# Modeling the implicit music representations in human brain with Deep Neural Networks

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science*
*in*
*Computer Science and Engineering by Research*

by

Ravinder Singh
20163052
ravinder.singh@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled ' Modeling the implicit music representations in human brain with Deep Neural Networks.' by Ravinder Singh, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Vinoo Alluri

To **family**, friends and

teachers

# Acknowledgments

I am deeply grateful to Prof. Vinoo Alluri for her continuous support and invaluable guidance throughout this journey. Her expertise, mentorship, and approachability have been invaluable assets. I am truly fortunate to have had the opportunity to learn from her and benefit from her constructive feedback and criticism, which have played a crucial role in improving my work.

I would also like to extend my heartfelt thanks to Dr. Petri Toiviainen, whose mentorship and guidance have been immensely valuable. His insights and expertise have contributed significantly to shaping this research and enhancing its quality.

I am incredibly grateful to my dear friend Rajat Agarwal, who has been my constant motivator. His unwavering support, willingness to listen, and insightful discussions have been invaluable. Without his presence and encouragement, this arduous but worthwhile research journey would not have been the same.

Furthermore, I would like to acknowledge the unwavering support and blessings of my family. My wife, Manender Kaur, and my sons, Mannandeep Singh and Sehajdeep Singh, have been my pillars of strength throughout this endeavor.. I would also like to express my gratitude to my parents, Jagmohan Singh and Jaswant Kaur & my sister Taranjeet Kaur, for their constant encouragement and belief in my abilities. Their belief in my potential and their continuous push for me to achieve my best work have been invaluable.

Lastly, I would like to extend my sincere appreciation to my closest friends for life, Kishore Gautam, Devendra Rawat, Vipin Sharma, Precious Kalia, Nakul and Arun Gupta. Their positive support and encouragement have been a constant source of motivation and inspiration.

"We are all born with a divine fire in us. Our efforts should be to give wings to this fire and fill the world with the glow of its goodness."

— A.P.J. Abdul Kalam, Wings of Fire

# Abstract

Music processing is a fascinating and intricate phenomenon that has garnered significant research interest in recent years. Understanding how the human brain represents various music features is crucial for unraveling the mysteries of music perception and cognition. In this study, our objective was to investigate neural representations of music using functional magnetic resonance imaging (fMRI) to analyze the blood-oxygen-level-dependent (BOLD) signal activations in selected regions of interest (ROIs) during a continuous listening task.

Additionally, we aimed to compare these neural activations with the hidden layer activations of a specific class of deep neural networks (DNNs). These DNNs, known as self-supervised models, have shown promise in capturing intricate patterns and encoding complex information. By utilizing representational similarity analysis (RSA), we aimed to explore the similarities and correlations between the neural representations of music features in the human brain and the hidden layers of the DNN encoder.

Our findings revealed a correlation between the low-level music feature encoding observed in two important brain regions, namely the Superior Temporal Gyrus (STG) and Heschl's gyrus (HG), and the hidden layers of the DNN encoder. This correlation provides evidence for the effectiveness of self-supervised DNNs as a reliable architecture for studying the domain of music processing. Importantly, this finding is particularly significant due to the limitations of naturalistic listening conditions in prior research studies.

By bridging the gap between the neuroscientific investigation of music processing and the computational power of self-supervised DNNs, our study contributes to the growing body of research aiming to uncover the underlying mechanisms of music perception and representation. The implications of this research extend beyond the field of music, as the insights gained from studying music processing can potentially shed light on broader topics such as auditory cognition, neural encoding of complex stimuli, and the applications of deep learning in cognitive neuroscience.

**Keywords**: music processing, neural representations, functional magnetic resonance imaging (fMRI), blood-oxygen-level-dependent (BOLD) signal, deep neural networks (DNNs), self-supervised models, representational similarity analysis (RSA), Superior Temporal Gyrus (STG), Heschl's gyrus (HG), naturalistic listening conditions, music perception, auditory cognition, computational neuroscience.

# Contents


Chapter Page

1 Introduction . . . 1
1.1 Motivation . . . 1
1.2 Research Objectives . . . 2
1.3 Key Contributions . . . 2
1.4 Thesis Roadmap . . . 3

2 Background . . . 4
2.1 Introduction . . . 4
2.2 Music Processing in the Auditory Pathways of the Human Brain . . . 5
2.2.1 Auditory Pathways and Music Perception . . . 5
2.2.2 System Neuroscience and the Study of Music Processing . . . 6
2.3 Computational Modeling and Music Processing . . . 6
2.3.1 Scope and Significance of Studying Music Processing . . . 6
2.4 Introduction to Neuroimaging . . . 6
2.4.1 Functional Magnetic Resonance Imaging (fMRI) in Music Cognition . . . 7
2.5 Understanding Musical Cognition with Deep Learning . . . 9
2.5.1 Auto Encoding . . . 10
2.5.2 Variational Autoencoder : Architecture, Functioning, and Mathematical Basis . . . 11
2.6 RSA for Studying fMRI with Computational Models . . . 12

3 Training and Evaluation of Autoencoders . . . 14
3.1 Introduction . . . 14
3.2 Method and Material . . . 15
3.2.1 Model Architecture overview . . . 15
3.2.2 Datasets . . . 16
3.2.3 Feature Extraction . . . 17
3.2.4 Method . . . 17
3.3 Results . . . 19
3.4 Conclusion . . . 20

4 Modelling the implicit music representation in brain using DNN . . . 23
4.1 Introduction . . . 23
4.2 Materials and Method . . . 23
4.2.1 Dataset . . . 23

4.2.2 Method . . . . . 24
4.3 Results . . . . . 26

5 Conclusion and Future Work . . . . . 30
5.1 Discussion . . . . . 30
5.2 Limitations . . . . . 31
5.3 Future Work . . . . . 31

Bibliography . . . . . 33

# List of Figures


Figure Page

2.1 Autoencoder Architecture . . . . . 10

3.1 VAE Encoder Architecture illustrating the hdden layer dimensions . . . . . 16
3.2 Classification Task Pipeline . . . . . 18
3.3 Confusion Matrix of GTZAN trained VAE classifier . . . . . 21
3.4 Inspection of reconstructed music stimuli from GTZAN trained VAE . . . . . 22

4.1 Schematic of pipeline to extract and process inputs and outputs for RSA . . . . . 25
4.2 Schematic to illustrate the RSA . . . . . 26
4.3 Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for non musician group . . . . . 27
4.4 Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for musician group . . . . . 28
4.5 Combined view of Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for both musician and non-musicians . . . . . 29

# List of Tables

Table Page

3.1 Datasets . . . . . 17
3.2 Average Classification Accuracies on reconstructred stimuli; Trained on custom datasets and tested on an unseen data by any model using our baseline CNN classifier. . . . . 20

*Chapter 1*

# Introduction

## 1.1 Motivation

The study of music perception has traditionally relied on behavioral experiments and subjective assessments (Peretz & Zatorre, 2003 ) [49]. While these approaches have provided valuable insights, they often fall short in capturing the intricate and implicit representations of music in the human brain. Recent advancements in neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), have opened up new possibilities for investigating the neural correlates of music perception (Zatorre & McGill, 2005) [63]. By examining the neural activity patterns associated with music processing, we can uncover the underlying representations and mechanisms at a more detailed level.

Understanding the neural basis of music perception and representation is crucial for gaining insights into how the brain encodes and processes musical information (Janata, 2009) [27]. By leveraging the power of deep learning techniques, such as variational autoencoders (VAEs), we can develop computational models that capture the implicit representations of music in the human brain (Herremans et al., 2017) [**?**] . These models have demonstrated remarkable capabilities in capturing complex patterns and generating meaningful representations from high-dimensional data (Knigam & Welling, 2013) [35]

Connecting the branches of neuroscience and machine learning presents both opportunities and challenges. While some progress has been made in exploring the neural correlates of music perception using fMRI, the current research has its limitations. Most studies have utilized have followed a block-based experimental design (Kuhl et al., 2013) [35]. One major limitation in recent study(Guculu et al., 2017) [22] used task-optimized Deep Neural networks classifiers with neuroimaging to interpret music processing in human brain using short music clips of 6 seconds. In this study, we aim to push the boundaries of music neuroscience research by addressing these challenges. By employing fMRI in naturalistic conditions and incorporating advanced machine learning techniques, we seek to capture the complexity of music processing in the human brain in a more ecologically valid manner. Our goal is to explore the uncharted

territories of music perception and representation, leveraging the latest advancements in machine learning to uncover new insights and refine our understanding of how the brain processes and represents music . This study aims to bridge the gap between neuroscience and machine learning, contributing to both fields while enhancing our understanding of music perception

## 1.2 Research Objectives

Our research was aimed at following objectives:

- Identify the best performing computational model from the different class of autoencoders via a music classification task.
- Investigate the neural representations of music features by examining fMRI activations and comparing them with hidden layer activations of the selected unsupervised deep neural network, specifically variational autoencoders (VAEs) using Representational Similarity Analysis(RSA).
- Unlike the prevailing experimental designs that rely on short music clips and a block-design approach, our approach utilizes natural stimuli, specifically real music clips of longer durations. Additionally, we capture fMRI data continuously throughout the entire song. By doing so, we can extract features based on natural music listening conditions, which enable a more accurate depiction of how music is processed in the brain.
- Gain insights into the correspondence between DNN representations and fMRI responses to better understand the neural processing and representation of music in the human brain.

## 1.3 Key Contributions

1. Adoption of an unsupervised learning approach: The research highlights the use of unsupervised deep neural networks, specifically variational autoencoders (VAEs), for modeling the complex learning representation of music. This approach overcomes the limitations of task-optimized DNNs and captures the intricate nature of music processing [22].
2. Investigation of neural representations: The study delves into the neural representations of music features by analyzing fMRI activations in selected brain regions and comparing them with hidden layer activations of VAEs. This analysis allows for a comprehensive exploration of the correspondence between learned representations and neural activations in response to music.

3. Broadening the hypothesis space: Unlike previous studies that focused on hand-designed features, this research utilizes complex and dynamic natural stimuli, specifically real music clips. By extracting features derived from VAE representations, the study broadens the hypothesis space and captures a probabilistic and generative representation of how music, as a complex stimulus, is processed in the human brain.

These key contributions contribute to a better understanding of music perception and representation in the human brain and pave the way for future studies in the field.

## 1.4 Thesis Roadmap

The thesis is constructed in the following way

- Chapter 2 deals with the background of Musical cognition, Neuroimaging & Representational similarity analysis, Computational Modeling and mathematical basis of special class of autoencoders called variational autoencoders(VAE)
- Chapter 3 deals with a study to identify the best-performing computational model in terms of class of deep learning autoencoder- Long Short-tem Memory based(LSTMs) v/s (VAE) based. This is achieved in a classification task experimentation of eight different computational models trained on diverse music dataset.
- Chapter 4 of this thesis focuses on the main study, which aims to investigate how music is processed in the human brain. This investigation involves modeling implicit music representations using functional magnetic resonance imaging (fMRI) and a specific class of deep learning autoencoder as identified in earlier experimentation. The chapter describes the implementation of Representational Similarity Analysis technique to examine the correlations of brain activations with autoencoder hidden layers
- Chapter 5 concludes with summary, limitations and future prospects of the studies.

