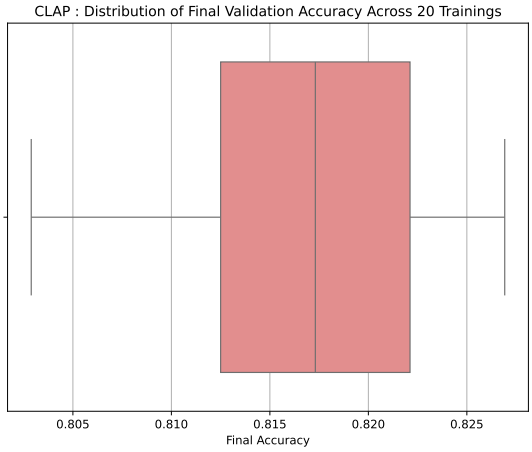
4.5 Classification Results Before Data Cleaning

The classification results for both CLAP and BYOL-A embeddings offer insightful comparisons on model performance across different organ registration classes before correcting the mislabeled data. The detailed performance metrics for each embedding type through all classes are summarized in Table 1.

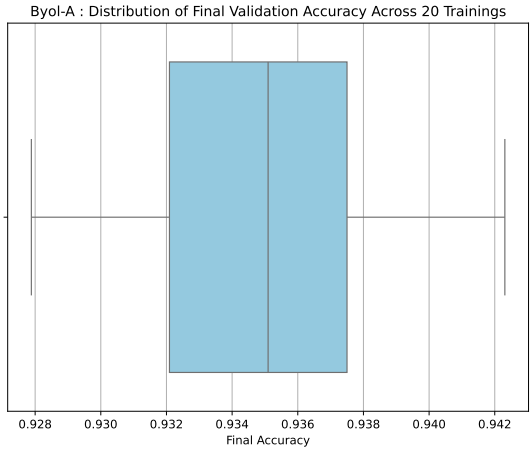
The boxplot for BYOL-A embeddings, illustrated in Figure 9b, reveals that BYOL-A method achieves a mean accuracy of approximately **93.47%**. This high mean is accompanied by a relatively low standard deviation of 0.0041, suggesting that the model's performance is both high and consistent across different samples.

In contrast, the boxplot for CLAP embeddings, depicted in Figure 9a, shows a mean accuracy of **81.67%**. The standard deviation of 0.0063 is higher than that of BYOL-A, reflecting greater variability in the performance of the CLAP embeddings. The accuracy range for CLAP embeddings is broader, spanning from 80.29% to 82.69%. This wider range indicates that CLAP embeddings exhibit more fluctuation in accuracy, demonstrating less consistency in performance compared to BYOL-A.

The results across all classes in Table 1 indicate that BYOL-A embeddings significantly outperform CLAP embeddings in classification accuracy, particularly in the context of class imbalance. The macro



(a) Boxplot for the model trained on CLAP Embeddings



(b) Boxplot for the model trained on BYOL-A Embeddings

FIGURE 9 – Boxplots comparing final accuracy of our model with CLAP and BYOL-A embeddings as input.

average F1-score for BYOL-A embeddings is 0.921, which is notably higher than the 0.779 achieved with CLAP embeddings. This suggests that BYOL-A embeddings offer a more nuanced and distinct representation of the audio data.

For individual classes, BYOL-A embeddings demonstrate superior precision and recall in many cases. For instance, the class *Plein Jeu* achieves a precision and recall of 0.989 with BYOL-A embeddings, compared to 0.902 and 0.968 with CLAP embeddings, respectively. Similarly, classes such as *Recit Cornet* and *Recit Cromorne* show improved performance metrics with BYOL-A embeddings, indicating enhanced differentiation capability for these classes.

As illustrated in Figure 8, the BYOL-A embeddings reveal more distinct clusters with less overlap compared to CLAP embeddings. This clearer cluster separation, demonstrated by t-SNE, likely contributes to the higher accuracy observed, especially in distinguishing between similar classes like *Basse de Trompette* and *Basse de Cromorne*, as well as between *Fond d'Orgue* and *Flûtes*, which are challenging due to similar timbres.

4.6 Classification Results After Data Cleaning

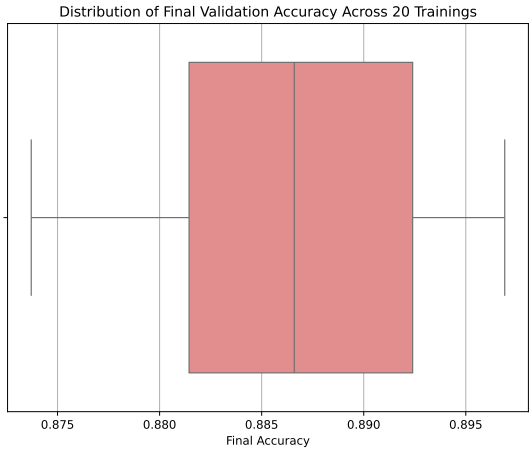
In this section, we provide a comprehensive comparison of the BYOL-A and CLAP embedding methods, focusing on their performance metrics before and after data cleaning. Our objective is to assess how each method performs in terms of accuracy and consistency, and to determine the impact of data cleaning on these performance metrics.

