Leveraging Diffusion Models For Predominant Instrument Recognition

Charis Cochran
Drexel University
Philadelphia, PA
crc356@drexel.edu

Yeongheon Lee
University of Pennsylvania
Philadelphia, PA
aidenlee@sas.upenn.edu

Youngmoo Kim
Drexel University
Philadelphia, PA
ykim@drexel.edu

Abstract

Predominant Instrument Recognition (PIR) remains a challenge in MIR primarily due to data limitations. Recent work suggests that generative diffusion models learn rich timbre representations from these limited sets, yet their utility for recognition tasks has not been explored. We present the first study probing intermediate diffusion features for PIR. Starting from a pretrained diffusion model, we fine-tune variants on IRMAS (the premier PIR dataset) and OpenPIR, a new metadataset of multi-predominant annotations for OpenMic that we introduce. We sample activations across noise levels and layers and evaluate them with lightweight classifier heads. Results show that low-noise bottleneck features are the most informative, and even simple Multi-Layer Perceptron (MLP) probes achieve promising results. Incorporating OpenPIR improves performance across models, with diffusion features rivaling baselines for certain instruments. These findings provide early evidence that audio diffusion models encode discriminative features, pointing toward the need for further research into unified diffusion-recognition frameworks.

1 Introduction

In Music Information Retrieval (MIR), tasks such as source separation, transcription, and recommendation benefit instrument labels. While models achieve good results for isolated instruments, it remains challenging for polyphonic mixtures. Predominant Instrument Recognition (PIR) addresses identifying the most salient instruments in a recording. Modeling predominant instrumentation is difficult because timbre is highly contextual and overlaps in time and frequency. Prior systems achieve moderate success but suffer from large disparities in class performance [5] and fragile feature representations [2]. Augmenting with synthetic data, either isolated notes [16] or generated mixtures [12], yields some gains but does not resolve class imbalance or performance plateaus.

Meanwhile, generative models have shown strong ability to represent and synthesize timbre. Diffusion and autoencoder approaches can generate realistic instrument tones and transfer timbre between sources, and recent advances suggest they can also be guided to emphasize specific instruments using attention or disentanglement methods [9]. In vision, unifying recognition and generation has improved both domains [6], and diffusion features have been shown to transfer effectively into discriminative models [10].

In this paper, we probe diffusion models trained under constrained conditions and show that their intermediate features contain salient information for PIR. Even simple classifier probes on these features produce competitive results. To better match real-world evaluation, we also extend IRMAS with a small meta-dataset of multi-predominant annotations. Together, these experiments explore whether diffusion models learn features that are useful for discriminative tasks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

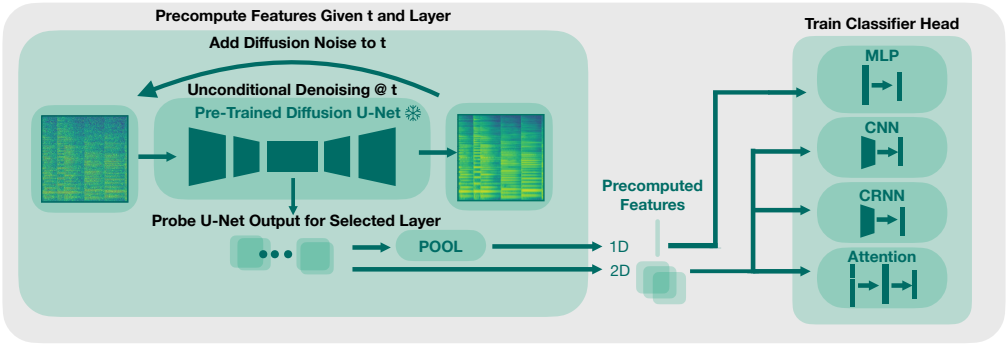

Figure 1: **Overview of proposed method**: Intermediate feature extraction from a diffusion model.

Model	Pre-Training	Epochs	Batch Size	Training Set	Input Length
Model 0	None	400	64	IRMAS & Solos [3]	1 sec
Model A	Model 0	100	64	IRMAS	2 sec
Model B	Model 0	100	64	IRMAS & OpenPIR	2 sec
Model C	Model A	100	64	IRMAS & OpenPIR	2 sec

Table 1: **Diffusion Model Training Parameters:** Outlined are the baseline model, Model 0, from [3] as well as three models trained in this study along with their specific training parameters. All models are trained on a single V100 GPU. The T5 text embedding model used as input to these models is frozen. Model 0 is only used in training diffusion models and not classifier heads. All models are 240M parameters.

