マルチラベルランキング学習による感情とジャンルを統合した音楽推薦
Music Recommendation Integrating Emotion and Genre via Multi-label Ranking Learning

姚 越成 (YAO Yuecheng)*

法政大学大学院 情報科学研究科 情報科学専攻 24T0031
yuecheng.yao.9n@stu.hosei.ac.jp

Abstract

Music emotion recognition (MER) is vital for recommendation systems, yet existing research often relies on non-commercial datasets and single-label classification, limiting reproducibility and the capture of complex emotional nuances. Furthermore, embedding spaces used for music exploration typically lack interpretability. This study addresses these issues by introducing a multi-label ranking and exploratory framework. First, a reproducible corpus, Melon-Emotion-14k (ME-14k), was constructed with ranked emotion tags. Second, a Learning to Rank approach using ListNet was employed to learn tag relevance. Third, a multi-branch model, QuadFusionNet, was proposed to integrate mel-spectrograms, MuQ embeddings, genre, and audio features. Fourth, a "Dynamic Axis Interpretation" mechanism was introduced to dynamically map abstract embedding axes to interpretable audio features. The model achieved an nDCG@3 of 0.426 on ME-14k, outperforming the random baseline. Moreover, a qualitative case study on the "Jazz & Romantic" subset demonstrated that system-generated axis labels aligned with human auditory perception. The contributions of this study include the construction of a reproducible dataset, a multi-branch ranking model, and an enhanced visualization interface that enhances the explainability of music exploration.

1 はじめに

音声信号から感情を推測することを目的とする音楽感情認識 (MER) は、アフェクティブ・コンピューティングと音楽情報検索の交差点に位置する中核的な研究分野である。近年の深層学習の進歩と大規模な音楽データセットの利用可能性が、この分野の進展を著しく加速させている。 MER は、ストリーミングプラットフォームにおける音楽推薦、ヘルスケアにおける気分調節など、幅広い応用を支えている。

最近の研究は主に 2 つのタスクに焦点を当てている： Valence–Arousal(VA) 平面 [11] に基づく連続的な感情回帰と、VA 平面の象限ベースの分割または離散的な感情分類法 [8] を用いたカテゴリ分類である。2020 年以降、 CNN （畳み込みニューラルネットワーク)、 LSTM （長・短期記憶ネットワーク)、 Transformer 、およびマルチモーダル融合モデルなど、広範な深層学習フレームワークが提案されており、これらはしばしばアテンションメカニズム、音源分離、および事前学習済み音声埋め込みを取り入れている。入力モダリティには通常、音声、歌詞、および MIDI が含まれる。しかし、過去 5 年間のデータセット使用状況に関する調査 [8] によると、最も頻繁に使用されているのは、独自に構築されたデータセット（11 研究)、 DEAM [1]（8 研究)、および MTG-Jamendo データセット（6 研究）であり、商用ストリーミングデータセットは依然として稀にしか使用されていない。この傾向は、既存のモデルの多くが、実世界の商用音楽シナリオにおいて一般化可能性と再現性を欠いていることを示唆している。

しかし、既存研究の多くは、音楽の感情を「単一カテゴリ」として扱う分類問題に主眼を置いてきた。実際の楽曲は「悲しい」かつ「孤独」であるなど、複数の感情が混在する性質を持つ。これらを無理に単一ラベルに集約することは、感情間の豊かな相関関係を無視することになる。

本研究では ListNet を用いた マルチラベルランキング学習 (Learning to Rank) を導入、楽曲の聴感的な類似性を保持した埋め込み空間を構築する。さらに、この空間上でクラスタリングを行うことで、タグ情報だけでは区別できない「サブカテゴリ」を生成し、多様なスタイルを提示可能な探索的推薦フレームワークを提案する。

