A Multi-Modal AI Framework for Emotion-Aware Music Recommendation and Mental Health Signal Detection through User Listening Behaviour

MSc Research Practicum
Data Analytics

Hariramakrrishnan Ramachandran
Student ID: x23413948

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Hariramakrrishnan Ramchandran
Student ID:	x23413948
Programme:	MSc in Data Analytics **Year:** 2025
Module:	Research Practicum-Part II
Supervisor:	Jorge Basilio
Submission Due Date:	11/12/2025
Project Title:	A Multi-Modal AI Framework for Emotion-Aware Music Recommendation and Mental Health Signal Detection through User Listening Behaviour

Word Count: 9240 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Hariramakrrishnan Ramachandran

Date: 10/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Multi-Modal AI Framework for Emotion-Aware Music Recommendation and Mental Health Signal Detection through User Listening Behaviour

Hariramakrrishnan Ramachandran
x23413948

Abstract

The study demonstrates a multi-modal AI system of emotion-aware music suggestion and mental health signal recognition that incorporates acoustic, lyrical and behavioral aspects of user engagement. The system consists of machine and deep learning technologies to identify audio characteristics, the mood of the lyrics and the user listening history, and can make suggestions that are adaptive to the emotional context. A sequential LSTM model represents behavioral drift, whereas a fusion layer integrates multi modular information into an understandable decision-making process. SHAP and gradient-based analysis are explainable AI techniques that can be used to improve the transparency and trust of the model. The construct illustrates the way user behavior modelling and emotion recognition can be used together to inform intelligent context-sensitive music systems. In addition to personalization, the paper addresses the possibility to connect emotion-aware recommendation with signs of mental health, which also leads to the creation of empathetic, morally responsible, and user-oriented AI in online music services.

1 Introduction

1.1 Background

Music is an inherent medium of emotional expression and control, and the development of artificial intelligence (AI) has changed the perception of the understanding of emotional stimuli in the digital world. Conventional music recommendation algorithms depend to a great extent on the user activity or popularity dynamics, and do not provide a significant degree of personalization in relation to the emotional intent (Ricci, Rokach and Shapira, 2015). The increasing sphere of affective computing has allowed machines respond more efficiently to human feelings, thus making it possible to apply affective computing to the music industry (Picard, 1997).

To determine emotional states, emotion-aware music recommendation is expected to combine audio attribute (tempo, loudness, rhythm) (Yang and Chen, 2012), lyrical sentiment (Malheiro et al., 2016), and user behaviour (Panda, Malheiro and Paiva, 2018) cues. These heterogeneous data sources can be combined into coherent models that can comprehend mood dynamics in a more holistic way due to the recent advances in multi-modal learning (Ngiam et al., 2011; Baltrusaitis, Ahuja and Morency, 2019). Not only do these systems increase the

personalization effect, but they also play a vital role in the emotional aspect of the experience by matching music suggestions with the current psychological conditions of the users and offers an empathetic and context-specific listening experience.

1.2 Research Problem and Motivation

The existing music recommendations are either based on collaborative filtering or content-based approaches, which are centered on the history of listening, the similarity of songs, or even the trends in popularity (Ricci, Rokach and Shapira, 2015). Although these methods are useful in forecasting the preferences of users, they do not take into consideration the emotional situation of the user when making selections and thus end up with stale and non-adaptive playlists. Music consumption has a strong emotional nature; when an individual hears a song, he/she tends to choose the one that corresponds or is capable of changing his/her mood, and emotional intelligence is critical to personalization (Yang and Chen, 2012). Nonetheless, the majority of the currently available systems only analyse audio, lyrics, or user behaviour, and thus cannot comprehend complex affective patterns (Delbouys et al., 2018).

This limitation is the reason why this research was conducted to create a multi-modal AI structure that will allow considering acoustic, lyrical and behavioral cues to recognize emotions and suggest adaptive music. The proposed system should produce interpretable and emotionally sensible recommendations with the help of deep learning as a sequential behavioral modelling model (Panda, Malheiro and Paiva, 2018) and explainable AI models to make the proposed system more transparent like SHAP (Lundberg and Lee, 2017). This not only goes forward to develop the precision of emotion-aware systems but it also leads to better mental health by providing individualized and empathetic musical experiences.

1.3 Research Aim and Objectives

This research will attempt to create a multi-modal AI platform incorporating acoustic signals, lyrical mood, and user active listening pattern in order to identify emotional patterns and output emotion sensitive music suggestions. The study attempts to go beyond the conventional systems of recommendations, and through behavioral drift analysis, the study attempts to capture real time emotional nuances and through explainable model outputs. This may allow music recommendation to be more individualized and interpretable, a more human interaction with AI (Baltrusaitis, Ahuja and Morency, 2019). To fulfil this purpose, the framework integrates classical machine learning, deep learning, and explainable AI methods, on various modalities. The system hopes to identify the dynamically changing emotional state of the listener by matching audio cues, lyrical semantics, and sequential data of behavioural reactions to propose appropriate songs (Ngiam et al., 2011). Affective computing principles direct the model design to be emphatic and contextually sensitive in automated recommendations (Picard, 1997).

Objectives:

1. To preprocess and combine multimodal data that incorporates a combination of acoustic, lyrical and behavioural features.
2. To model and test emotion recognition models of every modality.
3. To apply an emotion tracking LSTM based behavioural drift predictor.
4. To develop a fusion-based recommendation model, which can adjust to changing emotions.
5. To explain model choices with explainable AI models (SHAP and Gradients).

1.4 Research question

This paper examines the possibility of incorporating emotional knowledge into intelligent music recommendation algorithms by using multi-modal learning. The main question under discussion is:

How should a multi-modal AI model which incorporates acoustic attributes, lyrical emotion, and user listening behaviour be constructed to identify emotion pattern and create emotion sensitive music recommendation that responds to real-time emotional conditions in users?

This question is indicative of the necessity of emotionally adaptive systems converting audio, textual, and behavioural cues to better perceive the affective aspect of the need (Yang and Chen, 2012; Aljanaki, Yang and Soleymani, 2017). It also seeks to make certain interpretability and transparency in prediction with the help of explainable AI methods (Lundberg and Lee, 2017).

1.5 Limitations of the study

Despite the fact that the suggested multi-modal framework is effective in terms of integrating audio, lyrical, and behavioural modalities, it functions within the framework of research constraints. The datasets are a self-selected sample of interaction between the user and musical content, and this might not reflect the entire range of the emotion of people in real listening situations. The dynamics of emotion were established indirectly, based on the trend of user behaviour instead of using direct physiological or situational indicators. In addition, the models were tuned to be interpretable and computationally efficient and this can be a constraint to large-scale or live-streaming scenarios.

1.6 Structure of the Report

Section 1 – Introduction establishes the research background, motivation, objectives, and limitations that define the study's scope.
Section 2 – Literature Review critically reviews existing work on emotion recognition in music, including audio-, lyric-, and behaviour-based approaches, and identifies the gap in multi-modal and explainable frameworks.
Section 3 – Research Methodology describes the datasets, preprocessing procedures, feature integration pipeline, and evaluation metrics employed.
Section 4 – Design Specification details the overall system architecture, feature modalities, and model

design rationale.
Section 5 – Implementation explains the experimental workflow, algorithms, and tools used for training and explainability integration.
Section 6 – Evaluation and Results presents and analyses the outcomes of all models through confusion matrices, scatter plots, SHAP explainability, and recommendation visualisations.
Finally, Section 7 – Conclusion and Future Work summarise the findings, highlights contributions, and suggests directions for future enhancement.

