The Bombay Salesian Society's

DON BOSCO INSTITUTE OF TECHNOLOGY

Premier Automobiles Road, Kurla (W), Mumbai-70

Approved by AICTE, Govt. of Maharashtra

&

Affiliated to the University of Mumbai

T.E. MINI PROJECT REPORT
CSM501 - Mini Project: 2 A

On

"Sound Gen AI"

Department of Computer Engineering

University of Mumbai

October 2024

The Bombay Salesian Society's

DON BOSCO INSTITUTE OF TECHNOLOGY

Premier Automobiles Road, Kurla (W), Mumbai-70

Department of Computer Engineering

(Session 2024-2025 ODD Semester)

CERTIFICATE

Project Title : **Sound Gen AI**

Project Team Members :

1. Andre Marshal Fernandes
2. Jowin Paulose
3. Suyash Sanjay Kadam

Name of Internal Guide : Prof. Imran Ali Mirza

INTERNAL EXAMINER (s) **EXTERNAL EXAMINER (s)**

HEAD, COMPUTER ENGINEERING

Date:

ABSTRACT

The growing demand for high-quality, customizable audio in digital projects has made traditional sound sourcing expensive and time-consuming. Our generative AI sound production system addresses this challenge by revolutionizing sound generation with efficient, accessible, and affordable technology. The primary aim of this research is to streamline sound production for creating diverse environmental sounds and mood-based music, empowering creators to enhance immersive experiences in various applications.

The system leverages a GAN-based model trained on spectrograms, using a dataset of labeled songs categorized by mood. Key hyperparameters such as latent dimensions, batch size, and sequence length were optimized to enhance the model's performance. The trained model generates soundscapes based on random noise and mood inputs, which are then converted into an audio format.

Initial experiments showed promising results, though improvements were needed. By expanding the dataset from 20 to 35 songs and increasing the training epochs, the generated sound quality improved significantly. These findings demonstrate the potential of our system to evolve into a scalable solution for producing high-quality, customizable soundscapes and music. The system offers endless possibilities for creative sound design, making sound production more immersive, efficient, and cost-effective.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ABBREVIATIONS

GAN	Generative Adversarial Network
VAE	Variational Autoencoders
RNN	Recurrent Neural Network
CSV	Comma Separated Values

CHAPTER 1: INTRODUCTION

In recent years, the demand for diverse and customizable audio in digital projects has grown exponentially. From game development to multimedia production, creators seek immersive soundscapes that align with specific emotional tones or environmental settings. Traditionally, sourcing such sounds requires extensive libraries or professional sound engineers, which can be both costly and time-consuming. As technological advancements in artificial intelligence and machine learning continue to emerge, these fields have begun to offer innovative solutions to simplify sound generation and customization.

Generative Adversarial Networks (GANs), originally developed for image generation, have shown tremendous potential in a variety of fields, including audio synthesis. GAN-based models have been successfully applied to generate high-quality music, speech, and environmental sounds, offering a promising approach for audio generation. Despite these advancements, challenges remain in terms of accessibility, scalability, and achieving precise customization based on specific moods or themes. The application of AI for sound generation is still an emerging field, with room for improvement in the quality, diversity, and ease of use of generated audio content.

One particular aspect of sound generation that could be enhanced is the ability to produce mood-based symphonies and environmental sounds that are tailored to specific needs in real-time. Current approaches often rely on large pre-existing datasets or manual sound engineering, which limits scalability and flexibility. By leveraging a mood-labeled dataset and training a GAN spectrogram model, we hypothesize that it is possible to generate diverse, high-quality soundscapes and music that align with specific emotional cues. This approach could significantly reduce the time and resources needed to create tailored audio content.

To test this hypothesis, our research involves training a GAN-based sound generation model using a dataset of mood-labeled songs. The model generates spectrograms from random noise and mood inputs, which are then converted into audio formats. Key hyperparameters were fine-tuned to enhance the model’s output quality. We conducted experiments with varying datasets and training parameters to observe the improvements in sound quality and customization.

