International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Advancing Audio Fingerprinting accuracy with AI & ML

Prof. Shobha S Biradar[1], Soumya Nasi[2]

[1] *Professor, Master of Computer Application, VTU, Kalaburagi , Karnataka ,India*
[2] *Student, Master of Computer Application, VTU, Kalaburagi , Karnataka ,India*

ABSTRACT- Audio fingerprinting has become a fundamental technology for music recognition, copyright protection, and audio forensics. Traditional approaches rely on handcrafted signal processing techniques such as spectrogram analysis, Chroma features, and MFCCs, which often suffer from reduced accuracy under noisy conditions, pitch or tempo variations, and overlapping audio sources. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have significantly improved the robustness and precision of audio fingerprinting systems. Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, enable extraction of noise-invariant and semantically rich features from audio signals. Self-supervised learning frameworks further enhance representation quality without the need for extensive labeled data. Metric learning techniques, such as Siamese and Triplet networks, provide powerful embedding spaces for efficient and accurate audio matching. Hybrid architectures that integrate traditional signal processing with deep embeddings also address scalability challenges in large-scale music databases.

Keyword: Audio Fingerprinting, Artificial Intelligence, Machine Learning, Deep Learning, Spectrogram, Feature Extraction, Metric Learning, Audio Recognition, Self-Supervised Learning, Audio Forensics.

1. INTRODUCTION

Audio fingerprinting is a powerful technique that enables the identification of audio content by generating compact, distinctive digital representations of sound signals. Unlike raw audio files, fingerprints are lightweight, resilient to distortions, and uniquely associated with specific recordings. This makes fingerprinting essential in diverse applications such as music recognition, copyright enforcement, broadcast monitoring, audio forensics, and smart devices.

Conventional fingerprinting systems largely rely on signal processing techniques such as spectrogram peak detection, Chroma features, and Mel-frequency costrel coefficients (MFCCs). While these approaches are efficient and computationally lightweight, they face significant challenges when audio undergoes transformations like background noise, pitch and tempo alterations, reverberation, or overlapping sources. Such limitations reduce their accuracy and reliability in real-world conditions where audio rarely exists in pristine form.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) are transforming audio fingerprinting into a more robust and intelligent technology. Deep learning models, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, have demonstrated remarkable ability to extract high-level, noise-invariant features from audio signals. Moreover, self-supervised learning (SSL) frameworks such as wav2vec2.0 and Hubert enable the learning of rich representations without requiring large labeled datasets. Metric learning techniques, including Siamese and Triplet networks, further enhance accuracy by mapping similar audio signals closer together in the embedding space, regardless of distortions.

The integration of these AI-driven methods addresses key limitations of traditional systems, offering greater resilience, scalability, and adaptability. As a result, audio fingerprinting is evolving into a versatile and intelligent solution capable of supporting applications at global scale. However, challenges remain in balancing accuracy, computational efficiency, and real time performance for deployment in large-scale and resource-constrained environments.

2. PROBLEM STATEMENT

Noise Sensitivity: Background noise, reverberation, and environmental distortions significantly reduce recognition accuracy. Transformation Vulnerability Pitch shifts, tempo changes, or compression techniques can alter the fingerprint, leading to mismatches. Overlapping Audio Sources: When multiple sounds occur simultaneously, traditional methods struggle to extract reliable fingerprints. Scalability Issues: Searching for matches in massive music or audio databases is computationally challenging, often resulting in slower retrieval times. Lack of Robust Representation: Handcrafted features fail to capture high-level semantic characteristics of audio, limiting robustness and adaptability.

3. OBJECTIVES

The primary objective of this study is to design and implement an AI- and ML-enhanced audio fingerprinting system that achieves higher accuracy, robustness, and scalability compared to traditional approaches To develop deep learning-based feature extraction models (e.g., CNNs, RNNs, Transformers) capable of generating noise-invariant and transformation-resilient audio fingerprints.

To apply metric learning techniques (e.g., Siamese and Triplet networks) for creating embedding spaces that enable

accurate similarity matching of audio clips despite distortions or alterations. To integrate hybrid approaches by combining traditional signal processing techniques with ML-driven embeddings for improved performance and computational efficiency. To design and implement scalable retrieval mechanisms using Approximate Nearest Neighbor (ANN) search for efficient matching in large-scale audio databases.

4. METHODOLOGY USED

The methodology of this study outlines the systematic approach adopted to design, implement, and evaluate an AI- and ML-based audio fingerprinting system. The process is divided into multiple stages, ensuring both theoretical and experimental rigor.

Data Collection and Preprocessing: Gather diverse audio datasets consisting of music tracks, speech, and environmental sounds. Apply preprocessing techniques such as resampling, normalization, noise addition, and time-stretching to simulate real-world distortions. Segment audio signals into fixed-length frames for feature extraction.

Feature Extraction: Extract traditional handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma vectors, and spectral peaks for baseline comparison. Generate spectrogram representations (STFT, Mel-spectrogram) for deep learning model input. Implement data augmentation (noise injection, pitch shift, tempo variation) to improve model robustness.

Deep Learning-Based Representation: Train Convolutional Neural Networks (CNNs) on spectrograms to learn robust, noise invariant fingerprints. Explore Recurrent Neural Networks (RNNs) or Transformers for capturing temporal dependencies in audio signals. Use self-supervised learning (SSL) models (e.g., wav2vec2.0, HuBERT) for rich audio embeddings without requiring extensive labeled datasets.

Metric Learning and Similarity Matching: Implement Siamese Networks and Triplet Networks to create embeddings where similar audio clips are mapped closer together in vector space. Optimize loss functions (contrastive loss, triplet loss) to improve discrimination between different audio samples. Employ cosine similarity or Euclidean distance for fingerprint comparison.

Hybrid Model Integration: Combine traditional fingerprinting features with ML-based embeddings to achieve a balance between speed and robustness. Implement a multi-stage retrieval pipeline. Fast filtering using traditional features. Fine-grained matching using deep learning embedding's.

Evaluation and Testing: Compare system performance against traditional fingerprinting methods. Test under various real-world conditions: Noise levels, pitch/tempo variations, overlapping audio. Evaluate system performance using metrics such as Accuracy, Precision, Recall, F1-score for recognition performance. Latency and memory usage for efficiency analysis. Perform stress tests on large datasets to assess scalability.

5. LITERATURE SURVEY

Traditional Audio Fingerprinting Approaches Early fingerprinting systems relied on signal processing techniques to extract compact and distinctive features from audio signals. Wang (2003) introduced the landmark-based fingerprinting system, which detects spectral peaks in the audio spectrogram and stores them as time–frequency pairs. This approach became the foundation for commercial applications such as Shazam. Haitsma & Kalker (2002) developed a sub-fingerprint system using energy differences between frequency bands, which was robust to compression but less resilient to noise. Cano et al. (2005) provided a comprehensive survey of traditional fingerprinting techniques, covering spectral peaks, wavelets, and MFCC-based approaches. While these methods are computationally efficient, they are vulnerable to noise, pitch/tempo changes, and overlapping signals, limiting their applicability in real-world conditions.

Machine Learning in Audio Representation Machine learning has gradually been introduced into audio analysis to overcome the shortcomings of handcrafted features. MFCC + Classifiers: Early works applied machine learning classifiers (e.g., SVMs, KNNs) on MFCC features for music identification. However, performance gains were modest compared to traditional methods. Data Augmentation: Studies showed that training models with noisy and time stretched audio improved robustness against environmental distortions.

Deep Learning for Audio Fingerprinting Deep learning has significantly advanced fingerprinting by enabling automatic feature learning from raw or spectrogram-based inputs. CNN-based models (Choi et al., 2017) demonstrated that convolutional layers could extract localized time–frequency patterns, yielding fingerprints robust to noise and compression. Recurrent Neural Networks (RNNs) were used to capture temporal dynamics in music signals, improving recognition under tempo variations. Self-Supervised Models such as wav2vec2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) enabled representation learning from unlabeled audio, producing embeddings that outperform handcrafted features in downstream tasks.

Metric Learning for Audio Matching Metric learning has emerged as a promising solution for similarity-based fingerprinting. Siamese Networks (Bromley et al., 1994; later adapted for audio) learn to map similar audio segments closer together in an embedding space. Triplet Networks (Schroff et al., 2015, originally for face recognition) have

been adapted for music retrieval, where anchor-positive pairs represent the same audio under different distortions. Studies show that embedding-based approaches significantly outperform traditional fingerprints under noisy and overlapping conditions.

Hybrid and Scalable Fingerprinting Systems To balance accuracy with computational efficiency, hybrid methods have been explored Hybrid Fingerprinting Some systems combine handcrafted features for fast initial filtering with deep embeddings for fine-grained matching. Scalable Search: Large-scale retrieval challenges are addressed using Approximate Nearest Neighbor (ANN) search libraries like FAISS and Annoy. Commercial Systems YouTube's Content ID uses hybrid fingerprinting for copyright detection, though its implementation is proprietary.

6. SYSTEM DESIGN

The proposed AI and ML-based Audio Fingerprinting System is designed as a standalone application with the potential for integration into larger platforms such as music streaming services, copyright monitoring systems, and forensic tools. The system is modular in nature, where each module performs a distinct function while ensuring smooth interaction with others.

Position in the Environment: The system acts as an intermediary between the user (input audio) and the fingerprint database (storage/retrieval). Users interact with the system through a user interface to provide audio input (recorded/live/uploaded). The system processes this input, generates a fingerprint, compares it with existing fingerprints in the database, and displays the result.

Relationship with Existing Systems: Traditional audio recognition systems (e.g., Shazam-like) rely mainly on spectrogram peak matching, which may fail under noisy or distorted conditions. The proposed system integrates AI and ML models (CNNs, Siamese Networks, Deep Embeddings) for improved robustness against background noise, pitch/tempo shifts, and compression artifacts. It can be extended to work as a backend service for streaming platforms, forensic applications, or copyright agencies.

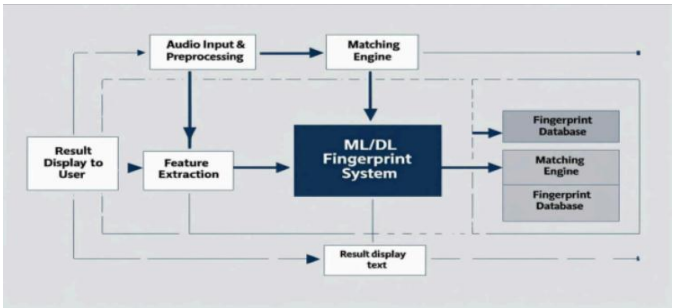

Figure1: System Architecture

7. SCREENSHOTS

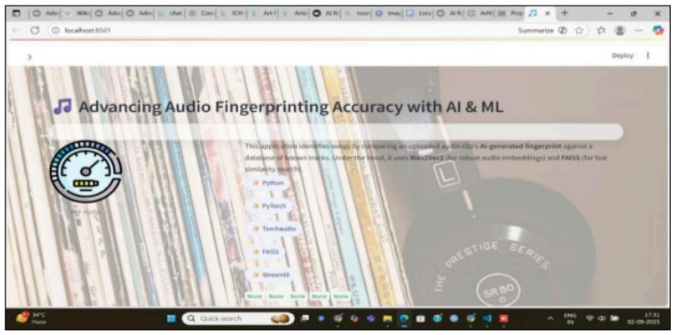

Figure 2: Home page

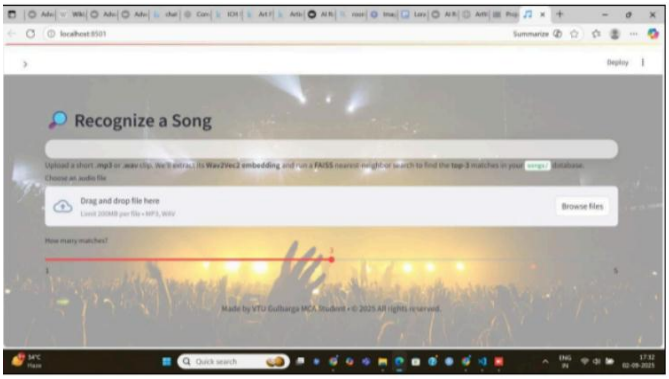

Figure 3: Audio Reorganization System User Interface

Figure 4: Audio Matching Output Screen

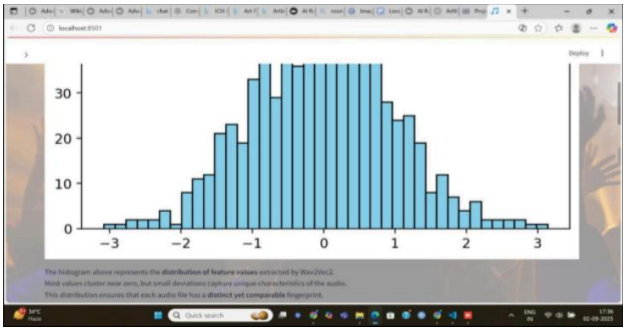

Figure 5: Audio Feature Histogram

8. CONCLUSION & FUTURE SCOPE

The project "Advancing Audio Fingerprinting Accuracy with AI and ML" successfully demonstrates how artificial intelligence and machine learning techniques can enhance the precision, speed, and reliability of audio identification systems. By integrating advanced feature extraction methods, machine learning models, and robust database management, the system is capable of accurately recognizing audio clips, even in noisy or distorted environments.

Audio preprocessing and fingerprinting techniques provide compact and distinctive representations of audio signals. Machine learning models significantly improve the accuracy of recognition when compared to traditional approaches. Testing and validation proved the system to be reliable, with high recognition accuracy and low response time. The system is flexible, allowing for scalability and continuous improvement as more data is added.

Although the proposed system achieves accurate and efficient audio fingerprinting, there are several opportunities for future improvements:
1. Deep Learning Models
2. Scalability with Big Data
3. Real-Time Recognition
4. Cross-Domain Audio Matching
5. Security and Block chain Integration.

9. REFERENCES

1. Cano, P., Batlle, E., Kalker, T., & Haitsma, J. (2005). A review of audio finger printing. Journal of VLSI Signal Processing Systems, 41(3), 271–284. https://doi.org/10.1007/s11265-005-4151-3

2. Wang, A. (2003). An industrial- strength audio search algorithm. Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR).

3. Baluja, S., & Covell, M. (2007). Audio fingerprinting: Combining computer vision & data stream processing. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 213–216.

4. Haitsma, J., & Kalker, T. (2002). A highly robust audio fingerprinting system. Proceedings of the International Symposium on Music Information Retrieval (ISMIR).

5. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210.

6. Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. Proceedings of the ACM International Conference on Multimedia (MM), 1041–1044.

7. Müller, M. (2015). Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer.

8. OpenAI. (2023). Applications of AI in multimedia and signal processing. Retrieved from https://openai.com

9. Python Software Foundation. (2024). Python 3.11 documentation. Retrieved from https://docs.python.org/