*Chapter 2*

# Background

## 2.1 Introduction

Music processing is a captivating and intricate phenomenon that has garnered significant research interest in recent years [38]. In this master's thesis, our focus is on exploring the neural representations of music features using functional magnetic resonance imaging (fMRI) activations in selected regions of interest during a continuous music listening task. We also aim to compare these activations with the hidden layer activations of a unique class of deep neural networks (DNNs). To achieve this, we employ representational similarity analysis (RSA) to investigate the correlation between low-level music feature encoding in the human brain regions, such as the Superior Temporal Gyrus (STG) and Heschl's gyrus (HG), and the hidden layers of unsupervised deep neural networks, particularly variational autoencoders (VAEs).

The theoretical framework underlying music perception and representation in the human brain has been extensively studied. Various investigations have employed deep learning techniques for music processing [52, 37, 34, 18, 41] , including the use of RSA to examine neural representations [10, 53, 62, 55]. RSA has proven applicable in studying music perception as it enables comparisons between neural responses and representations derived from computational models [10]. By adopting unsupervised learning approaches, such as VAEs, we aim to justify their usage in modeling the music learning representation, highlighting their advantages over task-optimized DNNs (Gómez-Herrero, Germán & Peeters, 2021) [21]

Our study places a significant emphasis on the use of unsupervised learning, particularly VAEs. Unlike supervised task-optimized DNNs that are trained for specific tasks, VAEs operate based on unsupervised learning principles, allowing them to capture the underlying structure and patterns in the data without the need for explicit labels [13]. This makes VAEs particularly well-suited for modeling complex and high-dimensional data, such as music features. By leveraging the latent representations learned by VAEs, we can explore the neural correlates of music processing in a more flexible and comprehensive manner [32].

In previous studies, the majority of research on functional auditory cortical representations has focused on hand-designed low-level and high-level features, which may introduce biases and limitations. However, in our investigation, we depart from this approach by utilizing complex and dynamic natural stimuli, specifically real music clips. The features extracted from these music clips are derived from the representations learned within a data-driven and unsupervised VAE. This approach allows us to broaden the hypothesis space and capture a more accurate representation of how complex stimuli, such as music, are processed in the brain.

Moreover, the utilization of DNNs, particularly VAEs, for probing neural representations has demonstrated remarkable success in visual neuroscience [62]. By extending this line of research to the domain of music, we aim to shed light on how the human brain responds to musical stimuli. The hierarchical structure of VAEs, with their encoder and decoder components, enables the exploration of representations at different levels of abstraction. Through the comparison of DNN representations and fMRI responses to the same sensory stimuli, we can gain valuable insights into the neural processing of music and its underlying representations.

In conclusion, this master's thesis investigates the neural representations of music features by analyzing fMRI activations and comparing them with hidden layer activations of unsupervised VAEs. By adopting an unsupervised learning approach, we aim to overcome the limitations of task-optimized DNNs and capture the intricate nature of music processing. The use of VAEs allows us to model the complex learning representation of music and explore the correspondence between the learned representations and neural activations. This research contributes to the understanding of music perception and representation in the human brain, paving the way for future studies in the field.

## 2.2 Music Processing in the Auditory Pathways of the Human Brain

The processing of music in the human brain is a fascinating and complex phenomenon that has garnered significant research interest in recent years. The auditory pathways in the brain play a crucial role in perceiving and understanding music, unraveling the mysteries of music perception and cognition [50].

### 2.2.1 Auditory Pathways and Music Perception

The auditory pathways in the human brain consist of a hierarchical network of structures involved in processing auditory information. The primary auditory cortex, located in the superior temporal gyrus (STG), serves as the initial processing stage for incoming sound signals. From there, information is transmitted to higher-level auditory areas, such as the secondary

auditory cortex, frontal cortex, and limbic system, which are responsible for more complex auditory processing, including music perception [45].

### 2.2.2 System Neuroscience and the Study of Music Processing

The study of music processing in the auditory pathways falls within the domain of systems neuroscience, which aims to understand the brain's organization and functioning at the system level. Systems neuroscience investigates how different brain regions interact and contribute to specific cognitive functions, such as music perception and processing [36].

## 2.3 Computational Modeling and Music Processing

Computational modeling provides a powerful tool for studying complex cognitive processes, such as music processing, in the human brain. By developing computational models that simulate the underlying neural mechanisms involved in music perception, researchers can investigate hypotheses and generate predictions about how the brain processes and represents music.

### 2.3.1 Scope and Significance of Studying Music Processing

Despite significant advancements in understanding music processing, many questions remain unanswered. The field of music neuroscience presents a vast and exciting frontier for further exploration. By studying music processing using computational modeling approaches, researchers can bridge the gap between neuroscience and computer science, uncovering the underlying mechanisms of music perception and representation. Moreover, insights gained from studying music processing can extend beyond the field of music itself, shedding light on broader topics such as auditory cognition, neural encoding of complex stimuli, and the applications of deep learning in cognitive neuroscience.

## 2.4 Introduction to Neuroimaging

Neuroimaging has revolutionized the field of neuroscience by providing powerful tools to investigate the inner workings of the human brain. Through non-invasive imaging techniques, researchers can observe and analyze brain activity, structure, and connectivity, offering unprecedented insights into cognitive processes, neurological disorders, and the mechanisms underlying human behavior.

Neuroimaging encompasses a wide range of methodologies, including magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), and magnetoencephalography (MEG). Each technique

offers unique advantages and captures different aspects of brain activity, allowing researchers to examine the brain at various spatial and temporal scales.

One of the most widely used neuroimaging techniques is functional magnetic resonance imaging (fMRI), which measures changes in blood oxygenation levels to infer neural activity [40]. By detecting regional blood flow, fMRI provides a detailed map of brain regions that are active during specific tasks or in resting state conditions [4]. This information helps researchers understand brain networks, functional connectivity, and how different regions collaborate to perform complex cognitive functions [7]..

The use of neuroimaging has significantly advanced our understanding of various cognitive processes, such as perception, attention, memory, language, and decision-making. Moreover, it has shed light on the neural mechanisms underlying neurological and psychiatric disorders, including Alzheimer's disease, schizophrenia, depression, and autism spectrum disorders. Neuroimaging studies have revealed structural and functional alterations in these disorders, contributing to the development of new diagnostic tools and treatment strategies.

Importantly, neuroimaging techniques have enabled researchers to investigate the brain *in vivo*, allowing for longitudinal studies and the examination of developmental changes across the lifespan. By tracking brain development from infancy to adulthood, researchers can uncover critical periods of neural plasticity, elucidate the impact of environmental factors on brain structure and function, and gain insights into the underlying mechanisms of neurodevelopmental disorders.

Furthermore, neuroimaging has fostered interdisciplinary collaborations between neuroscience, psychology, biology, medicine, and computer science. The integration of advanced data analysis techniques, such as machine learning and network analysis, has enhanced our ability to extract meaningful information from complex brain data and develop predictive models.

In conclusion, neuroimaging techniques have revolutionized the field of neuroscience by providing a window into the inner workings of the human brain. With their non-invasive nature, high spatial and temporal resolution, and ability to capture brain activity *in vivo*, these techniques have opened new avenues for understanding brain function, unraveling the mysteries of cognition and behavior, and advancing the diagnosis and treatment of neurological disorders. As technology continues to advance, neuroimaging holds immense promise for further discoveries that will deepen our understanding of the complex organ that is the human brain.

### 2.4.1 Functional Magnetic Resonance Imaging (fMRI) in Music Cognition

In the realm of music cognition, functional magnetic resonance imaging (fMRI) has gained significant prominence as a popular choice among researchers. This introduction highlights the rationale behind the widespread use of fMRI in studying music cognition and its unique advantages in this domain.

fMRI, a non-invasive imaging technique, provides valuable insights into the neural mechanisms underlying music processing [50]. By measuring changes in blood oxygenation levels, fMRI allows researchers to investigate brain regions and networks involved in various aspects of music perception, cognition, and performance [59]. Its popularity in music cognition research stems from several compelling reasons.

One crucial aspect is the exceptional spatial resolution of fMRI, enabling researchers to identify the specific brain regions engaged during music processing. With its millimeter-scale precision, fMRI helps localize neural activity associated with perceiving pitch, rhythm, timbre, and emotional responses to music. This spatial specificity is invaluable in unraveling the intricate neural architecture underlying music cognition [59, 58].

Additionally, the non-invasive nature of fMRI makes it an ethical and safe method for studying the neural processes involved in music cognition. Researchers can explore the musical experiences of individuals, including professional musicians and non-musicians, without the need for surgical procedures or exposure to harmful radiation. This accessibility allows for a comprehensive investigation of music perception and cognition across diverse populations [59].

fMRI's ability to capture neural activity in real-time contributes to its significance in music cognition research. By observing dynamic changes in brain activity, researchers can study the temporal unfolding of musical processes, such as melody processing, rhythm perception, and emotional responses to music. This temporal resolution provides a deeper understanding of how music engages and influences the brain over time [59].

The versatility of fMRI further enhances its utility in music cognition research. It can be integrated with various experimental paradigms, such as listening tasks, music performance, improvisation, and synchronization studies, allowing researchers to investigate different aspects of music cognition [59, 58]. By combining fMRI with behavioral measures, music theorists, psychologists, and neuroscientists can unravel the complex interplay between musical structure, emotion, and cognitive processes [59].

However, it is important to acknowledge the limitations of the traditional block-based design used in many fMRI experiments in music cognition [47]. This design involves presenting short music stimuli in discrete blocks, which may not fully capture the continuous and dynamic nature of music [64]. Music is often experienced as a continuous flow, and the block-based design may not capture the temporal nuances and structural complexity of music processing [39]. This limitation has led to calls for more naturalistic paradigms that simulate real-world music listening experiences, challenging the constraints imposed by block-based designs [28] [2].

Despite the challenges and potential inaccuracies associated with more naturalistic paradigms, pushing the boundaries of traditional fMRI designs is worth considering. By incorporating more ecologically valid and continuous tasks, researchers can explore a wider range of possibilities and gain a deeper understanding of music processing in the brain. These advancements may

come with their own set of technical and analytical challenges, but they offer opportunities for uncovering new insights into music cognition [59, 58].