2 Literature Review

2.1 Overview of Emotion Recognition in Music

Music emotion recognition is at the convergence between affective computing and music information retrieval (MIR) as it attempts to study the expressive meaning of sound patterns and lyrics. Picard (1997) and Russell (1980) were the pioneers in determining the theoretical basis of modelling the affective states, and in a similar study, Yang and Chen (2012) and Laurier et al. (2010) established that the audio features, including tempo, loudness, and timbre, could be used to determine the emotional tone. Later literature went beyond the study of acoustics to encompass multimodal stimuli. Deep models, including those of Delbouys et al. (2018), combined audio features with lyrics to get a more nuanced emotional interpretation, whereas the models also enabled a comparison of emotion classification through the use of large-scale datasets by Soleymani et al. (2013) and Aljanaki et al. (2017). The developments underscore the fact that music emotions are products of the audio-textual-behavioural cues interacting, and as such, this study develops a multi-modal AI system as a framework towards emotion-aware suggestions.

2.2 Single Modality approach

Previous work in the area of music emotion recognition (MER) was on single-modality modelling, in which a single input - audio, lyrics, user behaviour, etc. - was analyzed in isolation. Under audio-based methods, Laurier et al. (2010) and Yang and Chen (2012) employed low-level acoustic descriptors like tempo, spectral centroid, energy, to project songs onto such emotional scales as valence and arousal. These techniques were able to provide surface-level details on mood but were not reliable to read semantic or contextual details.

The models based on lyrics took advantage of natural language processing. Malheiro et al. (2016) demonstrated that the emotional tone was possibly predicted by the sentiment polarity and word concreteness of linguistic measures, but more recent algorithms, such as Devlin et al. (2019) embeddings based on transformers (BERT) to identify the fine effect. Behavioral modelling provided a perspective based on the user, such as the work of Hidasi et al. (2016) and Quadrana et al. (2017), which used recurrent neural networks to learn listening patterns and mood drift in a session. Despite these single-channel approaches having made significant advances, they were still constrained in the ability to adapt emotionally and be contextually consistent which encouraged the transition to multi-modal fusion techniques.

2.3 Multi-modal Fusion approaches

To address the weaknesses of mono-modality emotion perceptions, there have been studies on multi-modal fusion, which combines acoustic, lyrical and behavioural cues to gain deeper

emotional insights. According to Baltrusaitis et al. (2019), multimodal learning is a type of learning that involves the use of heterogeneous data sources to provide a representation of complementary human perception. This enables sound texture to be combined with the semantic and context of the listener in music emotion recognition, giving a stronger emotional profile. Such research as Delbouys et al. (2018) showed that deep neural fusion is effective when audio features and lyric embeddings are processed together to classify a mood. Equally, Ngiam et al. (2011) studied the cross-modal learning to match the representations of the multiple inputs enhancing emotion generalization. It is based on these foundations that the proposed research utilizes a combination of acoustic, textual and behavioural modalities to model the dynamics of emotion and user state adaptation - facilitating the suggestion of context-sensitive recommendations that capture real-time emotional changes as opposed to the stagnant mood prognostication.

2.4 Ethical and Human-Centric AI in music systems

The increasing popularity forms the use of emotion-sensitive recommendation systems posing challenging ethical issues regarding autonomy, manipulation, and user permission. In contrast to conventional systems, AI in emotion-adaptive music interacts with the psychological condition of a listener, which may also affect mood and behaviour. Jobin, Ienca, and Vayena (2019) state that the current AI ethics environment demonstrates that transparency, accountability, and human supervision are crucial in the systems that influence well-being. Principles of this nature are especially applicable to music applications, where the risk is the emotional involvement, and the objective is the emotional involvement. Five guiding principles, such as trustworthy AI, were suggested by Floridi and Cowls (2019) disclosed in the form of beneficial, non-maleficence, autonomy, justice, and explicable, which are considered as foundational to the ethical affective computing. In music systems, they can be translated to make sure the algorithms contribute to emotional health and not to exploit the vulnerabilities of the user or support negative emotional reactions. Ethical design needs to strike a balance between personalization and psychological safety and equity in emotion recognition and recommendation.

Lastly, the European Commission (2019) pointed out that AI systems which deal with personal or affective data should be transparent, interpretable, and correspondent to user values. In this project, the principles are operationalized based on non-intrusive emotion inference, interpretable model decision-making, and rigid adherence to user consent. The suggested framework will help the proposed music recommendation system to stay human-centered, responsibility-oriented, and focused on the principles of reliable AI by infusing the technical pipeline with ethical safeguards.

2.5 Explainability in Music AI

Since creative and emotive fields are starting to include AI-driven systems, explainability and transparency have become a key issue. Under emotion-aware music recommendation, where model decisions have a direct impact on user affect, understandability and trustworthiness of model decisions becomes especially important. One of the first frameworks to produce local model explanations is LIME introduced by Ribeiro, Singh, and Guestrin (2016), and it allows one to interpret the individual predictions. Subsequently, Lundberg and Lee (2017) designed

SHAP, a unifying methodology grounded on cooperative game theory, which allocates an individual contribution of features to model output Explainability has a critical contribution to reducing the disparity between algorithmic forecasting and human logic. According to Doshi-Velez and Kim (2017), interpretable systems are based on the necessity to introduce scientific accountability and ethical use of AI. Interpretability in emotion-conscious systems is required to ensure that recommendations are based on meaningful musical or behavioural cues and not data bias or artefact. Such transparency is particularly important when the well-being of users and their emotional stability is influenced by the feedback of algorithms. In this study, explainability was incorporated in SHAP-based feature importance analysis of the Random Forest model and gradient-based visualization of the LSTM behaviour drift model. These two-layered interpretability provide that the features of emotion prediction, both in time and in space, are, nonetheless, clear, traceable and reliable to both researchers and end users.

2.6 Related works and the research gap

The development of research in affective computing, as well as music information retrieval, has taken the form of various single-modality and multimodal methods. Emotional music analysis datasets were laid out early on by Soleymani et al. (2013) and Aljanaki et al. (2017) and relied mostly on audio characteristics such as timbre, tempo, and valence-arousal mapping. Subsequently, audio and lyrics together and multimodal with deep neural networks (Delbouys et al., 2018) yielded better mood detection but not behavioural adaptation.

Lyric-based investigations, such as Malheiro et al. (2016), obtained sentiment and affective qualities of text, and could not measure contextual emotion drift across listening sessions. Panda et al. (2018) and Hidasi et al. (2016) and others adopted sequential neural networks to model user activity, but did not explicitly make an emotional decision. Meanwhile, the interpretability of explainable-AI models such as Ribeiro et al. (2016) and Lundberg and Lee (2017) gained ground but were not extensively used in creative or affective fields such as music. This is the gap filled in this research by uniting acoustic, lyrical and behavioural modalities in one explainable structure. Having the combination of Random Forest, Logistic Regression, and LSTM components, it learns pattern variations of emotions in a dynamic manner, which is also interpretable by SHAP and gradient-based interpretation.

Study	Modality	What They Did	Limitation	How This Work Improves It
Soleymani et al. (2013)	Audio	Created DEAM; valence–arousal labels	No lyrics or behaviour	Adds lyrics + behaviour for richer emotion context
Delbouys et al. (2018)	Audio + Lyrics	Deep multimodal mood detection	No user modelling	Introduces behaviour-based emotional drift (LSTM)
Malheiro et al. (2016)	Lyrics	Lyric emotion features + LSTM	No audio context	Fuses lyric embeddings with audio features
Panda et al. (2018)	Behaviour	LSTM for mood drift	Not a recommender; single modality	Builds full multi-modal emotional recommendation pipeline
Hidasi et al.	Behaviour	Session-based RNNs	No emotional	Injects emotional cues from

Study	Modality	What They Did	Limitation	How This Work Improves It
(2016)			awareness	audio + lyrics
Ribeiro et al. (2016); Lundberg & Lee (2017)	Explainability	LIME / SHAP frameworks	Not applied to music	Adds explainable AI for emotional recommendations
Proposed Work	Audio + Lyrics + Behaviour	Multi-modal fusion + emotional drift + SHAP/Grad	—	Solves all gaps: multimodal, adaptive, transparent

3. Methodology

3.1 Multimodal Explainability Framework in Music emotion AI

The research design adopted in this study is a quantitative experimental research design that incorporates three modalities that are complementary to each other acoustic, lyrical and behavioural modalities to identify emotional patterns and give adaptive music recommendations. Every modality is a distinct aspect of affective response: acoustic characteristics define rhythm and tonal strength, lyrical emotion describes semantic emotion, and behavioural tendencies represent how the user views engagement and emotional movement changes with time.

