FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Performance Configuration Analysis in Portuguese Traditional Music: A Computational Approach

Nawaraj Khatri

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Masters in Multimedia

Supervisor: Prof. Gilberto Bernardes

July 31, 2025

Performance Configuration Analysis in Portuguese Traditional Music: A Computational Approach

Nawaraj Khatri

Masters in Multimedia

Approved in oral examination by the committee:

Chair: Prof. António Sá Pinto
External Examiner: Prof. Nádia Margarida Trindade Moura
Supervisor: Prof. Gilberto Bernardes de Almeida

July 31, 2025

Abstract

I present an analysis of performance configurations in Portuguese traditional music, using computational methods to process field recordings from the A Música Portuguesa A Gostar Dela Própria (MPAGDP) archive. Our approach employs YOLOv11s ("You Only Look Once"), a computer vision system that can detect and count performers in archival footage, allowing us to automatically classify performances into meaningful categories: solo, duo, small, and large ensembles. This computational classification method processed 8122 field recordings with 96% classification accuracy, enabling systematic examination of performance contexts that would be time-consuming through manual analysis. Our analysis reveals significant relationships between performance configuration and musical practice across Portuguese traditions. Solo performers, com- prising 48% of vocal recordings, predominantly appear in narrative and poetic traditions requiring individual expression. Large ensembles (21%) maintain collective practices like polyphonic singing tra- ditions. The geographic distribution shows regional traits—Alentejo features large-ensemble singing traditions, while northern regions favor solo performances. The temporal analysis traces how traditional forms maintain continuity through specific performance configurations, while contemporary adaptations emerge primarily in small group formats, illuminating the social dimensions of musi- cal transmission and adaptation in Portuguese traditional music.

Keywords: Performance Configuration Analysis, Portuguese Traditional Music, YOLOv11s, MPAGDP Archive, Ethnomusicology, Computational Musicology

Acknowledgements

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor, Prof. Gilberto Bernardes, for his continuous support, expert guidance, and patience throughout this project. His advice was invaluable at every stage of this research.

I also want to thank the Faculdade de Engenharia da Universidade do Porto for providing the academic environment and resources that made this thesis possible.

My sincere thanks go to Tiago Pereira and the team behind *A Música Portuguesa A Gostar Dela Própria (MPAGDP)* for their work in creating and maintaining the archive that served as the foundation for this study. Without this rich corpus, the research would not have been feasible.

I am grateful for the open-source community, especially the developers of the YOLOv11s model and the various tools used in this project, including OpenCV and yt-dlp, which made the large-scale data processing manageable.

Finally, I would like to thank my family and friends, especially Pedro, Mario, Shyane & João for their unwavering support and encouragement throughout my studies.

Nawaraj Khatri

"Until I began to learn to draw,
I was never much interested in looking at art."

Richard P. Feynman

Contents

List of Figures

List of Tables

Abbreviations

MPAGDP	A Música Portuguesa A Gostar Dela Própria
YOLOv11s	You Only Look Once (version 11, small variant)
HOG	Histogram of Oriented Gradients
CNN	Convolutional Neural Network
SVM	Support Vector Machine
IoU	Intersection over Union
mAP	mean Average Precision
FAIR	Findable, Accessible, Interoperable, Reusable
AI4Culture	Artificial Intelligence for Culture
OpenCV	Open Source Computer Vision Library

Chapter 1

Introduction

1.1 The Cultural Significance of Portuguese Traditional Music

Traditional music serves as a cornerstone of cultural heritage, embodying the narratives, social structures, and collective identities of communities across generations [3]. Primarily transmitted orally and deeply embedded within the fabric of social and communal life, these musical practices represent more than aesthetic expressions; they are forms of "humanly organized sound" that reflect and shape the patterns of human organization [3, p. 25]. Ethnomusicological scholarship has long advocated for a holistic analysis that considers the music inseparable from the "total event of producing and hearing" it, thereby acknowledging its profound connection to its cultural context [18]. In an era of increasing globalization, these unique cultural forms face a heightened vulnerability to dilution and disruption, making their documentation and study a critical task for cultural preservation [1, 6].

In Portugal, this musical heritage comprises a rich and varied tapestry of styles, each carrying distinct historical and regional significance. The scope of Portuguese traditional music is vast, illustrated by prominent examples such as *Fado*, *Cante Alentejano*, and *Chamarrita*. *Fado*, an urban tradition, is deeply connected to the affective politics and social life of cities like *Lisbon*, often expressing themes of longing and fate [9]. In contrast, *Cante Alentejano*, recognized by UNESCO as Intangible Cultural Heritage, is a polyphonic singing tradition performed by large, non-professional male choirs, reflecting the communal and agrarian lifeways of the *Alentejo* region [8]. Other forms, such as the narrative traditions of northern Portugal, often privilege solo performance, highlighting the importance of individual expression in storytelling contexts [14]. This diversity underscores how different musical practices are not merely artistic choices but are fundamentally interwoven with regional identity, social organization, and historical memory, making archives like *A Música Portuguesa A Gostar Dela Própria (MPAGDP)* essential for their continued preservation and scholarly inquiry [14].

1.2 The Challenge of Digital Archives: The *MPAGDP* Corpus

The proliferation of digital technologies has catalyzed a paradigm shift in the field of ethnomusicology, enabling the creation of large-scale archives dedicated to the preservation and dissemination of global musical traditions. For example, digital repositories, such as the UCLA Ethnomusicology Archive with its collection of more than 150,000 items, serve as critical safeguards for intangible cultural heritage, which is increasingly vulnerable in a globalized world[21]. However, the scale of these collections introduces significant logistical and analytical challenges. Although digitization offers unprecedented opportunities for access and research, the sheer volume of data often outpaces the capacity for manual cataloging and analysis, creating a bottleneck that can limit the scholarly utility of these invaluable resources [15], [7]. Similarly, *A Música Portuguesa A Gostar Dela Própria (MPAGDP)* stands as a vital initiative within this landscape, representing a comprehensive audiovisual archive of contemporary Portuguese traditional music and expressive culture. Founded by filmmaker Tiago Pereira, this non-profit project has amassed an extensive corpus of over 8,000 field recordings since its inception in 2011 [14]. The collection is a living repository documenting a wide array of cultural practices, including music, dance, poetry, and oral histories in Portugal and its diaspora. Each entry in the *MPAGDP* database is accompanied by rich descriptive metadata that document attributes such as genre, instrumentation, geographic location, and performers. This information is crucial because it provides the necessary context for retrieval and enables researchers to filter and navigate the vast collection.

Despite the richness of its existing metadata, the *MPAGDP* corpus exemplifies the central challenge of modern digital archives: its scale makes a comprehensive manual study infeasible. The creation of detailed metadata is a time-consuming and costly endeavor, often accounting for a significant portion of a digitization project's budget [15]. Consequently, while the archive's metadata provides an essential foundation, it cannot capture every salient feature of the recorded performances. The immense volume of audiovisual material necessitates the application of computational methodologies to unlock deeper musicological insights that lie dormant within the data, beyond what is explicitly described in the catalog. This analytical gap highlights the need for innovative approaches to explore latent patterns in the collection, thus transforming a passive repository into a dynamic site for scholarly discovery.

1.3 The Research Problem: Performance Configuration as an Analytical Lens

While the *MPAGDP* archive contains rich descriptive metadata concerning genre, instrumentation, and geography, a critical dimension of performance practice remains systematically uncaptured: the performance configuration. Defined as the number and spatial arrangement of performers within a musical event, this variable offers a powerful yet underexplored analytical lens for ethnomusicological inquiry. The configuration of an ensemble is not merely a logistical detail but a fundamental component of musical practice that can reveal significant insights into the social organization of musical transmission, regional performance conventions, and the intrinsic relationship between ensemble size and musical genre.

The musicological importance of this parameter is rooted in the understanding that music is "humanly organized sound," where patterns of human organization directly relate to the sonic structures produced [3, p. 25]. A systematic analysis of performance configurations across the *MPAGDP* archive allows for the formulation and testing of specific hypotheses about Portuguese traditional music. For instance, it can be hypothesized that certain narrative traditions, particularly those prevalent in northern Portugal, favor solo performances to emphasize the role of individual expression and storytelling [14]. Conversely, polyphonic singing traditions, such as the UNESCO-recognized *Cante Alentejano*, by their very nature, necessitate large ensembles, reflecting the communal and social functions embedded in the practice [8].

The absence of this data point within the existing metadata constitutes a significant research gap. Without a method to systematically classify and analyze ensemble size, researchers are limited in their ability to explore how musical practices are shaped by their social contexts across different Portuguese regions. Manually reviewing and annotating over 8,000 video recordings is a prohibitively time-consuming task, making a large-scale analysis infeasible. This challenge underscores the research problem: a crucial layer of musicological data is inaccessible due to the scale of the archive, necessitating a novel computational approach to automatically extract and analyze performance configurations, thereby unlocking deeper insights into the cultural fabric of Portuguese musical life.

1.4 Research Motivation, Questions, and Objectives

The central motivation for this research arises not solely from the scale of the *MPAGDP* digital archive, but from the limitations of prior computational studies that have attempted to analyze large folk music corpora. While the *MPAGDP* corpus offers an unprecedented resource for the study of Portuguese traditional music, its vastness of over 8,000 recordings, renders comprehensive manual analysis of its visual content impractical.

Previous efforts to apply computer vision for human detection have typically relied on generic object detection models or have been limited to small, curated datasets, often failing to account for the unique visual and contextual complexities of traditional music performance environments. For example, classical computer vision techniques such as Haar Cascades and HOG have struggled with occlusions, low-light conditions, and culturally specific staging, while early deep learning approaches have rarely been domain-adapted or validated at scale [23]. Consequently, these studies have not produced tools capable of robust, large-scale, and musicologically meaningful analysis of ensemble practices in living folk traditions. This research addresses these shortcomings by developing a domain-adapted, scalable performer detection system and applying it to the entirety of the *MPAGDP* archive, thereby facilitating new forms of quantitative, evidence-based musicological inquiry.

To resolve these challenges, the thesis adopts dual methodological frameworks; technical and musicological, through distinct research questions. From a technical perspective, it investigates how deep learning models, specifically the *YOLOv11s* architecture, can be fine-tuned to automatically and accurately classify performance configurations (solo, duo, small group, large group) within the complex visual environments of large-scale traditional music video archives. This will be achieved through the development of a specialized annotated dataset extracted from the MPAGDP archive, followed by iterative fine-tuning and evaluation of a YOLOv11s-based detection system using metrics including F1-score, mAP, and overall accuracy.

From a musicological perspective, the research examines quantitative relationships between performance configurations and musical genres within the *MPAGDP* corpus, assessing whether computational analysis confirms or extends qualitative understandings of ensemble size in traditions like narrative forms or collective polyphony. It further explores the existence of statistically significant regional patterns in performance configurations across Portugal, investigating how such distributions illuminate localized cultural practices, historical developments, and social organization of music-making. Methodologically, this involves applying computationally derived performance data to analyze genre-geography correlations and conducting statistical and interpretive analysis of regional and genre-based ensemble patterns.

To guide the empirical analysis, three hypotheses are hypothesized: first, that large ensembles (>5 performers) cluster significantly in Alentejo, reflecting polyphonic traditions like *Cante Alentejano*; second, that solo configurations dominate Northern Portugal's narrative traditions, aligning with its emphasis on individual expression; and third, that a domain-adapted computer vision model will achieve high accuracy in classifying *MPAGDP* performance configurations, enabling reliable large-scale musicological analysis. This integrated approach aims to transcend prior technical and disciplinary constraints, establishing a new paradigm for computationally driven inquiry into Portuguese traditional music.

1.5 Thesis Outline

This thesis is structured into six chapters to systematically present the research from its conceptual foundations to its musicological conclusions and future implications.

- **Chapter 1:** Introduction provides a comprehensive overview of the research, establishing the cultural context of Portuguese traditional music, the challenges posed by large-scale digital archives like *MPAGDP*, the specific research problem, and the study's core motivation, questions, and objectives.

- **Chapter 2:** Background and Related Work reviews the theoretical and technical literature that underpins this study. It situates the concept of performance configuration within ethnomusicological theory and provides a historical survey of computational object detection, culminating in the rationale for selecting the *YOLOv11s* model.

- **Chapter 3:** Corpus and Data Methodology details the research corpus derived from the MPAGDP archive. This chapter outlines the data collection process, characterizes the dataset's genre and geographic distributions, justifies the performance configuration classification system, and describes the frame extraction and sampling strategy.

- **Chapter 4:** Implementation of the Performer Detection System offers a detailed technical account of the model's development. It describes the iterative fine-tuning process, presents the final model's performance metrics, and validates the system's overall accuracy, confirming its suitability for large-scale analysis.