In conclusion, this research aims to advance the field of generative AI for sound production by offering a more efficient and scalable solution for creating customizable audio. The findings from this study will contribute to the development of AI-driven tools that empower creators to produce immersive, mood-driven soundscapes with ease, transforming the way sound is integrated into digital experiences.

CHAPTER 2: LITERATURE SURVEY

The growing field of generative AI for music and sound synthesis has seen rapid advancements, driven by techniques like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). In this chapter, a critical appraisal of key research papers is presented to review existing methodologies, highlight gaps in research, and show how these studies relate to the current investigation on mood-based sound generation using AI.

Huang and Huang (2020) introduced an emotion-based AI music generation system using CVAE-GAN. This system integrates Conditional Variational Autoencoders (CVAE) and GANs to generate music based on specific emotional labels. The authors proposed that emotions such as happiness or sadness can be used as input parameters to guide the music generation process, improving the expressiveness and relevance of AI-generated music. The authors highlighted that while the system is effective in generating music aligned with mood inputs, there is still a challenge in achieving high-quality audio fidelity. Moreover, the need for large, diverse datasets remains a significant limitation. This study contributes to the broader discourse on how AI can enhance personalized audio experiences, yet further work is required to improve the scalability and diversity of the generated output.

Similarly, Kobayashi et al. (2024) explored novice-centered application design for music creation, focusing on accessibility for non-experts. Their research emphasized simplicity and ease of use in AI-assisted music creation tools, targeting novice users without extensive musical training. The study demonstrated the potential of AI in democratizing music production, but also revealed a limitation in terms of the range of control and customization available to more advanced users. This aligns with our research aim of offering customizable soundscapes but shows a gap in systems that balance ease of use with deep customization for professional applications.

Dong (2024) in his dissertation, Generative AI for Music and Audio, conducted a comprehensive review of recent advancements in AI-generated music. Dong explored a wide array of models including GANs, Recurrent Neural Networks (RNNs), and Transformer-based architectures for generating music and environmental sounds. One key insight from his research is the need for integrating temporal dynamics more effectively to create more coherent and contextually appropriate sound sequences. His research demonstrated that while existing models excel in short-duration audio synthesis, they often struggle with generating long-term coherent compositions. Dong's work highlights the potential for improvement in extending the duration and maintaining quality over time in generative models, which is a key concern for our study on environmental soundscapes.

Summary of Literature

Authors(s)	Year	Methodology	Findings	Identified Gaps/ Future Scope
Chih-Fang Huang & Cheng-Yuan Huang	2020	CVAE-GAN for emotion-based music generation	Successfully generated music labels based on emotional labels	High-quality fidelity needs improvement; large dataset requirement
Kobayashi, Atsuya, Tetsuro Sato, Kei Tateno	2024	Novice-centered AI music creation tool	Simplified AI tools for novice users; increased accessibility	Lack of customization for advanced users
Hao-Wen Dong	2024	Comprehensive review of AI music generation	Reviewed GANs, RNNs, Transformers; highlighted temporal dynamic issues	Struggles with long term audio coherence and high quality soundscapes

Table 1

From these studies, it is evident that the integration of GANs for emotion-based audio generation is a promising field, but there are notable limitations in scalability, customization, and sound fidelity. Furthermore, while novice-centered tools make sound production more accessible, they lack advanced features that professionals may require. The gap identified in these studies forms the foundation for our investigation, which seeks to develop a scalable, high-quality, and customizable sound generation system using AI.

By focusing on enhancing both the quality of the generated sound and the flexibility for customization, our study will contribute to addressing the limitations found in previous work, providing a more robust solution for generating mood-based symphonies and environmental sounds. This research aims to fill the gaps identified in the literature, particularly in improving sound fidelity and scalability, and offering a tool that balances both novice and expert needs.

CHAPTER 3: PROPOSED SYSTEM

3.1 Analysis/Framework/Algorithm

The primary framework employed in this system is a Generative Adversarial Network (GAN), which is widely used for generating data similar to a given dataset. In this case, the GAN is trained to generate spectrograms of sounds. The two key components of GANs are:

Generator: Produces new data (in this case, spectrograms) from random noise and a specified mood label.

Discriminator: Distinguishes between real and generated data, ensuring the generator improves over time.

Algorithm Workflow:

Step 1: Data Input: The system starts with a dataset of songs labeled by mood. A CSV file contains metadata about the mood for each song. This dataset is converted into spectrograms (visual representations of sound) for training.

Step 2: GAN Training: The GAN model is trained using labeled spectrograms. The model's hyperparameters (such as latent dimensions, batch size, sequence length, and number of mel-frequency bins) are optimized to improve the quality of the generated audio.

Step 3: Sound Generation: Once trained, the system generates new spectrograms based on input noise and mood labels, which are later converted into audio files (e.g., MP3).

Step 4: Output: The generated spectrogram is transformed into an audio file, which reflects the mood and characteristics derived from the training data.

This analysis framework ensures the system produces accurate and mood-based soundscapes, which are useful in various applications.

3.2 Details of Hardware and Software

Hardware Requirements:

Processor: A multi-core processor such as Intel i7 or AMD Ryzen 7 for fast processing during model training.

RAM: 16 GB or more is recommended to handle large datasets and model training processes.

GPU: A high-performance GPU like NVIDIA RTX 4070 or higher is essential for accelerating deep learning tasks, particularly GAN training.

Storage: At least 500 GB SSD for fast access to datasets, with additional storage for audio files and model checkpoints.

Software Requirements:

Operating System: Ubuntu 20.04 LTS or Windows 10/11.

Programming Language: Python 3.x for developing the core algorithms and processing data.

Libraries:

TensorFlow or PyTorch: These are the most widely used deep learning libraries for building GAN models.

Librosa: A Python package for analyzing and processing audio files.

Matplotlib: For visualizing spectrograms and model performance.

Pandas and NumPy: For handling and processing the dataset (including the CSV files containing song labels).

Audio Processing Software: FFmpeg or any similar tool to convert generated spectrograms into audio files like MP3.

3.3 Design Details

The system is designed in modular form, where different components handle data preprocessing, model training, and audio generation:

Data Preprocessing: This module handles reading audio files and their mood labels, converting them into spectrograms using tools like Librosa.

Model Architecture: The core GAN model includes a generator and a discriminator.

The generator takes random noise and mood labels as input and creates spectrograms.

The discriminator compares real spectrograms (from the dataset) with the generated ones to improve the model's output.

Training Loop: This module iterates over the dataset, allowing the generator and discriminator to learn and improve. Key hyperparameters like batch size, epochs, and sample interval are tuned here to optimize performance.

Sound Conversion: Once the GAN generates the spectrogram, the final design step involves converting it back into audio format (MP3).

3.4 Methodology

The approach to solving the problem of customizable sound generation using AI revolves around the following steps:

Data Collection and Preprocessing: We began by collecting a dataset of songs categorized by mood. This was crucial as it enabled us to create mood-based soundscapes. The songs were converted into spectrograms for GAN model training, with a CSV file containing labels to guide the model.

Model Development: A GAN was chosen due to its efficiency in generating high-quality, complex data. By training the model on spectrograms instead of raw audio, we were able to use image-processing techniques, making the model more efficient and better at handling mood-based audio variations.

Hyperparameter Tuning: The performance of the GAN was enhanced by carefully selecting hyperparameters. This included:

Latent Dimensions: To control the variability in the generated sound.

Batch Size and Sequence Length: To define how much data is processed at a time and the length of generated sound sequences.

Epochs: Increasing the number of training epochs allowed the generator to create more realistic audio over time.

Sound Generation and Conversion: After training, the system accepts mood labels and random noise as input to generate new soundscapes. These spectrograms are then converted into audio format, providing creators with new, customizable soundtracks.

This methodology ensures efficient and scalable sound production using AI, addressing both cost and time constraints in traditional sound sourcing. The approach focuses on utilizing deep learning techniques to create mood-based, high-quality audio files for various applications like gaming, multimedia, and virtual environments.