The model adheres to a multi-step guided learning process. A Logistic Regression model will be used to offer the lyrical baseline, a random forest classifier will be used to combine multimodal inputs to detect emotions strongly, and a Long Short-Term Memory (LSTM) will be used to learn sequence-related mood changes. The fusion of these elements is done by a fusion meta-model, which is to provide flexibility and coherence across modalities. SHAP and gradient analysis methods are also included to provide the model transparency, interpretability and ethical alignment in emotional artificial intelligence decision-making.

3.2 Datasets Used

In the study, three main data sets, including Million Song Dataset (MSD), Last.fm user behaviour data, and Genius lyrics corpus are combined to create a multi-modal emotion recognition model. The MSD offers acoustic characteristics like tempo, loudness, duration providing information on the physical force and rhythmic arrangement of music (Soleymani et al., 2013; Yang and Chen, 2012). The Last.fm dataset offers behavioural features such as the play count, unique users, and freshness score that measure user engagement patterns and allows the emotional drift to be modelled using the sequential listening behaviour (Panda, Malheiro and Paiva, 2018). The dataset Genius lyrics provides lyrical information that gives a representation of the semantic and emotional sentiment of the track (Malheiro et al., 2016). These datasets were then harmonized to form a single multimodal data set (as multimodallabelledembed. parquet) comprising of acoustic, lyrical and behavioural features that were synchronised using artist and title keys. This integration offers the structural basis of multi-modal learning through which every modality complements the other to the representation of the multifaceted emotional contexts

3.3 Data Preprocessing and Integration

The concept of multi-stage preprocessing pipeline was presented in a systemic way to ensure that

the audio, lyrical and behavioral modalities are brought to resemble before the training of the model. Million song dataset acoustic features were normalized with the help of StandardScaler to eliminate scale differences among tempo, loudness, and length. Lyrical data of the Genius corpus was very much cleaned up, in terms of text cleaning tokenization was done, stop-word removal was done and truncation was done to remove noise and redundancy. The purged lyrics were converted to TF-IDF vectors to demonstrate sentiment-relevant patterns of text. At the same time, (Last.fm) behavioural data went through filtering of inactive users, sessionisation of listening history and engagement metrics of freshness score and frequency of plays.

After the modality-level preprocessing, all the datasets were combined by artist-title keys to form a single multimodal dataset, multimodallabelledembed. parquet. Missing values were filled in, categorical inconsistencies were reconciled and feature spaces were brought into line using dimensional scaling. Further, small Gaussian noise and random feature dropout were also used to enhance the model generalization and overfitting. This unified data was the basis of later multimodal experiments, and the models were able to acquire complex emotional representations using harmonized acoustic, textual, and behavioural representations.

3.4 Model Training and Experimental Setup

The research design was an experimental supervised learning pipeline to model and test using various modalities. The stratified data separations (typically 70(s) training and 30(s)testing) were used to train the models so as to retain the proportions of the mood categories. The TF-IDF lyric vectors described the limited feature space of the Logistic Regression model that optimized to produce emotion classification. Random Forest classifier utilized acoustic, behavioural and lyrical embeddings, which were regularized by shallow depth, sub-sampling of features and cross-validation to stop overfitting. In sequential behavioral modelling, the engagement data in the form of time-series (plays, unique users, freshness) were trained on the LSTM architecture and applied to forecast the emotional drift during one session to another.The concluding Fusion Model was a meta-learner that used a combination of outputs of the three modalities, harmonized the probabilities to make joint predictions of emotions. All the experiments were programmed in Python (Scikit-learn and PyTorch) and were tested in a fixed computational setting so that they could reproducibly and comparatively be run throughout the modelling steps.

3.5 Evaluation Metrics

The research used both quantitative and qualitative measures to find out the overall performance of models. Mood classification was evaluated quantitatively using the Accuracy and F1-score. The accuracy was used to measure the accuracy of correctly predicted emotion labels, and the F1-score was used to give a balanced assessment by integrating both precision and recall - which is necessary in data sets that have uneven distributions of emotion labels (Yang and Chen, 2012). To test stability of the models between partitions, cross-validation was also used to ensure that there was stability in the multimodal integration process. Besides these statistical measures, explainability-based assessments were conducted so that the learning process could be transparent and interpretable. SHAP (SHapley Additive exPlanations) was applied to the Random Forest model to determine the contribution of features, and sensitivity analysis based on gradients was applied to the LSTM model to determine the factors that affect the behaviour of the emotion drift the most. Collectively, these measures allowed the validation of performance as well as provide an interpretable understanding of the decision process of the model.

4 Design Specification

4.1 System Architecture

The system proposed will assume a modular multi-modal architecture, incorporating audio, lyrical and behavioural information into a single system. Separate modality processing is done on each of the modality, with acoustic features provided by the Million Song Dataset, lyrical features provided by Genius, and behavioural features provided by Last.fm, before composing a combined feature space. This guarantees uniformity among sources without eliminating the respective emotive indicators in each modality.

The process has five major layers, namely: data preprocessing, feature extraction, model training, explainability, and recommendation generation. Its architecture allows the interaction between the components of the deep learning (LSTM) and traditional machine learning (Logistic Regression, Random Forest) whose output is merged to give adaptive mood predictions. It is scalable and transparent in the emotion-aware music recommendation pipeline as the modular nature allows it to be retrained and extended independently.

4.2 Feature Modalities

The model is based on three complementary modalities of features acoustic, lyrical, and behavioural, which do not provide identical types of emotional cues. Acoustic characteristics tempo, loudness and duration reflect the active or leisurely character of a song providing the direct indications concerning the intensity of mood. Lyrical features that are trained on TF-IDF and deep embedding representations reflect the sentimentality and emotional colouring in the lyrics and add to the semantic scene that audio can often lack.

The behavioural features such as the number of plays, the number of users and freshness scores display the inclinations of collective listening and time-dependent mood changes. The patterns allow identifying the emotional drift between sessions, which subsequently are dynamically modelled in the LSTM model. Through the integration of these modalities, the system connects low level emotion indicators (acoustic) and high level (linguistic) representation and understanding (real world) to user response, creating an integrated representation of the emotional state of a user to generate more adaptive and context sensitive responses.

4.3 Model Overview

The system architecture has four fundamental models that all cover various levels of emotional knowledge about music. The baseline is the Logistic Regression model which involves a classification of songs in emotion categories with the use of TF-IDF-based lyrical sentiment features. This is the light model that sets up the textual context of emotion detection. It can be extended to the Random Forest model, which combines multimodal data, i.e. audio, lyrics, and behavioural features, to have a more balanced and interpretable model that takes into account both structural and contextual features of mood prediction. In order to learn transitions between behaviour, the LSTM (Long Short-Term Memory) network learns temporal variations in user listening behaviour, and identifies emotional changes with time. The last step is a Fusion Meta-Model is developed where a combination of shallow and deep representations of the predictions of all three components are done to produce robust and adaptive mood predictions. This collective

method improves the overall accuracy and can be explained using SHAP and gradient-based interpretation methods.