- **Chapter 5:** Musicological Analysis, Results, and Discussion presents and interprets the findings. This chapter analyzes the quantitative results for each performance configuration category, examines their regional distribution, and discusses how these patterns reflect the social structure, regional identity, and dynamics of cultural transmission in Portugal.

- **Chapter 6:** Conclusion and Future Work summarizes the study's key findings and contributions. It acknowledges the research's limitations and proposes concrete directions for future work, including multimodal analysis, model enhancements, and cross-cultural applications.

Chapter 2

Background and Related Work

2.1 Performance Configuration in Ethnomusicological Analysis

The concept of performance configuration, the number and spatial arrangement of performers within a musical event, has long been recognised as a critical analytical dimension in ethnomusicology [3], [18]. Foundational theorists have argued that ensemble size and structure are not merely logistical details but are deeply intertwined with musical form, social organisation, and cultural meaning. Blacking (1973) famously posited that there is a direct relationship "between patterns of human organisation and the patterns of sound produced" [3, p. 25], while Seeger (1987) emphasised the inseparability of musical sound from the total event of its production and reception. [18]

The profound cultural significance of performance configuration is exemplified within Portuguese traditional music, where ensemble size functions as a tangible manifestation of regional identity and social organisation. The *Cante Alentejano* tradition, recognised by UNESCO demonstrates this principle clearly: characterised by large, non-professional male choirs, this configuration is inseparable from both the polyphonic musical texture and the communal, agrarian heritage of the Alentejo region [8]. The collective nature of these performances embodies values of solidarity and shared history, with the choir's formal structure, comprising distinct vocal parts such as *ponto*, *alto*, and *baixo*, directly mirroring the social fabric of rural communities. Conversely, in the narrative and poetic traditions prevalent in Northern Portugal, solo performance constitutes the dominant configuration, emphasising individual expression and the clarity of storytelling [14]. Here, the solitary performer serves as both narrator and custodian of communal memory, with the solo format facilitating improvisational freedom and personal interpretation. These contrasting configurations—large choirs in the south and soloists in the north—illustrate how ensemble size in Portuguese traditional music represents a manifestation of broader cultural patterns and regional histories rather than arbitrary musical choices.

Despite the centrality of ensemble configuration to musical practice, as demonstrated by such ethnographic examples, quantitative, cross-tradition studies of the number of instrumentalists or performers remain remarkably scarce. Most ethnomusicological literature has focused on qualitative accounts or genre-specific descriptions, leaving a significant gap in systematic, data-driven analysis of ensemble size across diverse traditions. While standard reference works document typical ensemble sizes for Western classical music (e.g., string quartets, orchestras) and jazz (e.g., trios, big bands), comparable quantitative surveys are lacking for many folk and traditional musics. This scarcity is particularly evident in the context of large, heterogeneous digital archives, where manual annotation of ensemble size is prohibitively time-consuming [13].

Recent computational studies have begun to address this methodological gap by applying machine learning to the quantitative analysis of ensemble size. Grollmisch et al. (2019) developed a supervised deep learning approach to classify the number of performers in Colombian Andean string music, using a corpus of 150 annotated recordings and applying convolutional neural networks to spectrogram representations to distinguish between solo, duet, trio, and quartet performances. They reported up to 80.7% accuracy in file-wise classification, demonstrating the feasibility of computational ensemble size estimation even in challenging, non-Western traditional contexts [10]. This work stands out as a rare example of quantitative, tradition-specific analysis of ensemble size using modern computational methods.

Within Western art music, Davidson and Good (2002) conducted an in-depth observational study of string quartets to explore how social and musical coordination is shaped by fixed ensemble size. Their corpus consisted of video-recorded rehearsals and performances by established quartets, and their methodology combined qualitative analysis of rehearsal strategies with quantitative measures of coordination (e.g., timing, gesture). Their findings demonstrated that ensemble size—fixed at four—directly impacts both musical outcomes and the social dynamics of performance, with coordination and communication emerging as central to group success. While their approach is primarily qualitative, it underscores the musicological significance of ensemble size as a lens for understanding group interaction [5].

From a technical perspective, advances in computer vision provide important methodological precedents for large-scale performer detection, even when not directly applied to musicological contexts. Kim et al. (2020) systematically compared state-of-the-art object detection models—namely, Faster R-CNN, YOLO, and SSD—for real-time vehicle type recognition, evaluating these architectures on large-scale, real-world video datasets and assessing their accuracy, speed, and robustness under varied conditions. Their results demonstrated that YOLO-based models offered the optimal trade-off between detection speed and accuracy, making them particularly suitable for large-scale, real-time applications where computational efficiency is crucial [12]. Although their research focused on vehicle detection rather than musical performers, the technical insights and benchmarking they provide inform the selection and adaptation of object detection models for automated performer identification in traditional music video archives.

Despite these advances, there remains a marked scarcity of large-scale, quantitative studies of ensemble size in folk and traditional musics, especially within the context of digital archives.

Most prior work has concentrated on Western classical or popular music, with limited attention to the diversity of ensemble practices in regional and oral traditions. This gap is particularly significant, as ensemble size is often closely tied to genre, region, and social function—factors that are central to the musicological understanding of tradition [13]. The Portuguese examples cited above underscore how performance configuration serves as both a reflection of and influence upon the social dynamics of musical practice, yet such insights have rarely been investigated systematically at scale.

The present research directly addresses this gap by developing and applying a domain-adapted performer detection system to over 8,000 field recordings from the MPAGDP archive. By leveraging a fine-tuned YOLOv11s model, this study enables the first large-scale, quantitative analysis of performance configuration in Portuguese traditional music. This approach not only builds upon the methodologies of prior computational studies but also introduces a scalable, systematic framework for ensemble size analysis in a domain that has historically relied on qualitative or anecdotal evidence.

In summary, whilst the theoretical importance of performance configuration is well-established in ethnomusicology, systematic, quantitative studies of ensemble size, especially in folk and traditional contexts, are only beginning to emerge. This thesis contributes to filling this gap by integrating computational methods with musicological inquiry, providing new empirical insights into the social and cultural dynamics of Portuguese traditional music.

2.2 Technical Approaches to Performer Detection

The computational task of detecting performers in video has evolved substantially, moving from classical computer vision techniques to contemporary deep learning models. Early approaches to object detection relied on methods that used manually engineered features, such as Haar Cascades and the Histogram of Oriented Gradients (HOG). Haar Cascades, for example, use feature-based classifiers in a cascaded structure to rapidly discard background regions of an image and focus on object-like areas [23]. The HOG method operates by counting occurrences of gradient orientation in localized portions of an image, which proved effective for human detection [4]. However, these classical methods often falter when faced with the complexities of real-world recordings. They are particularly vulnerable to challenges common in the field recordings of the MPAGDP archive, including cluttered backgrounds, variable and poor lighting conditions, and frequent occlusions where performers are partially obscured.

The limitations of these traditional methods precipitated a shift towards deep learning, particularly the use of Convolutional Neural Networks (CNNs). CNNs revolutionized object detection by automating the critical process of feature extraction. Instead of relying on hand-crafted features, these networks learn a hierarchy of increasingly complex features directly from the data, which results in significantly improved accuracy and robustness, especially in unstructured and varied environments. Initial deep learning models for object detection, such as the two-stage detector

Faster R-CNN, achieved high accuracy by first proposing regions of interest and then classifying them [17]. While powerful, such models often demand significant computational resources, rendering them less practical for processing large-scale video archives.

In response to these computational demands, the YOLO (You Only Look Once) family of models was developed. YOLO models reframe object detection as a single regression problem, directly predicting bounding boxes and class probabilities from full images in one pass. This one-stage approach offers a state-of-the-art balance between speed and accuracy, a crucial consideration for large-scale analysis of video archives [12]. Each iteration of the YOLO architecture has introduced significant improvements. This research employs YOLOv11s, a recent advancement in this lineage. The YOLOv11 model incorporates enhanced spatial attention mechanisms, which improve its ability to detect objects in challenging conditions such as low light and cluttered scenes—environmental factors frequently encountered in the *MPAGDP* field recordings. Furthermore, its architecture offers improved feature extraction that better handles occlusions and overlapping performers, making it particularly well-suited for this study [22]. The choice of the lightweight "s" variant ensures efficient processing across the extensive archive without sacrificing significant detection capability.

While general object detection has advanced rapidly, relatively little research has specifically addressed performer detection within the distinct contexts of traditional music, where visual complexity and cultural specificity present unique challenges [20]. This underscores the need for specialized models and domain-specific datasets tailored to folk music performance environments.

2.3 Identifying the Research Gap

A review of the literature reveals a significant disconnect between the theoretical imperatives of ethnomusicology and the practical capabilities of computational musicology. While ethnomusicological theory has long recognized performance context as central to musical meaning [3],[18], the field has lacked the methodological tools to analyze this dimension at scale within large digital archives. The rise of computational methods offers a potential solution, yet a "semantic gap" often persists between the research questions posed by humanistic inquiry and the analytical capabilities of computer science [2]. This study is positioned at the intersection of these fields, addressing three specific, interrelated gaps.

First, there is a methodological gap within ethnomusicology. The discipline's traditional reliance on qualitative and manual methods of analysis, while providing deep contextual understanding, is insufficient for engaging with the sheer volume of data in archives like MPAGDP, which contains 8,122 recordings at the time of retrieval for this thesis in March 2025. This scale presents a "massive barrier for conducting computer-aided music research," where direct contact with data owners is often the only way to access and interpret information [11]. This research directly addresses this gap by introducing a scalable computational methodology that enables the systematic analysis of a core musicological concept—performance configuration—across an entire large-scale corpus.

And, there is a technical gap in computational musicology and computer vision. Most existing object detection models are generalist and have not been specifically adapted for the unique visual complexities of folk music performance. Research shows that models trained on specific-domain data outperform generalist models, particularly for culturally diverse or underrepresented genres [19]. Prior research has identified gaps in computational tools for analyzing Portuguese traditional musical components, particularly in visual domains [20]. This study addresses these gaps by fine-tuning YOLOv11s on a domain-specific dataset to detect performers in challenging conditions—such as occlusions, variable lighting, and culturally specific staging—thereby creating a specialized tool for musicological research.

Also, there is a consequential data gap. The development of specialized models is contingent on the availability of high-quality, annotated, domain-specific datasets. Such a resource for Portuguese traditional music performance does not currently exist. A significant part of this thesis's contribution is the creation of precisely such a dataset through an iterative and model-assisted labeling process. This not only facilitates the present study but also provides a foundational resource for future computational research in this area, aligning with the growing call for better data management and FAIR (Findable, Accessible, Interoperable, Reusable) principles in ethnomusicology.

By simultaneously addressing these methodological, technical, and data-related gaps, this thesis aims to demonstrate how a synergistic approach, integrating deep learning with ethnomusicological theory, can unlock new layers of understanding from cultural heritage archives.

2.4 Limitations of the YOLO Family for Performer Detection

Despite the advantages of the YOLO architecture, its application to performer detection in ethnomusicological video archives is not without limitations. Known failure modes include difficulty in detecting small objects—such as distant or partially occluded performers—and susceptibility to viewpoint bias, where the model's performance degrades with non-frontal or unconventional camera angles. These constraints are particularly salient in field recordings, where staging and camera work are highly variable. To mitigate these issues, this study implemented an active-learning loop: ambiguous cases and low-confidence predictions were iteratively identified, manually annotated, and reintroduced into the training set, thereby improving model robustness in challenging scenarios and reducing the impact of YOLO's inherent biases.

2.5 Ethical Considerations and Data Governance

The deployment of computer vision models in cultural heritage archives raises critical ethical questions concerning performer identity, consent, and algorithmic bias. Many recordings in the MPAGDP archive feature identifiable individuals, often in community or religious contexts, necessitating careful consideration of privacy and informed consent, particularly when automated systems are used for large-scale analysis. Furthermore, algorithmic bias—arising from imbalanced training data or cultural specificity—can result in systematic misrepresentation of certain groups or practices. In line with the AI4Culture guidelines, 2024, this research adheres to principles of responsible data governance: all data processing was conducted within the ethical framework established by the archive, with explicit attention to minimizing the risk of re-identification and ensuring that outputs are used solely for scholarly and preservation purposes.