2 Background and Related Work

Within the last decade, progress in Predominant Instrument Recognition (PIR) has moved from handcrafted features and SVMs [1] to deep learning methods such as CNNs and transfer/multi-task learning [5]. Despite improvements, state-of-the-art models still struggle from disparities in class performance. Later work explored synthetic pretraining data [16], feature learning [11], data augmentation [8], and ensemble or transformer-based systems [12]. These strategies yield modest gains but remain heavily reliant on synthetic data and still mirror class imbalance problems of previous models.

In parallel, diffusion models have demonstrated strong capabilities for modeling complex, high-dimensional data such as images, audio, and music [4]. They capture rich internal representations of timbre, generalize across playing styles, and enable controllable transformations. Recent work in computer vision has shown that diffusion models can support discriminative tasks either as frozen feature learners or through joint generative–discriminative training. RepFusion [14] dynamically selects layers and timesteps to distill features, while Mukhopadhyay et al. [10] pool and combine features across the denoising trajectory. These studies highlight the potential of intermediate feature extraction from diffusion models, but these applications remain underexplored in audio and MIR.

3 Dataset

We base our experiments on the IRMAS dataset [1], which contains audio clips labeled with one predominant instrument from 11 classes. Training samples are additionally annotated with one of five genres and about ten percent include drum presence/absence labels. IRMAS suffers from two major limitations: strong class imbalance and a mismatch between the training and test sets. While the training data contains only single-label annotations, more than three-quarters of the test samples include multiple predominant instruments. To mitigate this gap, we introduce OpenPIR, a complementary dataset of 1,234 multi-predominant annotations derived from OpenMIC [7], aligned with the IRMAS instrument taxonomy and containing predominant instrument, genre, and drum labels.[1]

[1]Dataset, code, and sound examples avaibale at https://github.com/charisrenee/DiffPIR/

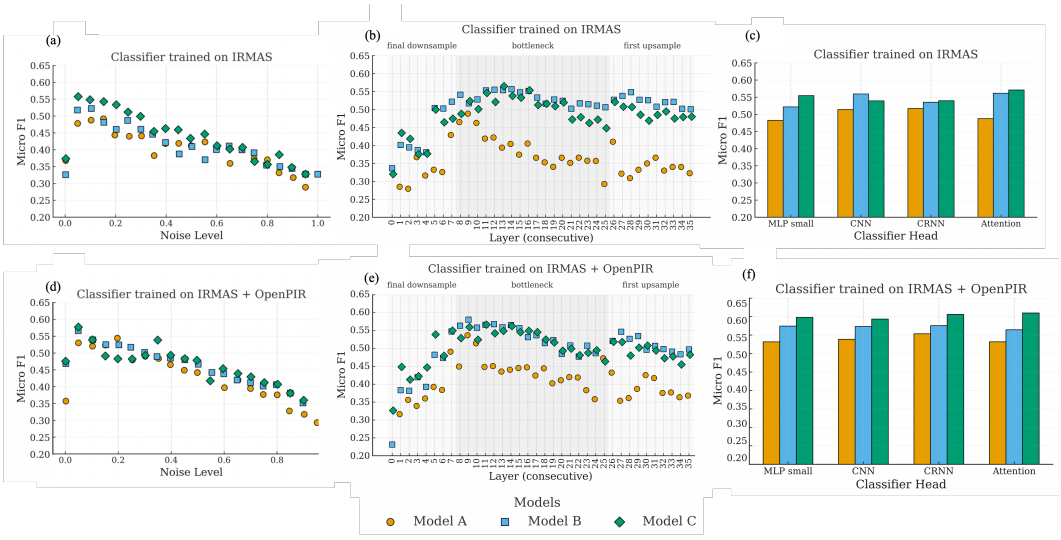

Figure 2: **Overview of Feature Selection Experiments:** Classifiers trained on IRMAS (top) and Classifiers trained on IRMAS & OpenPIR (bottom).