Following the data cleaning process, as shown in Figure 10, BYOL-A's performance improved significantly, with a mean accuracy of **98.77%** and a reduced standard deviation of 0.0036, highlighting enhanced accuracy and even greater consistency. Although CLAP also saw an improvement, with its mean accuracy rising to **88.69%** and a slight increase in standard deviation to 0.0068, it still lags behind BYOL-A in both accuracy and stability. This comparison underscores that while both methods benefited from the data cleaning process, BYOL-A consistently demonstrates superior performance and reliability compared to CLAP.

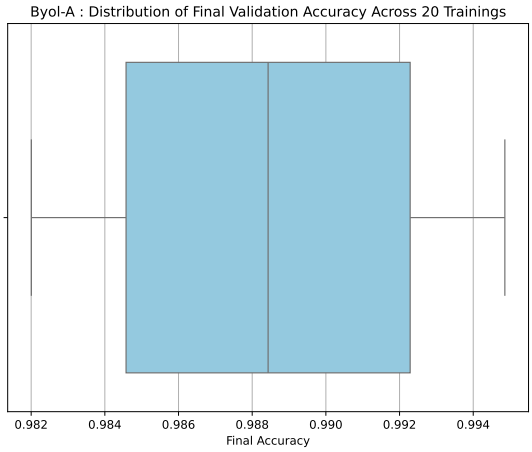
Examining the performance metrics in Table 2 for CLAP and BYOL-A embeddings after data cleaning reveals significant insights, particularly when compared to the pre-cleaning metrics presented in Table 1. Correcting the ground-truth has notably improved the performance of both embedding methods, but several classes still exhibit challenges. For instance, the *Récit de Cornet* class, which was problematic before cleaning, shows minimal improvement post-cleaning, with precision and recall remaining high for BYOL-A but still showing room for enhancement for CLAP. More prominently, the *Basse de Trompette* and *Basse de Cromorne* classes, which share similar timbre characteristics, continue to present difficulties. Before data cleaning, *Basse de Trompette* had a precision of 0.77 and recall of 0.72 for CLAP, compared

Class	CLAP / BYOL-A Precision	Recall	F1-Score	Support
Basse de Cromorne	0.82 / 0.94	0.56 / 0.94	0.67 / 0.94	16
Basse de Trompette	0.77 / 0.89	0.72 / 0.86	0.74 / 0.87	36
Cromorne En Taille	0.81 / 0.89	0.88 / 0.96	0.84 / 0.92	24
Duo	0.72 / 0.89	0.84 / 0.92	0.78 / 0.90	25
Flûtes	0.63 / 0.94	0.79 / 0.90	0.70 / 0.92	19
Fond d'Orgue	0.80 / 0.88	0.50 / 0.88	0.62 / 0.88	8
Grand Jeu	0.91 / 0.92	0.88 / 0.96	0.90 / 0.94	73
Plein Jeu	0.90 / 0.99	0.97 / 0.99	0.93 / 0.99	95
Recit de Cornet	0.73 / 0.95	0.84 / 1.00	0.78 / 0.97	19
Recit de Cromorne	0.93 / 1.00	0.78 / 0.78	0.85 / 0.88	18
Recit de Nazard	0.73 / 0.91	0.73 / 0.91	0.73 / 0.91	11
Tierce En Taille	0.85 / 0.96	0.87 / 0.96	0.86 / 0.96	45
Voix Humaine	0.86 / 0.92	0.67 / 0.89	0.75 / 0.91	27
Macro Avg	0.80 / 0.93	0.77 / 0.92	0.78 / 0.92	416
Weighted Avg	0.84 / 0.94	0.83 / 0.94	0.83 / 0.94	416

TABLE 1 – Performance metrics for different classes before data cleaning. Metrics are presented as CLAP value / BYOL-A value.



(a) Boxplot for the model trained on CLAP embeddings after correcting mislabeled instances



(b) Boxplot for the model trained on BYOL-A embeddings after correcting mislabeled instances

FIGURE 10 – Boxplots comparing final accuracy of our model with CLAP and BYOL-A embeddings as input.

to 0.89 and 0.86 for BYOL-A, respectively. After cleaning, the precision and recall for this class improved to 0.79 and 0.93 for CLAP and 0.96 and 0.93 for BYOL-A. The *Basse de Cromorne* class also exhibits notable issues, with pre-cleaning metrics showing CLAP at 0.82 precision and 0.56 recall, while BYOL-A achieved 0.94 and 0.94, respectively. Post-cleaning, these metrics for CLAP improved to 0.92 precision and 0.80 recall, while BYOL-A's metrics rose to 0.93 and 0.93, respectively. Despite these improvements, the continued difficulty in differentiating between these similar classes highlights the challenge of accurately classifying audio recordings with closely related timbres. This ongoing issue underscores the need for further refinement in both the data cleaning process and the embedding methods to enhance classification performance in such nuanced scenarios.