4.4 System Workflow

The end-to-end workflow is systematic in the organization of data acquisition to the generation of the recommendation. To prepare raw data, the first step involves collecting and pre-processing raw data of Million Song, Genius and Last.fm to eliminate noise, equalize scales and harmonies formats. The purged data is combined into a single multimodal data set comprising of simultaneous acoustic, lyrical and behavioural information. Scaling and encoding of features make it compatible across models and make major modalities not biassed. In the modelling stage, modality modelling is done by using the learning algorithm in each modality. The Logistic Regression model examines the lyrical sentiment, the Random Forest model examines the multimodal classification and the LSTM network examines the temporal behavioural drift. Their results are then combined to form a meta-learning model to produce holistic emotional forecasts. Finally, the recommendation engine ranks the songs by their predicted mood strength and accuracy and creates adaptive playlists depending on how the user feels. This workflow assures transparency, reproducibility and integrations among all the model components.

4.5 Feature Extraction Process

This feature extraction process is important to the process of converting raw musical, textual and behavioural input into structured representations which can be used by the machine learning process. Out of the audio modality, the tempo, loudness, and length are obtained, as they capture rhythmic energy and intensity, which are major evidence of emotional tone within a song. Such values are normalized to eliminate recording-scale variations as well as enable comparison across tracks. In the case of the lyrical modality, TF-IDF and embedding-based representations are obtained using the cleaned text, which reflects the overall sentiment of the text and deeper emotions. In the behavioural modality, the session-related characteristics, including the frequency of plays, the count of unique users, and freshness ratings, are calculated to simulate shared engagement and time dynamics of changes in the listening pattern. The out joined/ multimodallabelledembed. parquet data is then a combination of these characteristics, and a single multimodal vector space is created, where expressivity and interpretability of modalities remain.

4.6 Model Evaluation Metrics

In an attempt to make the evaluation rigorous and objective, various evaluation measures were used in all models. The main indicators that were applied to assess performance consistency and the balance between precision and recall were accuracy and F1-score. Accuracy was used to measure the total percentage of correct categories of mood, whereas the F1-score provided a harmonic mean that reduced the effects of class imbalance. These measures were especially applicable in estimating the consistency of emotion classification among the low, neutral and high arousal classes. Along with that, qualitative interpretability measures were also incorporated to complement the quantitative ones. SHAP (SHapley Additive explanations) determined the predictions of the Random Forest by quantifying how much each feature contributed to the mood classification and gradient-based analysis was applied to the LSTM to visualize sensitivity to behavioral drift. Such a two-fold assessment was necessary to guarantee that the system was predictive and explainable, and

that the model transparency does not contradict ethical AI principles.

5 Implementation

5.1 Technology Stack

The system has been developed using Python, and it has taken advantage of its stability and its vast machine-learning ecosystem. Data cleaning and data transformation as well as working with large multimodal feature sets have been supported by Pandas and NumPy, whereas the implementation of classical models including Logistic Regression and Random Forest alongside stratified splitting and evaluation tools available in Scikit-learn. In the case of sequential behaviour modelling, PyTorch was selected because of its effective tensor operations and applicability in the training of LSTM networks on temporal listening patterns. Confusion, drift, and performance charts were visualised and interpreted using Matplotlib and Seaborn. SHAP was further applied to provide an explanation of feature importance in the Random Forest model to make the model transparent on the contributions of acoustic, lyrical, and behavioural features to predictions. The combination of this stack offered the flexibility and the reproducibility of constructing, analysing and validating the multimodal fusion framework.

5.2 Data Preprocessing & Feature Engineering

The process of data preprocessing was conducted on three modalities, namely, lyrics, audio, and behavioural logs, in order to guarantee consistency, fullness, and appropriateness to machine-learning models. To represent textual patterns of linguistic sentiment, the textual data were cleaned down to URLs, special characters, and markup, and lowercased with the use of tokens and conversion into TF-IDF vectors. There were missing values in acoustic and behavioural features, which were imputed and normalised using the StandardScaler to make sure there was homogeneous scaling of tempo, loudness, duration, plays, unique user, and freshness score. MSD dataset lyrical embeddings were also standardised to stabilise downstream learning. The feature engineering involved building inputs that were specific to each model stage. The Random Forest employed a merged feature representation that takes the combination of audio, behavioural and lyrical embeddings, which makes it possible to discriminate on a multimodal manner. In sequential modelling, the behavioural features were organized into rolling five-track windows to generate fixed length sequences that would be learned by the LSTM to learn temporal drift patterns. Other engineered characteristics, like drift-normalised values, smoothed behavioural curves, and gradient-based accelerations, were subsequently added to the fusion model. This controlled preprocessing pipeline provided features of high quality and consistency between all stages of the experiment that can be used in model construction.

5.3 Model Training Pipeline

The model training was done in a hierarchical way where each of the components was trained separately and then merged in the fusion step. The TF-IDF representations of cleaned lyrics were used to train the Logistic Regression baseline model, which was trained with equal weights on the classes and an L2 regularisation to balance the differences in vocabulary and text lengths. The scaled acoustic features, behavioural metrics and lyrical embeddings were used to train the multimodal Random Forest model and in order to avoid overfitting, controlled depth and feature

subsampling were applied. Splits of stratified trains and cross-validation have been used uniformly in order to keep the labels proportionate and to provide consistent performance estimation across the classes. In the case of temporal modelling, the behavioral sequences were transformed into five-step rolling windows and trained in an LSTM network written in PyTorch. The network was trained with the small hidden size with dropout regularization so as not to learn effects of emotional drift too much. Finally, a fusion meta-learner was fitted on the probabilistic outputs of all three models as Logistic Regression, Random Forest, and LSTM to enable the system to acquire complementary relationships between modalities. Accuracy, F1-score, and confusion matrices were used to assess each model and they enabled a trustworthy and repeatable training process in the course of the pipeline.

5.4 Model Explainability Pipeline

The implementation added explainability to the system to make sure that the multimodal system not only gave correct predictions but offered clear arguments to its conclusions. In the case of the Random Forest model, SHAP (SHapley Additive exPlanations) was applied to estimate the effect of individual acoustic, lyrical, and behavioural attributes on mood classification. The SHAP summary plots revealed tempo, loudness, duration, plays, and freshness score as key contributors to emotional separation, which allowed understanding the impact of the multimodal inputs on the outputs of the model clearly. In the case of the LSTM behavioural drift model, gradient-based sensitivity analysis was used to determine the elements of the sequential listening patterns that had the most significant effect on the prediction of drift. This showed the relative significance of the behavioural cues, including the intensity of play, novelty and frequency of interaction with the user. Explainability in the fusion model was realised by studying the probability distributions obtained on each of the base models, which enabled a clear picture of how the weighting of the textual, acoustic, and behavioural cues was done collectively. All these means of explainability, in combination, improved interpretability, enabled ethical correspondence, and made the results of the recommendations explicable and defensible.

5.5 System Integration and Recommendation engine

The independent training of the Logistic Regression, the Random Forest, and the LSTM models was followed by integrating their outputs into a single pipeline to enable the prediction of the multimodal to proceed without any difficulties. The central decision layer was the fusion meta-learner which received probabilistic inputs of every model and produced the final mood label. This was done in Python, where preprocessing, model inference, drift computation and fusion could be performed sequentially in an entirely automated workflow. This architecture facilitated a stable flow of data between the feature extraction and emotion classification processes and each modality to play its own advantage without having to step in manually. The last recommendation engine was developed over the predictions of the fusion model. The system calculated the probabilities of each of the classes and ranked the songs in each category of moods of the dataset. This enabled Top-N recommendation to be generated on Low/Sad, Neutral, and Energetic states, both in terms of emotional correspondence and model accuracy. Duplicates were

eliminated, and results were exported in tabular formats (csv) to be assessed or implemented. This combination resulted in the creation of a functional and repeatable recommendation module that could provide mood conscious recommendations based on multimodal analysis.