Chapter 3

Corpus and Data Methodology

3.1 The MPAGDP Archive as a Research Corpus

The empirical foundation for this study is the comprehensive audiovisual corpus of the *A Música Portuguesa A Gostar Dela Própria* (MPAGDP) archive, a non-profit initiative dedicated to the documentation and preservation of contemporary Portuguese traditional music and expressive culture [14]. The dataset was constructed from the entire collection available at the time of research in March 2025, comprising 8,122 catalogued entries. Using the yt-dlp[1] Python library, 8,120 of these entries were successfully downloaded from their hosting on the Vimeo platform; two entries were excluded due to unavailability. The inclusion criteria were deliberately exhaustive, incorporating all audiovisual materials catalogued by MPAGDP—including musical performances, poetry recitations, dances, prayers, choral works, interviews, and other forms of expressive culture—irrespective of genre, style, or region. The collection spans from 2011 to 2025, offering a significant longitudinal perspective on contemporary Portuguese cultural practices. Each entry in the archive is accompanied by a rich set of descriptive metadata that includes genre classification, instrumentation, recording location, date, and performer information, which provides the necessary context for subsequent analysis. To characterize the corpus, we analyzed this metadata, revealing a highly heterogeneous collection that reflects both the diversity of Portuguese cultural heritage and the curatorial focus of the MPAGDP project.

[1] `https://github.com/yt-dlp/yt-dlp`, last accessed April 2025.

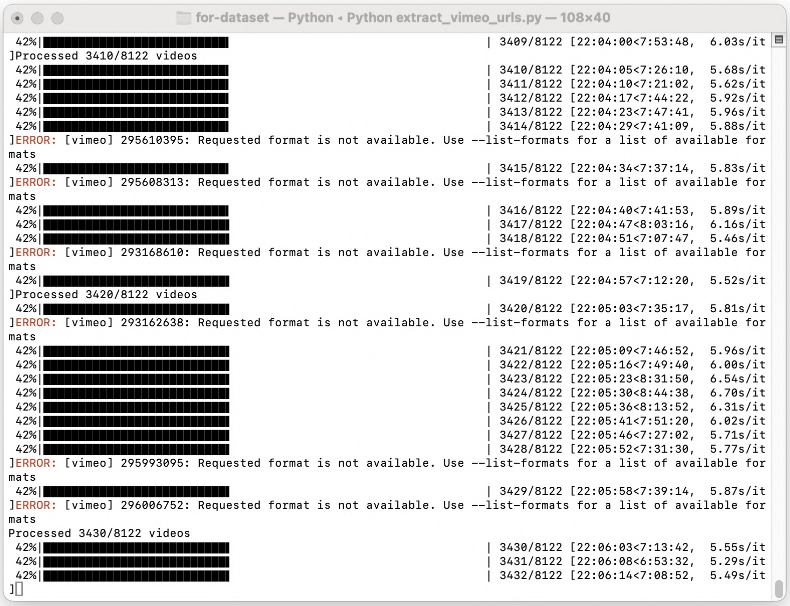

Figure 3.1: Downloading of Videos from Vimeo using *yt-dlp* python library

MPAGDP is not simply a static dataset but a dynamic, web-based digital archive that has, since its founding by Tiago Pereira in 2011, grown into one of the most comprehensive repositories of Portuguese music, encompassing over 8,000 catalogued field recordings from across Portugal and its diaspora. The archive's mission is not only to preserve but also to revitalize Portuguese musical traditions, making them accessible to both scholars and, crucially, the general population [14]. Designed as a living and continually expanding resource, MPAGDP reflects the evolving nature of Portuguese expressive culture and prioritizes public engagement and cultural transmission over academic specialization.

Access to the MPAGDP archive is provided through a public-facing website, which offers an intuitive and visually engaging interface. Users can browse and search the collection using a variety of filters, including genre (such as narrative music, secular/festive, instrumental, poetry, and dance), geographic region (covering all major areas of Portugal, as well as the Azores, Madeira, and select international locations), performer or group name, date of recording, and instrumentation. Each entry typically includes a video recording, hosted on Vimeo, accompanied by descriptive metadata. This metadata encompasses genre classification, performer and group names, instrumentation, recording location, and date, and often features supplementary notes that provide cultural or contextual background. The interface is designed to encourage exploration and discovery among a broad audience, supporting MPAGDP's mission to foster widespread appreciation and participation in Portuguese musical heritage.

While the breadth and richness of the MPAGDP archive make it an invaluable resource for ethnomusicological research, its structure as a public-facing, web-based service introduces several limitations for scholarly inquiry. The metadata, though extensive, is not always uniform or complete; entries can vary significantly in the detail and accuracy of their genre, instrumentation, and performer information. Certain variables of interest to researchers, such as precise performer counts, standardized instrument tags, or consistent genre taxonomies, are not systematically captured. Furthermore, the web interface is optimized for human browsing rather than bulk data extraction, and there is currently no public API or downloadable structured dataset. As a result, researchers must rely on custom scripts and manual workflows to collect and standardize data at scale, a process that is both time-consuming and prone to error.

It is important to note that MPAGDP maintains an internal Excel database containing structured metadata, including artist names, genres, locations, instruments used, and dates of recording. However, this internal database is not publicly available; researchers interested in accessing this resource must contact the MPAGDP team directly to request permission. In my case, with the support and guidance of my supervisor, Prof. Gilberto Bernardes, I was granted access to this internal Excel database. This privileged access greatly facilitated the data extraction and preprocessing stages of my research, enabling a more systematic and comprehensive analysis than would have been possible using only the public-facing website.

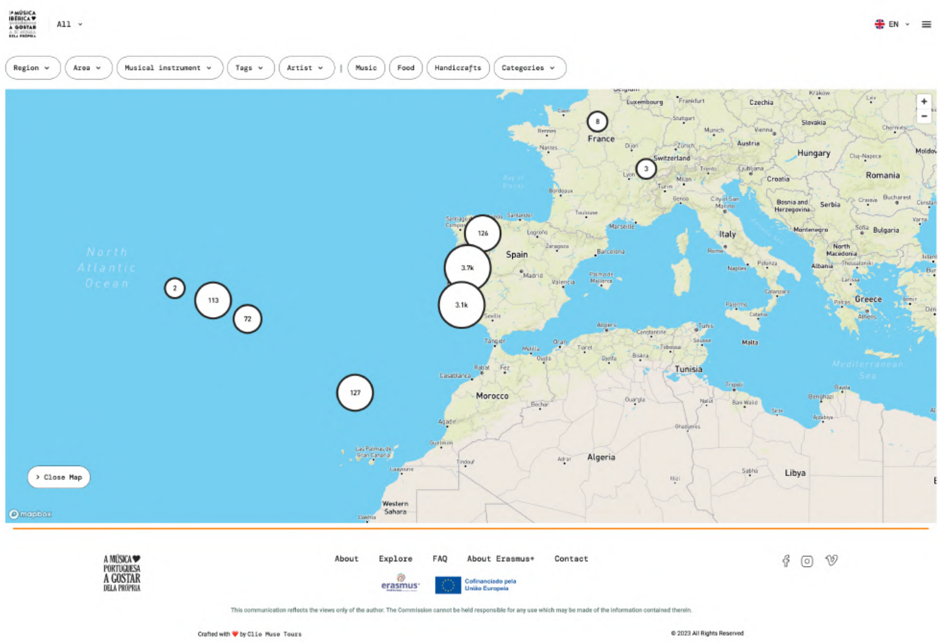

Figure 3.2: MPAGDP website's interactive map interface, illustrating geographic clustering of recordings and the available filters for exploring Portuguese traditional music by region, genre, instrument, and other metadata.

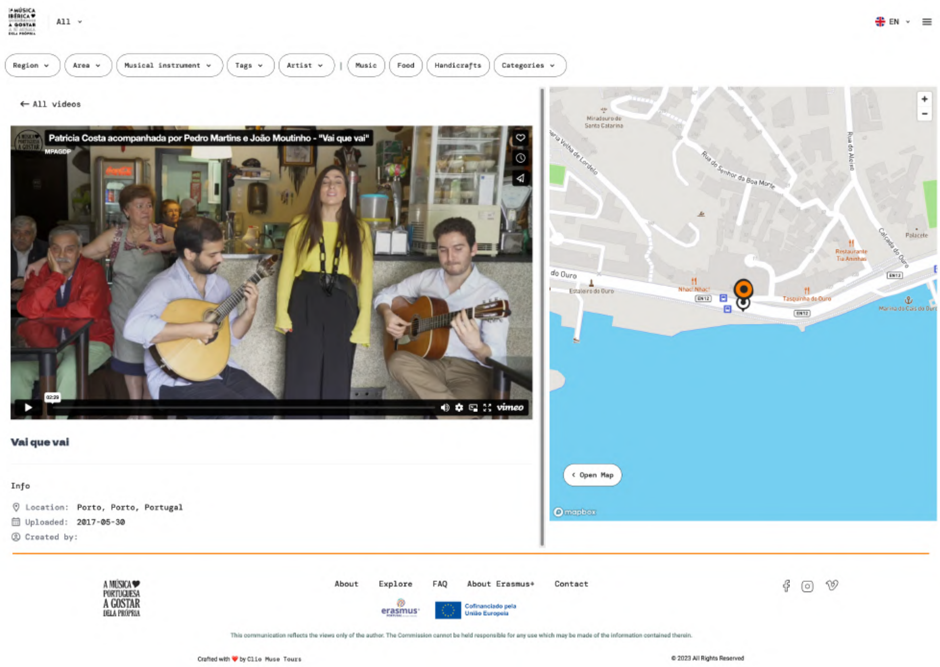

Figure 3.3: MPAGDP website's video page, illustrating how each recording is presented alongside detailed metadata and an interactive map showing the exact geographic location of the performance.

The archive's ongoing, community-driven curation also introduces potential biases in regional, genre, or performer representation, reflecting the priorities and fieldwork opportunities of its creators rather than a systematic sampling of Portuguese musical life. These characteristics underscore the need for more parsed and structured data formats to support rigorous computational and statistical research. The current embedding of metadata within web pages, rather than in machine-readable formats such as CSV or JSON, presents a significant barrier to reproducibility, scalability, and integration with computational tools. More structured data would not only facilitate automated and reproducible data extraction but also enable deeper and more systematic musicological inquiry, aligning the archive with FAIR (Findable, Accessible, Interoperable, Reusable) data principles. For large-scale, cross-sectional, or longitudinal studies—such as the present analysis of performance configurations—these improvements are essential to unlock the full potential of the MPAGDP archive as a foundation for computational ethnomusicology.

Each entry in the archive is accompanied by a rich set of descriptive metadata that includes genre classification, instrumentation, recording location, date, and performer information, which provides the necessary context for subsequent analysis. To characterize the corpus, we analyzed this metadata, revealing a highly heterogeneous collection that reflects both the diversity of Portuguese cultural heritage and the curatorial focus of the *MPAGDP* project.

3.1.1 Genre and Geographic Distribution

We conducted an analysis of the genre metadata which reveals a significant concentration in specific forms of expression. "*Música narrativa*" (narrative music) is the most prevalent genre, accounting for 3,721 entries, which underscores the importance of storytelling traditions within the corpus. Other prominent genres include "*Música secular/festiva/comunidade*" (secular/festive/-community music) with 793 entries, "*Música instrumental*" with 744 entries, "*Poesia*" (poetry) with 429 entries, and "*Dança*" (dance) with 409 entries. This distribution highlights the archive's focus on a wide range of community-based and individual expressive forms.

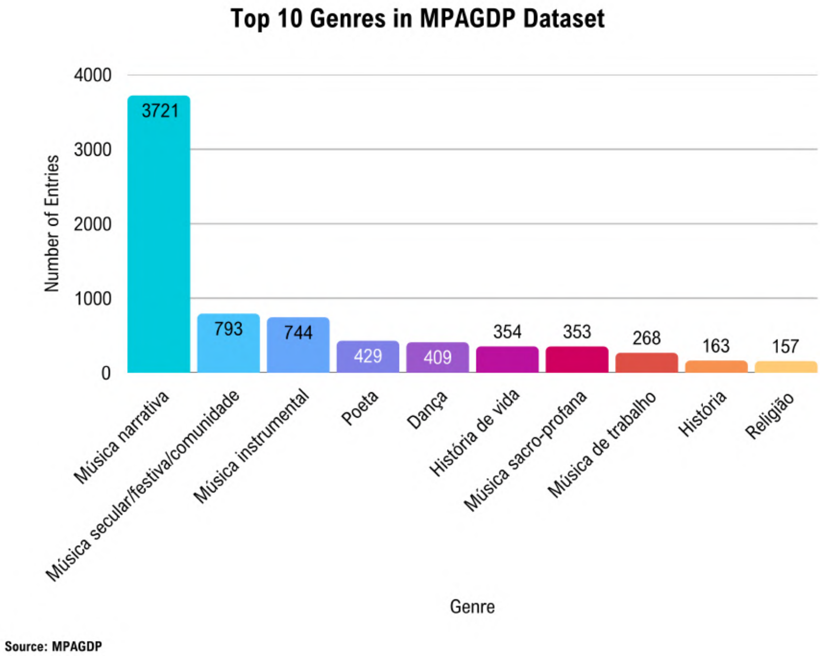

Figure 3.4: Top 10 Genres in MPAGDP Dataset

Geographically, the dataset is dominated by recordings from *Estremadura* (1,242), *Beira Litoral* (1,122), and *Alentejo* (1,037), with additional significant representation from *Centro*, *Baixo Alentejo*, *Trás-os-Montes*, *Alto Douro*, and other regions, including the Azores and *Madeira*. International recordings are also present, comprising about 0.5% (41) of the dataset (excluding Spain) and about 3.5% (291) of the dataset (including Spain). This comprehensive dataset provides the foundation for systematic analysis of performance configuration and its correlation with genre, region, and other contextual variables, allowing robust musicological inquiry into the dynamics of Portuguese traditional and contemporary expressive culture.

