A Machine Learning Approach for an End-to-End Music Inpainting Pipeline

Gustavo Romão[1], Elisa T. Albergaria[1], Diego R. C. Dias[2(✉)], Marcela Alves de Almeida[2], Vinícius F. S. Mota[2], and Leonardo Chaves Dutra Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ), São João del-Rei, Brazil
gugahr@aluno.ufsj.edu.br, {etuler,lcrocha}@ufsj.edu.br
[2] Universidade Federal do Espírito Santo (UFES), Vitória, Brazil
{diego.dias,marcela.almeida,vinicius.f.mota}@ufes.br

Abstract. Preserving musical memory is a constant challenge, as time is a key factor in the degradation of old scores. Over the years, stains, fading, and other physical damage erase notes and symbols, making reading difficult and jeopardizing the correct performance of these historical works. Among these symbols, musical clefs, reference markers that anchor pitch on specific staves, are particularly important, and as such, their absence makes the positions of subsequent notes semantically ambiguous. This study proposes a workflow that integrates an optical music recognition step based on YOLOv11 to detect clefs and note positions, statistical processing to convert the extracted staff position indices into compact meta-features of position distribution, and supervised classification for clef prediction. Using musical scores from the PRIMUS dataset, 950 images were selected and, according to three clef classes, the ability of the k-nearest neighbors, random forest, and gradient boosting models to predict missing clefs was compared, framing inpainting as a classification task. Gradient boosting showed the best performance, achieving an F1 score of 0.7242 on the test set. The results show that clefs can be predicted solely from positional statistics of pitch, which corroborates both the practical application of gap filling and an interpretable link between pitch patterns related to the register and the occurrence of clefs.

Keywords: Music Inpainting · Optical Music Recognition · YOLOv11 · Clef Prediction · Digital Preservation

1 Introduction

There are a number of original compositions confined to physical sheets of paper which are prone to continuous degradation. Stains, tears, ink fading, and chemical deterioration progressively compromise the legibility of notation, often leading to irreversible loss of musical content. Among the notational elements most

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 629–638, 2027.
https://doi.org/10.1007/978-3-032-30494-0_39

susceptible to such damage are musical clefs, foundational symbols that anchor the pitch reference of a staff. When clefs, in particular, become missing or illegible, the semantic interpretation of every subsequent note position is destabilized, rendering the correct reconstruction of the passage effectively ill-posed [1].

In order to mitigate this issue, musical restoration (or inpainting) has emerged as a research area that shifts the focus from the principle of visual restoration to the musical domain, allowing the reconstruction of missing elements while preserving coherence with the musical context. Previously, solutions to this problem depended on rigid rules based on musical theory. Today, the landscape has changed with the advancement of Artificial Intelligence, which allows models to learn musical styles and structures directly from data [4]. In this context, Optical Music Recognition (OMR) acts as an essential bridge, translating images of musical scores into a language that the computer can understand [5]. However, we still face a practical challenge: while most restoration research assumes that the data is already digitally organized, real-world preservation requires that we be able to read and extract information from worn images before even attempting any reconstruction.

In this work, we propose an approach that aims to address this gap by formulating a complete note-filling pipeline that integrates visual perception and predictive modeling. Specifically, we combine an optical clef recognition (OMR) stage based on the You Only Look Once (YOLOv11) algorithm to detect clefs and note positions in digitized scores. This stage consists of a statistical processing stage that converts the extracted pitch values into compact meta-features describing the pitch distributions, and a supervised classification stage that predicts the most plausible note, considering only these pitch-derived descriptors. In our experiments, we used musical scores from the PRIMUS dataset [2], with YOLOv11 trained to detect three classes of clefs and note positions with an input resolution of 640×640 pixels and a confidence threshold of 0.7. From the numerical representation of the resulting height (0-21), we constructed 10 descriptive statistical features per detected segment, resulting in a labeled dataset of 950 samples with naturally unbalanced class frequencies. We then compared three traditional machine learning families: k-nearest neighbors, random forest, and gradient boosting, using grid search fitting and 5-part stratified cross-validation, with the weighted F1 score as the main metric to account for class imbalance.