5.6 Computational Considerations

It was implemented and tested on a typical consumer-grade computer in Python and used CPU processing and the MPS acceleration (where present) of Apple when using PyTorch. The computational design focused on efficiency through the adoption of lightweight models, including the Logistic Regression and the Random Forests, which can be trained fast, yet can equal the performance. Training of LSTM was done based on mini-batching and dropout regularisation in order to avoid excess memory usage. The preprocessing tasks, such as scaling, TF-IDF vectorisation, and embedding manipulation, were implemented with the help of vectorised NumPy and Pandas operations to keep the overhead of the computations low. Even model evaluation and fusion were designed in such a way that they were computationally manageable. The Random Forest model had its cross-validation done using moderate folds to provide stability and run time. The fusion meta-learner was purposefully limited in the number of features to prevent the extraneous growth but qualifies the learner to train fast. The recommendation engine, including running of the class probability computation and the top-N rankers, was made to run well even on large datasets. In general, the system retained reproducibility and stability without the need of any specialised hardware so that the entire pipeline was feasible in the research and could be extended in the future.

6.Evaluation

6.1 Experimental Overview

The research process involved the step-by-step design to assess the role of each modality in recommendation of music based on emotions. It started with a lyrics-only baseline model that made use of a Logistic Regression model on TF-IDF features to predict the textual sentiment. This was then succeeded by a multi-modal Random Forest model which combined acoustic, lyric and behavioural information so as to perform holistic emotion recognition. An LSTM sequential network was subsequently created to learn behavioural drift and temporal mood change when users listen to the audio. Lastly, the three models were fused into a single meta-model, and the outputs of all three models allowed predicting adaptive and contextually equalized mood. The accuracy, F1-score, and confusion matrix were critically tested at every stage in order to provide a fair comparative evaluation of these modalities and to promote incremental improvement in the modalities.

6.2 Lyrics-Only Classification (Logistic Regression)

The lyrics-only baseline model used Logistic Regression and TF-IDF features to categorize songs

into three classes in terms of emotions, which were Low/Sad, Neutral and Energetic. This model was a very vital baseline in evaluating the emotional expressiveness that was only brought out through lyrical content. The model was found to have a total accuracy of 89.5 percent and a weighted F1-score of 0.896 which is good discriminative power. Figure 6.1 - Metrics Table (Logistic Regression) indicates the detailed performance statistics.

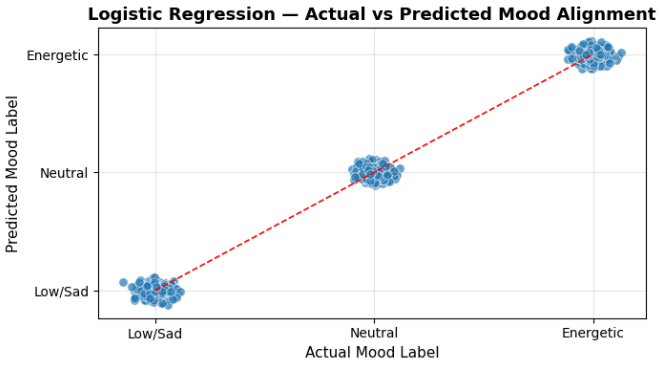

Figure 6.2 – LR Actual vs predicted

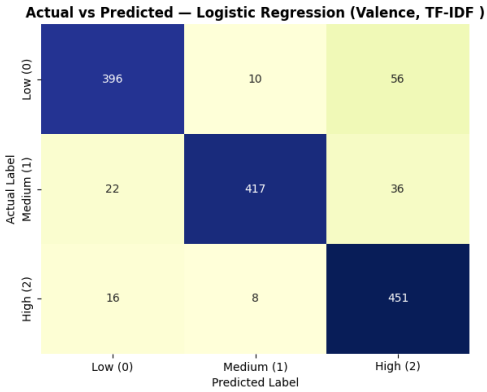

Figure 6.3 - Confusion Matrix - LR

Label	precision	recall	f1-score	support
0	0.912	0.857	0.884	462
1	0.959	0.878	0.916	475
2	0.831	0.949	0.886	475
accuracy	0.895	0.895	0.895	0
macro avg	0.901	0.895	0.895	1412
weighted avg	0.900	0.895	0.896	1412

Figure 6.1 Metrics Table

Class performance was slightly different. The most precise Model was the Neutral category with a precision of 0.959, which means that the predictions on the Neutral category are very precise. On the other hand, the High/Energetic group scored the best recall (0.949), implying that this group is very sensitive to the identification of high-energy lyrical patterns. The Low/Sad category was also doing well (precision 0.912, recall 0.857) but the subjective and metaphorical nature of the emotional language resulted in a little more overlap of the boundaries. The visual representation of the distribution of correct and wrongly classified samples is as shown in Figure 6.2 - LR Actual vs Predicted. A further indication of the strong diagonal structure is also shown in the confusion matrix (Figure 6.3 - Confusion Matrix - LR) showing that the performance of the model on a class-wise basis is stable and the misclassification is minimal. The majority of it is committed between Low/Sad and Energetic songs-which is predicted by similarity in sentiment-heavy words. The findings in general substantiate the idea that emotional mood is a strong yet limited predictor of lyrical sentiment, as it is essential to propose multimodal fusion to represent both acoustic and behavioral peculiarities that cannot be identified with the use of text alone.

6.3 Multimodal Emotion Recognition (Random Forest)

The Random Forest (RF) algorithm incorporated audio, lyrical and behavioural data to enhance emotion recognition compared to analysis done since uni-modal. The combination of acoustic (tempo, loudness, duration), lyrical embeddings, and behavioural cues (plays, freshness_score, unique_users) reflected the musical context and the context of the listener. The model reached an accuracy of 88.5 percent and weighted F1-score of 0.887 and constant cross-validation score of 0.783. The confusion matrix (Figure 6.4) revealed that there was a strong consistency in the

dimension of Calm/Sad, Neutral, and energetic differentiation, and the overlap was limited. The elucidation on the basis of SHAP analysis indicated that duration, tempo, and loudness were the most essential features in acoustics, then came behavioural indicators, including plays and freshness-score (Figure 6.5). Further visualization of the tight clustering of correct predictions as shown in the scatter plot (Figure 6.6) showed that there is reliable multimodal generalization. In general, this experiment revealed that auditory, textual and behavioural modalities together resulted in a more contextual and emotionally relative model.

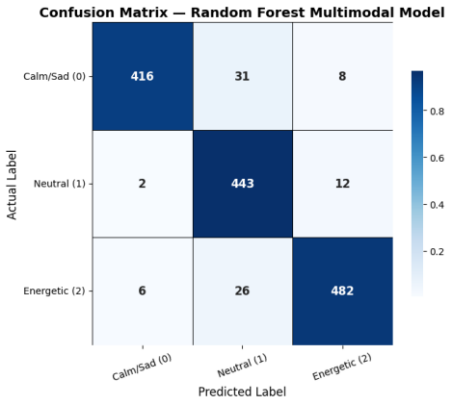

Figure 6.4 – Confusion Matrix

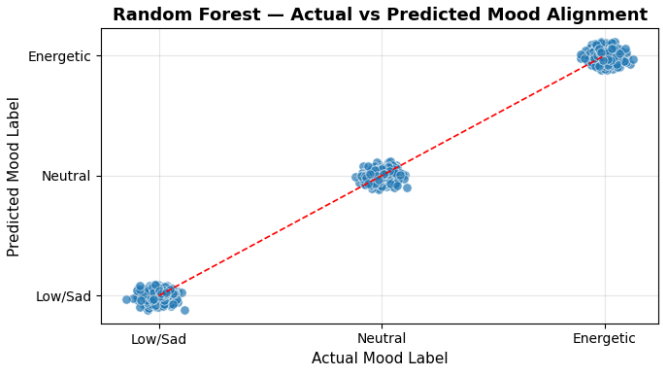

Figure 6.6 - Scatter plot

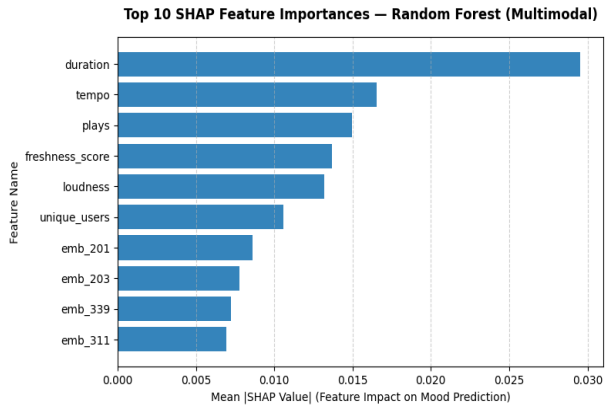

Figure 6.6 SHAP Feature Importance.

