**Reviewer #1:**

1. Include all important numerical results in the abstract.

We appreciate the reviewer’s suggestion. In the revised manuscript, we have **updated the abstract to include key numerical results** that highlight the performance of our proposed method.

1. Better highlight new contributions and novelty of your work in the introduction section.

We appreciate the reviewer’s suggestion. In the revised manuscript, we have enhanced the introduction to better highlight the novelty and contributions of our work. We have explicitly outlined the key contributions and novel aspects within the introduction itself for improved clarity. The unique aspects of our ensemble approach, particularly the integration of CNN and INN for instrument recognition, are now more clearly emphasized. Additionally, we have strengthened the discussion after literature review on how our method differs from previous approaches, specifically in terms of its ability to handle variable-length audio inputs, improve computational efficiency, and achieve state-of-the-art performance. These revisions ensure that our contributions are clearly articulated and effectively distinguish our work from existing research.

1. Enhance linkage to recent literature that demonstrates the great potential and usefulness of different machine learning models, such as the neural network (doi: 10.1108/AJEB-01-2024-0007; 10.18282/gfr.v6i1.3491; 10.1108/JM2-09-2023-0207) and Gaussian process regression (doi: 10.1177/03019233241254891; 10.1108/JM2-12-2023-0315; 10.1007/s42824-024-00123-y), for modeling complicated (nonlinear) patterns across a diverse variety of research fields to better motivate your present work's exploration of machine learning models.

We appreciate the reviewer’s suggestion and have incorporated recent literature to strengthen the motivation for our work. Specifically, we have expanded the related work sections to discuss the **usefulness of neural networks and Gaussian process regression** in capturing complex nonlinear patterns across various research domains. The recommended references have been reviewed and cited appropriately.

1. Provide a summary of previous studies in terms of what has been done, what is still missing, and correspondingly your contributions.

We appreciate the reviewer’s suggestion. In the revised manuscript, we have provided a more structured summary of previous studies in the **Related Work** section. This includes a discussion of existing approaches in music instrument recognition, highlighting their methodologies, strengths, and limitations. Specifically, we have outlined what has been done in prior studies, including the use of CNNs, recurrent architectures, and handcrafted feature-based methods. We then identify key gaps, such as the challenges in handling variable-length audio inputs, limitations in generalization to multiple predominant instruments, and computational inefficiencies. To address these gaps, we clearly articulate our contributions, emphasizing how our proposed ensemble method—integrating CNN and INN—improves accuracy, computational efficiency, and robustness compared to state-of-the-art approaches. These additions provide a clearer motivation for our work and establish its significance in the field.

1. Improve the quality of the figures to make them clearly visualized by using high-resolution formats. Improve the organization of the tables to make them clearly laid out. Add more explanations to the figures and tables to make them self-explanatory.Offer more detailed discussions of the results by linking to the figures and tables.

We appreciate the reviewer’s valuable feedback. In the revised manuscript, we have made significant improvements to the figures, tables, and discussion of results. Specifically, we have replaced all figures with **high-resolution versions** to enhance clarity and readability. The tables have been reorganized for **better structure and readability**, ensuring consistency in formatting and alignment. Additionally, we have **added detailed captions** to both figures and tables to make them more self-explanatory. To strengthen the discussion, we have revised the **Results and Analysis** section by explicitly referencing the figures and tables. These improvements enhance the overall presentation and ensure that the figures and tables effectively support the discussion.

1. Use more mathematical formulae and equations to demonstrate your proposed approach.

We appreciate the reviewer’s suggestion. In the revised manuscript, we have **incorporated additional mathematical formulations** to better illustrate our proposed approach.

1. Conduct further benchmark analysis to demonstrate the advantage or disadvantage of your method. More comparisons with some traditional methods would be nice to have to see how much the improvement is by adopting your proposed advanced methods.

We appreciate the reviewer’s suggestion. In response, we have conducted additional experiments using **handcrafted features** such as **MFCC, spectral centroid, zero-crossing rate, and chroma STFT**, which were applied to a **Deep Neural Network (DNN) framework**. Furthermore, we have experimented with **Support Vector Machines (SVM)** using these handcrafted features. The results of these additional experiments have been **tabulated in Table [6]** and compared against both our proposed method and state-of-the-art (SOTA) approaches. This expanded benchmark analysis provides a clearer understanding of the improvements introduced by our CNN-INN ensemble method in contrast to traditional handcrafted feature-based approaches. The discussion section has also been updated to analyze the performance differences and highlight the benefits of our proposed deep learning-based solution.