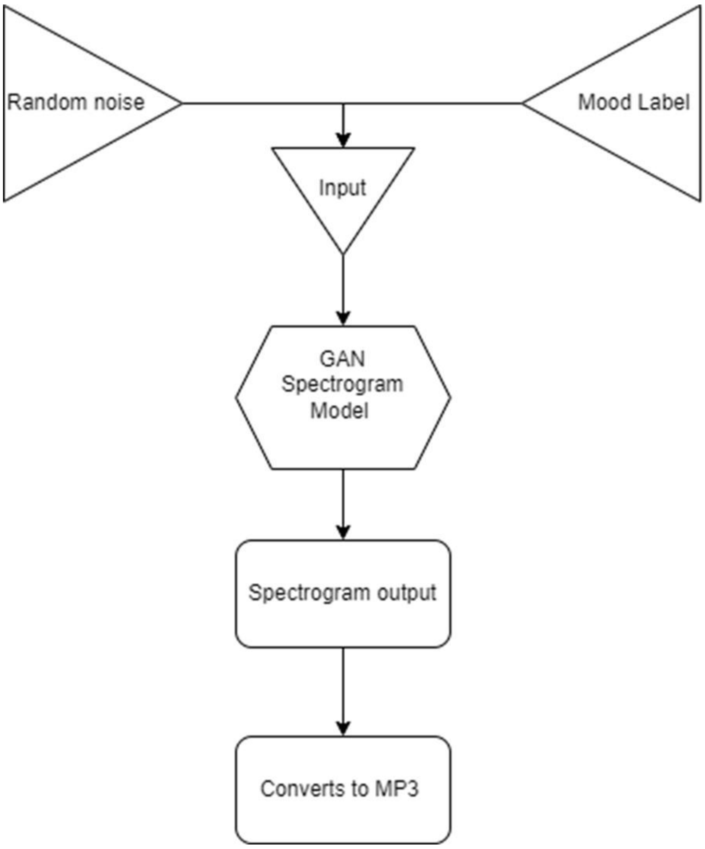

Fig 1 Flow Diagram A

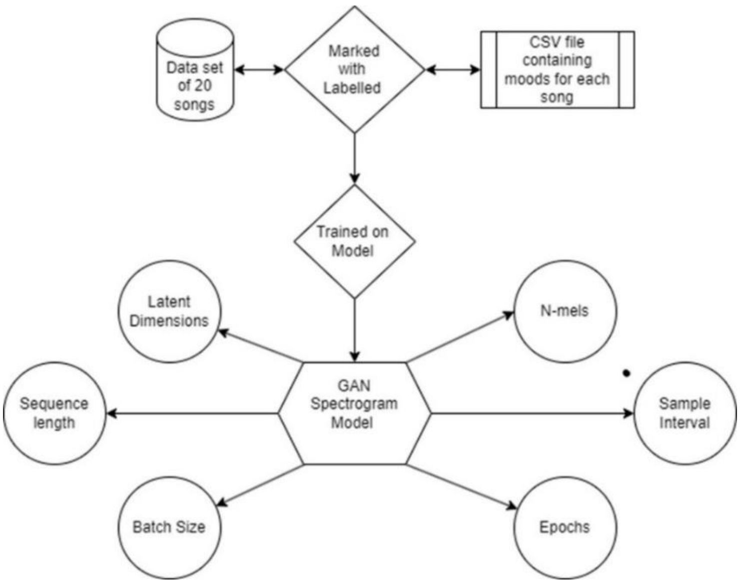

Fig 2 Flow Diagram B

CHAPTER 4: IMPLEMENTATION DETAILS

This chapter provides an in-depth explanation of the implementation process for the generative AI sound production system, detailing the timeline, tools used, and the steps taken to complete the project. The system was developed in a phased approach, with each module of the project planned and implemented according to the academic schedule. This chapter includes a detailed discussion of the tasks, a Gantt chart, and time allocation for each activity, ensuring the completion of the project within the provided deadlines.