In summary, the BYOL-A embeddings provide a more refined and accurate representation of organ registrations compared to CLAP embeddings, leading to improved classification performance across various metrics. This underscores the effectiveness of BYOL-A embeddings in capturing subtle distinctions between different organ sounds. Overall, the comparison between the two methods reveals that BYOL-A embeddings offer superior performance relative to CLAP embeddings. The higher mean accuracy and

Class	CLAP / BYOL-A Precision	Recall	F1-Score	Support
BasseCromorne	0.92 / 0.93	0.80 / 0.93	0.86 / 0.93	15
BasseTrompette	0.79 / 0.96	0.93 / 0.93	0.86 / 0.95	29
CromorneEnTaille	0.90 / 1.00	0.76 / 1.00	0.83 / 1.00	25
Duo	0.76 / 0.95	0.90 / 1.00	0.83 / 0.98	21
Flutes	0.78 / 1.00	0.74 / 1.00	0.76 / 1.00	19
FonddOrgue	1.00 / 1.00	0.86 / 1.00	0.92 / 1.00	7
GrandJeu	1.00 / 1.00	0.90 / 1.00	0.95 / 1.00	70
PleinJeu	0.96 / 1.00	0.99 / 1.00	0.97 / 1.00	93
RecitCornet	0.89 / 1.00	0.89 / 1.00	0.89 / 1.00	19
RecitCromorne	0.75 / 1.00	0.90 / 0.91	0.82 / 0.95	10
RecitNazard	0.89 / 1.00	0.73 / 1.00	0.80 / 1.00	11
TierceEnTaille	0.84 / 1.00	0.88 / 1.00	0.86 / 1.00	43
VoixHumaine	0.85 / 0.96	0.88 / 1.00	0.87 / 0.98	26
Macro Avg	0.87 / 0.99	0.86 / 0.99	0.86 / 0.99	388
Weighted Avg	0.90 / 0.99	0.89 / 0.99	0.89 / 0.99	388

Table 2 – Performance metrics for different classes using CLAP and BYOL-A embeddings after data cleaning. Metrics are presented as CLAP value / BYOL-A value.

lower variability for BYOL-A suggest that this approach provides not only better overall performance but also greater reliability and stability. The narrower range of accuracy values for BYOL-A embeddings further emphasizes their superior performance, making them a more robust choice compared to the more variable CLAP embeddings. For a deeper insight into the classification results, including the confusion matrix after correcting the ground-truth, please refer to the annex 8.

The substantial improvement in performance of BYOL-A embeddings over CLAP embeddings can be attributed to several key differences in their underlying learning frameworks and methodologies. BYOL-A (Bootstrap Your Own Latent Audio) utilizes a self-supervised learning approach that avoids the need for negative samples. Instead of contrasting positive samples against negative ones, BYOL-A focuses on maximizing the consistency between different augmented views of the same audio clip. This method leverages a "moving average" mechanism to stabilize the learning process, which helps in capturing more stable and discriminative features of the audio data.

It is noteworthy that BYOL-A, despite being a significantly smaller network compared to CLAP, demonstrates impressive classification performance. The relative compactness of BYOL-A, combined with the fact that it was not specifically trained on musical data—where musical data comprises only a minor portion of the training set—highlights its robustness and versatility. The model's ability to achieve high classification accuracy in this context is particularly remarkable, given the limited exposure to musical data during training. This performance underscores the effectiveness of BYOL-A's embedding approach and its potential for handling specialized tasks such as musical instrument classification, even when the training data is not predominantly composed of music-related content. This characteristic makes BYOL-A an appealing choice for applications requiring both efficiency and high performance. Another

In contrast, CLAP (Contrastive Language-Audio Pretraining) employs a contrastive learning framework that relies on negative sampling to differentiate between classes. While this approach can be effective, it introduces inherent variability and can be less stable, particularly in distinguishing subtle differences between similar audio classes. The reliance on negative samples in CLAP may also limit the quality of learned representations, as the model needs to balance between distinguishing true positives and avoiding false negatives, which can sometimes lead to less nuanced embeddings.

BYOL-A's self-supervised approach enhances the quality of embeddings by concentrating on internal consistency and self-prediction, leading to a higher mean accuracy, lower standard deviation, and more consistent performance across various audio classes. This results in better precision and recall, as observed in the confusion matrices and performance metrics. The model's ability to better differentiate between closely related audio classes and its effective handling of overlapping features contribute to its superior performance compared to CLAP embeddings.

4.7 Analysis of Misclassified Instances

In addition to the above results, a crucial aspect of future work will involve a detailed analysis of misclassified instances to identify potential mislabeled data. By examining cases where the models have made incorrect predictions, we can gain insights into common patterns or anomalies that may indicate errors in the dataset's labeling process. This analysis will help in pinpointing specific instances or classes that are prone to misclassification, which can then be reviewed and corrected. Improving the accuracy of the labels will not only enhance the performance of the models but also contribute to a more reliable and high-quality dataset. This iterative process of scrutinizing misclassifications and refining the dataset will ultimately lead to more accurate predictions and a deeper understanding of the underlying challenges in classifying musical registrations. Unfortunately, due to time constraints, we were unable to complete this analysis but it will be done in the near future.

5 Practical Applications

5.1 A Creative Experiment

In an attempt to further explore the capabilities of the model, a creative experiment was conducted to track the evolution of organ registrations over time. The goal was to analyze audio recordings and visualize how the registrations change during performance. This involved using embeddings from an audio model to predict registrations and generate time-series data.