3.2 Performance Configuration Classification

A central methodological contribution of this thesis is the development of a four-part classification system to systematically analyze ensemble size within the MPAGDP corpus. While Music

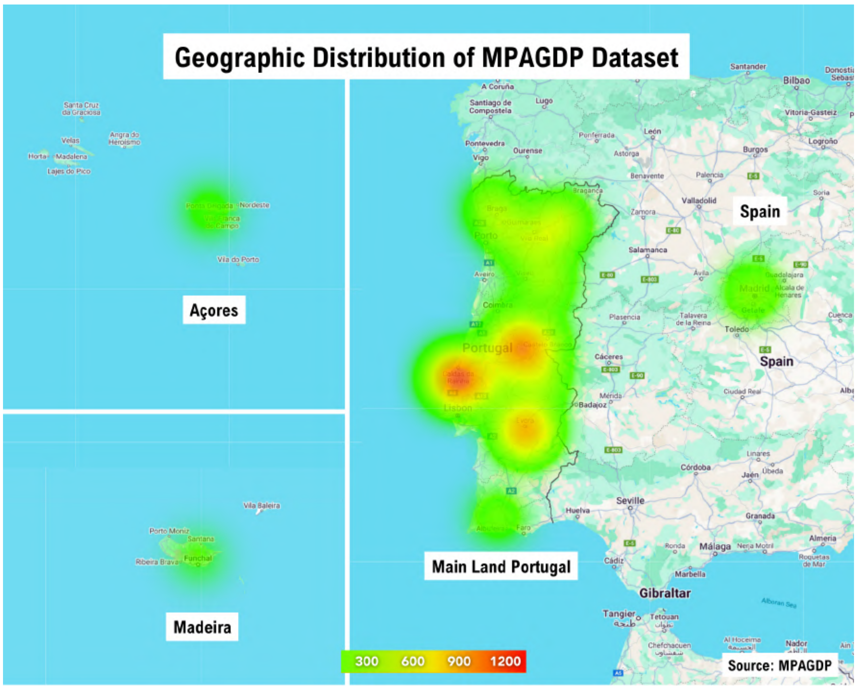

Figure 3.5: Geographic Distribution of MPAGDP Dataset

Performance Analysis has traditionally focused on parameters such as timing and dynamics, this study extends the analytical framework to include performance configuration as a critical variable that shapes musical expression and cultural transmission [13]. This taxonomy is not merely a quantitative grouping but is founded on musicological rationales that reflect distinct social and performance dynamics observed in Portuguese traditional music. The four categories are defined as follows:

- **Solo (1 performer):** This category represents the fundamental unit of individual expression. Solo performances are paramount in traditions where textual clarity and personal interpretation are prioritized, such as in narrative genres (*estória*, *poema*) and devotional practices (*oração*). The solo format allows for maximum improvisational freedom and a direct, intimate mode of communication with the audience, embodying the role of the individual as a vessel for cultural memory, as seen in the tradition of the solitary "*cantadeira*" [14].

- **Duo (2 performers):** The duo configuration marks the transition from individual to interactive music-making. It maintains the intimacy of solo performance while introducing harmonic, rhythmic, and dialogic complexity. This format is characteristic of harmonized singing (cantiga, fado) and is essential for competitive traditions like the desgarrada, where two performers engage in a musical dialogue of improvised verse. It represents a distinct social dynamic, balancing individual contribution with dyadic coordination.

- **Small Group (3–5 performers):** This category represents a significant threshold where collective coordination becomes more complex and musical roles often become more differentiated. Small groups allow for greater instrumental and textural diversity while retaining an

Figure 3.6: Frames extracted as a dataset using Open-CV

intimate character reminiscent of chamber music. Importantly, this configuration frequently serves as a flexible vehicle for the adaptation of traditional forms and for stylistic innovation, blending folk idioms with contemporary genres like rock. This aligns with studies of small ensembles that emphasize the importance of social coordination in performance [5].

- **Large Group (more than 5 performers):** This configuration is strongly associated with communal expressions of cultural identity, ceremonial functions, and formalized musical structures. Large groups are characteristic of collective vocal traditions (*coro*, *polifonia*) and community-based ensembles like *ranchos*. The UNESCO-recognized *Cante Alentejano* is a prime example, where the large, non-professional choir and its formal vocal structure (*ponto*, *alto*, *baixo*) embody the region's history of agrarian collectivism [8]. In this context, the performance configuration is a direct reflection of the social organization it represents [3].

This classification system provides a robust framework for systematically quantifying and analyzing performance contexts across the entire MPAGDP archive, enabling new insights into the relationship between ensemble size and musical practice that would be prohibitively time-consuming to identify through manual review.

3.3 Frame Extraction and Sampling Strategy

To facilitate computational analysis of performance configuration across the 8,120 videos in the corpus, a systematic frame extraction and sampling strategy was implemented. Given that analyzing every frame of every video would be computationally prohibitive and largely redundant, a representative sample was required for each recording. This process was automated using the OpenCV[2] Python[3] library, a standard tool for computer vision tasks. The sampling method was designed to capture a representative cross-section of each performance while mitigating potential biases. For each video, five distinct frames were extracted. A crucial constraint was imposed: each sampled frame had to be separated by at least 10 seconds of video footage. This temporal separation was essential for several reasons:

- Ensuring Representativeness: Sampling at intervals across the duration of a performance provides a more accurate snapshot of the typical ensemble configuration than analyzing frames clustered at the beginning or end of a recording.
- Mitigating Redundancy from Static Cameras: In many field recordings, the camera position is static for long periods. The 10-second interval ensures that the extracted frames are not nearly identical, thus providing more distinct data points for analysis.

[2] https://github.com/opencv/opencv, last accessed April 2025.
[3] https://www.python.org/, last accessed April 2025.

Figure 3.7: Samples of Frames from Dataset

- Capturing Dynamic Events: The strategy increases the likelihood of capturing variations in performance configuration that may occur within a single recording, such as a soloist being joined by other performers partway through a piece.
- Handling Camera Movement: For recordings that involve camera panning or movement, this spaced sampling method helps to capture different angles and compositions of the performance space.

Figure 3.7 illustrates a selection of frames extracted using this methodology, showcasing the diversity of visual contexts within the MPAGDP archive. This strategy yielded a total dataset of 40,600 frames (8,120 videos × 5 frames), creating a robust foundation for the training and application of the performer detection model.

In accordance with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and to respect the copyright of the MPAGDP archive and the privacy of the performers, the dataset of 40,600 frames is not published directly. Instead, to ensure full reproducibility, all resources required to programmatically reconstruct the dataset are made publicly available in the project's GitHub repository[4].

[4]The project repository containing all code and metadata for dataset reconstruction can be found at: https://github.com/hellonrk/performance-configuration-yolov11s

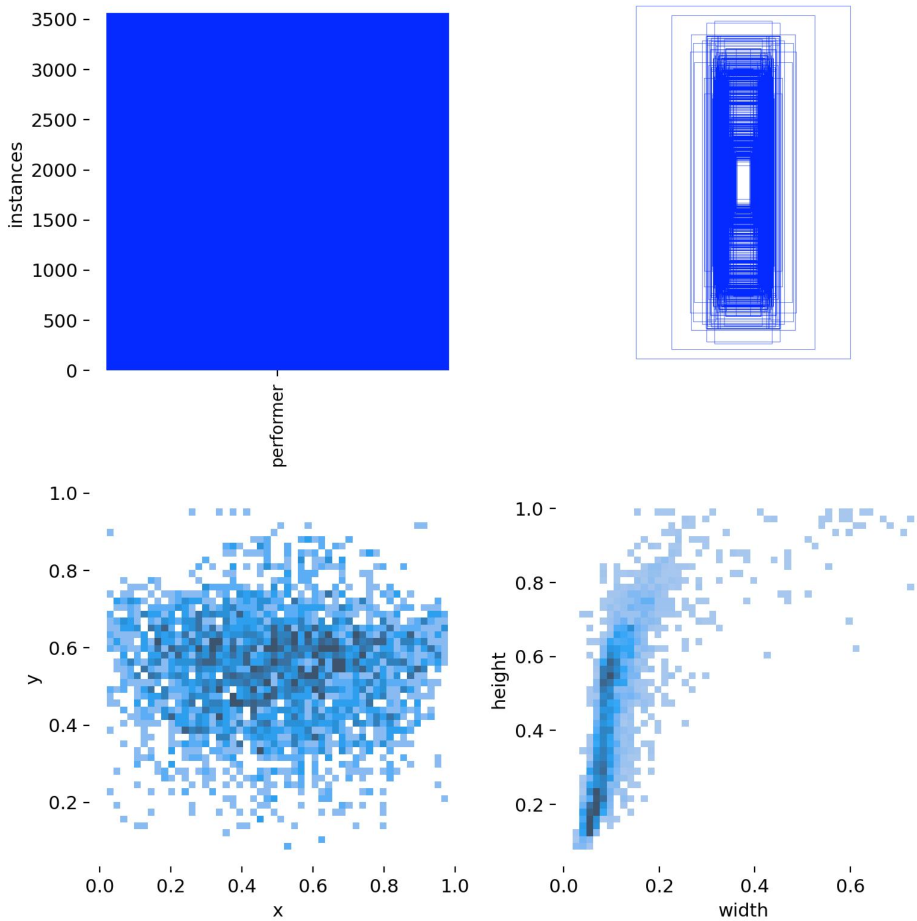

Figure 3.8: Distribution of annotated performer labels in the training dataset. The plots visualize the location of bounding box centers (x, y) and the distribution of their relative width and height, illustrating the visual characteristics of the data used for model training.

The visual characteristics of the resulting annotated dataset are summarized in Figure 3.8. The spatial distribution plot (bottom left) shows that while performers are most frequently located near the center of the frame, their positions are varied, reflecting the diverse and unconstrained nature of the field recordings. The width-height distribution (bottom right) reveals a strong correlation, with bounding boxes being consistently taller than they are wide. This confirms the expected vertical aspect ratio of human performers and attests to the quality and consistency of the manual annotations. Overall, this analysis confirms that the extracted dataset provides a robust and representative foundation for training the performer detection model.

Chapter 4

Implementation of the Performer Detection System

4.1 Model Selection and Fine-Tuning Rationale

The selection of an appropriate performer detection model was guided by both the specific challenges presented by the MPAGDP corpus and the evolving landscape of object detection research. As mentioned in Chapter 2, Traditional computer vision approaches such as Haar Cascades and Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVM) have been widely employed for human detection tasks [23], [4]. However, these classical techniques exhibit significant limitations when confronted with the complex visual environments characteristic of field recordings in ethnomusicological archives—namely, cluttered backgrounds, variable lighting, and frequent occlusions [12], [20]. Such constraints necessitate the adoption of more robust and adaptable methodologies.

Recent advances in deep learning, particularly the development of Convolutional Neural Networks (CNNs), have transformed object detection by enabling models to learn hierarchical feature representations directly from data, thereby substantially improving accuracy and resilience in unstructured settings [17],[16]. Among the state-of-the-art architectures, the YOLO (You Only Look Once) family has emerged as especially suitable for large-scale video analysis due to its real-time detection capabilities and its balance of speed and accuracy [12]. As illustrated in Figure 4.1, the YOLO architecture consistently demonstrates a superior trade-off between performance (mAP) and latency compared to other models. The latest iteration, YOLOv11, introduces improved spatial attention mechanisms and improved feature extraction, which are particularly advantageous for detecting human figures in low light and cluttered conditions, common features of the MPAGDP archive. [22].

[1] While these results provide a useful overview, they should be interpreted with caution, as they are not peer-reviewed and may reflect the developer's own benchmarking setup. Still, they offer a practical point of reference. Source: [22]

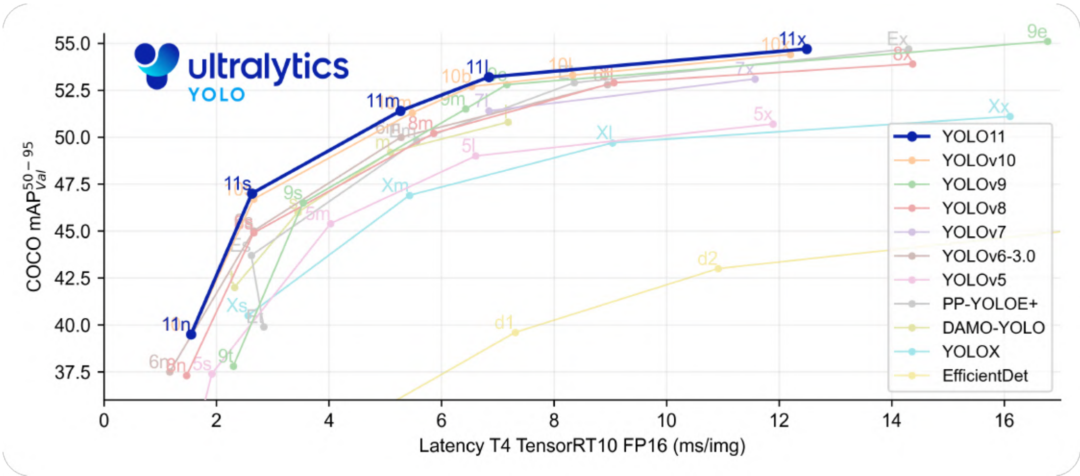

Figure 4.1: Performance comparison of object detection models based on COCO mAP (accuracy) and inference latency (ms/img), adapted from Ultralytics[1].

The choice of the variant *YOLOv11s* ("small") was motivated by its lightweight architecture, which facilitates efficient processing of the extensive MPAGDP dataset without significant compromise in detection accuracy [22]. As shown by the data points on the YOLOv11 performance curve in Figure 4.1, the smaller variants like "s" (small) and "n" (nano) offer substantial gains in speed for a modest trade-off in accuracy, making them ideal for large-scale projects such as this one. Furthermore, the model's ability to handle overlapping performers and partial occlusions is critical for the diverse ensemble configurations documented in Portuguese traditional music [20].

The entirety of the model training and inference was conducted on a MacBook Pro (M1, 2020 model), leveraging the Apple Silicon M1 chip's integrated GPU and unified memory architecture. This hardware configuration provided a balance between computational efficiency and accessibility, allowing for the execution of deep learning workflows without reliance on high-end dedicated GPUs or cloud-based resources. The use of the MacBook Pro M1, with its ARM-based architecture and optimized machine learning libraries, ensured that the iterative fine-tuning process—including model training, inference on over 40,000 frames, and active learning cycles—could be completed within reasonable time frames and with minimal thermal or power constraints. This choice underscores the feasibility of conducting advanced computational musicology research on consumer-grade hardware, broadening the accessibility of such methodologies to a wider range of scholars and practitioners.

Despite YOLOv11s's pre-trained human detection capabilities, domain adaptation was essential. Generalist models, trained on broad datasets, often underperform when applied to culturally specific contexts due to differences in attire, staging, and performance environments [19]. Thus, fine-tuning the model on a domain-specific dataset derived from the MPAGDP archive was necessary to address these idiosyncrasies and to achieve high detection accuracy for the unique visual signatures of Portuguese traditional music performances.

The fine-tuning process involved an iterative methodology: initial manual labeling of a random subset of frames, followed by model-assisted labeling and targeted retraining on samples with low confidence or prediction inconsistency. This approach aligns with best practices in transfer learning and active learning, allowing the model to progressively adapt to the most challenging cases in the dataset while maintaining generalizability [12]. The resulting system achieved robust performance, with a final accuracy of 96% as validated on a random sample of 500 videos, and strong evaluation metrics across precision, recall, F1 score, and mean Average Precision (mAP).

In summary, the selection and fine-tuning of *YOLOv11s* was driven by its architectural suitability for large-scale, real-world performer detection, its proven efficacy in challenging visual contexts, and the necessity for domain adaptation to the specificities of Portuguese traditional music, thereby enabling systematic and scalable musicological analysis of the MPAGDP corpus.

4.2 Model Design & The Iterative Fine-Tuning Process

The design and training of the performer detection system were structured to address the distinctive visual and contextual challenges posed by the MPAGDP archive, leveraging state-of-the-art deep learning while iteratively adapting the model to the specificities of Portuguese traditional music performance environments. The process unfolded in several interlocking stages, each informed by best practices in transfer learning and active learning within computer vision and music information retrieval [12], [22]. The complete source code, trained model weights, and scripts developed for this research are publicly available on GitHub to ensure full reproducibility and to encourage future work in this area.[1]

4.2.1 Initial Manual Labeling and Training

The process commenced with the manual annotation of 500 randomly selected frames, each drawn from a unique video in the corpus. This initial dataset was meticulously labeled to identify all visible performers, providing a domain-specific foundation for fine-tuning the YOLOv11s model, which, while pre-trained on general human detection, lacked adaptation to the complex visual environments of the MPAGDP archive.

The model was trained over 50 epochs as shown in Figure 4.3 on this initial set, allowing it to internalize the visual cues and ensemble configurations characteristic of traditional Portuguese music settings. The results of fifth and final training process are illustrated in Figure 4.2. The plots demonstrate a consistent decrease in the training and validation loss functions *(box_loss, cls_loss, and dfl_loss)*, indicating that the model was successfully learning from the data. Concurrently, the key performance metrics—precision, recall, and mean Average Precision (mAP)—steadily improved, confirming that the model achieved successful convergence without significant overfitting.

[1]The project repository can be accessed at: https://github.com/hellonrk/performance-configuration-yolov11s

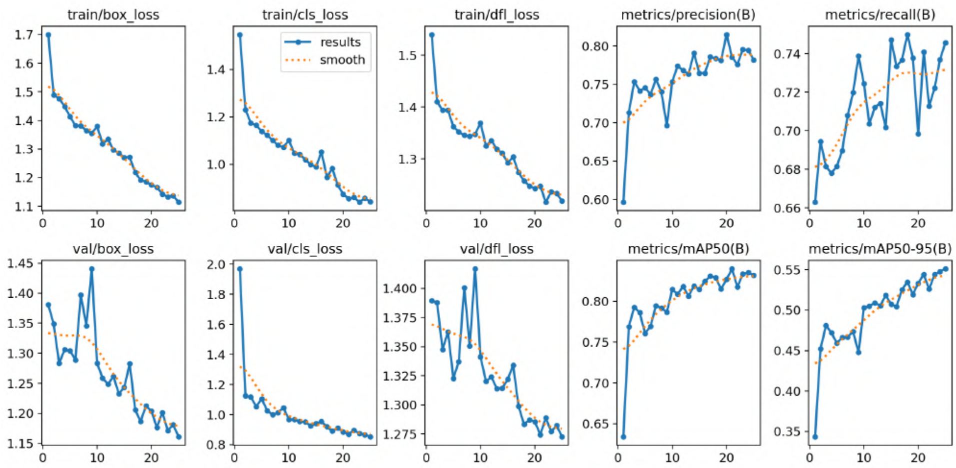

Figure 4.2: Training and validation metrics for the YOLOv11s model over final 25 epochs training. The plots show the decrease in box, classification (cls), and distribution focal loss (dfl), alongside the improvement in precision, recall, and mean Average Precision (mAP), demonstrating successful model convergence.

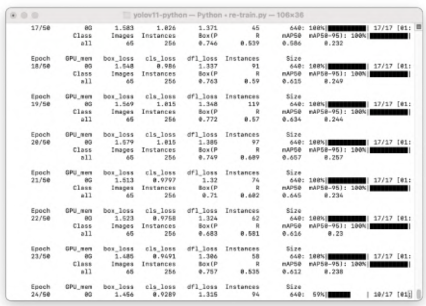

Figure 4.3: Training the YOLOv11s model with 50 epoch

4.2.2 Initial Inference and Consistency Scoring

The trained model was then run on all 40,600 frames (5 frames per video× 8,120 videos). For each frame, the model detected the number of performers and returned a classification label along with a confidence score (e.g. '`3 performers (small Group, conf: 0.82)`'). The confidence score indicates the certainty of the model in its detection, with higher values representing greater confidence. To assess prediction stability, we introduced a consistency check across the five frames sampled from each video:

- If all five predictions matched, the sample was marked "Consistent (100%)"
- If 3 or 4 out of 5 predictions matched, it was marked "Mostly consistent (60–80%)"
- If 2 or fewer matched, it was flagged as "Inconsistent – Manual review required" ($\leq$ 40%).

For example:

- 1, 1, 1, 1, 1 →100% Consistent
- 3, 4, 6, 5, 4 →40% Consistent – Manual Review Required

4.2.3 Model-Assisted Iterative Fine-Tuning

From the consistency evaluation, we selected the least confident frame of 250 least consistent samples and 250 random samples from the full dataset. These 500 images were labeled using a model-assisted labeling process, where initial labels were generated by the model and manually adjusted if needed. This dataset was then used to fine-tune the model for an additional 25 epochs.

- We repeated this iterative fine-tuning process five times:
- Run the model on all 40,600 frames
- Filter results based on consistency and confidence scores
- Select 250 images with the lowest confidence scores and 250 random samples
- Fine-tune the model on this new batch using model-assisted labeling

This approach allowed us to progressively improve the performance of the model by focusing on the most challenging cases in each iteration, while maintaining a diverse training set through random sampling.

4.3 Final Model Performance and Evaluation Metrics

The evaluation of the YOLOv11s model, fine-tuned for performer detection in the MPAGDP corpus, was carried out using a comprehensive suite of standard metrics to ensure both technical rigor and musicological relevance. The assessment focused on the model's ability to accurately identify and localize performers in a highly heterogeneous set of traditional music videos, reflecting the complex visual environments characteristic of field recordings in Portuguese expressive culture [12],[22].

4.3.1 Performance Metrics Overview

The final model, after the fifth iteration, demonstrated robust detection capabilities, as evidenced by the following key metrics:

Metric	Value
Box Precision	0.782
Recall	0.746
F1 Score	0.750
mAP50	0.832
mAP50-95	0.551

Table 4.1: Final YOLOv11s performer detection model evaluation metrics on the MPAGDP dataset

As shown in the Table 4.1,

- Box Precision (0.782) indicates that 78.2% of the performers identified by the model were correct detections, reflecting the system's effectiveness at minimizing false positives.
- Recall (0.746) shows that 74.6% of all actual performers present in the frames were successfully detected, underscoring the model's capacity to avoid false negatives and comprehensively capture ensemble configurations.
- F1 Score (0.750) provides a balanced measure of overall detection performance, harmonizing precision and recall to reflect the model's reliability in practical application contexts.
- mAP50 (0.832) measures mean Average Precision at a 50% Intersection over Union (IoU) threshold, signifying high detection accuracy in identifying the general location of performers.
- mAP50-95 (0.551) averages precision across multiple IoU thresholds from 50% to 95%, offering a more stringent assessment of bounding box placement. The lower value here indicates the inherent difficulty in exact boundary delineation under conditions of occlusion, complex postures, and variable lighting typical of traditional music field recordings.

To further illustrate the model's detection performance, the Precision-Recall (PR) curve is presented in Figure 4.4. This curve visualizes the trade-off between precision, which measures the accuracy of the performer detections, and recall, which measures the model's ability to identify all actual performers present in a frame. The area under the PR curve represents the mean Average Precision (mAP), a key indicator of overall model quality. As shown, the model achieves a mAP of 0.819 at an Intersection over Union (IoU) threshold of 0.5, visually confirming the strong detection performance reported in Table 4.1 and underscoring the model's effectiveness in balancing both precision and recall.

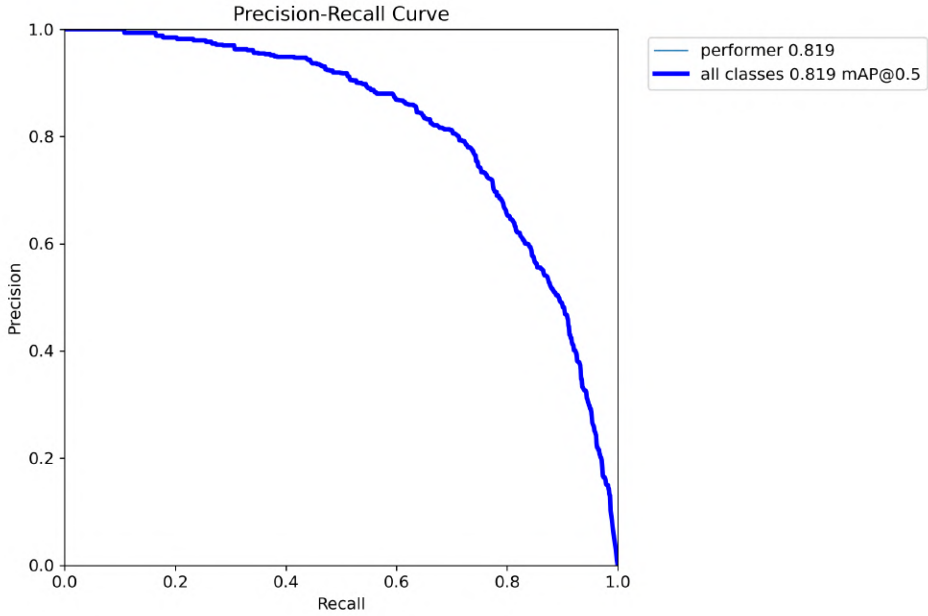

Figure 4.4: The Precision-Recall (PR) curve for the performer class. The area under this curve represents the mean Average Precision (mAP), which was 0.819 at an IoU threshold of 0.5, indicating strong detection performance.