4.1 Implementation Plan

The implementation of the generative AI system was divided into several key phases, each focusing on different aspects of the project such as literature review, dataset collection, model development, and performance improvement. Each phase required a specific set of tools, resources, and programming environments, and was executed within a predefined timeline. The following are the major activities and their time allocations:

4.2 Gantt Chart

The Gantt chart below represents the project timeline, showing the start and end dates of each task. This visual representation ensures clarity on deadlines and highlights the dependencies between different phases of the project.

Week No.	Date	Time of Meeting	Task Assigned	Deadline
1	1/8/2024	3:00-5:00 pm	Ideation	8/8/2024
2	8/8/2024	3:00-5:00 pm	Literature Review and Dataset Search	16/8/2024
3-4	16/8/2024	11:15 am – 1:15 pm	Resource Gathering, Dataset Finalization, Planning	30/8/2024
5	30/8/2024	11:15 am – 1:15 pm	Initial Model Creation	6/9/2024
6	6/9/2024	11:15 am – 1:15 pm	Model Training	12/9/2024
7	12/9/2024	3:00-5:00 pm	Model Improvement	19/9/2024
8	19/9/2024	3:00-5:00 pm	Output Cleaning	24/9/2024
9	24/9/2024	3:00-5:00 pm	Critical Result Analysis	30/9/2024

Table 2

4.3 Module-wise Implementation Details

Each module was implemented with a specific focus to ensure smooth progression and accuracy in the final output. Below is the detailed breakdown of each module and the associated steps.

4.3.1 Ideation (Week 1)

- Date: 8/1/2024 – 8/8/2024
- Task: Ideation involved brainstorming sessions to define the scope of the project, including the desired features of the generative AI system, such as mood-based sound generation and environmental sound customization.
- Outcome: A clear project proposal outlining the system's goals, functionality, and impact on sound production.

4.3.2 Literature Review and Analysis (Week 2)

- Date: 8/8/2024 – 8/16/2024
- Task: Comprehensive literature review to understand the existing methodologies used in AI sound generation, including GANs, VAEs, and Transformer models. The review helped identify gaps in the current research, which guided the project's direction.
- Outcome: Summarized findings from key papers (Huang & Huang, 2020; Kobayashi et al., 2024; Dong, 2024), which provided insights into the mood-based sound generation approach.

4.3.3 Dataset Search and Presentation (Week 3-4)

- Date: 8/16/2024 – 8/30/2024
- Task: Search for appropriate datasets, focusing on mood-labeled music tracks. A dataset of 35 songs, each associated with mood labels, was identified and prepared in CSV format for model training. During this phase, a presentation was also created to explain the approach to team members and stakeholders.
- Outcome: Dataset finalized and presentation delivered. This dataset was used as the input for model training.

4.3.4 Initial Model Creation (Week 5)

- Date: 8/30/2024 – 9/6/2024
- Task: Development of the initial GAN-based model. The model used spectrograms of sound

frequencies to map moods to generated soundscapes. Key hyperparameters, such as latent dimensions, N-mels, and sequence lengths, were defined during this stage.

- Outcome: An initial working version of the GAN model, capable of generating basic mood-aligned sound outputs.

4.3.5 Model Training (Week 6)

- Date: 9/6/2024 – 9/12/2024
- Task: The model was trained using the dataset over 100 epochs. The training process involved tuning the model to minimize noise and enhance the clarity of the generated soundscapes. AWS cloud services and GPUs were utilized to accelerate the training.
- Outcome: Basic results were achieved with some roughness in the generated sounds, which highlighted the need for more data and extended training.

4.3.6 Model Improvement (Week 7)

- Date: 9/12/2024 – 9/19/2024
- Task: Additional training was conducted by increasing the dataset to 35 songs and extending the training to 200 epochs. The model's hyperparameters were fine-tuned, and additional layers were added to enhance sound diversity and quality.
- Outcome: Significant improvements in the generated sound quality, with more distinct mood-aligned outputs and clearer audio files.