1. Elaborate on the potential limitations of your work, and correspondingly, the path to future studies in the conclusion. More carefully proofread your manuscript.

We appreciate the reviewer’s valuable feedback. In the revised manuscript, we have expanded the Conclusion section to discuss the potential limitations of our study and outline future research directions. Specifically, we have:

* Acknowledged the limitations of our approach, including possible constraints related to dataset diversity, generalization to unseen instrument combinations, and computational efficiency for real-time applications.
* Outlined potential future work, such as exploring more advanced feature fusion techniques, incorporating transformer-based architectures, improving robustness to noisy environments, and evaluating performance on larger and more diverse datasets.
* Revised the manuscript with careful proofreading, addressing grammatical inconsistencies and improving overall readability.

These revisions ensure a more balanced discussion of our work and provide a roadmap for future studies in this domain.

**Reviewer #4:**

1. The complexity challenges of all the investigated approaches could be better emphasized.

We appreciate the reviewer’s suggestion. In the revised manuscript, we have included a detailed complexity analysis of the investigated models. Specifically:

* A comparative table (Table 5) has been added, presenting the number of parameters ,training time and inference time of each model.
* We have discussed the trade-off between accuracy and computational cost, emphasizing how our proposed ensemble method effectively balances performance and efficiency compared to SOTA approaches.
* The discussion now highlights the computational feasibility of our method for real-world applications.

These additions provide a clearer understanding of the complexity challenges associated with different models and reinforce the strengths of our approach.

1. The authors should mention learning approaches when data scarcity is a problem (e.g. A Survey on Deep Learning Tools Dealing with Data Scarcity: Definitions, Challenges, Solutions, Tips, and Applications, Journal of Big Data, 2023).

We appreciate the reviewer’s suggestion regarding learning approaches for addressing **data scarcity**. In the revised manuscript, we have included a discussion in **Section X**, citing the suggested survey paper and relevant techniques as potential solutions for low-resource instrument classes and discussed as future directions.

1. To facilitate reproducible research, if possible, I suggest that the author to release the related source codes on github.com, the website of the authors' research group or similar website. This could have a positive impact on the academic community.

We appreciate the reviewer’s emphasis on **reproducibility** and the positive impact of open-source research. To facilitate further research in music instrument recognition, we plan to release the source code and trained models on **Github.** The link to the repository will be included in the final published version of the manuscript.

1. There are several inconsistencies in the reference section that should be corrected (e.g., the name of the authors are written in various forms - Han, Y., Jifeng Dai, reference [13] has no authors etc).

We appreciate the reviewer’s attention to detail regarding reference formatting. In the revised manuscript, we have carefully **standardized the author names** to ensure consistency throughout the reference section. Additionally, missing author details (e.g., in reference [13]) have been corrected, and all references have been formatted according to the required citation style

**Reviewer #5:**

In the manuscript, the authors combine the strengths of CNN and INN to propose an ensemble method, achieving state-of-the-art performance and reducing computational complexity. The core idea is to combine the global frequency patterns captured by CNN with the dynamic localized features extracted by INN. They also perform some real data to show the advantage of the proposed method. Overall, the manuscript is well-written, and the contributions are solid. Here are some suggestions.

* 1. Please give the pseudocode of the proposed method.

We added the pseudocode of the proposed method in Table3

* 1. In addition to the IRMAS dataset, please consider more datasets.

We specifically chose the IRMAS dataset because it is the only dataset that meets the unique requirements of our study. Unlike other datasets, IRMAS provides separate training and testing data, both of which are polyphonic. However, the training data consists of single-labeled audio samples with a fixed length of 3 seconds, while the testing data is multi-labeled with multiple predominant instruments and variable lengths ranging from 5 to 20 seconds. This setup presents a significant challenge, as the model must generalize from fixed-length, single predominant instrument training data to recognizing an arbitrary number of multiple predominant instruments in variable-length unseen test data. Moreover, all the existing works we cited and compared in our study also used the IRMAS dataset, further establishing it as the standard benchmark for this specific problem. Unlike prior methods that rely on sliding window analysis and aggregation strategies to address this challenge, our approach eliminates the need for such techniques, directly processing entire test files to improve efficiency and accuracy. Additionally, we validate our model using the designated training set and evaluate its performance on the unseen test set, ensuring a fair and rigorous assessment of our method.

* 1. In addition to CNN, INN and Han model, there are a large number of models to identify the predominant instrument in polyphonic music. Please compare the proposed method with some state-of-the-arts.