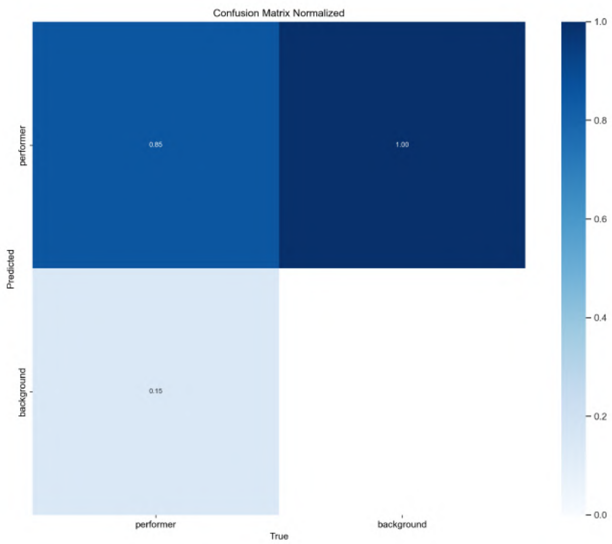

Figure 4.5: Normalized confusion matrix for the performer detection task.

The model's classification accuracy is further detailed in the normalized confusion matrix, shown in Figure 4.5. The matrix confirms the model's high precision; when it predicts a 'performer', it is correct 85% of the time. Furthermore, the model perfectly classifies 'background' instances, indicating that it does not generate false positives on frames that contain no performers, a crucial feature for the subsequent quantitative analysis.

4.3.2 Confidence Threshold Optimization

The optimal balance between precision and recall was achieved at a confidence threshold of 0.47, as demonstrated by the F1-confidence curve, which peaked at an F1 score of 0.75 as you can see in the Figure 4.6. This threshold was adopted for final predictions, ensuring that the model's outputs were both reliable and interpretable for subsequent musicological analysis.

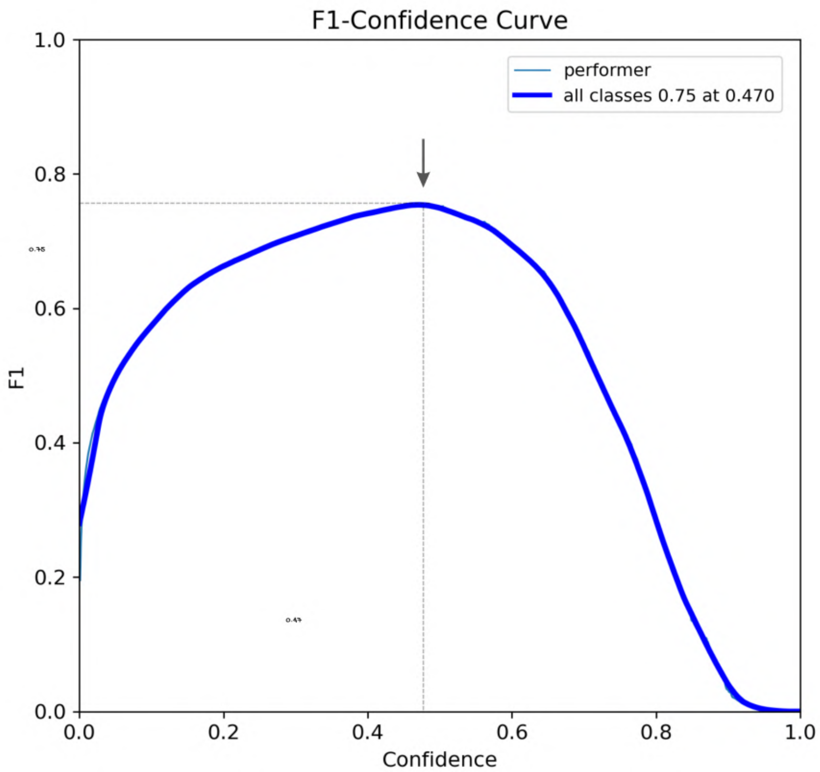

Figure 4.6: F1-Confidence curve for the YOLOv11s model on the performer detection task.

4.3.3 Prediction Consistency Analysis

The consistency of the performer count predictions across the five sampled frames per video was systematically analyzed to assess the stability of the model's classifications:

Table 4.2: Prediction Consistency Analysis: Distribution of MPAGDP videos by the consistency level of performer count predictions using the YOLOv11s

Consistency Level	**Number of Videos**	**Percentage**
Fully consistent (100%)	4,616	56.8%
Mostly consistent (60-99%)	2,072	25.5%
Needs manual review (<60%)	1,395	17.2%
No frames found	37	0.5%

Table 4.2 presents the results of the prediction consistency analysis. The table lists four categories: "Fully consistent (100%)" with 4,616 videos, corresponding to 56.8% of the total; "Mostly consistent (60-99%)" with 2,072 videos, representing 25.5%; "Needs manual review (<60%)" with 1,395 videos, accounting for 17.2%; and "No frames found" with 37 videos, which is 0.5% of the total.

A high rate of full or mostly consistent predictions (over 82%) attests to the robustness of the model even in challenging visual environments, while flagged cases for manual review reflect a systematic approach to quality assurance in ambiguous or low-confidence scenarios.

4.3.4 System Accuracy Validation

To rigorously evaluate the performance of our model, we performed a comprehensive validation process using a random sample of 500 videos from the complete MPAGDP dataset of 8,120 videos. For each video in this validation set, we examined the model's final performance configuration prediction (solo, duo, small group or large group) against manual verification.

The validation process considered the five sample frames extracted from each video along with their associated confidence scores and consistency metrics. A prediction was deemed correct only if the model's final classification accurately represented the actual number of performers in the video. This approach ensured that our validation accounted for the complete performance context rather than isolated frames. The results demonstrated the robust performance of the model, with 485 correct classifications out of 500 videos, which produced an overall accuracy of 96%. This high accuracy rate confirms the effectiveness of our YOLOv11s implementation for performer detection and classification in the *MPAGDP* archive cluster, even in diverse performance settings, lighting conditions, and recording environments.

This validation process provides strong evidence that our computational approach can reliably analyze performance configurations across the entire MPAGDP archive, enabling systematic musicological analysis that would be prohibitively time-consuming through manual review alone.

4.3.5 Interpretation and Context

These results are particularly noteworthy given the environmental challenges inherent to the MPAGDP archive, including variable lighting, occlusions, complex backgrounds, and culturally specific staging. The difference between mAP50 and mAP50-95 underscores the difficulty of precise boundary placement in such contexts, yet the system's overall accuracy and consistency rates demonstrate its practical viability for systematic musicological inquiry. A qualitative review of the model's predictions, as shown in 4.7, further illuminates its behavior. The figure compares the ground truth labels with the model's predictions on a sample validation batch, showcasing numerous correct detections, often with high confidence (e.g., `performer 0.9`). It also reveals common challenges, such as lower confidence scores in large, occluded groups (e.g., `performer 0.3`) and occasional missed detections, which aligns with the quantitative limitations discussed. This visual validation confirms the model's effectiveness while also highlighting areas for future improvement.

Figure 4.7: Qualitative analysis of model performance on a sample validation batch. The top panel shows the ground truth labels, while the bottom panel shows the corresponding predictions from the YOLOv11s model, including confidence scores. This comparison highlights successful detections as well as challenges such as missed performers in dense crowds and lower confidence scores in complex scenes.

In summary, the final YOLOv11s model achieved a robust balance of precision, recall, and consistency, enabling scalable and reliable detection of performance configurations across a vast and diverse corpus of Portuguese traditional music recordings. This provides a methodological foundation for subsequent musicological analysis at a scale and granularity previously unattainable through manual review alone.

4.4 Model Final Readings

Table 4.3: Final Readings of Overall Model

Category	Number of Videos	Percentage of Dataset
Solo	3,946	48.58%
Duo	1,312	16.15%
Small Group	1,225	15.08%
Large Group	1,600	19.70%
Unknown	39	0.48%

The final readings of the YOLOv11s-based performer detection system are presented in Table 4.3. The table enumerates the total number of videos classified into each performance configuration category as determined by the model. Specifically, the "Solo" category comprises 3,946 videos, representing 48.58% of the dataset; the "Duo" category includes 1,312 videos, accounting for 16.15%; the "Small Group" category contains 1,225 videos, corresponding to 15.08%; and the "Large Group" category encompasses 1,600 videos, which constitutes 19.70% of the total. And for 39 videos, there was no detectable frames which comprise of 0.48% of total dataset. These results reflect the model's capacity to systematically categorize ensemble size across the entire MPAGDP archive, providing a quantitative foundation for subsequent musicological analysis.

4.5 Further Analysis

The model's classification outputs were further analyzed by integrating the predicted performance configuration categories with existing MPAGDP metadata, focusing primarily on geographic location and genre. This cross-referencing enabled a detailed examination of correlations between ensemble size, regional practices, and musical genres, with the comprehensive analysis of these relationships presented in Chapter 5

Chapter 5

Musicological Analysis, Results, and Discussion

5.1 Analysis of Performance Configurations by Category

The systematic computational analysis of the MPAGDP archive reveals distinct patterns in the prevalence and function of each performance configuration—solo, duo, small group, and large group—across Portuguese traditional music. This section presents a quantitative breakdown and a qualitative interpretation of each category, elucidating their musicological significance and contextual roles within the dataset corpus.

5.1.1 Solo Performances

Solo performances constitute the most prevalent configuration, accounting for 3,946 occurrences (48.58% of the dataset). This predominance underscores the centrality of individual expression within Portuguese traditional music, especially in narrative and poetic traditions. The solo format is particularly prominent in genres such as *estória* (107 occurrences) and *poema* (271 occurrences), where the unaccompanied voice functions as a primary vehicle for communal memory and literary expression. The tradition of the solitary *cantadeira* (90 occurrences) in northern Portugal exemplifies the role of women in maintaining melodic lineages through individual performance, often characterized by melismatic ornamentation and free rhythmic interpretation. Additionally, solo performances are integral to religious practices, including *oração* (53 occurrences) and *responso* (15 occurrences), reflecting the deeply personal dimension of sacred expression and the continuity of medieval modal characteristics in contemporary practice. The high concentration of *tradição oral* (143 occurrences) associated with solo performances further highlights their function in preserving intangible cultural knowledge through individual embodiment.

5.1.2 Duo Performances

Duo configurations appear in 1,312 recordings (16.15% of the dataset), representing an intermediate category that bridges individual and collective musical expression. Duos frequently feature complementary musical roles, enabling harmonic and rhythmic interplay while retaining the intimacy of solo performance. Notably, harmonized singing is prevalent in duo performances of *cantiga* (34 occurrences) and *fado* (47 occurrences), where the dialogic relationship between voice and *viola acústica* (256 occurrences) exemplifies the urban tradition's characteristic interplay of text and accompaniment. The duo format is also essential in competitive traditions such as *desgarrada* (6 occurrences) and *despique* (4 occurrences), which are structured around poetic improvisation and call-and-response exchanges. Furthermore, the frequent appearance of interviews (*entrevista*, 66 occurrences) in duo configurations illustrates the format's utility in capturing conversational dynamics and contextual information about performance practices.

5.1.3 Small Group Performances

Small group configurations, defined as ensembles of three to five performers, account for 1,225 occurrences (15.08% of the dataset). This category marks a threshold where musical roles become more differentiated and collective coordination intensifies. Small groups are notable for their instrumental diversity and adaptability, serving as a primary site for stylistic innovation and cross-genre fusion—evident in the presence of "folk" (45 occurrences) and "rock" (25 occurrences) performances. Instrumental combinations such as *bombo* (134 occurrences), *caixa* (94 occurrences), and *gaita de foles* (82 occurrences) create distinctive rhythmic and melodic textures, supporting both dance and song. Small group vocal ensembles, including *cante* (49 occurrences), often preserve the essential polyphonic structure of traditional singing while allowing for individual expressivity and flexibility. This configuration is particularly prominent in urban centers, where traditional and contemporary forms intersect.