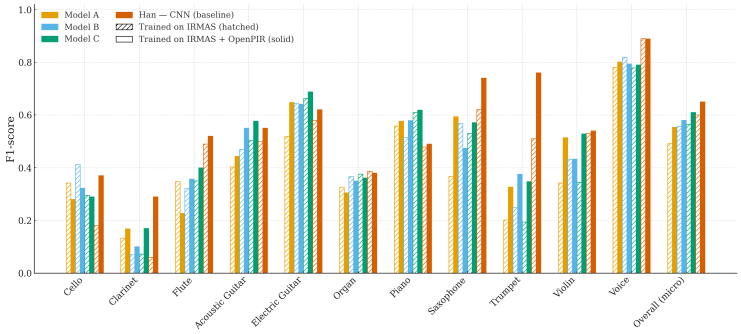

Figure 3: **Class-wise and overall Micro F1:** Results for the optimal Noise, Layer and Classifier parameters for each training data and diffusion model combination as well as the Han et.al. baseline.

4 Diffusion Model Architecture and Training

Our diffusion backbone adapts components from a-unet and audio-diffusion-pytorch [13]. The DDIM network contains three up/downsampling stages and a bottleneck with [128, 256, 512, 1024] filters. Instrument conditioning is injected via pre-trained T5 embeddings. Models operate on 16 kHz Mel spectrograms (window size 640, hop 320, 128 bins) and are decoded into audio using a frozen SoundStream vocoder [15]. We initialize from a pretrained baseline (Model 0) trained on IRMAS combined with solo instrument data, as presented by Cochran et al. [3]. From this baseline, we train three variants: Model A, fine-tuned on IRMAS only; Model B, fine-tuned on IRMAS plus OpenPIR; and Model C, trained sequentially by continuing Model A training with IRMAS and OpenPIR together. Each model is trained for 100 epochs. Model details can be seen in Table 1.

5 Experiments

We aim to answer whether learned diffusion representations are informative for PIR, by investigating whether intermediate features extracted from a pretrained diffusion model are suitable representations for this downstream task as it has been in vision tasks [10]. This technique exploits the class information preserved in denoising trajectories, and utilizes this as the input to small classifier head. The overview of this setup can be seen in Figure 1. To extract features, we first corrupt an input with noise at a chosen timestep t using the diffusion scheduler. A single denoising step is then performed

without text conditioning, after which intermediate activations are collected from the target layer L. These activations are either globally pooled or passed directly into classifier heads.

We evaluate four lightweight classifiers. The multilayer perceptron (MLP) has a single hidden layer of 128 units with approximately 130K parameters. The CNN contains a single 2D convolutional layer with 128 filters and about 160K parameters. The CRNN uses a GRU with hidden size 128 for a total of 200K parameters. The attention-based classifier employs four attention heads with hidden size 128, totaling 330K parameters. The MLP is training on the Global Average Pool of the feature outputs (1024 maps for the bottleneck and 512 for the up/downsample), and all other models are trained on the unpooled 2D feature maps.

We run a systematic parameter search to investigate how noise level, layer choice, and classifier architecture affect PIR performance. First, we vary the noise level t using the MLP head and a bottleneck ResNet block. Once the best noise level is identified, we probe features across the final downsampling block, the bottleneck, and the first upsampling block. With noise and layer fixed, we then evaluate each classifier head. This procedure is repeated for models Model A, Model B, and Model C, trained on both IRMAS alone and IRMAS with OpenPIR. Results are compared against a re-implementation of the Han et al. CNN baseline [5] which has approximately 1.2M parameters.

6 Results

The overall results of the feature selection experiments are shown in Figure 2, while Figure 3 presents class-wise performance for the best parameters of each model and training configuration and comparison to the baseline. The trends align with findings in computer vision: low-noise conditions and bottleneck layers yield the most informative features. This follows the intuition that excessive noise erases class information, while insufficient noise leads the model to focus primarily on denoising rather than reconstructing timbre. Notably, the second and third bottleneck layers consistently provide the strongest features across models. Classifier comparisons reveal that more sophisticated architectures achieve higher accuracy, yet it is striking that even a simple global average pooling followed by a small MLP can produce a capable PIR model. The benefits of incorporating OpenPIR are also clear. Diffusion model features trained on OpenPIR (Model B, Model C) outperform the IRMAS-only baseline (Model A), and adding OpenPIR at the classifier training stage further boosts performance across all three models.