However, the experiment encountered a significant issue due to the network not being online. Specifically, integrating a Recurrent Neural Network (RNN) could have provided the necessary temporal context for the analysis, but this approach was not feasible with the time left for this study.

5.2 A Tool for Organists and Scholars

The development of this tool for organists and scholars seeks to address a significant challenge in the study and performance of historical organ music : the lack of explicit registration instructions. Many compositions, particularly from Italian and German traditions, do not specify the registrations to be used. This omission places the onus on the performer to choose the stops, requiring a profound understanding of the organ's capabilities and the stylistic practices of the period. This gap in documentation often results in varied interpretations and significant personal expression, as performers navigate the absence of detailed guidance.

To compensate for this issue, the tool we tried to develop provides a means to explore and understand the practice of organ registration more deeply. By analyzing audio recordings and generating detailed visualizations of how registrations change over time, this tool could have help bridge the gap left by missing registration instructions. It would have offered insights into both historical practices and contemporary performances, thereby enhancing interpretative accuracy and preserving the legacy of organ music. By integrating these insights into educational resources, scholars and performers can gain a better understanding of historical registration practices and their effects, enriching both the study and performance of organ music. This not only aids in preserving the tradition but also provides new perspectives for modern interpretations, ensuring that the rich heritage of organ music continues to be explored and appreciated in meaningful ways.

6 Conclusion

In the context of low data availability and class imbalance, this research delves into innovative methods to tackle the inherent challenges of predicting organ registrations and analyzing the distinctive characteristics of organ sounds. Historical organ compositions, particularly from Italian and German traditions, often lack specific registration instructions, leaving the choice of stops to the performer. This absence of explicit guidance requires organists to possess a deep understanding of the instrument and its stylistic practices, leading to substantial variability between performances and allowing for considerable interpretative freedom.

Our study addresses these challenges by introducing a novel dataset and developing a specialized machine learning model designed to predict organ registrations even in the absence of detailed instructions. This model not only aids in the accurate reproduction of historical performances but also enhances our understanding of the unique sonic qualities associated with different organ stops. By predicting registrations based on historical context and performance practice, the model provides valuable insights for organists, helping them make informed decisions about registration choices.

The implications of this research extend beyond mere technical achievement. By bridging the gap between historical performance practices and modern technology, we contribute to the preservation and enrichment of organ music heritage. This approach not only facilitates a more accurate recreation of historical interpretations but also supports organists in achieving a deeper connection with the stylistic and historical context of the compositions.

In the long term, the vision for this research is to foster a greater appreciation of organ music's rich tradition and to enhance the quality of musical performances. By preserving the nuanced details of historical performances and offering practical tools for modern organists, this work contributes to a more informed and authentic interpretation of organ music. Future research will focus on expanding the dataset to include a broader range of compositions and exploring advanced machine learning techniques to further refine registration predictions. Ultimately, this endeavor aims to enrich musical interpretations and ensure that the legacy of organ music continues to be celebrated and understood in its full historical and stylistic depth.

7 Future Work

7.1 Data Augmentation and Class Balancing

Future work will emphasize enhancing the robustness of models through advanced data augmentation techniques and class balancing strategies. The challenges posed by low data availability and class imbalance necessitate the use of methods such as synthetic data generation and oversampling to improve model performance and generalizability. These techniques aim to create a more balanced and comprehensive dataset, thereby allowing for more accurate predictions and insights. By increasing the diversity and quantity of training data, the models will be better equipped to handle various scenarios and reduce biases that arise from imbalanced classes.

7.2 Model Refinement and Understanding

Continuous improvement of prediction accuracy and robustness will be prioritized. Efforts will be directed towards refining the models through iterative testing and incorporating feedback from real-world applications. This will involve adopting the latest advancements in machine learning techniques, including those that enhance the models' precision and reliability. Regular updates and refinements will ensure that the models remain state-of-the-art and capable of delivering high-quality predictions in various practical scenarios. A more in-depth analysis of the differences between CLAP and BYOL-A could be conducted. Specifically, examining how each model handles various aspects of audio data, such as feature extraction, embedding spaces, and the alignment between predicted labels and ground-truth labels, would be insightful. Additionally, exploring how these models react to different types of mislabeling, as well as their performance on challenging datasets, could provide valuable insights. This deeper analysis could lead to the development of more sophisticated techniques for detecting and correcting mislabeling, potentially improving the accuracy and fairness of the dataset further.

7.3 Broader Applications

Future research will explore the application of the developed models to other musical traditions and instruments. By adapting the techniques and models to different contexts, the insights gained from this research can contribute to a broader understanding of musical registration and performance practices across various genres and instruments. This will include investigating how the methods used for organ music can be translated to other types of musical compositions and instruments, thereby expanding the impact and relevance of the research.

Contributions

In the execution of this project, responsibilities were divided as follows :

— **Peter van Kranenburg** : Peter was responsible for the data set preparation and ground truth cleaning. His meticulous work in preparing the data and ensuring the accuracy of the ground truth was crucial to the project's success.
— **Pablo Dumenil** : I handled all aspects beyond data preparation and ground truth cleaning.

This collaborative approach leveraged our respective expertise and led to a successful completion of the project.