2 Theoretical Framework

In this section, we establish the theoretical framework necessary to bridge the gap between raw digitized scores and predictive modeling, beginning with an exploration of musical reconstruction as a data-driven reconstruction challenge. Thus, in this study we examine the role of musical record reconstruction (OMR) and the advantages of single-stage object detection, specifically YOLOv11, in extracting structured data from degraded manuscripts. Finally, we discuss the use of pitch distribution meta-features as compact and interpretable representations that capture the statistical signatures of musical register and tessitura.

Musical restoration can be understood as something similar to the task of visual restoration, since, given a partially corrupted artifact, the goal is to reconstruct the missing content while preserving coherence with the context [1]. In the musical domain, this problem is contextual, considering that the plausibility of a reconstructed element depends not only on the local notation but also on the structural regularities that govern the melodic contour, harmonic function, rhythm, and stylistic constraints. Initially, computational solutions attempted to repair sections of musical scores using manual rules based on traditional music theory. Over time, the field shifted direction: instead of programming each rule, we began using machine learning so that the models themselves could discover musical patterns directly from the data. Currently, research ranges from classical methods to cutting-edge technologies, such as generative models (VAEs) and Transformers, which can understand the structure of a piece of music in much longer and more complex contexts [4].

Within this context, clef reconstruction is a separate problem. Unlike a note or an accidental, which have their own meaning, the clef is what names the positions on the staff; without it, we don't know the actual pitch of any note that follows. Therefore, recovering a clef is, in fact, a work of deduction: we analyze the remaining notes to understand in which region of the scale the music fits. This makes the problem ideal for statistical methods that look at the distribution of notes (such as the mean and pitch variation), since these characteristics reveal the musical notation even when individual notes are blurred or damaged (Fig. 1).

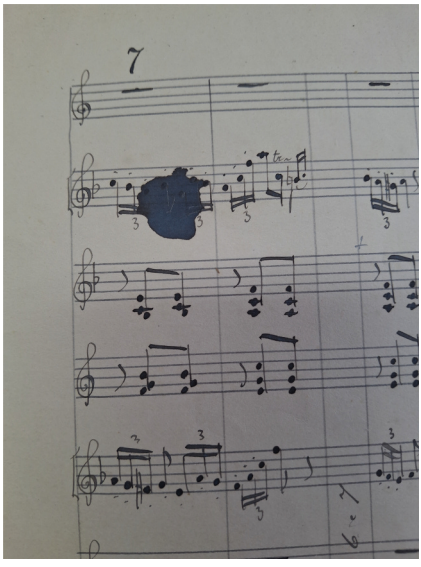

Fig. 1. An example of a music score with missing elements.

OMR is the tool that allows us to extract organized data from a musical score image. Previously, this process followed very rigid steps: it was necessary to erase the staff lines, isolate each symbol, and only then attempt to classify them one by one. Although it worked under ideal conditions, this method easily failed when faced with poor images or manuscripts worn by time. Today, the use of neural networks and object detectors has changed this scenario, allowing the computer to learn to identify symbols directly from the image, without needing those intermediate manual steps [6].

In this work, we use YOLOv11 as the visual basis of the system, treating musical notation as an object detection problem: clefs and notes are located and classified within identification frames. The efficiency of this type of technology for recognizing musical symbols has been proven in recent studies, showing that modern models are very robust at detecting small, close-up details, something common in musical scores [7]. Figure 2 illustrates YOLO being used to find and identify several musical symbols simultaneously.

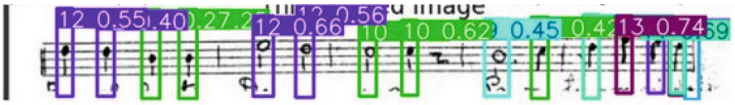

Fig. 2. YOLO applied to OMR: bounding-box detection of musical symbols (e.g., notes, clefs, accidentals) within a score.

Mapping symbolic music into informative numerical representations is a long-standing theme in computational musicology and Music Information Retrieval. Beyond note-by-note encodings, *distributional descriptors* such as mean, variance, skewness, and quantiles have been widely employed to characterize musical style, register, and authorship, capturing invariants that remain meaningful across notational variations.