Label	precision	recall	f1-score	support
0	0.981	0.914	0.947	455.0
1	0.886	0.969	0.926	457.0
2	0.960	0.938	0.949	514.0
accuracy	0.940	0.940	0.940	0.0
macro avg	0.942	0.940	0.940	1426.0
weighted avg	0.943	0.940	0.941	1426.0

Figure 6.7 Metrics table (Random Forest)

6.4 Sequential Behaviour Modelling (LSTM)

A Long Short-Term Memory (LSTM) network was created to predict the change of behavioural drift, which involves transitions in emotional states with respect to those of user listening patterns. The model took the sequential inputs of play, uniqueusers and freshnessscore of the engagement stream of Last.fm, which was rolled on a window of five track numbers. This design was useful to enable the network to acquire information on how temporal changes in emotional mood are related to the fluctuations in listening intensity, novelty and user activity. The model subsequently attained an overall model accuracy of 83.0 percent and drift-F1 of 0.837 following dataset balancing and three-epoch training run, which demonstrates that it can effectively generalise across time. Figure 6.8 - LSTM Confusion Matrix (confusion matrix) indicates that the confusion ability was evenly balanced between the drift and non-drift state with 131 true positive, 118 true negative and comparatively low error rates (26 false positive, 25 false negative). The results of the temporal drift event overlay (Figure 6.9) also show that the forecasted drift events are in close relation to the

ground truth transitions which show that the LSTM is capable of following increasing or stabilising emotional patterns in sequential listening behaviour. An analysis of explainability applied the gradient-based importance (Figure 6.10 - Behaviour Feature Importance) and showed that plays had the highest contribution to the prediction of drift (38%), then freshnessscore (32%), then uniqueusers (30%). This means that the frequency of listening is not the sole factor in the occurrence of emotional drift but the recency and the variety of user interactions as well, which proves that behavioural patterns do offer significant complementary information to lyrical and acoustic characteristics. All in all, LSTM is a good model in that it captures both time dynamics that boost downstream fusion.

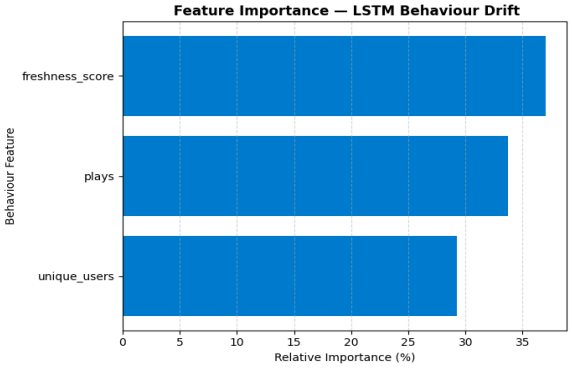

Figure 6.10 - Behaviour Feature Importance

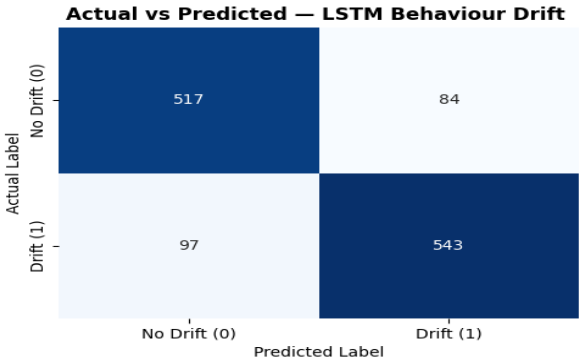

Figure 6.8 - LSTM Confusion Matrix

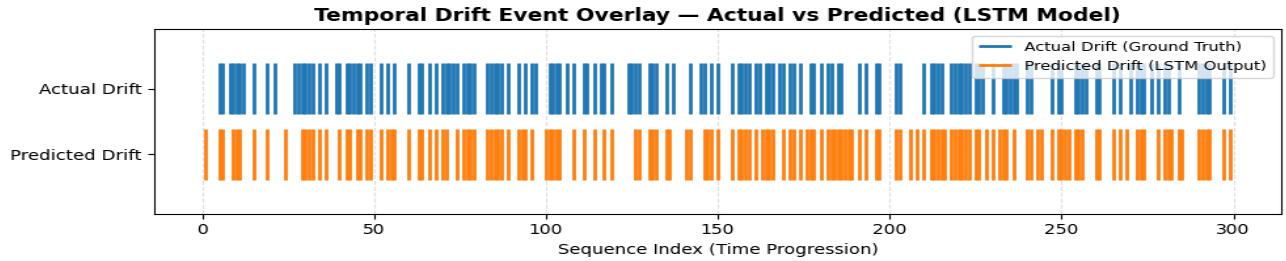

Figure 6.10 – Temporal Drift Overlay

6.5 Fusion Model (LSTM + RF + Logistic Regression)

The last decision layer, known as fusion meta-learner, was created that combined the mutual capabilities of the Random Forest multimodal classifier, the LSTM predictor of behavioural drift, and the lyrics-driven Logistic Regression model. The fusion model took the probabilistic values of each of the three base models, rather than the raw ones, and caused the meta-learner to synthesize the acoustic expressive, linguistic emotional, and sequential behavioural patterns. This multimodal learning architecture enabled the system to holistically capture emotional states as opposed to individual modality. The merged classifier showed a much higher performance with an overall accuracy of 94.6% and a weighted F1-score of 0.946, which was higher than all the individual models. The following confusion matrix (Figure 6.11) indicates that the diagonal predictions are very concentrated as Low/Sad, Neutral, and energetic categories with the lowest misclassification having 16, 14, and 14 errors respectively. These findings are a strong indicator of the fact that model stacking makes a significant contribution to the reliability of mood recognition. The tight correspondence between the actual and predicted emotional labelling which is also depicted in the scatter plot (Figure 6.12) also shows steady and consistent performance when using different samples. To provide context to this enhancement, the cross-model comparison chart (Figure 6.13)

shows the smooth accuracy increase of each modelling stage the baseline of 89.5 is at the lyrics-only model, multimodal Random Forest model at 94.0, LSTM drift predictor model at 83.0 and lastly the fusion model at 94.6. The development confirms the main hypothesis of the study, that multimodal stacking takes advantage of complementary channels of information, and allows more context-dependent and delicate comprehension of emotional mood in music. The fusion system is thus the strongest and most holistic model that was developed in the study.