8 Annex

8.1 Mathematical computation of t-SNE and PCA

8.1.1 T-Stochastic Neighbor Embedding

T-SNE is a dimensionality reduction technique that helps visualize high-dimensional data by reducing it to a lower-dimensional space. The algorithm aims to preserve the pairwise similarities between data points as much as possible.

- Probability Distribution in High Dimensions

For a data point $\mathbf{x}_i$ and its neighbor $\mathbf{x}_j$, the conditional probability $p_{j|i}$ is defined as :

$$p_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_i^2}\right)}$$

where :

— $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the squared Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$,

— σ_i is the bandwidth of the Gaussian kernel for point $\mathbf{x}_i$,

— The denominator normalizes the probabilities so that $\sum_j p_{j|i} = 1$.

The bandwidth σ_i is chosen to ensure a specific number of neighbors are considered, controlled by the perplexity parameter. The perplexity $\mathcal{P}$ is related to the effective number of neighbors and is given by :

$$\mathcal{P}(\sigma_i) = 2^{H(P_i)}$$

where $H(P_i)$ is the Shannon entropy of the distribution $p_{j|i}$:

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

- Probability Distribution in Low Dimensions

In the low-dimensional space, we define the conditional probability $q_{j|i}$ using a Student's t-distribution with one degree of freedom (Cauchy distribution) :

$$q_{j|i} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}$$

where :

— $\mathbf{y}_i$ and $\mathbf{y}_j$ are the low-dimensional representations of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

Objective Function

The objective of T-SNE is to minimize the Kullback-Leibler (KL) divergence between the high-dimensional distribution $p_{j|i}$ and the low-dimensional distribution $q_{j|i}$:

$$\mathrm{KL}(P\|Q) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

This KL divergence is minimized using gradient descent, adjusting the positions $\mathbf{y}_i$ of the low-dimensional data points to make $q_{j|i}$ as close as possible to $p_{j|i}$.

Following the extraction of embeddings, we first apply T-distributed Stochastic Neighbor Embedding (t-SNE) for visualizing high-dimensional data [9]. T-SNE facilitates the reduction of dimensionality, allowing us to observe clusters, patterns, and relationships between data points in a 2D or 3D space. The algorithm computes pairwise similarities between data points using a Gaussian kernel, constructs a probability distribution based on these similarities, and optimizes the low-dimensional embedding to minimize the KL divergence between the high-dimensional and low-dimensional distributions.

8.1.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is another dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving as much of the variance in the data as possible. Unlike t-SNE, which focuses on preserving local structure, PCA emphasizes capturing the global structure of the data by identifying the directions (principal components) along which the variance is maximized.

- Computation of Principal Components

PCA works by computing the eigenvectors (principal components) and eigenvalues of the data's covariance matrix. For a dataset with n data points, where each data point $\mathbf{x}_i$ is a vector in $\mathbb{R}^d$:

1. **Mean Centering** : Subtract the mean μ of the data from each data point to obtain a mean-centered dataset.
$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mu$$
2. **Covariance Matrix** : Compute the covariance matrix $\mathbf{C}$ of the mean-centered data.
$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T$$
3. **Eigen Decomposition** : Perform eigen decomposition on the covariance matrix to obtain the eigenvectors $\mathbf{v}_k$ and corresponding eigenvalues λ_k.
4. **Projection** : Project the data onto the first k principal components (eigenvectors corresponding to the largest eigenvalues) to reduce the dimensionality.
$$\mathbf{y}_i = \mathbf{W}^T \tilde{\mathbf{x}}_i$$
where $\mathbf{W}$ is the matrix of the top k eigenvectors.

In practice, we often first apply PCA to reduce the dimensionality of the data, making it more computationally feasible for t-SNE to operate. By reducing the dimensions to a smaller number (e.g., 50), PCA retains most of the data's variability while discarding noise. Subsequently, t-SNE is applied to this lower-dimensional representation to visualize the data in a 2D or 3D space, revealing clusters, patterns, and relationships that may not be apparent in the original high-dimensional data.

8.2 Confusion Matrix before Data Cleaning

The confusion matrix for CLAP embeddings is presented in Figure 11.

The confusion matrix reveals several key insights :

— **Class Overlap :** There is notable overlap between the *Basse de Trompette* and *Basse de Cromorne* classes, reflecting their similar musical characteristics. This is consistent with the t-SNE visualization. Additionally, *Voix Humaine* and *Recit de Cromorne* are frequently confused, indicating similarities in their musical contexts.

— **Problematic Classes :** Classes such as *Fond d'Orgue*, *Flûtes*, and *Recit de Nazard* show frequent misclassifications. These classes may share common features, making them difficult to distinguish.

— **Well-Defined Clusters :** In contrast, the classes *Plein Jeu*, *Grand Jeu*, *Tierce en Taille*, and *Cromorne en Taille* exhibit clear separation, indicating they are well-defined and easily distinguishable from one another.

The confusion matrix for BYOL-A embeddings is shown in Figure 12. It demonstrate generally strong performance, as reflected in the classification metrics :

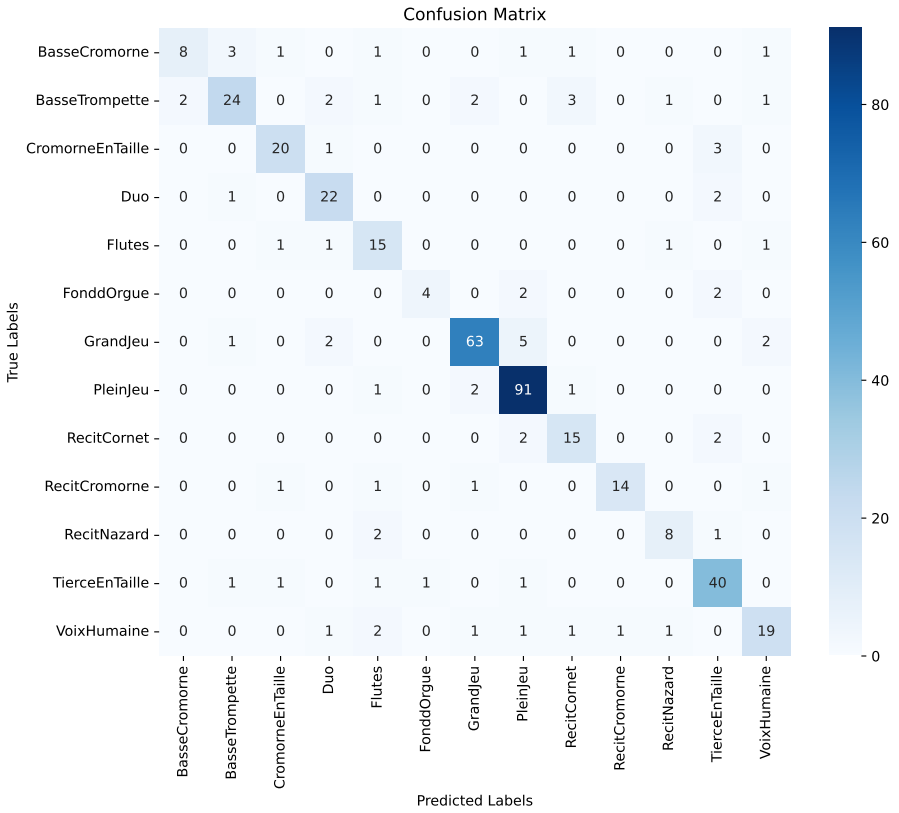


FIGURE 11 – Confusion Matrix for CLAP embeddings.

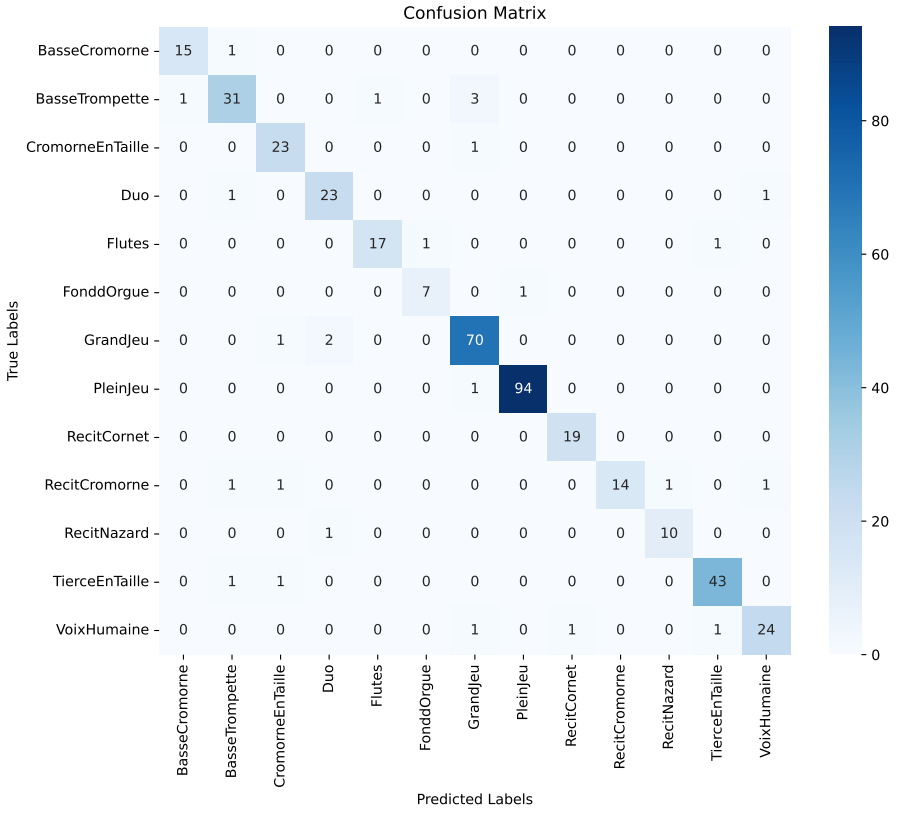


FIGURE 12 – Confusion Matrix for BYOL-A embeddings.

Comparison with CLAP Embeddings Comparing the BYOL-A embeddings with CLAP embeddings reveals several differences :

— **Performance Improvement :** BYOL-A embeddings generally exhibit higher precision and recall across most classes. For instance, the F1-score for *Plein Jeu* significantly improves from 0.934 with CLAP embeddings to 0.989 with BYOL-A embeddings, as supported by the confusion matrix.
— **Class Distinction :** BYOL-A embeddings provide better separation between classes, particularly in distinguishing between similar timbres. This includes a clearer distinction between *Basse de Trompette* and *Récit de Cromorne*.
— **Misclassification Patterns :** While both models struggle with classes featuring similar timbres, such as *Fond d'Orgue* and *Flûtes*, the BYOL-A model demonstrates fewer overall misclassifications, indicating improved handling of overlapping features.

8.3 Confusion Matrix after Data Cleaning

The confusion matrix for BYOL-A reveals several important insights :

— **High Accuracy in Certain Classes** : The BYOL-A model demonstrates strong performance in classes such as *Grand Jeu*, *Plein Jeu*, and *Recit Cromorne*. For example, *Grand Jeu* has all instances correctly classified with 70 true positives, and *Plein Jeu* also shows excellent performance with 93 true positives.
— **Challenges with Similar Classes** : There is some confusion between classes with similar timbres, like *Basse de Cromorne* and *Basse de Trompette*. A few instances are misclassified between these classes, but overall, BYOL-A performs relatively well in distinguishing these similar classes.
— **Performance in Low-Support Classes** : For classes with fewer instances, such as *Recit Cornet* and *Voix Humaine*, BYOL-A shows a good balance with minimal misclassifications.

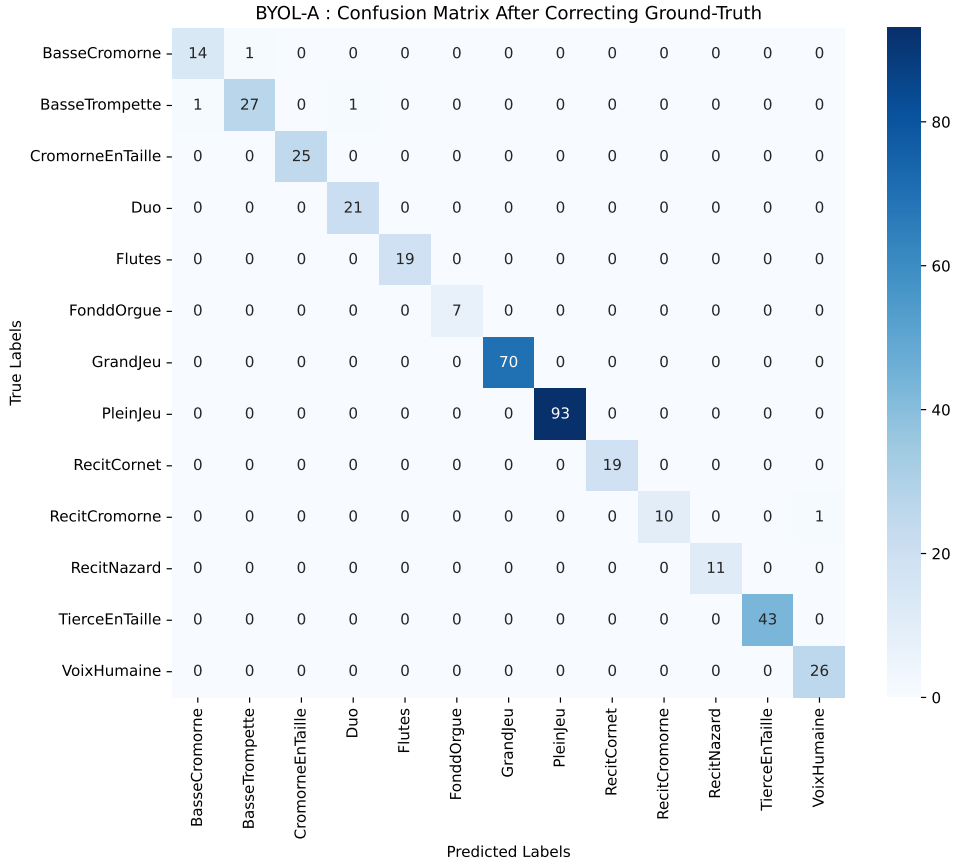


FIGURE 13 – Confusion Matrix for BYOL-A embeddings after correcting ground-truth.

The confusion matrix for CLAP shows different performance characteristics :