Accordingly, this study paves way for clef inpainting through statistical meta-features extracted from OMR-derived positional values. Rather than modeling the full symbolic sequence, we represent each staff segment by compact distributional summaries (e.g., median, dispersion, and upper/lower quartiles). This design emphasizes interpretability through data, by making use of meta-features that have direct musical meaning (register, spread, asymmetry) and allow the learned decision function to be inspected in terms of musically plausible evidence. Prior work in mathematical and statistical perspectives on music supports the notion that pitch distributions encode discriminative signatures related to register and tessitura, enabling meaningful differentiation even without explicit symbolic reasoning about staff semantics [3].

Together, these foundations motivate an end-to-end formulation in which OMR extracts positional evidence from images, distributional meta-features summarize that evidence in a compact and interpretable form, and supervised learning performs clef inference as a classification task grounded in register-related statistics.

3 Methodology

We operationalize musical clef inpainting as a supervised classification task grounded on pitch evidence extracted directly from score images. The proposed architecture follows a sequential four-stage pipeline, as illustrated in Fig. 3.

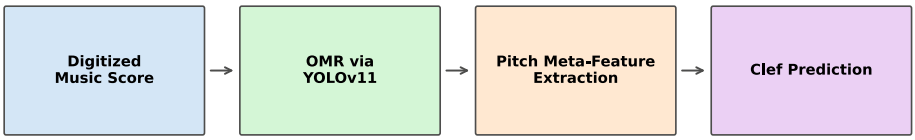

Fig. 3. Diagram illustrating the proposed end-to-end clef inpainting pipeline.

The Optical Music Recognition (OMR) stage is not treated as an external prerequisite, but as an integral component of the pipeline that converts unstructured visual notation into structured, machine-learning-ready tabled data. The downstream classifier is intentionally constrained to use only statistical descriptors computed from extracted pitches, enabling both end-to-end feasibility and interpretability in terms of register and pitch dispersion.

For our experiments, we used the PRIMUS dataset, a large collection of sheet music images widely used in OMR research. Each example in the dataset is identified with one of the three main clefs: *Csol*, *Cdo*, and *Cfa*. The goal is to test whether we can predict which clef should be there using only the statistics of the remaining notes. If this works, we will have a reliable method to "fill in the gap" in cases where the original clef has disappeared or is illegible due to wear and tear on the sheet music.

We processed the sheet music images using the YOLOv11 model, trained to identify the clefs (*Cdo*, *Cfa*, *Csol*) and the position of the notes. We configured the detector with a resolution of 640 pixels, a balance we found to ensure the computer can see small musical symbols without requiring excessive processing. For each item found, the model generates a location frame and a confidence score; to avoid errors and false positives, we only accepted detections where the system was at least 70% certain.

Our approach transforms the position of the notes found into numbers from 0 to 21, which represent the pitch of each note on the staff. This conversion transforms the image into a list of numerical values, retaining only the pitch information necessary to deduce the clef. It is important to note that, at this stage, the term "pitch" refers only to the physical position of the note on the lines and spaces; the actual musical meaning (whether the note is an A or a C) will only be defined after we discover the correct clef.

For each section of the score, we take the positions of the notes and calculate 10 statistical measures that summarize how they are distributed. These characteristics (listed in Table 1) serve as a summary of the notes' behavior: we include the extremes (*minimum* and *maximum*), the center of the distribution (*mean*

and *median*), how much they vary (*standard deviation* and *range*), as well as indicators of note volume and asymmetry. We use standard mathematical definitions to ensure that the calculations are accurate and easy to reproduce.

The resulting dataset contains 950 labeled samples and exhibits an imbalanced class distribution typical of real-world scores: *Csol* (67.7%), *Cdo* (28.6%), and *Cfa* (3.7%).

Table 1. Calculated statistical meta-features from pitch indices.