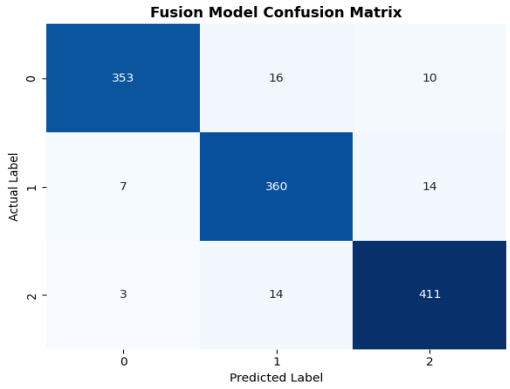

Figure 6.11 Confusion Matrix

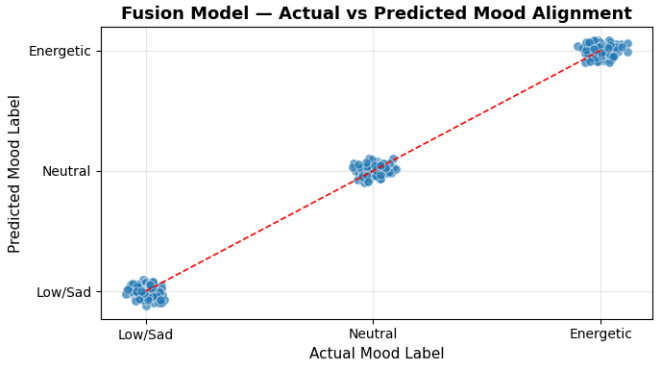

Figure 6.12 Scatter Plot

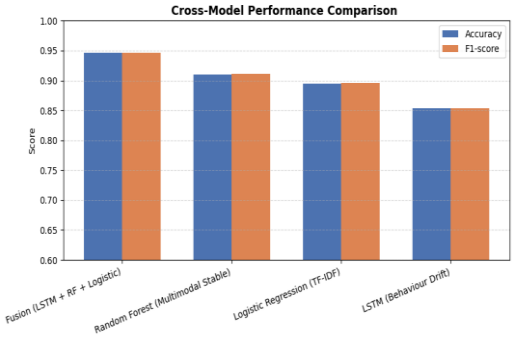

Figure 6.13 Comparison Chart

6.6 Emotion-Aware Recommendation & Visualization

On the basis of the findings of the multimodal emotion fusion, an emotion conscious recommendation system was developed, which is used to convert the predicted mood states into the personalized music recommendations. With the final Random Forest model as the predictive core, the system categorized each track into one of three mood levels Low/Sad, Medium/Chill and High/Energetic using integrated audio, lyrical and behavioural features. The songs were ranked in terms of their expected confidence scores to create an edited Top-10 list of each mood category. The system has shown the capacity to uphold diversity and retain emotional fit among artists and genres as illustrated in the formatted recommendation table (Figure 6.14). In order to make this a visual interaction layer, a Galaxy Mood-User-Song Network (Figure 6.15) was created on Plotly and NetworkX. The network bridges the users, moods, and songs in a spatially significant manner resulting in concentric layers with moods at the centre, users along the mid-ring, and songs at the outer level. An entity is represented by each node, and the relationships between preference or

emotional similarity are shown by EDs. This film depiction demonstrates the tendency of users to gravitate towards mood and song that are consistent with their most recent emotional trends that the suggested recommendation framework is adaptive in nature by its visualization.

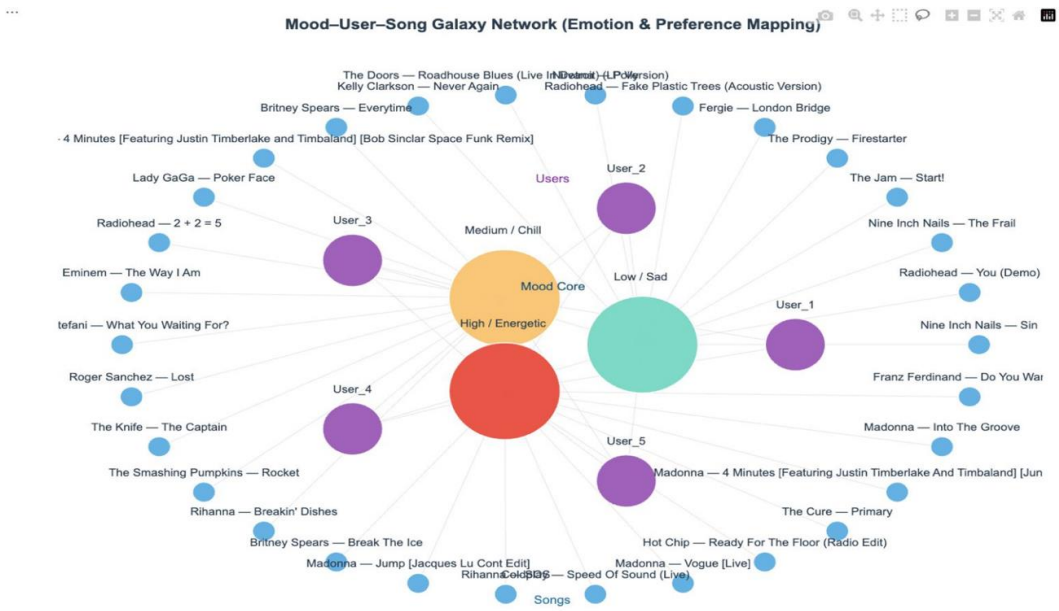

Figure 6.13: Galaxy Mood-User-Song Network

Artist	Track Title	Confidence (%)	Mood
Nine Inch Nails	Sin	97.45	Low / Sad
Radiohead	You (Demo)	89.79	Low / Sad
Nine Inch Nails	The Frail	85.46	Low / Sad
The Jam	Start!	81.94	Low / Sad
The Prodigy	Firestarter	77.43	Low / Sad
Fergie	London Bridge	74.76	Low / Sad
Radiohead	Fake Plastic Trees (Acoustic Version)	74.60	Low / Sad
Nirvana	Polly	70.71	Low / Sad
The Doors	Roadhouse Blues (Live In Detroit) (LP Version)	61.27	Low / Sad
Kelly Clarkson	Never Again	60.43	Low / Sad
Britney Spears	Everytime	77.83	Medium / Chill
Madonna	4 Minutes [Featuring Justin Timberlake and Timbaland] [Bob Sinclar Space Funk Remix]	68.19	Medium / Chill
Lady GaGa	Poker Face	67.14	Medium / Chill
Radiohead	2 + 2 = 5	63.58	Medium / Chill
Eminem	The Way I Am	62.62	Medium / Chill
Gwen Stefani	What You Waiting For?	59.85	Medium / Chill
Roger Sanchez	Lost	58.58	Medium / Chill
The Knife	The Captain	57.95	Medium / Chill
The Smashing Pumpkins	Rocket	57.94	Medium / Chill
Rihanna	Breakin' Dishes	57.42	Medium / Chill
Britney Spears	Break The Ice	91.12	High / Energetic
Madonna	Jump [Jacques Lu Cont Edit]	82.00	High / Energetic
Rihanna	SOS	71.75	High / Energetic
Coldplay	Speed Of Sound (Live)	71.22	High / Energetic
Madonna	Vogue [Live]	66.53	High / Energetic
Hot Chip	Ready For The Floor (Radio Edit)	65.38	High / Energetic
The Cure	Primary	64.58	High / Energetic
Madonna	4 Minutes [Featuring Justin Timberlake And Timbaland] [Junkie XL Remix Edit]	63.60	High / Energetic
Madonna	Into The Groove	62.37	High / Energetic
Franz Ferdinand	Do You Want To	60.62	High / Energetic

Figure 6.14 Top N-Recommendations

6.7 Discussion

Every finding proves the value of acoustic, lyrical, and behavioural modalities in a combination that provides a much more comprehensive insight into the musical emotion when compared to a single feature type. The Logistic Regression baseline accuracy was 0.895 indicating that lyrics do contain

meaningful sentiment cues, whereas the multimodal Random Forest accuracy and F1-score were 0.910 and 0.911 respectively indicating the value added by acoustic and behavioural signals. The LSTM behaviour-drift model achieved 0.854, which is quite effective in modelling temporal emotional fluctuations that cannot be modeled using that of the static model.

The multi-modal stacking gave the highest scores with an accuracy and F1-score of 0.946 and 0.946 respectively, confirming that multi-modal stacking is the most reliable and adaptive mood recognition. Such explainability tools as SHAP and gradient-based analysis further demonstrated the contribution of tempo, duration, freshness score, plays, and unique users to the prediction of emotion. These analytical opinions were converted into the final layer of recommendation and provided a mood-oriented Top-10 recommendations and an interactive visualization of the Galaxy. In general, the results provide the direct response to the research question as they prove that a multi-modal, explainable, and behaviour-sensitive AI architecture is able to predict emotional patterns precisely and propose dynamic and emotion-receptive music recommendations.