5.1.4 Large Group Performances

Large group performances, comprising more than five performers, are observed in 1,600 recordings (19.70% of the dataset). This configuration is emblematic of the collective dimension of Portuguese musical expression, particularly in choral and polyphonic singing traditions. The strong association with *coro* (254 occurrences), *coral* (116 occurrences), and *polifonia* (376 occurrences) highlights the importance of communal vocal practices. The UNESCO-recognized *Cante Alentejano* exemplifies this tradition, featuring a formal structure with distinct vocal parts (*ponto*, *alto*, *baixo*) that mirror social organization within the community [8]. Large ensembles are also integral to ceremonial and communal events, as seen in *rancho* (66 occurrences) and *baile* (20 occurrences), where performance is intertwined with choreographed movement and local identity. The regional concentration of large group performances in Alentejo contrasts with the solo-dominant traditions of northern Portugal, reflecting divergent historical and social trajectories.

Table 5.1: Summary Table: Performance Configuration Distribution

Configuration	Occurrences	Percentage	Key Genres/Functions
Solo	3,946	48.58%	Narrative, poetic, religious, oral tradition
Duo	1,312	16.15%	Harmonized singing, fado, musical dialogue, interviews
Small Group	1,225	15.08%	Instrumental diversity, folk, rock, polyphonic singing
Large Group	1,600	19.70%	Choral, polyphonic, ceremonial, community-based

This analysis demonstrates that performance configuration is not merely a logistical parameter but a critical lens for understanding the social structure, genre conventions, and regional identities embedded in Portuguese traditional music. The computational approach employed here enables the systematic quantification and interpretation of these patterns, providing new insights into the dynamics of musical transmission and adaptation within the corpus.

5.2 Regional Distribution of Performance Configurations

The systematic computational analysis of the MPAGDP archive uncovers noticeable regional patterns in performance configurations, reflecting the interplay between ensemble size, social organization, and cultural identity in Portuguese traditional music. By correlating performer counts with geographic metadata, it is possible to discern how historical, economic, and social factors have shaped distinctive musical practices across Portugal's diverse regions.

5.2.1 Northern Portugal: The Primacy of Individual Expression

In the northern regions, particularly in districts such as Vila Real and Vimioso, there is a marked predominance of solo performances. For example, Vila Real registers an extraordinary concentration of 729 solo performances with virtually no large ensembles, while *Vimioso* exhibits an even more extreme pattern, with 999 solo performances and 219 duos but almost no large group events. This solo-dominant landscape is rooted in the region's historical isolation, rugged terrain, and the persistence of oral traditions, where individual storytelling and musical self-sufficiency are valorized. The prevalence of solo and duo formats is further reinforced by the region's instrumental traditions, such as the *gaita transmontana* (bagpipe) and *cavaquinho*, which are well-suited to intimate, domestic, or small community contexts. Miranda do Douro, by contrast, presents a more balanced distribution (158 solo, 94 small group, and 34 large group performances), reflecting its status as a cultural crossroads and the influence of the *pauliteiros* stick dance tradition, which necessitates coordinated group performance.

5.2.2 *Alentejo*: The Embodiment of Collective Voice

The Alentejo region stands in stark contrast, epitomizing the collective dimension of Portuguese musical expression. In municipalities such as Serpa and Castro Verde, large group performances are dominant: *Serpa* documents 308 large group performances compared to only 144 solos, and *Castro Verde* records 248 large groups to 101 solos. This pattern is directly linked to the region's UNESCO-recognized *Cante Alentejano*, a polyphonic singing tradition performed by large, non-professional choirs whose formal structure (*ponto*, *alto*, *baixo*) mirrors the communal organization of agrarian life [8]. The prevalence of large ensembles in *Alentejo* is not simply a musical preference but a social practice, emerging from agricultural labor and the need for collective solidarity and identity. *Viana do Alentejo*'s robust *choral* tradition (248 large group performances) further highlights the role of communal singing in religious festivities and community celebrations, reinforcing social cohesion across generations.

Table 5.2: Performer Count and Composition Summary by District & Autonomous Region

District/Autonomous Region	Solo	Duo	Small Group	Large Group	Unknown
Lisboa	1,200	501	610	252	20
Beja	884	240	100	1,220	7
Bragança	1,304	260	151	147	8
Coimbra	889	205	273	121	11
Castelo Branco	722	284	202	196	2
Vila Real	785	55	83	318	2
Viseu	134	43	118	531	0
Santarém	450	85	122	96	6
Porto	243	51	120	185	2
Évora	248	57	18	51	9
Leiria	209	42	51	85	0
Faro	150	80	60	40	0
Setúbal	120	40	50	25	1
Guarda	80	25	35	70	0
Viana do Castelo	50	20	30	109	0
Madeira	153	20	10	0	0
Portalegre	90	15	20	45	0
Braga	75	30	25	20	0
Aveiro	120	35	45	50	0
Açores	30	10	5	0	0

5.2.3 Urban Centers: Crucibles of Musical Diversity

Urban centers such as *Lisboa* and *Coimbra* function as dynamic musical ecosystems, where all performance configurations coexist and interact. *Lisboa*, for example, demonstrates the most diverse distribution: 1,048 solos, 569 small groups, 486 duos, and 208 large groups. This diversity is a product of the city's historical role as a cultural crossroads, attracting regional traditions from

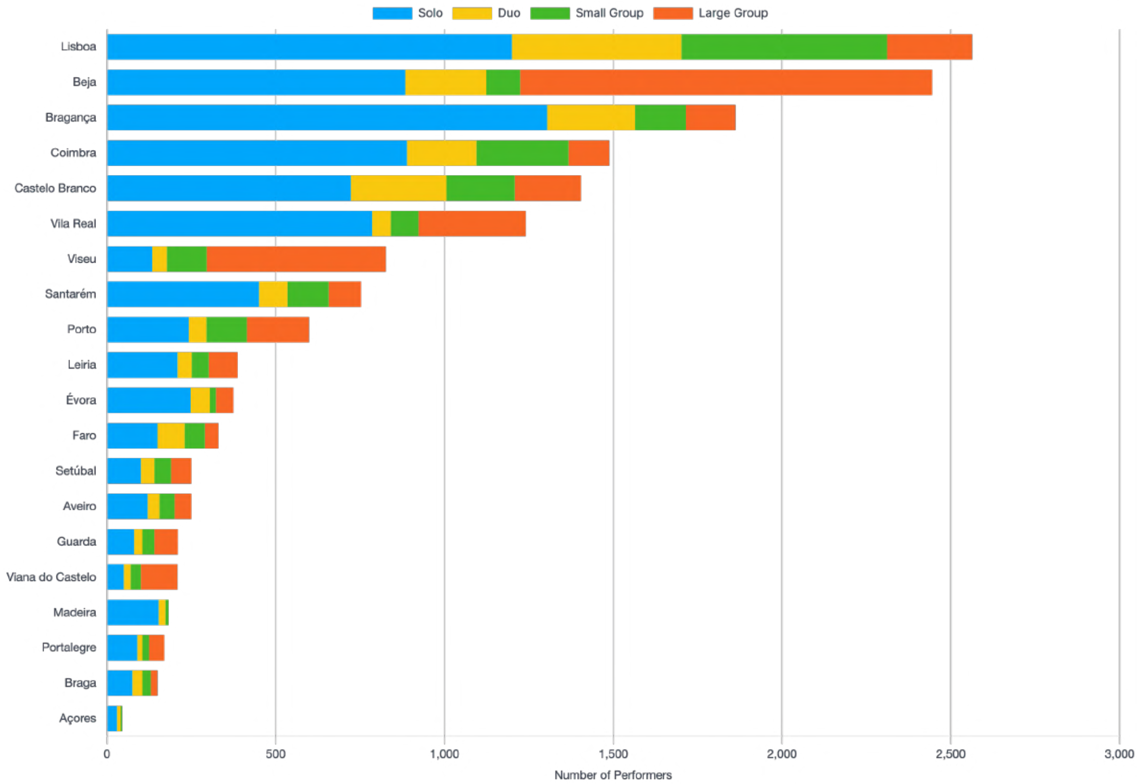

Figure 5.1: Regional Distribution of Performance Configurations by District

across Portugal and facilitating cross-pollination with international influences. Fado houses in *Lisboa* host solo and duo performances, while cultural centers accommodate both traditional *ranchos folclóricos* and contemporary fusion ensembles. *Coimbra*, while smaller in scale, also exhibits significant diversity (121 solos, 74 small groups, 72 duos, 8 large groups), reflecting its academic musical identity and the tradition of *Coimbra fado*, which typically bridges solo and small group categories.

5.2.4 Interpretive Summary

These regional patterns demonstrate that performance configuration in Portuguese traditional music is not merely a technical parameter but a visual and structural embodiment of regional cultural values and historical trajectories. The predominance of solo and duo formats in the north reflects a tradition of individual narrative and instrumental virtuosity, shaped by geographic and social isolation. In contrast, the large ensemble practices of *Alentejo* manifest collective identity and social solidarity, rooted in agrarian communalism and ritual. Urban centers, meanwhile, serve as laboratories of musical innovation, where the coexistence of all configurations fosters both preservation and transformation of traditional forms.

This analysis affirms that computational approaches, when integrated with musicological frameworks, can illuminate the complex interdependence between geography, social organization, and musical practice in the Portuguese context, providing a scalable methodology for future comparative studies in ethnomusicology.

5.3 Discussion

The computational analysis of the MPAGDP archive demonstrates that performance configuration—defined by ensemble size and arrangement—serves as a powerful visual indicator of the underlying social structures, regional identities, and mechanisms of cultural transmission in Portuguese traditional music. By leveraging a fine-tuned YOLOv11s model with 96% accuracy, this research systematically classified 8,120 recordings into solo, duo, small group, and large group categories, revealing that solo and duo performances predominate in northern Portugal, reflecting traditions of individual expression and narrative storytelling, while large ensembles are central to *Alentejo*'s communal polyphonic singing, embodying collective identity [8],[14]. Urban centers like *Lisboa* and *Coimbra* exhibit the greatest diversity of configurations, functioning as hubs for both preservation and innovation. These findings confirm that performance configuration is not merely a technical parameter but a visual and structural embodiment of how musical practices are shaped by, and in turn shape, the social and cultural fabric of Portuguese communities, thus providing new musicological insights that were previously unattainable at scale through manual analysis [12],[13],[22].

Chapter 6

Conclusion and Future Work

6.1 Summary of Findings and Contributions

This thesis demonstrates that computational musicology, specifically through the application of a fine-tuned YOLOv11s deep learning model, can systematically and accurately analyze performance configurations in large-scale audiovisual archives of Portuguese traditional music, achieving a 96% classification accuracy and robust evaluation metrics. By processing 8,120 field recordings from the MPAGDP corpus, the research reveals that ensemble size—categorized as solo, duo, small group, or large group—is not a trivial parameter but a visual and structural embodiment of social organization, regional identity, and cultural transmission within Portuguese musical traditions [3],[8]. From the findings, we can learn that Solo performances dominate in northern Portugal, reflecting a tradition of individual narrative and oral transmission, while large group configurations are central to *Alentejo*'s polyphonic singing practices, embodying communal identity and collective memory. Urban centers such as *Lisboa* and *Coimbra* display the greatest diversity of configurations, serving as hubs for both preservation and innovation. The methodological advances—developing a scalable, domain-adapted performer detection system and introducing a four-part ensemble taxonomy—enable new forms of quantitative and comparative musicological analysis that were previously unattainable at this scale. This research not only streamlines archival workflows and enhances access to traditional music collections but also provides unprecedented empirical insight into the dynamics of musical practice, adaptation, and cultural continuity in Portugal, thus establishing a foundation for future multimodal, cross-cultural, and instrument-specific research in computational ethnomusicology [13], [20].