#	Feature	Description
1	maximum	Maximum pitch index in the segment
2	minimum	Minimum pitch index in the segment
3	mean	Arithmetic mean of pitch indices
4	median	Median of pitch indices
5	standard_deviation	Dispersion of the pitch distribution
6	pitch_count	Number of detected pitches (n)
7	range	Difference between maximum and minimum
8	q1	First quartile (25%)
9	q3	Third quartile (75%)
10	skewness	Asymmetry of the pitch distribution

3.1 Preprocessing and Train–Test Partitioning

Before model training, feature vectors are standardized using StandardScaler. For each feature dimension x, scaling is computed as

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ are estimated on the training split only, preventing information leakage. Data is partitioned into 80% training and 20% test sets using stratification to preserve the empirical class proportions.

Three supervised learning models commonly used for tabular classification are compared in k-nearest neighbors, random forest and gradient boosting. Hyperparameters are selected via grid search on the training data. Specifically, k-nearest neighbors explores neighborhood sizes $k \in \{3, \ldots, 15\}$ and alternative distance metrics; random forest varies the number of trees, maximum depth, and split criteria; and gradient boosting evaluates combinations of learning rate, tree depth, and number of estimators. This procedure ensures that each model is evaluated under a competitive configuration rather than an arbitrary default.

We performed the model selection using stratified 5-fold cross validation with shuffling enabled and a fixed random state (42) to ensure reproducibility. Because

the dataset is markedly imbalanced, we adopt the weighted F1-score as the primary evaluation metric. We denoted $F1_c$ as the class-wise F1-score and w_c as the support (relative frequency) of class c; the weighted F1 we computed as

$$\sum_c w_c \, F1_c,$$

providing an effective assessment of performance, but without overestimating the accuracy of the majority class. For greater comprehensiveness, we also report accuracy on the test set and training time in the results section to characterize the relationship between accuracy and efficiency among the methods.

4 Discussion of Results

According to our analysis, the grid search results revealed a superiority of gradient boosting compared to the other algorithms tested. As shown in Table 2, gradient boosting achieved a weighted F1-score of 0.7242 on the test set, surpassing k-nearest neighbors (0.6950) and random forest (0.6898). This improvement is also reflected in the higher test accuracy (0.7316), although with a higher training cost (61.24 s). Overall, we believe these results indicate that additive sets of shallow decision trees provide an effective inductive bias for key inference under height distribution meta-features, capturing non-linear decision boundaries induced by quartiles, dispersion, and central tendency statistics.

Given the grid search results, the best configuration we found for gradient boosting was a learning rate of 0.1, a maximum depth of 3, a minimum sample split of 2, 50 estimators, and full subsampling (1.0), suggesting a conservative capability that balances expressiveness of learning and prevention of overfitting. This is consistent with the task structure: although clef selection is strongly linked to musical register, its discriminative cues rarely reduce to a single threshold of pitch values and instead rely on distributional signatures that benefit from non-linear combinations. Figure 4 complements Table 2 illustrating the comparative behavior of the models with respect to the reported metrics.

Table 2. Comparative results of machine learning models for clef prediction.

Model	F1-Score (CV)	F1-Score (Test)	Accuracy (Test)	Time (s)
Gradient Boosting	0.6802	**0.7242**	**0.7316**	61.24
KNN	0.6954	0.6950	0.7105	3.05
Random Forest	0.6886	0.6898	0.7105	40.54

In addition to being efficient, this model allows us to understand exactly what it takes into account to decide which clef it is. By analyzing the importance of the features (Table 3), we note that the third quartile (q_3) is the most influential

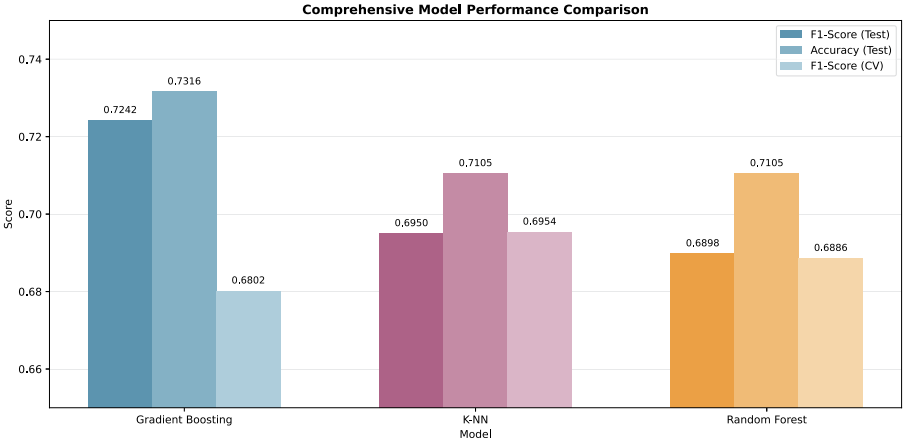

Fig. 4. Comparison of performance metrics in clef prediction.