2 関連研究

2.1 ラッセルの感情円環モデル

ラッセルの感情円環モデル [11] は、直交する Valence（快・不快）と Arousal （覚醒・鎮静）の軸で定義される円形の二次

* 指導教員：伊藤克亘 教授

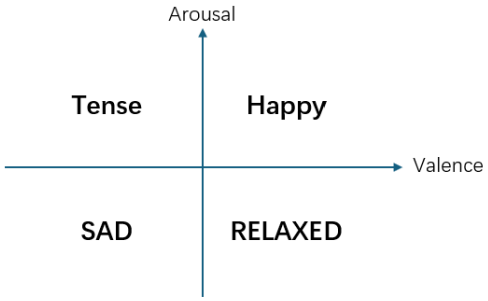

図 1　VA 平面に基づく 4 象限の感情分類

元平面上に感情語を配置する。この構造は経験的研究によって広く支持されており、例えば、「幸せ」は高い Valence と高い Arousal に対応し、「悲しい」は低い Valence と低い Arousal に対応する。MER では、この平面は一般的に 4 つの象限に離散化され、「幸せ」「穏やか」「緊張」「悲しい」といった基本的な感情クラスを形成する。

2.2　DEAM Dataset

DEAM [1] は、1,802 曲の非商用な 45 秒間の抜粋に対して、連続的および平均的な VA アノテーションとジャンルラベルを提供する。しかし、その非商用という性質は、商用音楽への一般化可能性を制限する可能性がある。

2.3　Melon Playlist Dataset

Melon Playlist Dataset[5] は、DJ によってキュレートされた 148,826 のプレイリストと 649,091 の商用トラックで構成されており、各トラックは 20 秒から 50 秒のメルスペクトログラムセグメント、30,652 のクラウドソーシングによるタグ、およびジャンルラベルによって表現される。タグ予測やプレイリスト継続といったタスクをサポートしている。生の音声の代わりにメルスペクトログラムを公開することで、著作権の制約下での大規模な実験も容易にしている。しかし、感情のアノテーションは含まれていない。

2.4　JoyfulJuneS16k Dataset

JoyfulJuneS16k[7] は、NetEase Cloud Music の感情テーマのプレイリストから導出された 12 の感情カテゴリへのマルチラベル割り当てを持つ約 16 万トラックを含んでいる。著者らは、ジャンルと感情の間に有意な関連があること（p 値分析による）を報告しており、これがジャンル情報の組み込みを動機付けている。API アクセスにより特徴抽出用の 30 秒プレビューが可能だが、データセットにはジャンルのアノテーションが欠けている。そのため、本研究では Melon と交差させ、整列した感情ラベルとジャンルラベルを取得する。

2.5　Essentia

Essentia[2] は、音楽情報検索のためのオープンソースライブラリであり、信号処理および統計アルゴリズムのセットを提供する。本研究では、後続の感情分類のために、手作業で設計された (handcrafted) 音声記述子を抽出するために使用される。

2.6　MuQ

MuQ[14] は、メルスペクトログラムを入力とし、Conformer バックボーンを用いたマスク予測フレームワークを使用して訓練された自己教師あり音楽表現モデルである。音楽タギング、楽器や旋法の認識、感情予測など、様々な MIR タスクに効果的に転移学習できる。EmoMusic や MARBLE などのベンチマークデータセットにおいて、MuQ は Valence/Arousal の R^2 スコアが 0.6〜0.75 の範囲で、最先端の性能を達成している。層ごとの分析は、MuQ が意味的（例：ジャンル、構造）および音響的（例：ボーカリスト、楽器）な手がかりの両方を捉えていることを示唆している。

2.7　MLP-SMOTETomek

Ospitia-Medina ら [9] に従い、まず動的な VA アノテーションをウィンドウ内で平均化し、次に各トラックで平均化して曲ごとの VA 値を取得し、それを 4 つの VA 象限ラベルに離散化する。結果として得られるクラス分布は不均衡である。著者らは、層化 80/20 訓練テスト分割の下で SVM、ランダムフォレスト、MLP（多層パーセプトロン）分類器を比較し、いくつかのリサンプリング戦略を評価している。MLP については、SMOTE オーバーサンプリングと Tomek リンククリーニングを組み合わせたハイブリッド SMOTETomek アプローチが最高の性能を達成し、Macro-F1 スコアは 0.50 であった。しかし、少数派クラス（Q2/Q4）の性能は依然として最適とは言えない。