In conclusion, fMRI has emerged as a popular and valuable tool in the study of music cognition. Its non-invasiveness, exceptional spatial and temporal resolution, and versatility make it an ideal choice for investigating the neural underpinnings of music perception, cognition, and performance [50, 59]. By shedding light on the intricate processes involved in music processing, fMRI contributes to our understanding of the profound impact of music on the human brain. As music cognition research continues to progress, it is important to explore alternative paradigms that challenge the limitations of traditional block-based designs and pave the way for more ecologically valid investigations using fMRI [2].

## 2.5 Understanding Musical Cognition with Deep Learning

Deep learning has emerged as a powerful and versatile approach in the field of artificial intelligence, revolutionizing various domains such as computer vision, natural language processing, and robotics [37]. Autoencoders, a type of neural network architecture, have proven to be valuable tools in deep learning research [23]. Autoencoders aim to learn efficient representations of input data by reconstructing it from a compressed latent space, making them useful for tasks such as dimensionality reduction, data denoising, and anomaly detection.

In the context of the research at hand, autoencoders offer a promising avenue for analyzing and understanding musical cognition [57]. Music cognition involves studying how the human brain perceives, processes, and responds to music [48]. Autoencoders can extract meaningful representations from musical data, revealing intricate patterns, musical motifs, and relationships between musical elements.

Autoencoders have practical applications in music analysis tasks, including music generation, recommendation, and style transfer [11]. By training autoencoders on large musical datasets, researchers can learn compact representations that capture the essence of musical compositions. These representations can be used to generate new musical pieces, recommend music based on user preferences, or transform music from one style to another while preserving its fundamental structure [25].

Additionally, autoencoders can help uncover hidden factors and latent variables in music [25]. By encoding musical data into a low-dimensional latent space, autoencoders reveal underlying musical concepts and dimensions. This insight enhances our understanding of how different musical elements contribute to the overall perception and emotional response to music [5]. Such information has significant implications for music composition, cognitive psychology, and the development of intelligent music systems [5].

In conclusion, deep learning and autoencoders provide exciting opportunities for exploring the intricacies of musical cognition [5]. By leveraging their ability to learn expressive repre-

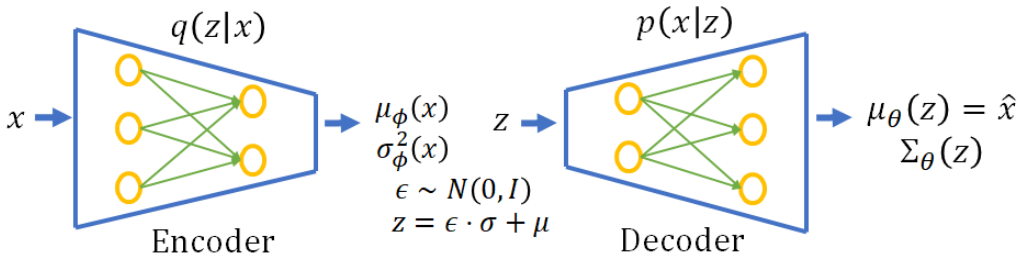



**Figure 2.1** Autoencoder Architecture

sentations from raw musical data, researchers can analyze, generate, and understand music at a deeper level [5]. Deep learning techniques, including autoencoders, contribute to unlocking new frontiers in music research and enhancing our understanding of the complex relationship between the human brain and music.

### 2.5.1 Auto Encoding

Autoencoders (AEs) are a neural network architecture used mainly to compress data or perform dimensionality reduction. They consist of two neural networks: an encoder $f_\phi(x)$ that encodes the original data $x$ into a latent space $z$, and a decoder $g_\theta(z)$ that reconstructs the data from the latent space. The encoder is parameterized by the weights $\phi$, and the decoder is parameterized by $\theta$. The overall structure is illustrated in Figure 2.1.

To optimize the neural network, the typical approach is to calculate the L2 norm of the difference between the original data $x$ and the reconstructed data $\hat{x} = g(f(x))$. This is achieved by evaluating the loss function as the squared Euclidean distance: Loss $= \|x - \hat{x}\|_2^2$. Common optimization techniques such as Gradient Descent (e.g., SGD, ADAM, etc.) are then used to update the weights of the neural network. Once the weights have been optimized, the outputs of the encoder and decoder become fixed: for a given input $x$, the encoder will consistently produce the same latent representation $z = f(x)$, and the decoder will consistently generate the same reconstructed data $x = g(z)$.

In the music domain, Long Short Term Memory(LSTM)-based autoencoders demonstrate effectiveness for several reasons supported by prior research. Firstly, LSTMs excel at capturing long-term dependencies in sequential data, a critical aspect in music where melodies, rhythms, and structures span extensive time frames [24]. Secondly, by integrating LSTMs into the autoencoder framework, hierarchical representations of music data are learned. Lower layers capture local patterns, while higher layers encapsulate more abstract musical features [8]. Lastly, LSTM-based autoencoders can generate coherent and expressive music sequences that encapsulate the stylistic elements of the input data. Through precise reconstruction of

input data during training, the acquired latent space facilitates the generation of new music compositions with characteristics akin to the training data [17].

It might seem intuitive to use this framework for generating new music by sampling random points in the latent space $z$ and passing them through the decoder [51]. However, this approach often yields suboptimal results. The reason behind this is that the optimization process of autoencoders focuses solely on reconstructing the actual data, disregarding the organization and structure of the latent space [51]. Consequently, the latent space can become disorganized and fail to provide meaningful representations for music generation [14, 51].

To address this limitation, Variational Autoencoders (VAEs) introduce probabilistic modeling to the latent space [51]. Rather than directly encoding music into a deterministic latent representation, VAEs map the music into a distribution over the latent space. This allows for the generation of new music by sampling from the latent distribution and decoding the samples through the decoder network [51]. By modeling the latent space as a probability distribution, VAEs enable the exploration and interpolation of different music styles and structures, leading to more diverse and realistic music generations [14, 51]. This probabilistic nature of VAEs provides a clear advantage over traditional autoencoders, such as LSTM-based autoencoders, which lack explicit modeling of uncertainty in the latent space [31]. Consequently, VAEs tend to produce more diverse and realistic music samples, making them a superior choice for tasks requiring generative modeling and latent music structures in the music domain [14].

### 2.5.2 Variational Autoencoder : Architecture, Functioning, and Mathematical Basis

Variational Autoencoders (VAEs) are a type of generative model that learns to encode high-dimensional data into a lower-dimensional latent space, allowing for the generation of new data samples [31]. VAEs are composed of two main components: an encoder network and a decoder network. The encoder network maps the input data to the latent space, while the decoder network reconstructs the original data from the latent space.

The architecture of a VAE typically consists of multiple layers of neural networks. The encoder network takes the input data and gradually reduces its dimensionality, mapping it to the mean and variance parameters of the latent space distribution. The decoder network, on the other hand, takes samples from the latent space and reconstructs the original data by upsampling it to the original dimensionality.

The encoder network often employs convolutional or fully connected layers to capture and abstract the features of the input data. These layers gradually reduce the dimensionality of the data, resulting in a bottleneck layer that represents the latent space. The decoder network then uses upsampling or deconvolutional layers to reconstruct the original data from the latent space representation.

To train a VAE, a combination of a reconstruction loss and a regularization term is used. The reconstruction loss measures the similarity between the original data and the reconstructed data, encouraging the VAE to capture the essential features of the input data. The regularization term, typically the Kullback-Leibler (KL) divergence, encourages the latent space distribution to resemble a predefined prior distribution, often a multivariate Gaussian.

During the training process, the VAE aims to minimize the sum of the reconstruction loss and the regularization term. This is achieved through gradient-based optimization methods such as stochastic gradient descent. By iteratively adjusting the weights of the encoder and decoder networks, the VAE learns to encode the data into a meaningful latent space and generate new data samples by sampling from the latent distribution.

The mathematical basis of VAEs relies on variational inference, which approximates the true posterior distribution of the latent variables given the observed data. VAEs introduce an inference model (encoder) and a generative model (decoder) to jointly approximate the posterior distribution. The inference model learns to encode the data into a latent space distribution, while the generative model learns to decode the latent samples back into the data space. The training objective of VAEs involves maximizing the evidence lower bound (ELBO), which decomposes the log-likelihood of the data into a reconstruction term and a regularization term. By maximizing the ELBO, the VAE learns to capture the essential features of the data and generate new samples from the learned latent space distribution.

In summary, VAEs provide a powerful framework for learning meaningful representations of high-dimensional data and generating new samples. By leveraging the encoder-decoder architecture and variational inference, VAEs enable the exploration and interpolation of data in a structured latent space. In the context of music processing, VAEs offer valuable tools for analyzing and generating music, facilitating research in music cognition, composition, and intelligent music systems.

## 2.6 RSA for Studying fMRI with Computational Models

The examination of neural representations using functional magnetic resonance imaging (fMRI) has provided valuable insights into how the human brain processes various stimuli, including music [10, 53]. To further our understanding of music perception and representation, it is essential to employ advanced analytical techniques that enable comparisons between neural responses and representations derived from computational models. One such technique is Representational Similarity Analysis (RSA), which has proven to be applicable in studying music perception and has demonstrated success in other domains such as visual neuroscience [62, 55].

RSA allows for the quantification of the similarity between neural response patterns and representations obtained from computational models. By comparing the neural responses recorded during fMRI experiments with the representations learned by computational models, researchers

can gain valuable insights into the underlying mechanisms of information processing in the brain. This approach has been employed in several studies to investigate the correspondence between brain activations and representations derived from deep neural networks (DNNs) in various domains [10, 53, 62, 55].

One important advantage of RSA is that it enables a flexible exploration of neural representations. It does not rely on predefined regions of interest or handcrafted features but instead allows for a data-driven examination of the similarity structure in the neural response patterns. This makes RSA particularly well-suited for studying complex and high-dimensional stimuli such as music [10, 53].

In the context of this thesis, we aim to utilize RSA to investigate the neural representations of music features. Specifically, we focus on examining the blood-oxygen-level-dependent (BOLD) signal fMRI activations in selected regions of interest during a continuous listening task. Additionally, we seek to compare these activations with the hidden layer activations of deep neural networks, particularly variational autoencoders (VAEs), which have shown promise in capturing the underlying structure of complex data [62, 55].

The application of RSA in the study of music perception and representation has been supported by previous research. For example, Chen, Penhune, and Zatorre (2008) employed RSA to investigate the brain network involved in auditory-motor synchronization during rhythmic tasks [**?**]. Santoro, Daudet, and Sandler (2014) utilized RSA to examine the rhythmic patterning in auditory cortical responses [53]. These studies demonstrate the applicability of RSA in understanding music-related neural processing.

Moreover, RSA has been widely used in visual neuroscience to explore the correspondence between DNN representations and neural responses. Yamins et al. (2014) and Seibert et al. (2016) utilized RSA to predict neural responses in higher visual cortex based on hierarchical models [62, 55]. These studies highlight the success of RSA in uncovering the hierarchical structure of representations in the brain.

In conclusion, RSA provides a powerful tool for studying fMRI data in conjunction with computational models. By employing this technique, we aim to uncover the neural representations of music features and examine their similarity to the representations learned by VAEs. This research contributes to our understanding of music perception and representation in the human brain and opens avenues for further investigations in the field.

*Chapter 3*

# Training and Evaluation of Autoencoders

## 3.1 Introduction

This chapter presents the method of selecting the best computational model from a diverse set of autoencoders. This is achieved by training and evaluating eight autoencoder models, including four Long Short Term Memory (LSTM)-based autoencoders and four variational autoencoder (VAE) models. These models are trained on four different custom datasets: Bollywood, Western Instrumental, Indian Devotional, and the GTZAN dataset. The goal of this study was to investigate the effectiveness of these autoencoders in preserving music structure and their impact on music classification accuracy using a vanilla Convolutional Neural Network (CNN) based classifier. We discuss the training procedure, evaluation metrics, and the results obtained from the classification experiments and explain the basis of the computational model identified for our core study.

The selection and evaluation of autoencoder models can be supported by previous works in the field of machine learning (Hinton  Salakhutdinov, 2006 [23]). The use of Long Short-Term Memory (LSTM) in autoencoders has been proposed by (Graves, 2013) [19]), offering an improved capability to capture temporal dependencies in sequential data. Additionally, recent studies, such as 'Brains on Beats' by Güçlü et al. (2016 [22]), have highlighted limitations in supervised deep learning models based on AlexNet architecture for modeling music processing. These limitations suggest the potential for alternative approaches, such as LSTM-based autoencoders and variational autoencoders (VAEs), to better capture the complexities of music cognition. LSTM-based autoencoders can effectively model sequential music data, capturing long-term dependencies and structural patterns (Roberts et al., 2018 [51]). Furthermore, VAEs have shown promise in preserving music structure and generating coherent music sequences (Roberts et al., 2018 [51]). The probabilistic nature of VAEs enables them to capture uncertainty in music data, making them a powerful tool for understanding music cognition (Kingma Welling, 2013 [31]). Therefore, both LSTM-based autoencoders and VAEs emerge as poten-

tially good choices for evaluating the efficacy of a deep learning model in the study of music cognition.

## 3.2 Method and Material

### 3.2.1 Model Architecture overview

To capture the musical characteristics of the different genres, we designed four LSTM-based autoencoder architectures and four VAE-based autoencoder architectures. Each autoencoder architecture was specifically tailored to the characteristics of the respective dataset. We provide a detailed description of the architectures and discuss the rationale behind their selection. Additionally, we describe the pre-processing of the custom datasets, ensuring diversity and high-quality samples.

The LSTM-based autoencoder architecture consists of three layers of 1-dimensional convolutions followed by an encoder LSTM layer and an asymmetric decoder LSTM layer. The convolutions capture local patterns in the input sequence, while the encoder LSTM layer encodes the information into a fixed-length latent representation. The asymmetric decoder LSTM layer reconstructs the input sequence in an asymmetric manner. This architecture combines convolutional and LSTM layers to capture both local and long-term dependencies in the music features.

The VAE (Variational Autoencoder) architecture comprises an encoder network, a latent space, and a decoder network. The encoder network takes the input data and maps it to a latent space representation. The latent space serves as a compressed representation of the input data. The decoder network takes a sample from the latent space and reconstructs the original input data. The VAE is trained to minimize the reconstruction loss while simultaneously regularizing the latent space using a KL divergence term. This architecture allows for the generation of new data samples by sampling from the latent space. The VAE enables the learning of meaningful representations and provides a probabilistic framework for data generation and reconstruction.

Figure 3.1 illustrates the architecture of a VAE Encoder with layer wise dimensions. It is inspired by Alexnet architecture with following modifications:

- The number of convolutional kernels was halved compared to AlexNet.
- The convolutional and pooling kernels, as well as the strides, were flattened. This transformation involved changing an $n \times n$ kernel to an $n^2 \times 1$ kernel and an $m \times m$ stride to an $m^2 \times 1$ stride.
- Local response normalization, as used in AlexNet, was replaced with batch normalization [26].

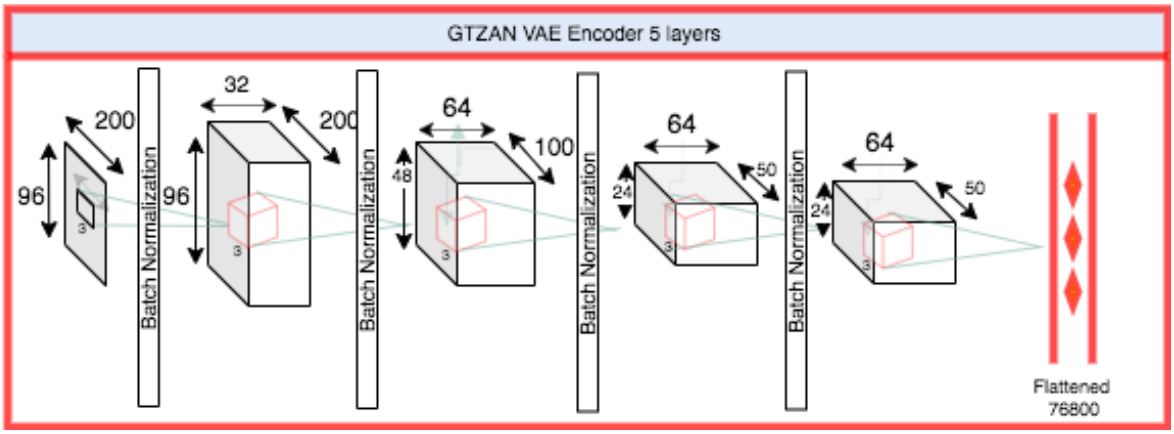



**Figure 3.1** VAE Encoder Architecture illustrating the hdden layer dimensions

- Rectified linear units (ReLUs) commonly used in AlexNet were replaced with parametric softplus units with initial parameters $\alpha = 0.2$ and $\beta = 0.5$ [16].
- Softmax units employed in AlexNet were substituted with sigmoid units.

For training, we utilized the Adam optimizer with learning rate $\alpha = 0.0002$, decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a small epsilon value $\epsilon = 1 \times 10^{-8}$ [30]. Additionally, a mini-batch size of 36 was employed during training [30]. The model was trained using Keras [12], with the final model selected based on the epoch in which the validation performance peaked.

It's important to note that the artificial neurons in the convolutional layers performed local filtering of their inputs through 1D convolutions, followed by non-linear transformations, resulting in temporal representations per stimulus. These representations were subsequently processed by averaging them over time. In contrast, the artificial neurons in the fully-connected layers conducted global filtering of their inputs via dot product operations, followed by non-linear transformations, and returned scalar representations per stimulus.

### 3.2.2 Datasets

In order to incorporate a wide range of diverse and complex stimuli, we carefully curated three musical datasets consisting of real musical clips. Additionally, we included the GTZAN Western music dataset [60] as a fourth dataset. The GTZAN dataset comprises 1000 music excerpts, each lasting 30 seconds and categorized into 10 different genres, with 100 examples in each genre. These genres encompass blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock.

To ensure consistency and maximize the number of data samples available for training and testing, we further divided the music clips into 10-second excerpts. This resulted in obtaining

**Table 3.1** Datasets

| Dataset Name | Samples | Genres, Styles |
|---|---|---|
| Bollywood | N=1500, duration=20 secs | Multi language,Mixed instruments<br>Varied genres including pop, rock, classical, instrumental |
| Indian Devotional | N=500, duration=60 secs | Indian Instrumental, Lyrical- Punjabi and Hindi, Sufi |
| Western Instrumental | N=1500, duration=20 secs | All genres including Classical, Rock, Hip-hop, Blues, Jazz |
| GTZAN | N=1000, duration=30 secs | 10 genres including blues, classical, country,<br>Vdisco, hip-hop, jazz, metal, pop, reggae, and rock |

a total of 3000 samples for each of the four datasets, maintaining a uniform number of data samples across all datasets. Table 3.1 summarize the datasets.

### 3.2.3 Feature Extraction

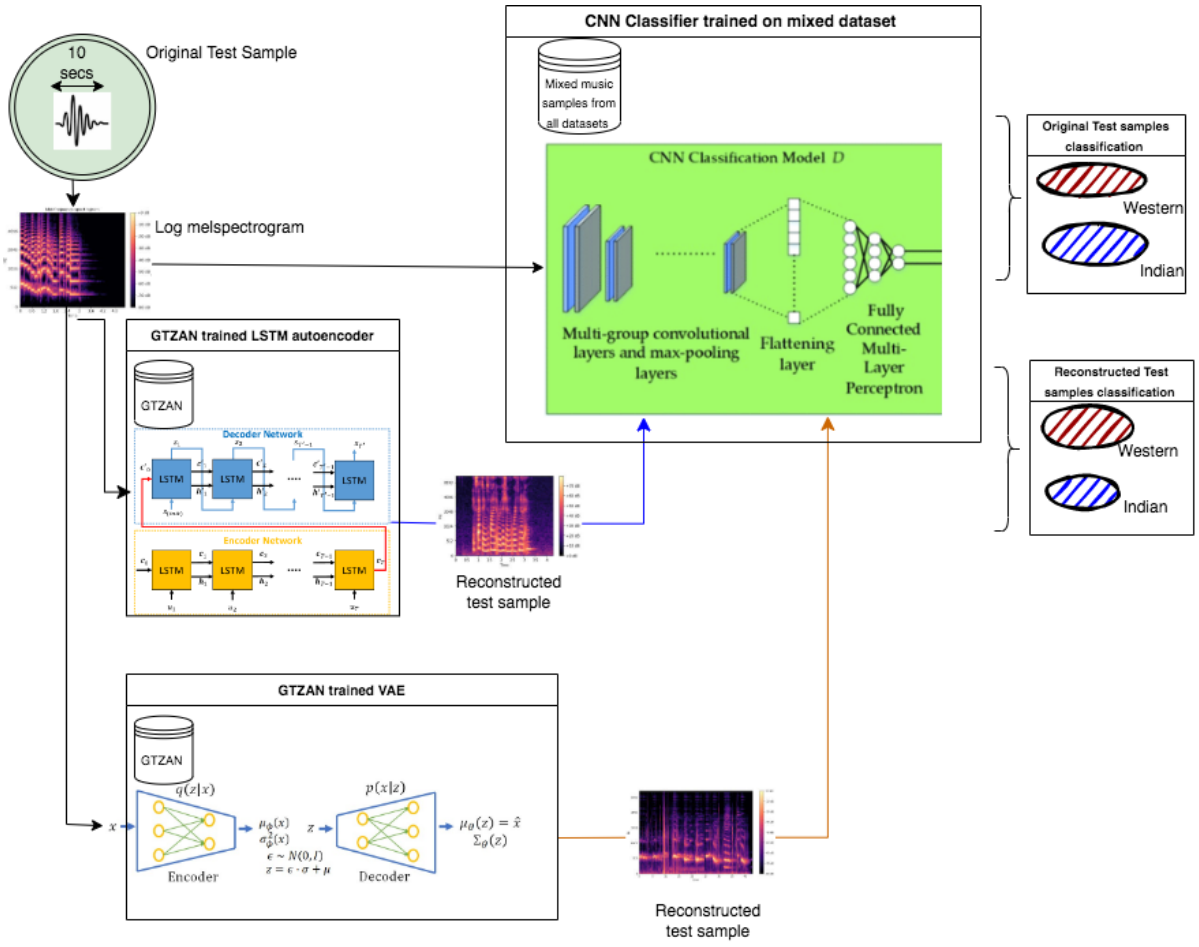
For feature extraction, we utilized the log melspectrogram, which is a representation of audio signals in the frequency domain. The mel spectrogram captures the distribution of energy across different frequency bands. We experimented with various mel bins and found that using 96 mel bins provided the most representative results for capturing music features. The input feature was the log mel spectrogram, which enhances the perceptual relevance of the mel spectrogram. We used a window size of 100 ms and a hop size of 50 ms, which determines the temporal resolution of the spectrogram. These settings allowed us to accurately capture the temporal dynamics of the music signals and extract meaningful features for further analysis and processing

### 3.2.4 Method

The autoencoder models were trained using the respective datasets to learn compressed representations of the music features. The figure 3.2 illustrates the pipeline for this experimentation.