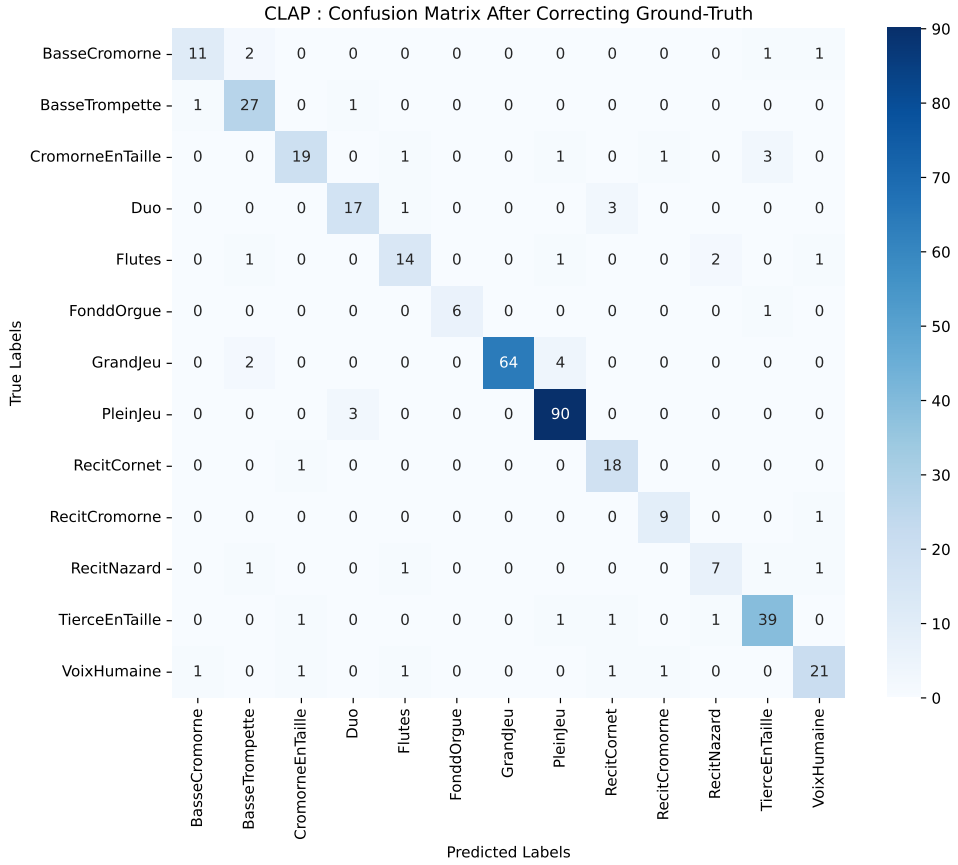


FIGURE 14 – Confusion Matrix for CLAP embeddings after correcting ground-truth.

— **Consistency in Some Classes** : CLAP performs reasonably well in classes such as *Grand Jeu* and *Plein Jeu*, with 64 and 90 true positives respectively. However, there are several misclassifications, indicating that CLAP is less consistent compared to BYOL-A.

— **Significant Misclassifications** : CLAP struggles more with distinguishing between similar classes. For example, there are notable misclassifications between *Basse de Cromorne* and *Basse de Trompette*, reflecting challenges in differentiating these classes.

— **Performance in Low-Support Classes** : CLAP shows more misclassifications in low-support classes compared to BYOL-A. For instance, *Recit Cornet* and *Voix Humaine* have some misclassified instances, showing room for improvement.

BYOL-A demonstrates superior performance compared to CLAP, particularly in accurately classifying challenging and similar classes. While both models benefit from the data cleaning process, BYOL-A's consistent performance and lower misclassification rates make it the more robust embedding method for this classification task. CLAP, on the other hand, shows more variability and struggles with certain classes, highlighting areas where further improvements are needed.

Références

[1] R. Gao et et al. « Instrument Classification with Pre-trained Audio Neural Networks ». In : *Proceedings of the European Signal Processing Conference (EUSIPCO)*. 2020.

[2] Stuart Geman, Elie Bienenstock et Richard Doursat. « Neural networks and the bias/variance dilemma ». In : *Neural Computation* 4.1 (1992), p. 1-58.

[3] C. Hung et et al. « Frame-level Instrument Recognition by Timbre and Pitch ». In : *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2018.

[4] Sergey Ioffe et Christian Szegedy. « Batch normalization : Accelerating deep network training by reducing internal covariate shift ». In : *International conference on machine learning*. PMLR. 2015, p. 448-456.

[5] D.P. Kingma et J.B. Ba. « Adam : A method for stochastic optimization ». In : *arXiv preprint arXiv :1412.6980* (2014).

[6] Q. Kong et et al. « PANNs : Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition ». In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020).

[7] F. Korzeniowski et et al. « Timbre Recognition and Classification in Music Using Attention Mechanisms ». In : *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2019.

[8] V. Lostanlen et et al. « Deep Convolutional Networks for Large-Scale Audio Classification ». In : *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016.

[9] L.J.P. van der Maaten et G.E. Hinton. « Visualizing High-Dimensional Data Using t-SNE ». In : *Journal of Machine Learning Research* 9 (2008), p. 2579-2605.

[10] R. Niizumi et et al. « BYOL-A : Bootstrap Your Own Latent for audio ». In : *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. 2021.

[11] J. Pons et et al. « Timbre Analysis Using Convolutional Neural Networks ». In : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[12] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. « Learning representations by back-propagating errors ». In : *Nature* 323.6088 (1986), p. 533-536.

[13] E. Schubert et al. « DBSCAN revisited, revisited : why and how you should (still) use DBSCAN ». In : *ACM Transactions on Database Systems (TODS)* 42.3 (2017), p. 19. doi : 10.1145/3068335.

[14] Nitish Srivastava et al. « Dropout : A simple way to prevent neural networks from overfitting ». In : *The Journal of Machine Learning Research* 15.1 (2014), p. 1929-1958.

[15] Y. Wu et et al. « CLAP : Contrastive Language-Audio Pretraining ». In : *Proceedings of the International Conference on Machine Learning (ICML)*. 2024.