6.2 Limitations of the Study

Despite the significant methodological and empirical advances presented in this thesis, several limitations must be acknowledged that affect both the technical robustness and musicological gen-

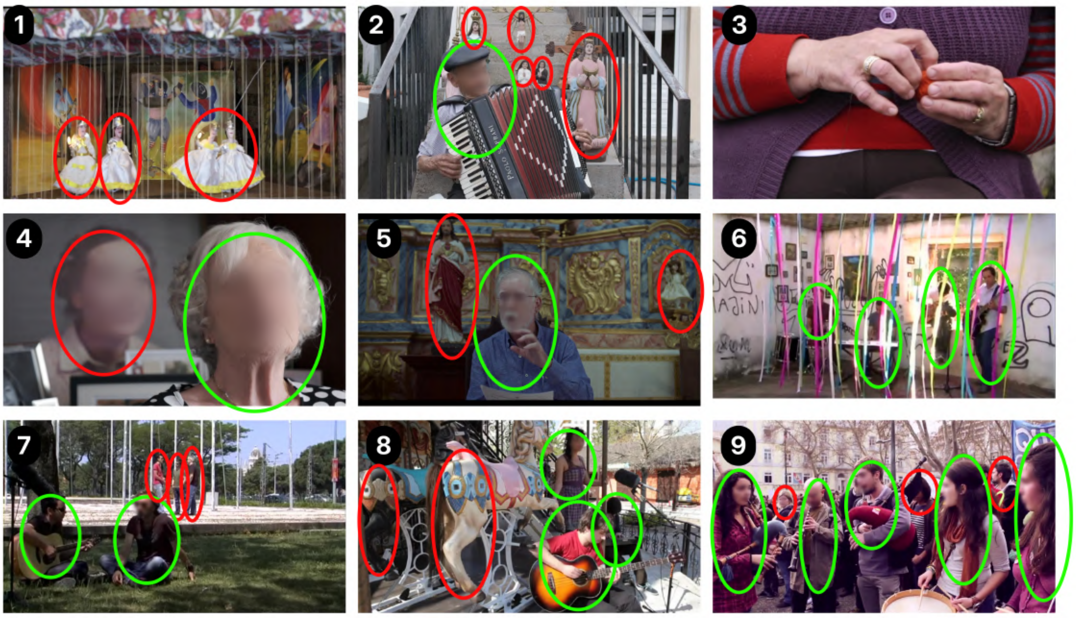

Figure 6.1: Qualitative Examples of the model's limitations and failure modes. Green ovals indicate correctly identified performers, while red ovals highlight model errors, such as the misclassification of non-performers and inanimate objects.

eralizability of the findings. First, the automated performer detection system, though highly accurate overall, is challenged by the inherent complexities of field recordings in the MPAGDP archive. Occlusions—where performers are partially hidden by instruments, other musicians, or environmental elements—frequently impede accurate detection, particularly in large ensemble settings or dynamic performances. Similarly, variable and low-light conditions, common in outdoor or informal venues, can degrade model performance, as deep learning models trained on standard datasets may not generalize optimally to such environments [12],[16].

These challenges are illustrated qualitatively in Figure 6.1. A significant issue arises from the misclassification of inanimate figures as performers, a problem particularly prevalent in visually complex settings where the model incorrectly identified dolls, religious statues, wall portraits and other objects (Panels 1, 2, 4, 5, and 8). Furthermore, the model's accuracy is frequently impeded by occlusions, where performers are partially hidden by instruments or other people, and by the difficulty of distinguishing performers from audience members in dense crowds (Panels 6, 7, 8, and 9). The system also shows limitations in detecting performers due to lighting or those partially obscured by environmental elements (Panel 7). Finally, the performance of the model can be degraded closeup images of instruments or human body parts, which impacts its ability to make accurate detections (Panel 3).

Another limitation arises from the static frame-based sampling strategy. By extracting five frames per video at fixed intervals, the approach may fail to capture the full temporal dynamics of performances, especially in cases involving camera panning, performer movement, or evolving ensemble configurations. This limitation can lead to under-or over-estimation of performer counts, particularly in videos where the visible ensemble changes over time. Moreover, the model occasionally misclassifies non-performers, such as statues, audience members, or images in religious settings, as performers, reflecting the challenge of distinguishing between animate and inanimate figures in complex visual scenes.

A significant limitation of this study stems from the dynamic and curated nature of the MPAGDP archive. As a living repository, the MPAGDP is continuously updated by its founder, Tiago Pereira, who actively shapes the archive's content through ongoing fieldwork and personal curatorial decisions. This means the archive's coverage at any given time reflects not only the diversity of Portuguese musical traditions but also the evolving priorities, geographic focus, and selection criteria of its creator, rather than a systematic or representative sampling of Portugal's musical life. Consequently, the patterns identified in this thesis, such as regional distributions of performance configurations or the prevalence of certain ensemble types, may be influenced as much by the archive's current composition and collection strategies as by the actual landscape of Portuguese traditional music. As new videos are added and geographic or genre gaps are addressed, the empirical basis for musicological claims derived from the archive may shift, potentially introducing bias and limiting the generalizability of the findings. Therefore, any analysis based on the MPAGDP corpus should be regarded as provisional and contingent, reflecting the state of a continually expanding and curated resource rather than offering a definitive or exhaustive account of Portugal's ethnic musical practices.

From a data perspective, the absence of a large, pre-existing, domain-specific annotated dataset for Portuguese traditional music necessitated the creation of a bespoke training corpus, which, despite iterative refinement, may still reflect biases related to regional representation, genre diversity, or recording conditions. The computational demands of large-scale deep learning inference also imposed practical constraints, limiting the feasibility of real-time processing or more granular video-based analysis.

Finally, while the present study focuses on visual performer detection, it does not integrate audio features or multimodal analysis, which could enhance the accuracy of performer identification and enable more nuanced interpretations of ensemble function and musical interaction. The absence of instrument-specific detection further limits the granularity of musicological insights, as the current system cannot distinguish between different instrumental roles within ensembles.

6.3 Directions for Future Work

Building on the methodological and empirical advances of this study, several promising directions emerge for future research in computational ethnomusicology and digital heritage analysis. First, integrating multimodal analysis, combining audio features with visual performer detection, offers a pathway to more nuanced and robust classification of performance configurations, as sonic cues can help resolve ambiguities arising from occlusions, low-light conditions, or the presence of non-performers in the visual field. The development of models capable of jointly analyzing audio-visual data would facilitate the identification not only of ensemble size but also of musical roles, interaction patterns, and instrument-specific contributions, thereby deepening musicological insight [13].

Also, enhancing model architecture and training protocols remains a priority. Future work should focus on improving detection in particularly challenging environments—such as performances with significant motion blur, complex staging, or rapid camera movement—by leveraging advances in transformer-based architectures or self-supervised learning [22], [12]. Transitioning from a static, frame-based sampling strategy to a video-based or temporal analysis approach would allow for the dynamic tracking of ensemble changes throughout a performance, capturing the full temporal evolution of musical events.

Expanding the scope of research to cross-cultural and cross-archival applications is essential for testing the generalizability of the proposed methodology. Applying the fine-tuned performer detection system to other traditional music archives—whether Iberian, Mediterranean, or from entirely different cultural contexts—would both validate and refine the approach, while potentially revealing universal and culture-specific patterns in ensemble practice and social organization. This comparative perspective is crucial for advancing computational musicology as a global, rather than regionally bounded, discipline.

Also, increasing the granularity of analysis by developing models capable of instrument-specific detection and classification would enable more detailed studies of musical arrangement, orchestration, and performance practice [20]. Such advancements could illuminate the interplay between vocal and instrumental forces, the distribution of musical roles within ensembles, and the evolution of instrumental traditions within Portuguese and other folk music.

Finally, future research should address the ethical and curatorial dimensions of automated archival analysis, ensuring that computational systems are culturally sensitive and that their outputs are interpretable and actionable for both scholars and community stakeholders. This includes enhancing validation mechanisms, reducing dataset biases, and fostering collaborative practices between technologists, musicologists, and tradition bearers.

In summary, the next phase of research should pursue multimodal integration, model enhancement, cross-cultural validation, instrument-level detection, and ethical curation, all of which will further elevate the role of artificial intelligence in the preservation, analysis, and celebration of traditional music heritage [13],[22],[20].

6.4 Final Thoughts

The present study demonstrates the efficacy and scholarly value of computational methods in the systematic analysis of performance configurations within large-scale audiovisual archives of traditional music. By leveraging a domain-adapted YOLOv11s model, this research has established a scalable framework for detecting and classifying ensemble sizes, thereby enabling the quantitative investigation of musicological questions that have historically relied on labor-intensive manual annotation. The findings elucidate how ensemble configurations are not merely logistical details but are deeply embedded in the social organization, regional identity, and cultural transmission of Portuguese musical traditions. The predominance of solo and duo performances in northern Portugal, contrasted with the collective large-ensemble practices of *Alentejo* and the diversity found in urban centers, underscores the extent to which musical practice is intertwined with broader patterns of community life and historical development. Moreover, the integration of computational analysis with ethnomusicological frameworks opens new avenues for comparative and cross-cultural research, suggesting that similar methodologies may be fruitfully applied to other musical traditions and archival contexts. And, this thesis affirms that the convergence of artificial intelligence and musicological inquiry not only augments the capacities of archival research but also deepens our understanding of the cultural logics that shape musical expression and its enduring significance within society.

References

[1] Md. Alam. The impact of globalization on indigenous cultures: Preservation and adaptation. *Research Journal of Humanities and Social Sciences*, 4(1):1–9, 2025.

[2] C. Biemann, G. R. Crane, C. D. Fellbaum, and A. Mehler, editors. *Computational Humanities—Bridging the gap between Computer Science and Digital Humanities*. Lorentz Center, 2017. Dagstuhl Seminar Report.

[3] John Blacking. *How Musical is Man?* University of Washington Press, Seattle, 1973.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[5] Jane W. Davidson and John M. M. Good. Social and musical coordination between members of a string quartet: An exploratory study. *Psychology of Music*, 30(2):186–201, 2002.

[6] Thi Minh Duc. The impact of globalization on traditional vietnamese culture. *International Journal of Religion*, 5(11):5660–5669, 2024.

[7] Brittan Dunham and Candace Ming. Challenges in delivering digital content in archival and production environments. *Journal of Digital Media Management*, 7(3):268–278, 2019.

[8] Karolina Golemo. Multi-part voices of the portuguese south: Cante alentejano in contemporary interpretations. *RM*, 8(1):150–169, 2024.

[9] L. E. Gray. Fado resounding: Affective politics and urban life, 2004. https://www.academia.edu/5426761/Fado_Resounding_Affective_Politics_and_Urban_Life.

[10] Sascha Grollmisch, Estefanía Cano, Fernando Mora Ángel, and Gustavo López Gil. Ensemble size classification in colombian andean string music recordings. In *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Marseille, France, 2019. CNRS - Laboratoire de Mécanique et d'Acoustique. Part of the ACMus research project, supported by the German Research Foundation (BR 1333/20-1).

[11] A. Hofmann, T. Miksa, P. Knees, A. Bakos, H. Sağlam, A. Ahmedaja, B. Yimwadsana, C. Chan, and A. Rauber. Enabling fair use of ethnomusicology data – through distributed repositories, linked data and music information retrieval. *Empirical Musicology Review*, 16(1):47–64, 2021.

[12] Jae-A Kim, Jin-Young Sung, and Seung-Ho Park. Comparison of faster-rcnn, yolo, and ssd for real-time vehicle type recognition. In *2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4. IEEE, 2020.

[13] Alexander Lerch, Curtis Arthur, Anjali Pati, and Siddharth Gururani. An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval*, 3(1):221–245, 2021.

[14] MPAGDP. A música portuguesa a gostar dela própria, n.d. http://amusicaportuguesaagostardelapropria.org/.

[15] Margot Note. Metadata for archival collections: Challenges and opportunities, July 2018.

[16] Noel O'Mahony, Sean Campbell, Ana Carvalho, Shayan Harapanahalli, German Velasco-Hernandez, Lenka Krpalkova, Daniel Riordan, and John Walsh. Deep learning vs. traditional computer vision. In Kohei Arai, Supriya Kapoor, and Rahul Bhatia, editors, *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC)*, pages 128–144. Springer, 2019.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 91–99, 2015.

[18] Anthony Seeger. *Why Suyá Sing: A Musical Anthropology of an Amazonian People*. Cambridge University Press, Cambridge, 1987.

[19] D. F. Silva, A. C. M. Silva, L. F. Ortolan, and R. M. Marcacini. On generalist and domain-specific music classification models and their impacts on brazilian music genre recognition. In *Anais do XVIII Simpósio Brasileiro de Computação Musical*, pages 60–67. Sociedade Brasileira de Computação, 2021.

[20] Luis A. Silva, Héctor Sánchez San Blas, Ana Sales Mendes, Álvaro Lozano Murciego, and Daniel M. Jiménez-Bravo. Automated detection system for portuguese musical instruments. In M. A. Gonçalves, M. T. Baptista, and H. K. Fonseca, editors, *Proceedings of the International Conference on Cultural Heritage*, pages 15–28. Springer, 2024.

[21] UCLA Ethnomusicology Archive. Ucla ethnomusicology archive, 2021.

[22] Ultralytics. Yolov11 - overview, 2024. https://docs.ultralytics.com/models/yolo11/.

[23] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511–I–518. IEEE, 2001.