We utilized a combination of objective functions, such as mean squared error (MSE) loss and Kullback-Leibler (KL) divergence, to optimize the training process. After training, we reconstructed the original music stimuli using the trained autoencoders.

To assess the impact of the autoencoders on music classification, we employed a vanilla CNN classifier as the baseline model to evaluate the efficacy of our different autoencoders and therefore help us select the autoencoder which can most reliably reconstruct the input stimuli. The selection basis of the autoencoder architecture assured that the underlying structure of the best performing model is the one where music stimuli is learnt relatively better than any other models. The vanilla CNN classifier was trained on a custom mixed dataset and it was ensured that this dataset is not seen by any of the autoencoder during their training phase.




**Schematic of classification task Pipeline**

Illustrates the flow for classifying test samples in Indian & Western music using original test stimulus and reconstructed versions of the test stimulus via trained CNN classifier

**Figure 3.2** Classification Task Pipeline

The dataset comprised of 10 secs music excerpts representing songs from all the 4 datasets. The dataset used for training and validation had in total 2000 songs with duration of each song as 10 secs. The test set contained 1000 samples representative of all datasets and unseen by any autoencoder.

To ensure robustness in training and validating the Autoencoder (AE), we implemented a 5-fold cross-validation strategy, dividing the data into training and validation sets. This approach introduces a level of randomness that is averaged across the folds, exemplified by how samples from different time segments may be allocated to either set. While song excerpts may display some similarity, they are not identical; however, this slight variation is inconsequential, particularly given the potential randomness introduced by cross-validation and the non-continuous nature of the excerpts. For the test dataset, it encompasses samples not only from the GTZAN dataset but also from mixed datasets. While there may be instances of samples from the same song in the test dataset, they originate from different excerpts. Moreover, the dataset includes a majority of samples from a custom dataset unseen by the GTZAN-based AE. Consequently, the impact of having excerpts from the same song in the test dataset is minimized by the inclusion of various other unseen samples. Given that the primary objective of training the AE is to learn representations of a predominant dataset, rather than optimizing for discrimination or classification efficiency, the implications of the dataset split are not significant.

## 3.3 Results

The overall results of the various autoencoders are shown in Table 3.2 . For the best performing GTZAN-trained VAE, we also manually inspected limited reconstructed samples as illustrated in figure 3.4. The reconstructed samples are highly smooth versions of the original mel spectrogram and give an indication that while it is not able to fully reconstruct the spectrogram, it is able to reliably learn the energy gradients albeit quite smooth.

In general, the accuracy of classification for reconstructed stimuli shows a decrease, notably with instances of Indian music being misclassified as Western. The average accuracies in some instances are below chance level as shown in Table 3.2. The confusion matrix for GTZAN based AE as shown in figure 3.3 illustrates the accuracies for Indian and Western music. As seen in the confusion matrix, the accuracy for western music is far greater (62.5%) than Indian music identification(42%) with average accuracy close to 52%. Several factors could contribute to this observed discrepancy. Firstly, the intricate nuances inherent in Indian music might require more intricate model architectures and a larger dataset volume to achieve precise classification. Moreover, the pervasive influence of Western music on Indian musical traditions, particularly notable in Bollywood songs characterized by Western musical elements, might also influence mis-classifications. However, given our primary focus on developing a straightforward generative self-supervised model for fMRI analysis, we chose to limit our investigation of this phenomenon.

**Table 3.2** Average Classification Accuracies on reconstructred stimuli; Trained on custom datasets and tested on an unseen data by any model using our baseline CNN classifier.

| **Training Dataset** | **LSTM based Autoencoder** | **Variational Autoencoder** |
|---|---|---|
| Bollywood Music | 0.32 | 0.42 |
| Western Custom | 0.37 | 0.46 |
| **GTZAN** | 0.42 | **0.52** |
| Indian Devotional | 0.45 | 0.47 |

Nevertheless, this discovery presents an intriguing avenue for further exploration, potentially warranting a separate dedicated study.

The results revealed that the VAE-based autoencoders retained the structure of the music stimuli more effectively compared to the LSTM-based autoencoders. Furthermore, the autoencoder trained on the GTZAN dataset showed the highest classification accuracy of 52% when using the reconstructed music stimuli. This indicates that the GTZAN-trained autoencoder was successful in preserving relevant features for music classification.

## 3.4 Conclusion

In this chapter, we trained eight autoencoder models on four different custom datasets and evaluated their impact on music classification accuracy. The VAE-based autoencoders demonstrated better preservation of music structure compared to the LSTM-based autoencoders. Additionally, the autoencoder trained on the GTZAN dataset yielded the highest classification accuracy when using the reconstructed music stimuli. These findings highlight the potential of autoencoders, particularly VAE-based architectures, for preserving musical features and improving music classification accuracy. It is therefore the selection of our computational model to perform the core study.