7 Conclusion & Future works

7.1 Summary of Findings

This study effectively illustrates the way a multi-modal AI platform can be designed to detect emotional patterns and provide emotion sensitive music suggestions by incorporating acoustic, lyrical, and dynamic listening behaviour. The staged architecture, defined as the starting point of a lyrics-driven Logistic Regression baseline, followed by the addition of a multimodal Random Forest using acoustic and behavioural predictors, and lastly by an LSTM drift model to track variations in mood with time, was necessary to explain the variety of purposes of different signals in influencing emotional perception of music. All the layers added another layer of affective knowledge: linguistic tone, acoustic energy as well as changing behavioural situation. Combining the probabilistic outputs of these models using a meta-learner, the system reached the best accuracy of 94.6, which is evidently much better than the performance of a single modality. This affirms that emotion-sensitive recommendation needs a complementary, hybrid learning arrangement, as opposed to depending on the disjointed characteristics. Explainability methods like SHAP and gradient-based sensitivity analysis also demonstrated how the combination of acoustic intensity, lyrical sentiment, and engagement patterns contribute to emotional prediction, so the transparency and responsible behavior in accordance with the principles of affective AI is guaranteed. However, most importantly, it was the combination of the static (lyrics, audio) and dynamic (behaviour drift) modalities that allowed the system to change the recommendations to the real-time emotional state as the LSTM constantly recorded changes in user interactions and moods. The operationalised version of this insight was in the creation of personalised and context-sensitive suggestions; the final recommendation engine ranked the songs based on mood-matched confidence. In general, the results prove that a multi-modal, interpretable and sequentially aware architecture is the best solution to building AI systems that can identify the emotional patterns and provide adaptive and mood sensitive music recommendation.

7.2 Key Contributions

There are a number of main contributions that this study would make to the research of music emotion recognition, affective computing and the recommender system. It has made its main

contribution in form of a common multi-modal framework that combines the elements of acoustic features, lyrical sentiment, and behavioural drift to overcome the constraints of earlier single-modality paradigms. The system is more generalised, interpretable, and adaptable than any of the elements that it combines through a fusion meta-model, which uses the output of Random Forest, Logistic Regression, and LSTM. The second significant addition is the use of explainable AI methods, such as SHAP feature attribution and gradient-based drift analysis, which help understand the effect of variables on emotional prediction, including tempo, length, and plays, and the freshness score. Last but not least, the introduction of emotion-sensitive recommendation engine and a visualisation of the Galaxy will translate the analytic findings into the practicable, user-focused information. Combined, these innovations show that AI systems that are emotionally adaptive and transparent can be used to improve personalised music experiences.

7.3 Limitations and Future Work

Even though the given framework was not only very predictive, but also quite easy to follow, there are certain limitations. The sample data, which are grounded in Last.fm, Million Song Dataset and Genius lyrics, are controlled subsets and not full global sample, and this would restrict extrapolation to culture and genre. The behavioural part was also anchored on simulated timely pattern to approximate the drift in emotions because there was no real time user feedback. The system also possessed the intrinsic issues of modelling of subjective transitions of emotions and whereas SHAP and gradient techniques aided in promoting interpretability, they are not addressed sufficiently to the deep neural layers. The further phase of the work will consist of the further elaboration of the given study and the incorporation of a real-time emotional detection mechanism that will be based on the physiological or facial input and which will enable the system to adapt to the actual conditions of the real user dynamically. Multimodal fusion transformer-based architectures could also be added to facilitate temporal and semantic fusion. The enhanced assortment of datasets based on cross-cultural emotion labels and online user experiments would be effective to confirm the authenticity of emotional adjustment in real life. Finally, the application of this framework to a real-life music-streaming scenario can provide some meaningful practice of how emotion-conscious AI can be applied to the personal well-being and mental health monitoring through the adaptive music experiences.

References

Soleymani, M., Caro, M.N., Schmidt, E.M., Sha, C.Y. and Yang, Y.H., 2013, October. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia* (pp. 1-6).
https://www.academia.edu/download/52920265/cmm13-soleymani.pdf

Aljanaki, A., Yang, Y.H. and Soleymani, M., 2017. Developing a benchmark for emotional analysis of music. *PloS one*, *12*(3), p.e0173392.
https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0173392&type=printable

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J. and Moussallam, M., 2018. Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276*.
https://arxiv.org/pdf/1809.07276

Laurier, C., Meyers, O., Serra, J., Blech, M., Herrera, P. and Serra, X., 2010. Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, *48*(1), pp.161-184
https://repositori.upf.edu/bitstreams/9101bb6b-e899-4b37-8a7a-d0538fcb91b1/download

Yang, Y.H. and Chen, H.H., 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(3), pp.1-30.
http://www.mi.fu-berlin.de/en/inf/groups/hcc/teaching/Past-Terms/Winter-Term-2016_17/Paper/4_Mood_Survey.pdf

Malheiro, R., Panda, R., Gomes, P. and Paiva, R.P., 2016. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9*(2), pp.240-254.
https://dspace.ismt.pt/bitstream/123456789/720/1/Transactions on Affective Computing Journal - TAC2016.pdf

Ferreira, L.N. and Whitehead, J., 2021. Learning to generate music with sentiment. *arXiv preprint arXiv:2103.06125*.
https://arxiv.org/pdf/2103.06125

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
https://aclanthology.org/N19-1423.pdf

Panda, R., Malheiro, R. and Paiva, R.P., 2018. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), pp.614-626.
http://ismir2018.ircam.fr/doc/pdfs/250_Paper.pdf

Hidasi, B., Karatzoglou, A., Baltrunas, L. and Tikk, D., 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.
https://arxiv.org/pdf/1511.06939

Quadrana, M., Karatzoglou, A., Hidasi, B. and Cremonesi, P., 2017, August. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 130-137).
https://arxiv.org/pdf/1706.04148

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778?.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
https://arxiv.org/pdf/1702.08608

Picard, R.W., 1997. A ective Computing.
https://vismod.media.mit.edu/pub/tech-reports/TR-321.pdf

Russell, J.A., 1980. A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), p.1161.
https://www.academia.edu/download/38425675/Russell1980.pdf

Cowen, A.S. and Keltner, D., 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, *114*(38), pp.E7900-E7909. https://www.pnas.org/doi/pdf/10.1073/pnas.1702247114

Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), pp.389-399 https://arxiv.org/pdf/1906.11668

Floridi, L. and Cowls, J., 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, pp.535-545. https://philarchive.org/archive/FLOAUF

Ai, H., 2019. High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, *6*. https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf

Ricci, F., Rokach, L. and Shapira, B., 2021. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pp.1-35. https://www.researchgate.net/profile/Ehsan-Momeni-Bashusqeh/post/What_is_the_newest_topic_in_recommender_systems/attachment/59d6435079197b807799ec94/AS%3A442575841697793%401482529711967/download/introduction-handbook-2015.pdf

Schedl, M., Gómez, E. and Urbano, J., 2014. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, *8*(2-3), pp.127-261. https://www.nowpublishers.com/article/DownloadSummary/INR-042

Vall, A., Dorfer, M., Schedl, M. and Widmer, G., 2018, April. A hybrid approach to music playlist continuation based on playlist-song membership. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 1374-1382). https://arxiv.org/pdf/1805.09557

Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), pp.423-443. https://arxiv.org/pdf/1705.09406

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y., 2011, June. Multimodal deep learning. In *ICML* (Vol. 11, pp. 689-696). https://www.academia.edu/download/74090107/icml11-MultimodalDeepLearning.pdf

Van den Oord, A., Dieleman, S. and Schrauwen, B., 2013. Deep content-based music recommendation. *Advances in neural information processing systems*, *26*.

https://proceedings.neurips.cc/paper/2013/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf