Decentralized Optical Music Recognition Using YOLO and FedGP for Music Education

Andrea Marselletti
Universidad de Extremadura
Mérida, Spain
amarsell@alumnos.unex.es

Elia Pacioni
HES-SO Valais-Wallis
Universidad de Extremadura
elia.pacioni@hevs.ch

Francisco Fernández de Vega
Universidad de Extremadura
Mérida, Spain
fcofdez@unex.es

Davide Calvaresi
HES-SO Valais-Wallis
Sion, Switzerland
davide.calvaresi@hevs.ch

***Abstract*—Federated learning (FL) offers a privacy-preserving paradigm for collaborative model training, where data remain on local devices. Leveraging this capability, we explore privacy-aware Optical Music Recognition (OMR) by coupling a state-of-the-art YOLOv9c detector with FedGP, a genetic-programming-based aggregation strategy tailored to highly non-IID client distributions. We cast OMR as the end-to-end transcription of printed and handwritten four-part harmony scores into structured MusicXML, a task complicated by symbol variability, staff-line distortions, and non-musical artifacts. To support this formulation, we assembled a hybrid corpus of $1\,810$ page images – 810 augmented handwritten exercises and $1\,000$ automatically annotated digital scores—comprising $112\,024$ bounding-box annotations across 166 symbol classes. The pages are deliberately partitioned in a non-IID manner among ten virtual clients. Training proceeds by freezing the first 22 layers of the pre-trained YOLOv9c backbone and fine-tuning the detection head for 160 local epochs per client. Parameter updates are transmitted every 20 communication rounds and aggregated over 500 global epochs. Comprehensive experiments show that FedGP consistently outperforms the canonical FedAvg baseline. At communication round 8, FedGP attains a mean $mAP50$ of 0.6775 ± 0.0179, significantly exceeding FedAvg's 0.6467 ± 0.0095. These findings demonstrate that genetic-programming-driven aggregation mitigates client heterogeneity while preserving data locality and imposing modest computational demands on edge devices. Overall, the study confirms the viability of FL for large-scale, privacy-conscious OMR and establishes FedGP as a robust alternative to standard aggregation schemes under challenging non-IID conditions.**

***Index Terms*—Optical Music Recognition, 4-part harmonization, Federated Learning, FedAVG, FedGP, YOLOv9**

I. Introduction

Digital support tools are pervading modern teaching and are boosting content personalization and adaptation to students' pace and level. Many areas benefit from e-learning systems and domain-specific applications, In this context, we address the teaching of 4-part harmony for conservatory students. Such a subject is compulsory, and to date, its *correction* is handled manually by the professors – entailing a long waiting time often resulting in significant delays and a lack of immediate feedback.

This work was partially supported by the Validate-H (PInter 14-2025), the Spanish Ministry of Economy and Competitiveness (PID2020-115570GB-C21, PID2023-147409NB-C22), funded by MCIN/AEI/10.13039/501100011033, and the Junta de Extremadura (GR24142)

Fernández et al. developed Sharpmony [1] [1], [2], an application that supports students in composing and correcting sheet music. Such an application can harmonize 4-part harmony scores by employing evolutionary algorithms [3]. To date, it can produce error-free scores using a Machine Teaching and Human Learning approach combined with local search, yet relying on digitally composed music sheets [4].

Students asked for the evaluation of handwritten music sheets. To enable users to write music sheets on multiple supports (e.g., paper, MuseScore [2], etc.) and digitize them within the platform for automatic correction, we need to include OMR within Sharpmony. The OMR problem is widely explored in the literature [5]–[7]. Although the results obtained for digital sheet music are excellent, the recognition quality of handwritten scores is still an open challenge.

To address this limitation, this paper proposes implementing and validating an FL OMR system capable of recognizing both digital and manuscript scores. Adopting a distributed approach is particularly advantageous, as it allows students to organize training data and decreases the workload on our servers. The problem of data shortage for model training is solved with an automatic labeling technique based on student data and data augmentation on existing data. In addition, we analyze and compare two aggregation methods for FL: FedAVG (standard FL-aggregator) and FedGP (innovative Genetic Programming (GP)-based aggregator). The latter, recently introduced by Pacioni et al. [8], [9], has been tested only on small neural networks. This paper evaluates its performance by applying it to the YOLO [10] algorithm, thus expanding its validation scope.

The remainder of this paper is organized as follows: Section II outlines the theoretical background and technologies relevant to the addressed problem. Section III presents the problem and the challenge to be addressed. Section IV details the methodology, the dataset creation, data distribution, and the process followed to adapt FedGP to the new project. Section V describes design and implementation. Experimental results are discussed in Section VI, comparing FedGP and FedAVG in this new context and analyzing the performance of the obtained model on digital and manuscript datasets. Section VII presents

[1] https://sharpmony.unex.es
[2] https://musescore.org/en

the study's limitations. Finally, Section VIII summarizes the findings, highlights contributions, and proposes future research methods.

II. State of the art

Recent research in OMR for music education has progressively refined techniques to support digital and handwritten score recognition [5], [6], [11]. Novotný and Pokorný [12] outline a four-phase OMR pipeline—preprocessing, symbol segmentation, object recognition, and semantic reconstruction—highlighting challenges like staff-line skew, symbol fragmentation, and high manuscript variability. They also emphasize the scarcity of large, heterogeneous ground-truth corpora, and the lack of standardized evaluation metrics hinders reproducibility and benchmarking in OMR research.

In parallel, YOLO-based approaches have been validated for their high precision in detecting objects as outlined in studies that report excellent detection metrics [10]; Zong et al. [13] provide a comprehensive survey of the YOLO series from v1 through v8, detailing key architectural innovations such as Darknet backbones, anchor-based vs. anchor-free heads, multi-scale training, and novel loss functions. The paper also critically assesses YOLO's challenges with small objects and crowded scenes; meanwhile, research by Garrido-Muñoz et al. [7] has advanced a holistic image-to-graph methodology to manage the complexities of manuscripts better, as it is still an open challenge in the OMR field. An interesting development in addressing this issue was the introduction of FL, which has enabled the decentralization of model training to safeguard data privacy and reduce server load, particularly in heterogeneous environments. In this framework, the FedGP aggregator developed by Pacioni et al. [8], [9] has been shown to outperform traditional FedAVG by dynamically adapting aggregation functions to non-IID data. This synthesis of advanced recognition, distributed learning, and adaptive aggregation techniques could pave the way for scalable, robust, and privacy-preserving OMR systems that address the evolving demands of modern music education while tackling the open challenges of this complex and extensive field of study.

III. The Challenge

The challenge of this field of research refers to the technology and process of automating the interpretation and conversion of musical notation found in scanned images, printed scores, or manuscripts into a structured digital format. This process involves analyzing the visual elements of the music notation and translating them into a digital music format. It deals with a distinct set of symbols, including note types, rests, key signatures, and other musical markings, each carrying a specific meaning. The precise location of each symbol on the staff is crucial; even a slight misplacement can alter the intended pitch or rhythm. Moreover, differences in notation styles and the condition of original scores can lead to variations in how symbols appear and are arranged, further complicating recognition and interpretation. These discrepancies create additional challenges in accurately capturing the subtle nuances of musical expression, as seen in Figure 1. Thus, making the automatic digitization of music scores a research field essential for transforming musical notation into a structured digital format, enabling more straightforward analysis, editing, and integration with modern music technology. OMR systems address these challenges by incorporating robust preprocessing pipelines—such as staff-line detection and removal, adaptive thresholding for binarization, skew and distortion correction, and noise reduction—to standardize input quality before symbol detection. Advanced segmentation algorithms must then distinguish overlapping elements (e.g., beams spanning multiple voices, ledger lines, and tuplets) and isolate individual symbols within dense polyphonic passages. Beyond basic note and rest detection, accurate recognition also requires handling higher-level notational constructs—ornamentations (trills, mordents), articulations (staccato, accents), and expressive markings (dynamics, tempo indications)—which may exhibit significant stylistic variability in both printed editions and manuscripts. Handwritten scores, in particular, introduce further hurdles: inconsistent stroke thickness, variable spacing, ink smudges, and non-musical artifacts (text annotations, fingerings) demand context-aware filtering and error-correction strategies. Finally, converting the detected symbols into interoperable formats (e.g., MusicXML, MEI) while preserving spatial relationships, rhythmic alignment, and semantic context is crucial for downstream applications, such as automated playback, analysis, and harmonization. Maintaining low symbol- and measure-level error rates across heterogeneous datasets remains an active area of research, motivating the integration of machine-learning methods with musicological grammar rules and user-in-the-loop correction mechanisms.

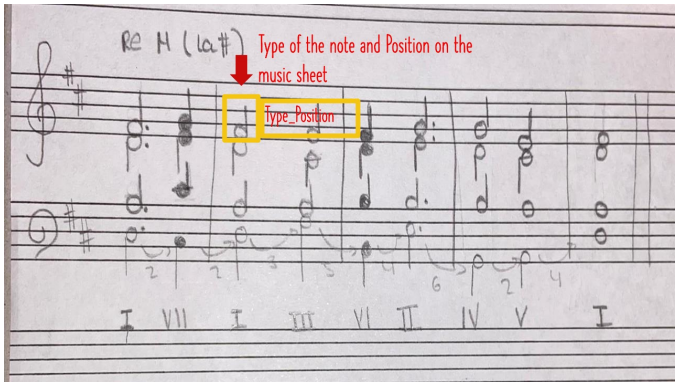


Fig. 1: Handwrittendataset'ss labels challenge

IV. Methodology

The methodology employed in this study comprises three key steps. S1: Dataset creation to train the neural network by combining handwritten and digital sheet music images. Given the limited availability of manually annotated manuscript scores (approximately 100 photos in our students' exercise data), the dataset is enhanced with data augmentation techniques and 1,000 digital scores generated by students using

the Sharpmony application. The digital sheet music are automatically annotated using a developed algorithm designed to process the original MusicXML files, reading the information and coordinates of each element in the sheet music, as shown in Figure 2. This process results from a hybrid dataset consisting of both types of sheet music. S2: Addition of the FedGP code and the entire FL pipeline and incorporation of the YOLO algorithm within the peers. Each client fine-tunes a frozen-backbone YOLOv9c head on its private data, and FedGP evolves aggregation functions to handle non-IID updates. S3: Experimentation, comparative analysis, and validation of the models obtained by FedAVG and FedGP. The evaluation metric is mAP50, which allows us to assess detection performance across different settings.

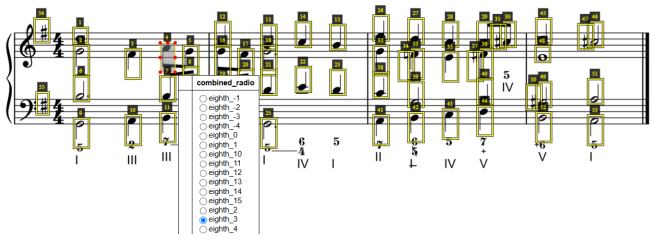


Fig. 2: Digital dataset's labels

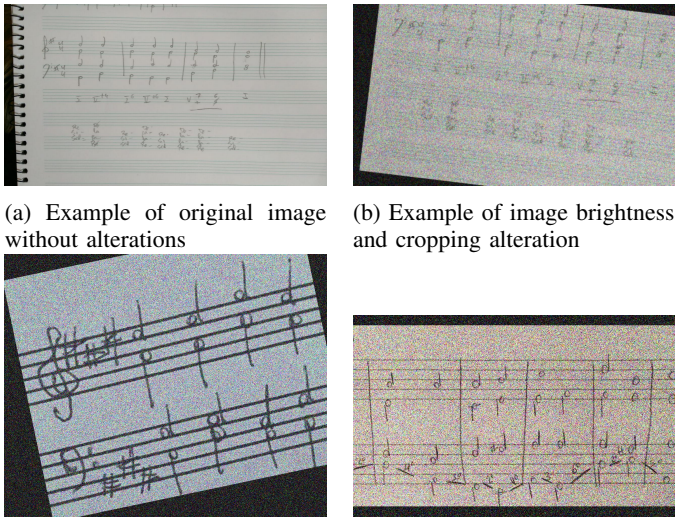
V. Design and Implementation

Our design and implementation leverage a modular approach, integrating multiple technologies for music notation analysis and federated model training. The project is developed entirely in Python, utilizing libraries such as PyTorch, Ultralytics (YOLOv9), and DEAP. In the dataset construction phase, the work is divided into two parts. First, data augmentation is applied to handwritten scores by performing transformations (e.g., ShiftScaleRotate, RandomBrightnessContrast, HueSaturationValue, GaussNoise, Blur), explained in Table I, that expand the dataset from 135 images to 810. This augmentation process not only enlarges the dataset but also introduces a variety of conditions to enhance the model's robustness against diverse handwriting styles and imaging conditions as shown in Figure 3.

TABLE I: Data Augmentation Transformations Parameters

Transformation	**Parameters**
HorizontalFlip	p = 0
ShiftScaleRotate	shift_limit = 0.1, scale_limit = 0.1, rotate_limit = 15, p = 0.5
RandomBrightnessContrast	p = 0.5
HueSaturationValue	p = 0.5
GaussNoise	p = 0.3
Blur	blur_limit = 7, p = 0.3

Second, to automatically label digital sheet music, we use MuseScore to obtain uncompressed MusicXML files that provide detailed information on music symbols and their locations through the coordinates given in the XML files. A



(a) Example of original image without alterations

(b) Example of image brightness and cropping alteration

(c) Example of image blur, saturation, and cropping alteration

(d) Example of image brightness and noise alteration

Fig. 3: Original and altered handwritten images used for the dataset

custom parser converts these MusicXML files into 1024-pixel-wide images and generates annotations in the VGG Image Annotator format. This procedure yields a comprehensive hybrid dataset comprising 1,810 images and 112,024 elements. The dataset comprises 166 classes. Figure 4 illustrates the distribution of 8 labels, chosen as examples, among the clients and the server, which characterizes the dataset's non-IID nature. It specifically shows how quantity and label skew are present. A Python script randomly splits the dataset. However, a second script checked for quantity and label skew. If there were no unbalance, the script would randomly move images from one client to another and proceed with verification again. An additional challenge is posed by the random distribution of manuscript and digital scores, which makes it impossible to control the number of manuscripts each client receives.

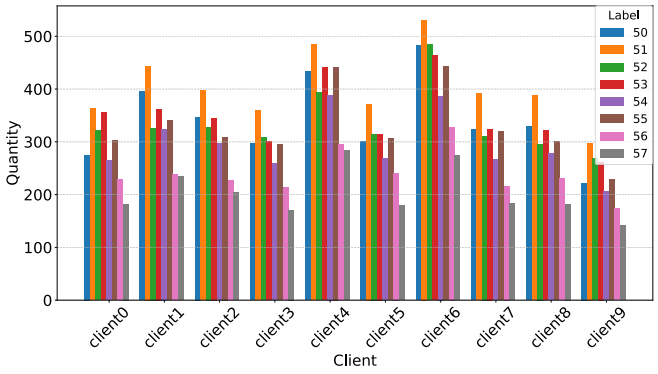


Fig. 4: NON-IID distribution of 8 labels among clients.

Table II explores label distortion in greater depth. No client has information about all classes, nor does the server. This

better represents the real-world situation in FL.

The same dataset is used for all clients during the validation phase. This ensures that all clients are evaluated equally. The same applies to the evaluation of the aggregate model. The validation dataset contains occurrences of all 166 classes. About 80% of the dataset is split between the client and server for training, and 20% is used for validation.

Device	Classes
Server	102
Client0	127
Client1	128
Client2	125
Client3	124
Client4	128
Client5	120
Client6	132
Client7	131
Client8	126
Client9	123

TABLE II: Label distribution on each device. Highlight the Label Skew issue.

Our experiments use the YOLOv9c architecture, with 25.5 million parameters (102.8 billion FLOPs), pretrained on the COCO dataset; it achieves a mean average precision on a server and 10 clients, each equipped with a separate dataset study, we simulate the FL environment by running all the code on the server while isolating the clients as if they were in a real application. We use YOLO's internal parameter optimizer to select batch size, optimizer, learning rate, and momentum. The selection of global and local epochs, along with the aggregation frequency, was determined through an empirical study:

- Global Epochs: 500.
- Local Epochs: 160.
- Aggregation Frequency: 20.
- Early Stopping: true.
- Global Patience: 20.
- Local Patience: 10.

We apply transfer learning techniques to enhance the efficiency of our training process further. Specifically, we freeze the first 22 layers of the network, thereby preserving the learned feature representations from the pre-trained model. Only the final layers—the head of the network—are retrained.

Two aggregation methods were tested: FedAvg and FedGP. In the case of FedGP, we employ 50 individuals with 10 generations, set elitism to 1, and use a 40% mutation probability and 60% crossover probability, as shown in the Table III. The initial population is generated using the *Ramped Half-and-Half* strategy, which combines the "grow" and "full" methods to promote structural diversity. During evolution, *Random Mutation* is applied as the mutation operator—replacing a randomly chosen subtree with one of uniformly distributed size—while *Tournament method* drives tournament-based parent selection. Crossover occurs randomly at a single cutting point.

In its original formulation as mentioned in [9], FedGP relied on a core repertoire of eight primitives, as indicated in

TABLE III: GP Parameters

Description	**Value**
Population size	50
GP generations	10
Elitism	1
Mutation probability	0.4
Crossover probability	0.6
Pop. creation	Ramped Half-and-Half
Selection	Tournament
Crossover	One Point

the Function set 1, to construct the update function. To broaden its expressive capacity—particularly for capturing non-linear interactions and richer functional forms—we have now introduced four additional primitives: `torch.pow`, `torch.log`, `torch.sin`, and `torch.cos`.

$$F = \left\{ \begin{array}{l} \text{torch.add, torch.sum, torch.sub, torch.mul,} \\ \text{torch\_protected\_div, torch\_mean,} \\ \text{torch\_median, torch.abs, torch\_protected\_sqrt} \\ \text{torch.pow, torch.log, torch.sin, torch.cos} \end{array} \right\} \quad (1)$$

In the Terminal set 2 we see the terminals used by FedGP, where each CLIENT_i placeholder is replaced at evaluation time by the original tensor.

$$T = \left\{ \mathrm{CLIENT}_1, \mathrm{CLIENT}_2, \ldots, \mathrm{CLIENT}_n \quad \right\} \quad (2)$$

VI. Results and Analysis

The experimental evaluation indicates that the overall system performance is satisfactory across various music scores. This allows us to state that the results are promising and that if implemented in production, the system can progressively improve during use in real-world scenarios. Additionally, the conducted experiments further validate the system's robustness. The FedGP aggregation method has demonstrated superior performance with respect to FedAVG. By dynamically evolving aggregation functions to better handle non-IID data across clients, FedGP produces a more accurate global model in each test.

Each experiment was repeated 4 times with 8 rounds of aggregation.

Table IV shows the mean and standard deviation for FedGP and FedAVG. We can see that at round eight, FedGP gets a maximum value of 0.6775 against 0.6467 obtained by FedAVG.

The results shown in Figure 5 show that FedGP consistently performs higher than FedAVG in each round of aggregation. Although FedGP's standard deviation increases in round eight, it remains higher than FedAVG.

Looking specifically at the difference between digital and manuscript music sheets, we note that with digital music sheets the system achieves excellent quality (Figure 6). Conversely, handwritten scores (while more challenging) are still processed with good results (Figure 7). Specifically, in Figure 6 we see how FedGP labels 7 more notes than FedAVG, while

TABLE IV: Performance metrics across aggregation rounds for FedGP and FedAVG

R	FedGP Av.	FedGP St. dev	FedAVG Av.	FedAVG St. dev
1	0,4772	0,0077	0,4382	0,0051
2	0,5607	0,007	0,5165	0,0075
3	0,6087	0,0086	0,5715	0,0078
4	0,6385	0,0034	0,6062	0,0065
5	0,6512	0,0066	0,6177	0,0067
6	0,6612	0,0114	0,6307	0,0054
7	0,67	0,0161	0,641	0,0085
8	0,6775	0,0179	0,6467	0,0095

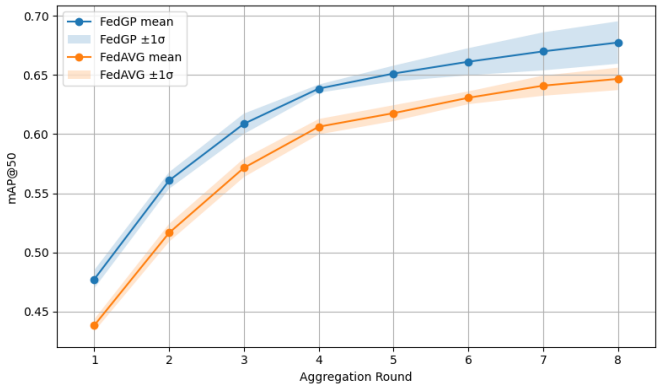
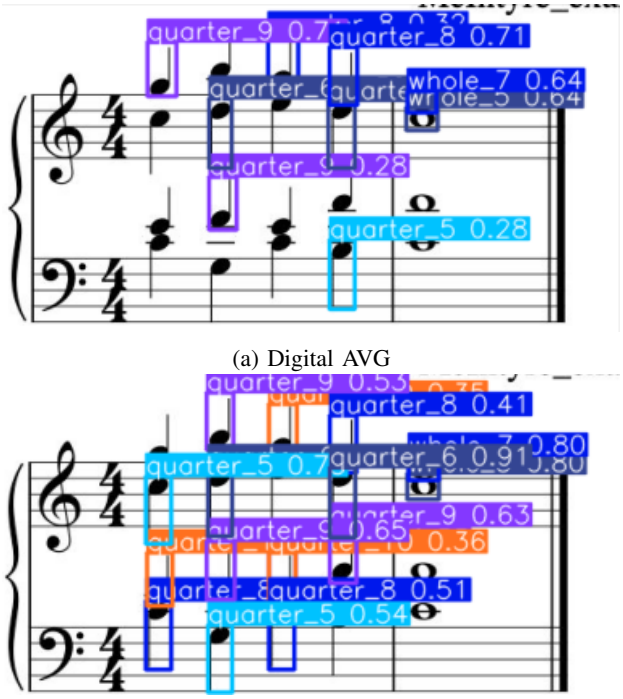


Fig. 5: Experiments with distribution of non-IID data. The blue represents the FedGP, while the orange represents the FedAVG.



(a) Digital AVG

(b) Digital GP

Fig. 6: Inference on digital sheet music.

consistently assigning high confidence levels to all detected notes. It is important to note that FedAVG has been observed to mislabel various notes by assigning erroneous positional identifiers. For instance, it inaccurately labels a note as quarter_5 rather than the correct designation of quarter_8, among other discrepancies. A similar situation occurs in Figure 7 for manuscript scores, with 3 more notes for FedGP, again demonstrating clear superiority with respect to the confidence levels assigned to its note detections.

To evaluate whether the performance difference between FedGP and FedAVG is statistically significant, we conducted a paired-sample t-test across eight communication rounds. Each round's value corresponds to the mean performance obtained over four independent runs per method, under identical experimental conditions.

Before applying the test, we verified that the assumptions required for the paired t-test were satisfied. The observations were independent, as each method was evaluated across separate runs. Moreover, we confirmed that the distribution of the paired differences between FedGP and FedAVG performances did not significantly deviate from normality, as indicated by the Shapiro–Wilk test (p = 0.370). These conditions justify the application of the t-test.

The test revealed a highly significant difference between the two methods ($p < 0.001$), supporting the conclusion that FedGP consistently outperforms FedAVG under the given non-IID data setting. This result reflects a systematic performance advantage rather than random fluctuation and reinforces the robustness of FedGP in heterogeneous data scenarios.

These trends confirm that dynamically evolving the aggregation function via GP captures client heterogeneity under non-IID conditions more effectively.

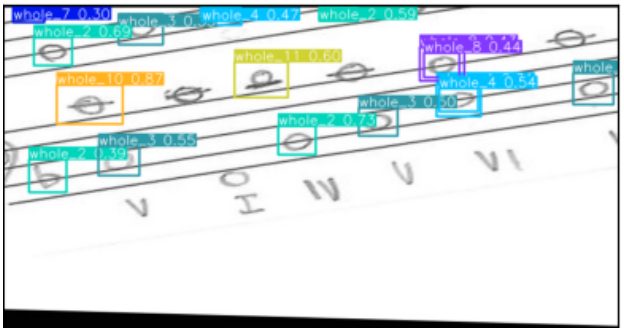
VII. Limitations & Challenges

Despite the promising results, our work faces several limitations that guide ongoing improvements.

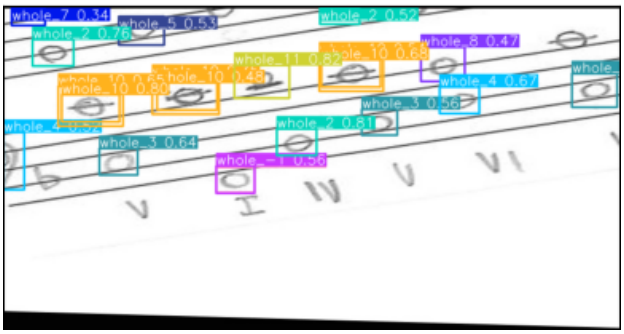
First, the manuscript subset of our custom hybrid dataset remains relatively small, comprising only 135 annotated images. This limited quantity may constrain the model's generalization to diverse handwriting styles: variations in stroke weight, slant, engraving conventions, and even writing instruments (pen versus pencil) can dramatically alter symbol appearance. So we are actively expanding the corpus by sourcing additional manuscripts and streamlining the manual annotation process. By increasing our handwritten data quantity and stylistic breadth, we expect YOLO to learn richer, more generalizable visual features, improving mAP on manuscript scores and narrowing the gap between printed and handwritten inputs.

Another limitation is FedGP's execution time, which is significantly more time-consuming than FedAVG's. We want to work on transfer learning within GP to improve this.

Finally, to make the experiments more realistic, we aim to simulate client churn (i.e., arrival and death) and data dynamics (i.e., arrival and partial deletion), which are common in real-world applications.



(a) Handwritten AVG



(b) Handwritten GP

Fig. 7: Inference on handwritten sheet music

VIII. Conclusions

This paper addresses the OMR challenge using a hybrid dataset built on Sharpmony student data in an FL setup. In particular, an FL approach is used to train the YOLO model, testing FedAVG and FedGP as aggregation methods.

The current performance (excellent for digital scores and promising for handwritten ones) demonstrates the system's potential to serve as a reliable tool for music notation analysis and 4-part harmonization support. With FedGP's superior performance over FedAVG, the FL component has proven effective in handling heterogeneous datasets while preserving data privacy and progressively improving the model with each round of aggregation.

Future work includes addressing the two main bottlenecks identified in Section VII: dataset size and runtime. We will expand and diversify the handwritten corpus by sourcing additional student manuscripts and streamlining the annotation workflow. Next, we plan to add an active-learning loop to the Sharpmony platform, allowing students to flag misrecognized symbols and help refine the model. We will pilot a classroom once these enhancements are complete to bridge the gap between lab experiments and real-world teaching practice. The crucial step of our future work envisions an integrated pipeline that combines FL with Active Learning. In this enhanced system, students will continuously improve the model by uploading their scores. Additionally, they can actively participate in the development process by reporting recognition errors and providing feedback on labels, which will be used to further refine the training process. With the help of Sharpmony's robust community of 4,000 users, this approach paves the way for large-scale testing and continuous system enhancement, ultimately ensuring that the tool evolves concurrently with user needs and emerging technological advancements.

References

[1] F. Fernandez de Vega, J. Alvarado, and M. Morita, "Caepia-app competition: Sharpmony a computational intelligence based tool for 4-part harmony," in *Proceedings XIX CAEPIA*, 2021, pp. 1009–1012.

[2] F. Fernández De Vega, J. Alvarado, A. Sánchez, M. Serrano, and E. Pacioni, "Evolutionary algorithms: A new hope for the future of music teaching," in *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, 2023, pp. 65–66.

[3] F. Fernandez de Vega, "Revisiting the 4-part harmonization problem with gas: A critical review and proposals for improving," in *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE Press, 2017.

[4] E. Pacioni and F. Fernández De Vega, "Combining local search and directed mutation in evolutionary approaches to 4-part harmony," in *Proceedings of the 27th European Conference, EvoApplications 2025, Held as Part of EvoStar 2025*, 2025.

[5] M. Mortia-Hernández, F. Fernández De Vega, and J. Villegas Cortez, "Aplicación de técnicas de aprendizaje profundo al reconocimiento óptico de partituras satb," in *Conferencia de la Asociación Espanola para la Inteligencia Artificial*, 2021, pp. 411–416.

[6] F. F. De Vega, J. Alvarado, and J. V. Cortez, "Optical music recognition and deep learning: An application to 4-part harmony," in *2022 IEEE Congress on Evolutionary Computation (CEC)*, 2022, pp. 01–07.

[7] C. Garrido-Munoz, A. Rios-Vila, and J. Calvo-Zaragoza, "A holistic approach for image-to-graph: application to optical music recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 25, no. 4, pp. 293–303, 2022.

[8] E. Pacioni, F. Fernández De Vega, and D. Calvaresi, "Towards a meaningful communication and model aggregation in federated learning via genetic programming," in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, INSTICC. SciTePress, 2025, pp. 1427–1431.

[9] E. Pacioni, F. Fernández de Vega, and D. Calvaresi, "Fedgp: Genetic programming for evolutionary aggregation in federated learning with non-iid data," in *Proceedings of the 27th European Conference, EvoApplications 2025, Held as Part of EvoStar 2025*, 2025.

[10] C.-Y. Wang and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," 2024.

[11] A. Pacha, J. Hajič, and J. Calvo-Zaragoza, "A baseline for general music object detection with deep learning," *Applied Sciences*, vol. 8, no. 9, 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/9/1488

[12] J. Novotný and J. Pokorný, "Introduction to optical music recognition: Overview and practical challenges," in *CEUR Workshop Proceedings*, vol. 1343, 01 2015, pp. 65–76.

[13] C. Zong, K. Meng, J. Sun, and Q. Zhou, "Real time object recognition based on yolo model," in *2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2023, pp. 197–202.