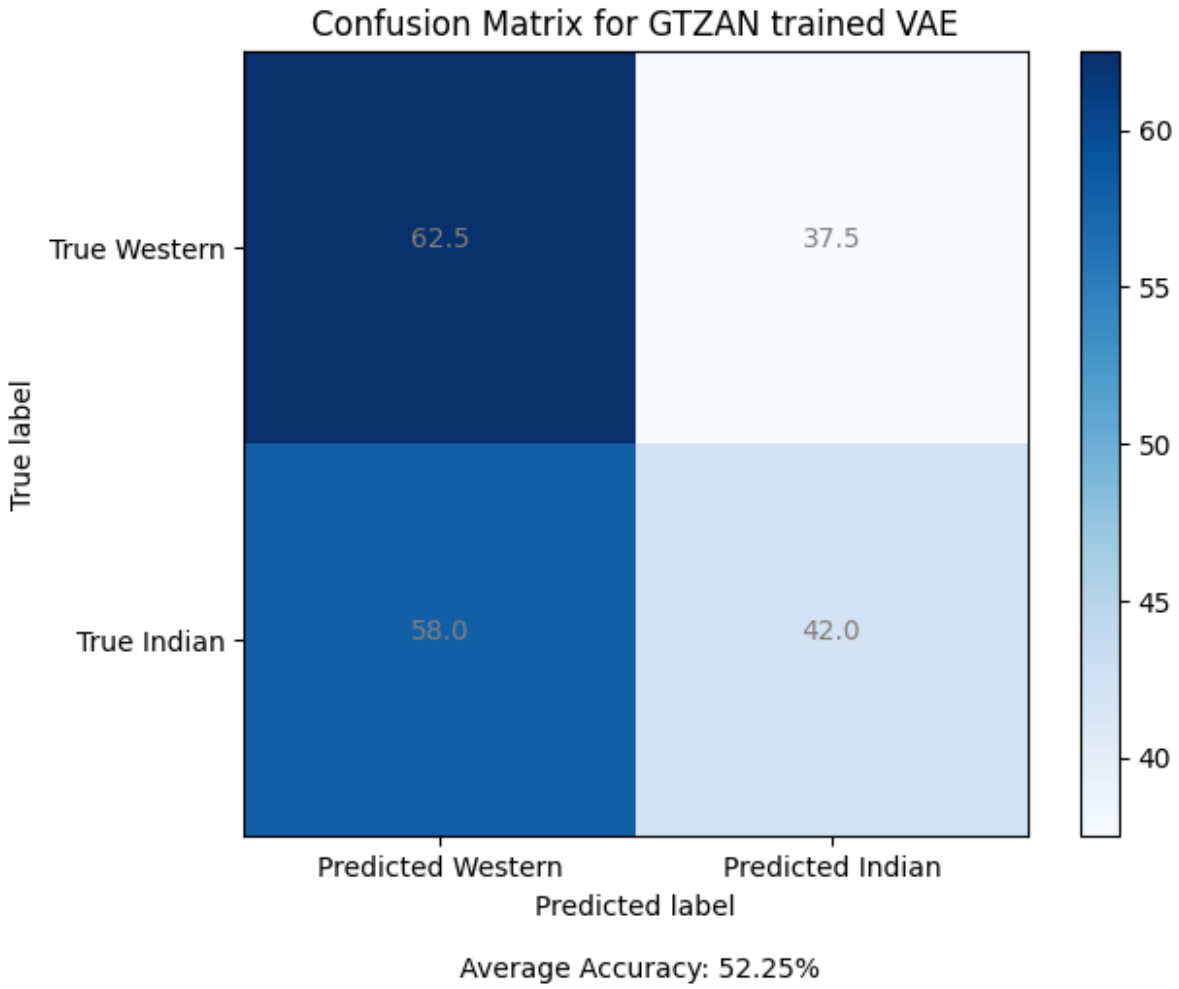



**Figure 3.3** Confusion Matrix of GTZAN trained VAE classifier

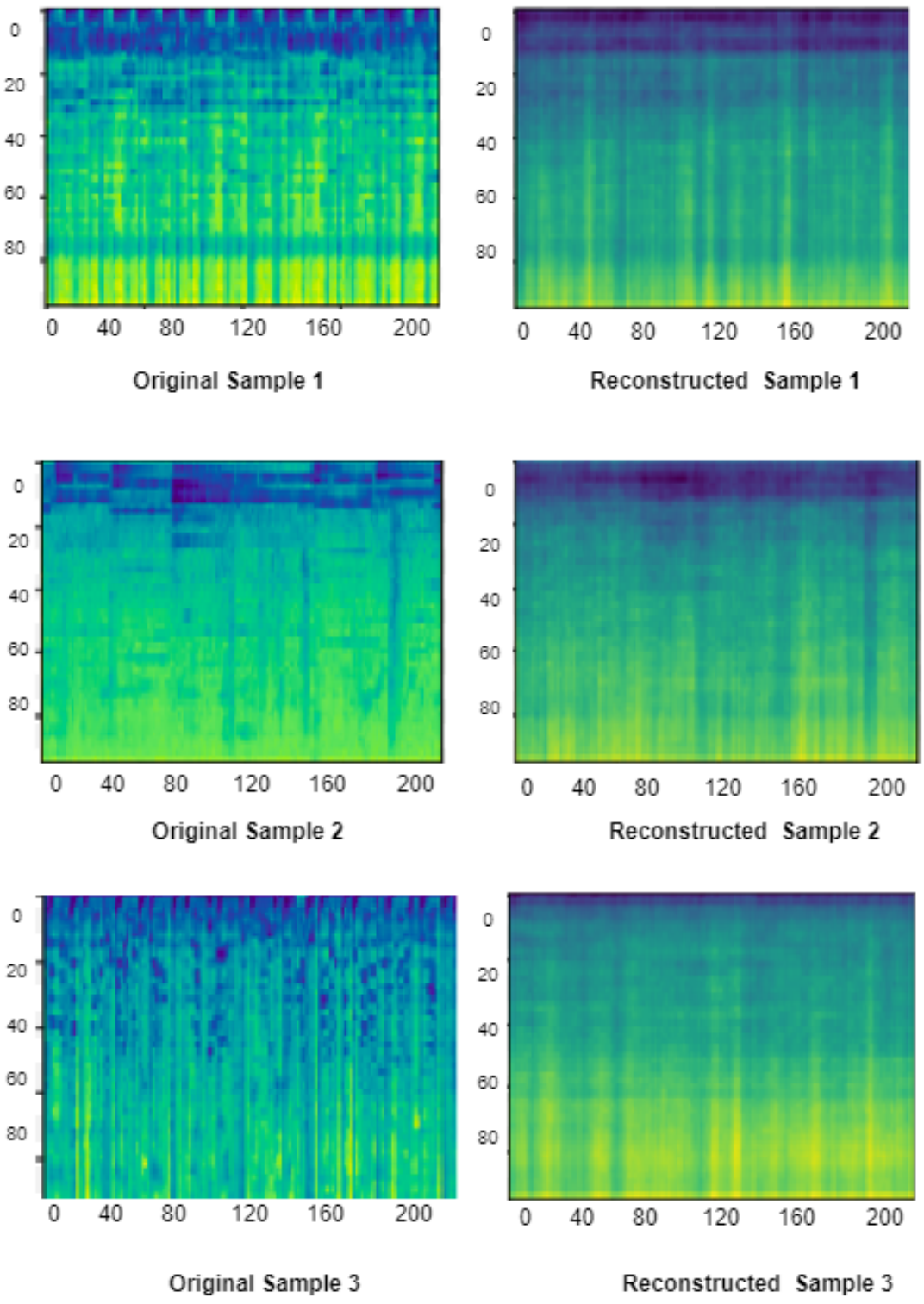



**Figure 3.4** Inspection of reconstructed music stimuli from GTZAN trained VAE

*Chapter 4*

# Modelling the implicit music representation in brain using DNN

## 4.1 Introduction

In this study, we aimed to investigate the potential of unsupervised deep neural networks, specifically the variational auto-encoder (VAE), to learn latent music representations and examine brain activity patterns in a continuous music listening task. Our hypothesis is that generative computational models with a probabilistic basis, such as VAE, are better suited for capturing the complexity of neural anatomy. Previous studies by Schaefer et al. [54] and Gómez-Herrero et al. [21] have shown that using probabilistic generative models allowed for a more accurate and comprehensive understanding of brain networks and music processing. To examine the correspondence between the computational model activity patterns and fMRI patterns, we utilized Representation Similarity Analysis (RSA), a method previously employed in similar studies [33], albeit in the visual modality. In addition, owing to the differences in the way musicians and non-musicians perceive and process music [1, 46], we also examine differences in musicians and non-musicians.

## 4.2 Materials and Method

### 4.2.1 Dataset

In this study, we used the GTZAN Western music dataset [60] which contains 1000 music excerpts, each lasting 30 seconds and divided into 10 different categories with 100 examples in each. The genres include blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. To process the audio data, we employed log mel-spectrograms to emulate the frequency representation of human auditory perception [42, 61]. Specifically, we set the window length to 100 ms and the hop size to 50 ms, and utilized 96 mel frequency bins based on established

heuristics in the literature [**?**]. This resulted in the creation of input feature maps which were subsequently used in the DNN model.

For the brain data, we utilized the dataset previously used [1, 44, 9], which included fMRI recordings from 20 musicians and 18 non-musicians while they were engaged in a continuous listening task. The task involved listening to three 8-minute long musical stimuli, namely, "Stream of Consciousness" by Dream Theater [15], "Adios Nonino" by Astor Piazzolla [3], and "Rite of Spring" by Stravinsky [56]. The two participant groups were matched for gender, age distribution, cognitive abilities, and socioeconomic status to ensure comparability.

Participants' brain responses were acquired while listening to the music delivered via MR-compatible insert earphones while keeping their eyes open. Thirty-three oblique slices (FoV: 192mm x 192 mm, 64 x 64 matrix, interslice skip = 0mm) were acquired every 2 sec, with echo time = 32ms and voxel size = 2 x 2 x 2 mm3 using a single-shot gradient echo-planar imaging (EPI) sequence, providing whole-brain coverage for each participant. The data was preprocessed in the same manner as previous studies, using identical steps including preprocessing in Matlab using SPM8 and VBM5. Normalization to MNI segmented tissue template was carried out. After regressing the components related to head movement, spline interpolation and temporal smoothing were applied.

### 4.2.2 Method

We used the best performing encoder architecture in a classification task based on our earlier study. The selected computational model is a 5-layer convolutional neural network (CNN) derived from the AlexNet architecture [34] , implemented as a variational autoencoder (VAE), to emulate the encoding of neural responses to music stimuli. The hidden layer activations of the DNN encoder were obtained after presenting the three test stimuli.

We applied Representation Similarity Analysis (RSA) [33] to compare the representational structures of the DNN model layers with those of the response patterns in selected Regions of Interest (ROI). In RSA, neural or model representations are quantified as n $\times$ n representational dissimilarity matrices (RDMs), with elements representing the dissimilarity between pairs of stimuli. Comparing the overlap between model and neural RDMs can reveal how well a model explains response patterns in a particular brain region. To investigate this, we created one target RDM for each group (averaged across musician and non-musician subjects) and five candidate RDMs corresponding to each DNN encoder layer. We then correlated the upper triangular parts of the target RDM with the candidate RDMs for each group, using Spearman correlation [43]. Benjamin Hochberg correction [6] was applied to correct for multiple comparisons. We restricted our analysis to the primary auditory cortex region, specifically STG and HG. Figure 4.1 illustrates the method and transformations applied. Figure 4.2 provides another schematic to illustrate the RSA process.