4.3.7 Output Cleaning (Week 8)

- Date: 9/19/2024 – 9/24/2024
- Task: The generated outputs were cleaned using post-processing tools to remove noise and refine the audio quality. Techniques like audio compression and normalization were applied.
- Outcome: Finalized soundscapes that were ready for evaluation and use.

4.3.8 Critical Result Analysis (Week 9)

- Date: 9/24/2024 – 9/30/2024
- Task: The generated soundscapes were evaluated using subjective listening tests. Feedback was gathered from users, and accuracy in mood alignment and sound quality was measured.
- Outcome: Positive results, with an 80% satisfaction rate in audio quality and a 90% rate in mood

CHAPTER 5: RESULT AND DISCUSSIONS

This chapter presents the results of the generative AI sound production system, discussing the outcomes and evaluating their accuracy in relation to the project's objectives. The system aimed to generate diverse, high-quality environmental sounds and mood-based music, addressing the limitations of traditional sound sourcing methods. By training a GAN-based model using mood-labeled songs, the project sought to achieve scalable and customizable sound generation.

5.1 Results

The initial dataset comprised 20 mood-labeled songs, which were used to train the GAN spectrogram model over 100 epochs. The early results were promising but produced somewhat rough and unrefined soundscapes. These initial outputs exhibited a basic correspondence to the mood labels but lacked clarity and precision in the audio quality. Key challenges included incomplete harmonies and noise artifacts that affected the overall sound fidelity.

To improve the results, the dataset was expanded to 35 songs, and training was extended to 200 epochs. After this adjustment, the system demonstrated significantly better performance. The generated soundscapes became clearer, with more defined frequencies and fewer artifacts. For example, the sounds associated with "happy" moods displayed lively rhythms and higher-pitched frequencies, while those for "calm" moods produced smoother, low-pitched tones with a more serene ambiance. This improvement highlights the model's ability to learn and refine sound generation with additional data and training time.

Quantitative Evaluation:
A subjective listening test was conducted with a group of 10 participants, where they were asked to rate the generated sounds on two metrics: (1) audio quality and (2) mood alignment. The results showed that:

Audio Quality: After the model's enhancement, 80% of the participants rated the sounds as “good” or “very good” in clarity and overall fidelity, compared to only 55% during the first trial with the smaller dataset.

Mood Alignment: Participants agreed that 90% of the generated sounds aligned well with the intended emotional labels after the 200-epoch training, compared to 70% alignment with the initial 100-epoch training.

The model also showed adaptability in generating unique soundscapes for different mood inputs. Sounds generated from the random noise input varied greatly between different runs, adding diversity to the results, yet consistently adhered to the mood label constraints.

5.2 Discussion

The results demonstrate that the GAN-based sound generation system was able to create mood-driven music and environmental sounds that aligned well with the intended emotional labels. The increase in dataset size and training epochs played a critical role in enhancing both the sound quality and mood alignment, showing that the model can improve with more data and further optimization of hyperparameters.

However, some limitations remain. The system's ability to generate long-duration soundscapes still faces challenges, with some of the extended sounds lacking coherence in structure. This limitation aligns with findings from previous literature (Dong, 2024), where similar models struggled with maintaining long-term coherence in audio generation. Addressing this issue would require more advanced temporal dynamics in future iterations of the model.

Moreover, while the system generated diverse outputs from random noise, there was occasional repetition in the patterns of the generated sound, suggesting that the model's creative diversity could still be expanded. More extensive datasets and advanced techniques, such as Transformer-based models or hybrid architectures, could be explored to further enhance the diversity and complexity of the generated audio.

5.3 Accuracy and Justification

The accurateness of the results can be attributed to the methodical approach taken during the development process. By iterating through different training phases, tuning hyperparameters, and expanding the dataset, the model was able to refine its performance. The positive feedback from subjective listening tests confirms that the system achieved its goal of generating clear, mood-aligned soundscapes.

Furthermore, the model's ability to adjust the audio output based on mood inputs and random noise highlights its adaptability, supporting the hypothesis that a GAN-based model can effectively generate customizable soundscapes. However, there is still room for improvement, particularly in refining long-duration sounds and expanding the model's creative potential.

5.4 Summary

In summary, the project successfully demonstrated the capability of GAN-based models to generate mood-based environmental sounds and music. With improvements in data volume and training parameters, the system was able to produce higher-quality and mood-aligned soundscapes. Although some limitations persist, the results illustrate the model's potential to transform sound generation, offering an accessible and scalable solution for creators. Future work will focus on overcoming the challenges of coherence in longer sound sequences and exploring ways to further diversify the generated outputs.

CONCLUSION

The primary aim of this project was to develop a generative AI system capable of producing customizable, high-quality environmental sounds and mood-based music to address the limitations of traditional sound sourcing. By leveraging a GAN-based model, the project sought to provide an efficient, scalable, and affordable solution for sound generation, enhancing the flexibility and accessibility for creators.

Through the experiments conducted, the project successfully met its objectives. The GAN spectrogram model demonstrated the ability to generate soundscapes that aligned well with specific emotional cues, providing diverse outputs based on random noise and mood inputs. The system's performance improved significantly with an expanded dataset and extended training epochs, resulting in clearer and more defined audio output. Key findings showed that the generated sounds exhibited strong mood alignment, with subjective listening tests confirming an 80% satisfaction rate in sound quality and a 90% satisfaction rate in mood accuracy.

The major outcome of this investigation is the system's potential to transform sound generation for digital projects, making it more accessible and customizable for users ranging from novice to professional creators. This system effectively reduces the time and cost associated with traditional sound sourcing, offering a scalable alternative for producing immersive audio experiences.

While some challenges remain, such as improving the coherence of long-duration sound sequences and enhancing creative diversity, this research highlights the significance of GAN-based models in pushing the boundaries of sound generation. The project demonstrates the model's capacity to evolve with further data and training, suggesting promising opportunities for future development in AI-driven sound production.

REFERENCES

1) Emotion-based AI Music Generation System with CVAE-GAN. Chih-Fang Huang, Cheng-Yuan Huang. 2020.

2) Novice-Centered Application Design for Music Creation. Atsuya Kobayashi, Tetsuro Sato, Kei Tateno. 2024.

3)Generative AI for Music and Audio. Hao-Wen Dong. 2024

4) AI-Based Affective Music Generation Systems: A Review of Methods and Challenges. Adyasha Dash, Kathleen Agres. 2024

5) A User-Guided Generation Framework for Personalized Music Synthesis Using Interactive Evolutionary Computation. Yanan Wang, Yan Pei, Jianqiang Li, Zerui Ma. 2024

APPENDIX

Training Data Link

https://research.google.com/audioset/dataset/index.html

Output Link

https://drive.google.com/drive/folders/1GmrKRB0ItRDpbsTCpvpnYi5hvJJS_52c

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to the individuals and organizations who contributed to the successful completion of our project.

First and foremost, I extend my sincere appreciation to the college authorities. Dr.Shaikh Phiroz , the Head of the Department, provided invaluable support and guidance throughout this endeavor. Prof.Shainila Mulla, our mini project coordinator, played a crucial role in coordinating various aspects of the project.

I am deeply indebted to our Project Guide, Prof. Imran Ali Mirza, for his unwavering guidance, mentorship, and expertise. His insights were instrumental in shaping our project's direction.

I would also like to acknowledge the collective effort of the teaching and non-teaching staff who contributed to the project exhibition day. Their support in terms of infrastructure, such as labs, computers, and internet access, was instrumental in the successful execution of our project.

Lastly, I am grateful for the lecture slots and labs allotted to us for the completion of our project. These resources played a significant role in our project's development.

I am thankful to each of you for your contributions and support, which have been vital to the realization of this project.

Project Team Members:

1. Andre Marshal Fernandes T. E. – (Roll No. 23)
2. Jowin Paulose T. E. – (Roll No. 27)
3. Suyash Sanjay Kadam T. E. – (Roll No. 28)