Thank you for your suggestion. In our initial submission, we had experimented with the Hann model and compared its performance with our proposed CNN, INN, and Ensemble methods. Specifically, we trained using fixed-length single predominant files and tested on variable-length single predominant files. .In the revised manuscript, we have further extended our analysis by including multiple predominant instrument recognition results and comparing our approach with additional state-of-the-art (SOTA) methods.

* 1. Please give the time cost of the proposed method and other comparison methods.

Thank you for your suggestion. In the revised manuscript, we have included a detailed analysis of the time cost for our proposed method as well as the Han model. Specifically, we report both training and inference times to provide a comprehensive evaluation of computational efficiency. These results help demonstrate the trade-off between accuracy and efficiency for each method. The updated manuscript now includes a table summarizing the time cost across all methods for better clarity. We appreciate your feedback in improving the completeness of our study.

* 1. Please give the robustness analysis of the proposed method.

Thank you for the valuable suggestion. We have added a comprehensive robustness analysis section in the revised manuscript, which includes:

1. Varying Input Lengths: We tested the model on different input lengths and compared the performance with SOTA methods. The results, presented in Table 6, confirm that our model maintains stable accuracy across varying audio durations.
2. Confusion Matrix Analysis: A confusion matrix has been provided in Figure highlighting instrument-wise misclassifications. We analyzed misclassification patterns and discussed possible reasons, such as spectral similarities between certain instrument pairs.
3. Computational Resource Evaluation: We measured the computational cost of our method compared to baseline models in terms of inference time and memory consumption, as shown in Table 5. Our method achieves competitive efficiency while maintaining high recognition accuracy.

**Reviewer #6:**

This study presents a novel approach to identifying the predominant instrument in polyphonic music. By combining the strengths of Convolutional Neural Networks (CNN) and Involutional Neural Networks (INN) through an ensemble method, the proposed approach achieves state-of-the-art performance while reducing computational complexity. Unlike traditional methods that rely on sliding window and aggregation strategies, the proposed approach directly learns to recognize individual instruments from variable-length polyphonic audio. The proposed ensemble model, using soft voting, effectively uses the global frequency patterns captured by CNN and the dynamic localized features extracted by INN. Evaluations of the IRMAS dataset demonstrate improved recognition accuracy and efficiency, making the proposed approach suitable for real-world music information retrieval applications. However, some problems should be solved.

1. There are some grammatical errors in this paper.

Comment : Thank you for your feedback. We have carefully reviewed the manuscript and corrected all grammatical errors. Additionally, we have improved the overall readability and clarity of the paper to ensure a more polished presentation.

1. Each variable in the formula should have an explanation.

Thank you for your valuable suggestion. We have carefully revised the manuscript to provide a clear explanation of each variable used in the formulas. The updated descriptions ensure that readers can fully understand the mathematical expressions and their relevance to our proposed approach.

1. Why use Involution Neural Networks?

Thank you for your insightful question. Involutional Neural Networks (INNs) were chosen for this study due to their ability to efficiently capture localized and dynamic feature variations in polyphonic audio signals. Unlike traditional CNNs, which use fixed spatial kernels, INNs leverage dynamic, data-dependent kernels, allowing for better adaptability in recognizing subtle instrument characteristics within overlapping audio sources.

By integrating INNs into our ensemble model, we enhance feature extraction capabilities, particularly in complex musical compositions where instruments may exhibit high variability in timbre and temporal structure. The combination of CNNs and INNs enables our approach to balance global frequency representation with localized feature refinement, ultimately improving recognition accuracy while maintaining computational efficiency.

1. Please give the runtime in the experiment.

Thank you for your suggestion. We have now included the runtime details for our experiments in the revised manuscript. Specifically, we report the training and inference times for our proposed ensemble model, along with a comparison to baseline methods. This addition provides a clearer understanding of the computational efficiency of our approach.

1. Please provide all the parameters of the proposed algorithm.

Thank you for your suggestion. We have now included a detailed description of all the parameters used in our proposed algorithm. This includes the hyperparameters for both CNN and INN models, such as learning rate, batch size, number of layers, kernel sizes, activation functions, and optimization settings.

1. More STOA method should be compared to show the advantages of the proposed method.

Thank you for your suggestion. In our initial submission, we had already experimented with the Han model and compared its performance with our proposed CNN, INN, and Ensemble methods. Specifically, we trained using fixed-length single predominant files and tested on variable-length single predominant files. .In the revised manuscript, we have further extended our analysis by including multiple predominant instrument recognition results and comparing our approach with additional state-of-the-art (SOTA) methods.

This expanded evaluation provides a more comprehensive comparison, demonstrating the advantages of our method in terms of recognition accuracy and generalization across different musical scenarios.