factor, followed by the standard deviation and the median. This shows us that, to identify the clef, the model prefers to observe where the highest part of the notes is concentrated and how much they vary, rather than focusing only on extreme values or an isolated average.

The maximum and minimum values of the notes had almost no importance (0.0093 and 0.0003), which proves that a single very high or low note is not sufficient to reliably predict the clef. Because of this, we removed these two characteristics before the final training, simplifying the model without losing quality. This result makes perfect sense from a musical point of view: clefs serve to indicate the region where the music "lives" (the register and the tessitura), something that becomes much clearer through the concentration of the highest

Table 3. Feature importance in the Gradient Boosting model.

Rank	Feature	Importance
1	q3	0.2940
2	standard_deviation	0.1837
3	median	0.1288
4	mean	0.1222
5	skewness	0.1097
6	q1	0.0908
7	pitch_count	0.0368
8	range	0.0244
9	maximum	0.0093
10	minimum	0.0003

notes (q_3) and their variation than through extreme notes that appear only occasionally.

5 Conclusion and Future Work

In this article, we tested whether it was possible to recover musical clefs by combining computer vision and data analysis. We created a workflow that first uses YOLOv11 to detect clefs and notes, transforms these images into numbers, and summarizes each section of the score with 10 simple statistical measures. To test whether these numbers were sufficient to predict the clef, we evaluated three algorithms (k-NN, Random Forest, and Gradient Boosting), focusing on the F1 score to ensure fair results, since some clefs appear much more frequently than others in our dataset.

Our analyses showed that Gradient Boosting was the most efficient (F1 of 0.7242), confirming that notes alone already give very strong clues about which is the correct clef. In addition to correctly predicting the result, the model gave us a logical explanation: it prioritizes the region where the highest notes are concentrated and how much they vary, rather than focusing on isolated extreme notes. This makes perfect musical sense, since the clef serves precisely to define the region (or tessitura) in which the music should be read.

For the future, we have three main paths. First, we want to expand the system to identify other essential symbols, such as key signatures and time signatures. Second, we intend to test the model on real historical manuscripts, which are much more challenging than digital images. Finally, we want to test whether including small melodic or rhythmic excerpts can help the model become even more accurate, combining the statistics we already use with the musical context of each passage.

Acknowledgment. The authors acknowledge partial funding from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the National Council for Scientific and Technological Development - Brazil (CNPq) - Funding Codes 442344/2023-6, 444602/2024-0, and 407466/2025-8; the Espírito Santo Research and Innovation Support Foundation (FAPES)–Call No. 02/2026 (Participation in Technical-Scientific Events)–Grant No. 428/2026 and project #2024-9G1WP BPC; and the Minas Gerais Research Funding Foundation (FAPEMIG)

References

1. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp. 591–596 (2011). https://ismir2011.ismir.net/papers/OS6-1.pdf
2. Calvo-Zaragoza, J., Rizo, D.: End-to-end neural optical music recognition of monophonic scores. Appl. Sci. **8**(4), 606 (2018)

3. Chew, E.: Mathematical and Computational Modeling of Tonality: Theory and Applications. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-9475-1
4. Huang, C.Z.A., et al.: Music transformer: generating music with long-term structure. In: International Conference on Learning Representations (ICLR) (2019). https://openreview.net/forum?id=rJe4ShAcF7
5. Pacha, A., Jr., J.H., Calvo-Zaragoza, J.: A baseline for general music object detection with deep learning. Appl. Sci. **8**(9), 1488 (2018). https://doi.org/10.3390/app8091488
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). https://openaccess.thecvf.com/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html
7. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475 (2023). https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html