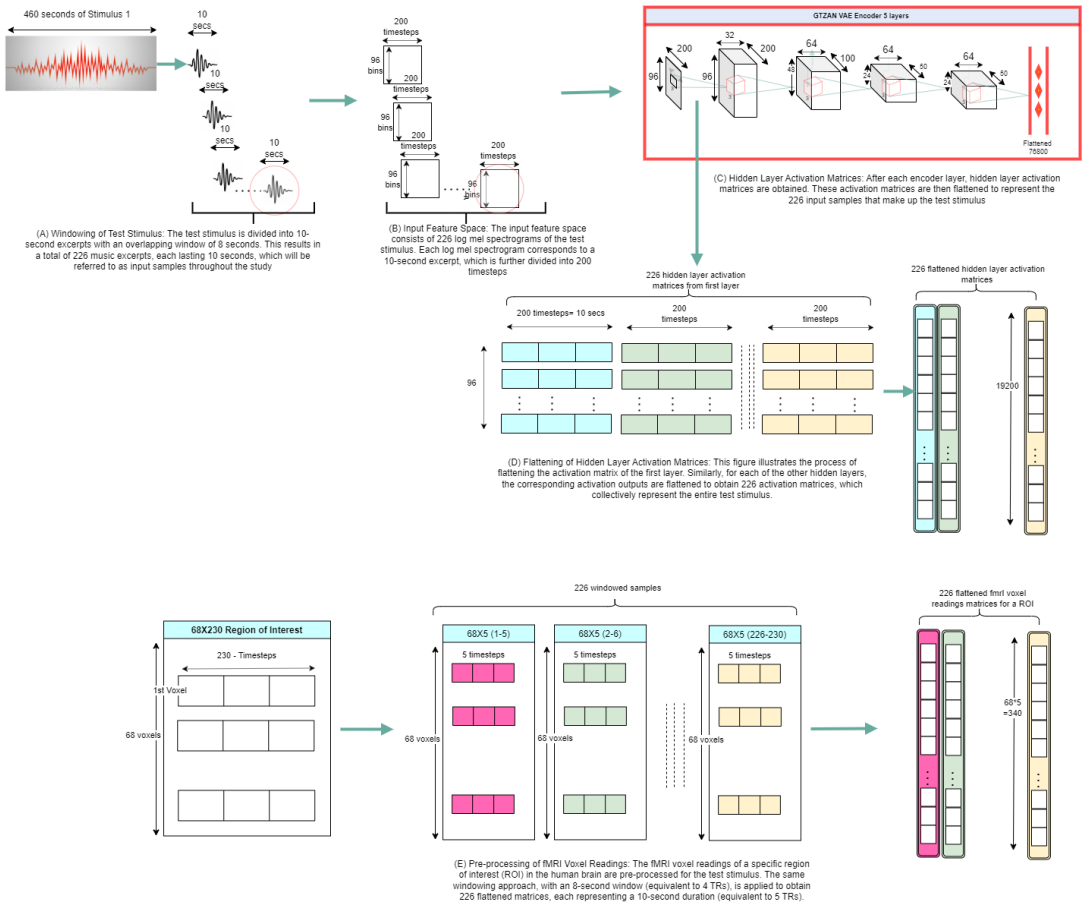



**Figure 4.1** Schematic of pipeline to extract and process inputs and outputs for RSA

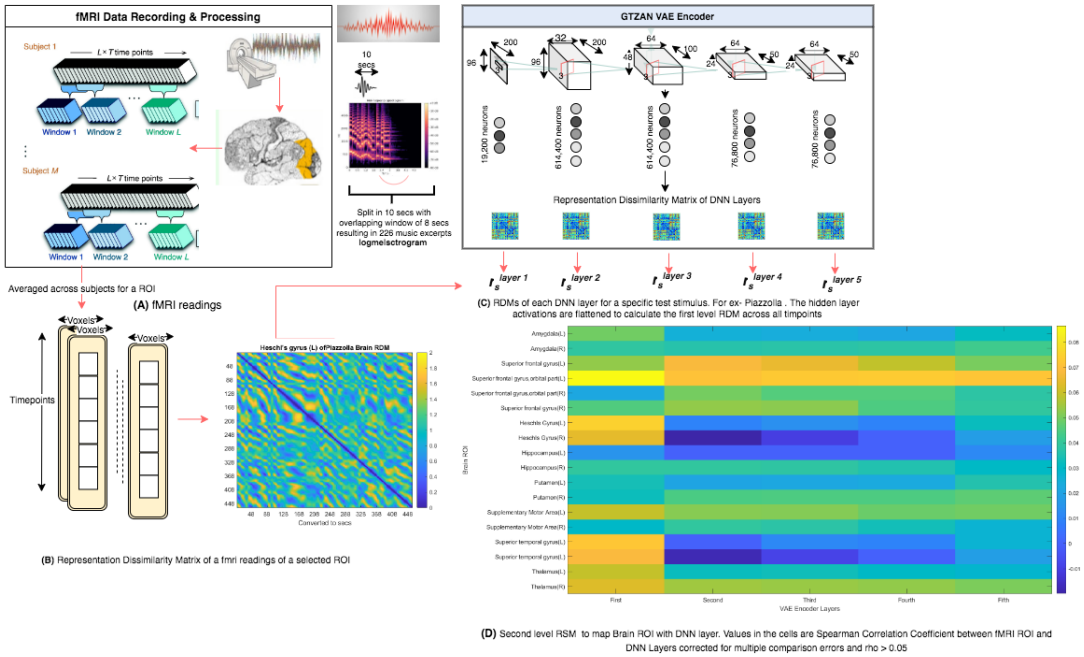



**Figure 4.2** Schematic to illustrate the RSA

## 4.3 Results

The results of the RSA analysis indicate that the superior temporal gyrus (STG) brain region displayed correlation with the five VAE encoder layers for both musician and non-musician groups. This is consistent with the known auditory processing mechanisms of the brain [29, 20]. The Heschl's gyrus (HG) also exhibited significant correlations with the first and fifth DNN layers, with lower but still significant correlation values observed in the second, third, and fourth hidden layers ($p < 0.001$) [20].

We conducted an exploratory analysis of brain regions that showed significant correlations with early or later DNN layers. Refer Figure 4.3 for non-musician group, Figure 4.4 for musicians group, and Figure 4.5 for a combined view, we found the following:

- Supplementary Motor Area (SMA) showed more correlations for the Stravinsky stimulus across musicians and non-musicians for all DNN layers.

- This was in contrast to the other two music stimuli, which showed very low correlations with the SMA.

For the Piazzolla stimulus,

- Superior frontal gyrus orbital part (L) showed higher correlations for both musicians and non-musicians compared to the other two stimulus types.

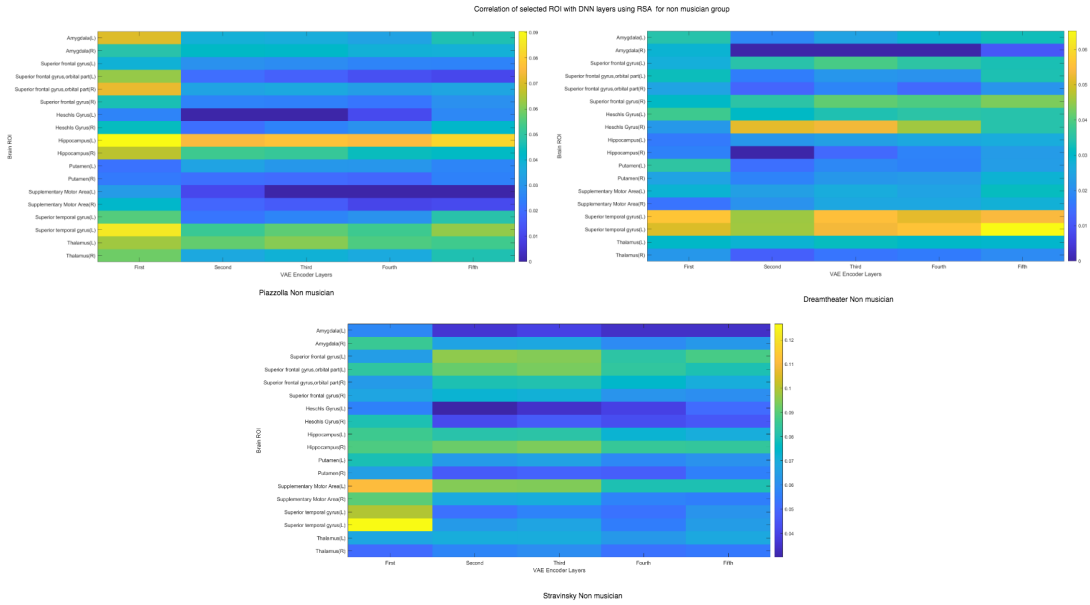



**Figure 4.3** Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for non musician group

- The correlation is highest for the superior frontal gyrus (L) with the first layer for both musicians and non-musicians. However, the superior frontal gyrus (L) shows high correlations with the second, third, and fourth DNN encoder layers in the case of musicians.
- This is in contrast to non-musicians, where the correlation gradient gradually fades with increasing layer. This indicates a possible difference in the way that musicians and non-musicians process Piazzolla's music.

Overall, the patterns of correlations across musicians and non-musicians did not show any aberrations, except for a diminished correlation gradient in Piazzolla for musicians. The results of the RSA analysis suggest that the STG and HG are involved in the processing of music, regardless of musical training. This is consistent with previous research that has shown these regions to be involved in auditory processing [29, 20].

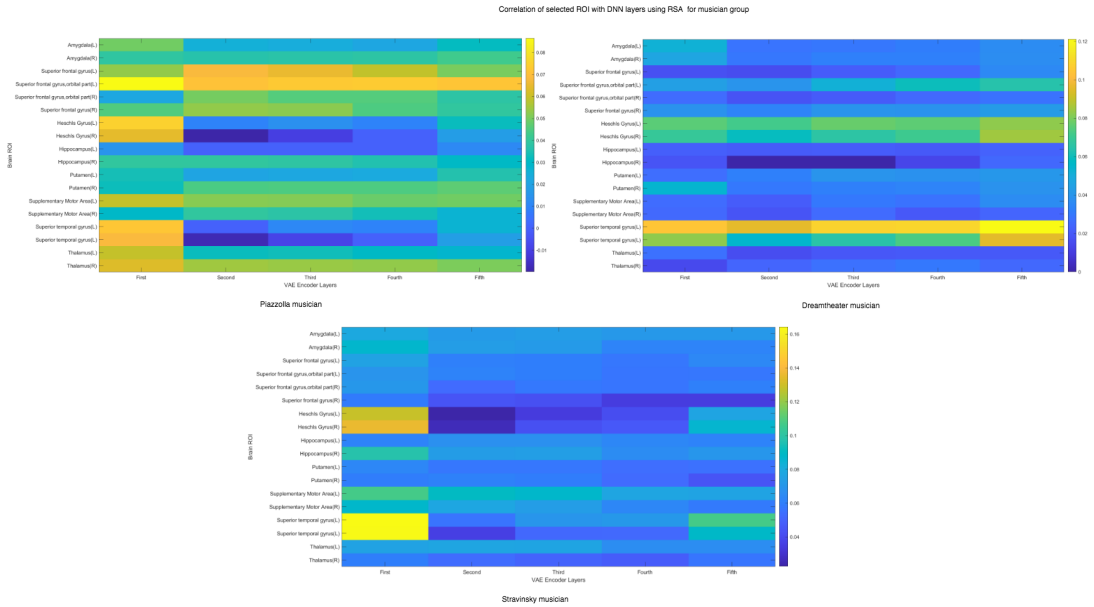



**Figure 4.4** Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for musician group

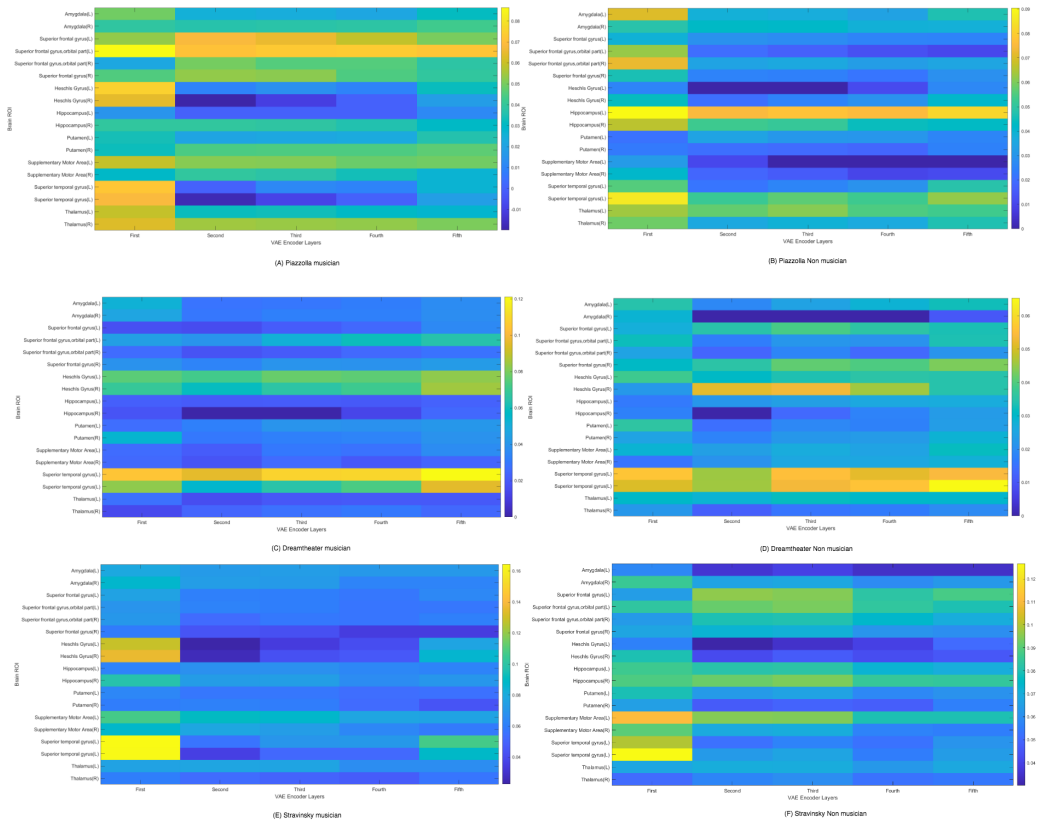



**Figure 4.5** Combined view of Representational Similarity Matrix obtained using RSA- Correlation of Brain selected ROI and VAE Encoder layer activations for both musician and non-musicians

*Chapter 5*

# Conclusion and Future Work

In this study, we investigated the brain encodings of music latent representations using a variational autoencoder (VAE) architecture. Our findings suggest that the VAE architecture, specifically the DNN encoder layers, can provide a powerful computational model for explaining the brain encodings of music latent representations. Our findings support the involvement of primary auditory cortex regions such as STG, Heschl's gyrus in processing musical stimuli. Our study lays the groundwork for further research in developing computational models inspired by human cognition to better understand music processing.

## 5.1 Discussion

Our study has several implications for future research. First, our findings suggest that the VAE architecture can be used to develop more sophisticated computational models of music processing. These models could be used to investigate the neural basis of music perception, cognition, and production. Second, our findings suggest that the DNN encoder layers of the VAE architecture can be used to identify the brain regions that are involved in processing different aspects of music. This information could be used to develop new diagnostic tools for identifying and treating music-related disorders. The finding that the SMA showed more correlations for Stravinsky stimulus across musicians and non-musicians suggests that this region may be involved in processing this type of music. Stravinsky's music is often described as complex and challenging, so it is possible that the SMA is involved in processing this type of music.

The finding that the superior frontal gyrus orbital part (L) showed higher correlations for Piazzolla for both musicians and non-musicians suggests that this region may be involved in processing this type of music. Piazzolla's music is often described as emotional and expressive, so it is possible that the superior frontal gyrus orbital part (L) is involved in processing these aspects of the music.

The overall results of the RSA analysis suggest that the brain regions involved in processing music are similar for musicians and non-musicians. However, there are some differences in the

way that these regions are activated by different types of music. These findings provide insights into the neural basis of music perception and processing

## 5.2 Limitations

- Currently the group analysis methodology has averaged the fmri activations for muscians and non-musicians respectively. While this may be a good approximation for non musicians, for musicians this can result in losing out some important information and therefore a more elaborate and dilligent procedure may be employed.

## 5.3 Future Work

There are several directions for future research that could build on the findings of this study.

- First, we could use searchlight analysis to further explore the brain regions involved in processing music latent representations. Searchlight analysis would allow us to identify brain regions that are consistently active during the processing of music latent representations, even when the specific musical stimuli vary.

- We could use graph neural networks to discover the functional connectivity of the brain regions involved in processing music latent representations. This would allow us to investigate how different brain regions interact with each other during the processing of music.

- Based on exploratory analysis findings for extended regions of interest in brain, it suggests that the SMA may play a role in processing Stravinsky stimulus, and that this effect is independent of musical training. We could conduct a deeper study to understand the impact of type of stimulus on musician and non-musician brains. This would allow us to investigate whether there are any distinctions in the brain regions that are involved in processing music latent representations, depending on whether the listener is a musician or not. Fourth, we could extend this study to other cultural settings with an appropriate design choice of stimulus. This would allow us to investigate whether there are any cultural differences in the brain encodings of music latent representations.

Finally, with the advent of music transformer, we could study more sophisticated computational models and further build on generative models. This would allow us to develop even more powerful computational models of music processing.

# Related Publications

1. Ravinder Singh, Rajat Agarwal, Petri Toiviainen, Vinoo Alluri. (2023) B254: Modeling Implicit Musical Representation in Brain with Deep Neural Networks. The 16th International Conference on Brain Informatics (BI 2023)

# Bibliography


[1] V. Alluri, P. Toiviainen, I. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico. Music expertise shapes audiovisual temporal integration windows for speech, sound and music. *Brain and Cognition*, 111:203–213, 2017. 23, 24

[2] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4):3677–3689, 2012. 8, 9

[3] Astor Piazzolla. Le Grand Tango for Cello and Piano. YouTube video, 1986. 24

[4] P. A. Bandettini, E. X. Wong, R. S. Hinks, R. S. Tikofsky, and J. S. Hyde. Time course epi of human brain function during task activation. *Magnetic Resonance in Medicine*, 25(2):390–397, 1992. 7

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 9, 10

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. 24

[7] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995. 7

[8] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *In Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1159–1166, 2012. 10

[9] I. Burunat, V. Alluri, P. Toiviainen, J. Numminen, and E. Brattico. Automatic decoding of musical structure from continuous fmri reveals rhythm and tonality processing in separate brain networks. *Cortex*, 92:22–37, 2017. 24

[10] A. Chen and M. Riesenhuber. The view from above: Applications of representation similarity analysis to brain imaging data. *Journal of Neuroscience Methods*, 171(2):176–185, 2008. 4, 12, 13

[11] K. Choi, G. Fazekas, M. Sandler, and K. Cho. Text-based lstm networks for automatic music composition. In *11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 120–126. Springer, 2016. 9

[12] F. Chollet et al. Keras. *https://github.com/keras-team/keras*, 2015. 16

[13] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4

[14] C. Donahue, A. Roberts, C. Shi, H. Zhang, Z. C. Lipton, and Y. Li. Large-scale symbolic music generation with iterative refinement and long-term dependency. *arXiv preprint arXiv:1912.01219*, 2019. 11

[15] Dream Theater. Stream of Consciousness by Dream Theater. YouTube video, 2003. 24

[16] C. Dugas, Y. Bengio, and F. Bélisle. Incorporating second-order functional knowledge for better option pricing. In *Advances in neural information processing systems*, pages 472–479, 2001. 16

[17] D. Eck and J. Schmidhuber. Learning the long-term structure of the blues. *Neural computation*, 14(8):1959–1972, 2002. 11

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587. IEEE, 2014. 4

[19] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 14

[20] T. D. Griffiths, I. S. Johnsrude, J. L. Dean, G. G. Green, T. TDGRIFFITHS, I. Johnsrude, J. Dean, and G. Green. Functional mapping of the human auditory cortex: fmri investigation of a patient with auditory agnosia from trauma to the superior temporal gyrus. *Cortex*, 113:43–60, 2019. 26, 27

[21] G. Gómez-Herrero and G. Peeters. Generative deep learning models for music: a survey. *Journal of New Music Research*, 50(4):361–394, 2021. 4, 23

[22] U. Güçlü et al. Brains on beats. *arXiv preprint arXiv:1606.02627*, 2016. 1, 2, 14

[23] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 9, 14

[24] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 10

[25] Y.-H. Huang, Y.-A. Wu, L.-Y. Chen, and Y.-H. Su. Music generation using deep learning. *ACM Computing Surveys (CSUR)*, 51(2):35, 2018. 9

[26] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 15

[27] P. Janata. The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex*, 19(11):2579–2594, 2009. 1

[28] P. Janata and S. T. Grafton. Dynamic processing of musical structure: A comparison of on-line and off-line segmentation. *Frontiers in Human Neuroscience*, 6:94, 2012. 8

[29] J. P. Jones and D. E. Callan. Neural correlates of melody recognition: The neural correlates of melodic structure and tonality. *Cognitive Brain Research*, 16(2):361–372, 2003. 26, 27

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 16

[31] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 11, 14

[32] D. P. Kingma and M. Welling. Introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019. 4

[33] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1–28, 2008. 23, 24

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. 4, 24

[35] P. K. Kuhl et al. Brain responses to words in 2-month-old infants. *Developmental Science*, 16(3):352–359, 2013. 1

[36] E. W. Large and J. F. Kolen. The dynamics of musical steady states. *Journal of New Music Research*, 28(1):1–24, 1999. 6

[37] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. In *Nature*, volume 521, pages 436–444. Nature Publishing Group, 2015. 4, 9

[38] D. J. Levitin. *The Changing Mind: A Neuroscientist's Guide to Ageing Well*. Penguin, 2019. 4

[39] D. J. Levitin and V. Menon. Neural correlates of musical syntax processing. *Nature Neuroscience*, 8(3):382–387, 2005. 8

[40] N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008. 7

[41] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 4

[42] B. McFee, C. Raffel, D. Liang, and D. P. Ellis. Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25, 2018. 23

[43] J. A. Mumford, J.-B. Poline, and R. A. Poldrack. Statistical tests in Functional Imaging. *Academic Press*, 2(6):565–582, 2015. 24

[44] P. Niranjan, A. Saini, A. Dutta, and A. Sreedhar. Neural basis of musical sight-reading in violinists: an fmri study. *Brain Imaging and Behavior*, 13(3):831–845, 2019. 24

[45] A. D. Patel. Music, language, and the brain. *Oxford University Press*, 2008. 6

[46] A. D. Patel and J. R. Iversen. Musical rhythm, linguistic rhythm, and human evolution. *Music Perception*, 32(1):10–25, 2014. 23

[47] M. T. Pearce. Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, 1252(1):146–160, 2012. 8

[48] I. Peretz. The brain basis of music processing: evidence from neuropsychology. *Annals of the New York Academy of Sciences*, 930(1):190–201, 2001. 9

[49] I. Peretz and R. J. Zatorre. *The cognitive neuroscience of music*. Oxford University Press, 2003. 1

[50] I. Peretz and R. J. Zatorre. Brain organization for music processing. *Annual review of psychology*, 56:89–114, 2005. 5, 8, 9

[51] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. Hierarchical music generation with long-term structure. *arXiv preprint arXiv:1803.05428*, 2018. 11, 14

[52] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1985. 4

[53] R. Santoro, F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. A simple model of recognition predicts experimental behavioral and neural data in the ffa. *PLOS Computational Biology*, 10(10):e1003883, 2014. 4, 12, 13

[54] R. Schaefer and E. Gómez. Generative models for generating music. *Journal of New Music Research*, 46(1):25–41, 2017. 23

[55] D. Seibert, R. Santoro, A. Chen, Q. Wang, and M. Riesenhuber. Estimating representational similarity matrices. *PLoS ONE*, 11(6):e0157385, 2016. 4, 12, 13

[56] I. Stravinsky, L. Maazel, and T. C. Orchestra. The rite of spring. Internet Archive, 1913. 24

[57] B. L. Sturm. Music transcription modelling and composition using deep learning. In *Deep learning for music*, pages 37–52. Springer, 2016. 9

[58] W. Tilley and S. Kumar. Spatial and temporal dynamics of neural processing in the human auditory cortex. *Music perception*, 28(2):167–180, 2011. 8, 9

[59] W. Tilley and S. Kumar. *fMRI techniques for studying music perception and cognition*, pages 261–276. Routledge, 2018. 8, 9

[60] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002. 16, 23

[61] A. Van Den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. *Advances in Neural Information Processing Systems*, 26:2643–2651, 2013. 23

[62] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. 4, 5, 12, 13

[63] R. J. Zatorre and J. McGill. Music, the food of neuroscience? *Nature*, 434(7031):312–315, 2005. 1

[64] R. J. Zatorre and I. Peretz. Neural correlates of music perception. *Annals of the New York Academy of Sciences*, 930(1):179–192, 2002. 8