The class-wise breakdown highlights remaining challenges and promise: these models remain below both the Han CNN benchmark and the state-of-the-art ($F_1 = 0.65$ [16]), though in certain cases, such as electric guitar, the diffusion-derived features surpass the Han baseline. Overall, the results demonstrate that even constrained generative diffusion models learn features with meaningful discriminative power, though they have yet to close the gap with the best specialized recognition models.

7 Conclusion & Future Work

This work provides initial evidence that diffusion models, even under constrained conditions, learn features relevant for PIR. With the introduction of OpenPIR we have shown that this additional multi-predominant instrument training data helps in all classifier head cases, but interestingly the largest improvements in the parameter search are seen when adding this data to the diffusion models and improving the learned features. This additionally points to the possibility of leveraging foundational models with more advances representation spaces as a starting point for this sort of feature extraction and downstream classification task. Overall, these preliminary results point to the utility of looking into ways to leverage the gains in music generation in recent years to inform discriminative systems.

References

[1] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 552–564, Porto, Portugal, 2012. ISMIR. doi: 10.5281/zenodo.1416076.

[2] Charis Cochran and Youngmoo Kim. Deepdream applied to an instrument recognition cnn. In *22nd Int. Society for Music Information Retrieval Conf*, 2021.

[3] Charis Cochran and Youngmoo Kim. Modeling predominant insrtumentation with diffusion. In *International Society of Music Information Retrieval (ISMIR), Late-Breaking/Demo Session, In-Person*. International Society of Music Information Retrieval, 2024.

[4] Zach Evans, C. J. Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast Timing-Conditioned Latent Audio Diffusion, February 2024. URL http://arxiv.org/abs/2402.04825. arXiv:2402.04825 [cs, eess].

[5] Yoonchang Han, Jaehun Kim, and Kyogu Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):208–221, 2017. doi: 10.1109/TASLP.2016.2632307.

[6] Runhui Huang, Jianhua Han, Guansong Lu, Xiaodan Liang, Yihan Zeng, Wei Zhang, and Hang Xu. Diffdis: Empowering generative diffusion model with cross-modal discrimination capability. (arXiv:2308.09306), August 2023. doi: 10.48550/arXiv.2308.09306. URL http://arxiv.org/abs/2308.09306.

[7] Eric J. Humphrey, Simon Durand, and Brian McFee. Openmic-2018, September 2018. URL https://doi.org/10.5281/zenodo.1432913.

[8] Agelos Kratimenos, Kleanthis Avramidis, Christos Garoufis, Athanasia Zlatintsi, and Petros Maragos. Augmentation methods on monophonic audio for instrument classification in polyphonic music. *CoRR*, abs/1911.12505, 2019. URL http://arxiv.org/abs/1911.12505.

[9] Yin-Jyun Luo, Kin Wai Cheuk, Tomoyasu Nakano, Masataka Goto, and Dorien Herremans. Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds. In *ISMIR*, pages 700–707, 2020.

[10] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Tianyi Zhou, Jun Ohya, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? *ECCV 2024*, (arXiv:2311.17921), September 2024. doi: 10.48550/arXiv.2311.17921. URL http://arxiv.org/abs/2311.17921. arXiv:2311.17921 [cs].

[11] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet, 2019. URL https://arxiv.org/abs/1808.00158.

[12] Lekshmi Chandrika Reghunath and Rajeev Rajan. Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music. *EURASIP J. Audio Speech Music Process.*, 2022(1), May 2022. ISSN 1687-4714. doi: 10.1186/s13636-022-00245-8. URL https://doi.org/10.1186/s13636-022-00245-8.

[13] Flavio Schneider. Archisound: Audio generation with diffusion. *arXiv preprint arXiv:2301.13267*, 2023.

[14] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 18892–18903, October 2023. doi: 10.1109/ICCV51070.2023.01736. URL https://ieeexplore.ieee.org/document/10377906/.

[15] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec, 2021.

[16] Lifan Zhong, Erica Cooper, Junichi Yamagishi, and Nobuaki Minematsu. Exploring isolated musical notes as pre-training data for predominant instrument recognition in polyphonic music. In *APSIPA ASC*, 2023. URL https://arxiv.org/abs/2306.08850.