2.8　音楽探索と可視化インターフェース

音楽推薦において、ユーザーが明確な検索意図を持たない「Open Mindset」にある場合、単なるリスト提示よりも、探索的なインターフェースが有効であることが示されている [6]。特に、楽曲の埋め込み表現（Embeddings）を 2 次元平面に投影する可視化手法は、ユーザーが自身のコレクションを再発見（Rediscovery）し、楽曲間の未知の関係性を探索する上でツールとなる [12]。本研究では、これらの知見に基づき、ランキング学習によって獲得された埋め込み空間を、探索的ナビゲーションの基盤として利用する。

3　データ準備と探索的研究

Melon Playlist Dataset[5] と JoyfulJuneS16k[7] にわたるアイテムを、曲名とアーティスト名を正規化（小文字化、スペースと句読点の削除、括弧内のコンテンツの除去、および「feat.」などのエイリアス処理）することによって整列させた。これにより、メルスペクトログラム特徴、ジャンルラベル、感情ラベルを持つ約 14,000 の整列済みトラックが得られた。この整列されたコレクションは、以降の実験の主要なデータセットとして機能し、本論文では Melon-Emotion-14k（ME-14k）と呼ばれる。

3.1　ME-14k 構築の必要性

既存の公開データセットは、本研究が目指す「ジャンルと感情を統合した探索」や「最新の表現学習モデルの適用」におい

て、以下の 3 点で制約がある。そのため、これらを解消する独自のデータセット（ME-14k）の構築が不可欠であった。

第一に、ラベル情報の相補性である。Melon Playlist Dataset[5] は商用楽曲のジャンルラベルを持つが感情次元が欠如しており、一方で JoyfulJuneS16k[7] は詳細なマルチ感情ラベルを備えているがジャンル情報が欠けている。これら 2 つのソースを紐付け統合することで初めて、ジャンルをコンテキストとして同種の感情における微細な聴感の違いを区別することが可能となる。

第二に、特徴抽出の柔軟性と最新モデルへの適応である。既存のデータセットの多くは、著作権の制約等により固定パラメータの特徴量（Mel-spectrogram 等）のみで提供されており、MuQ[14] のような生波形を入力とする最新の事前学習モデルには適用できない。また、本研究で提案する「動的意味解釈」機能の実装には、特定の音響特徴量を柔軟に抽出できる環境が必要であるため、API を通じて 30 秒間の生音声を確保する必要があった。

第三に、システム実装と検証における要件である。収集された楽曲の約 65% が複数の感情タグを保持している事実（後述の 3.2 節で詳述）は、単一ラベル分類ではなくマルチラベルランキング学習の必要性を裏付けている。加えて、提案する探索型 UI において、ユーザーがモデルの妥当性を直接聴感で検証するためには、特徴量だけでなく実際の音声データがシステム内に統合されていることが必須条件となる。

3.2 感情タグの順序付きマルチラベル特性

構築された ME-14k データセットは、各トラックに対して適合度の高い順に複数の感情タグが付与されている。例えば、ある静かな楽曲において「 Peaceful 」が最も支配的な感情であり、次いで「 Romantic 」の要素が含まれる場合、タグはその重要度順に列挙される。従来の DBSCAN 等による単一ラベル化や通常のマルチラベル分類では、このようなタグ間の順序関係や強弱のニュアンスが捨象されてしまう。そのため、本研究ではタグを単なる集合としてではなく、順序構造を持ったリストとして保持し、上位のタグほど高い関連度を持つランキング学習の正解データとして利用する。

3.3 VA 象限境界に関する予備研究

実際の感情カテゴリに対する VA 平面象限の識別能力を調査するため、DEAM データセットから 390 の ロックジャンルのサンプルを使用し、 Happy / Tense / Sad / Relaxed の 4 つのカテゴリを用いて手動での聴取とアノテーション研究を実施する。 VA 平面上で、 KDE （カーネル密度推定）を用いて分布を平滑化し、「高密度領域」を経験的な境界として抽出する。結果は、異なる感情の高密度領域が象限の境界を広く横断していること（例：図 2 では、 Rock–Tense 領域が複数の象限にまたがっている）を示しており、象限の閾値がジャンル内の感情境界を確実に線引きするものではないことを示している。したがって、本論文の後続のモデリングでは、離散的な感情ラベルを主要なタスクとして採用する。

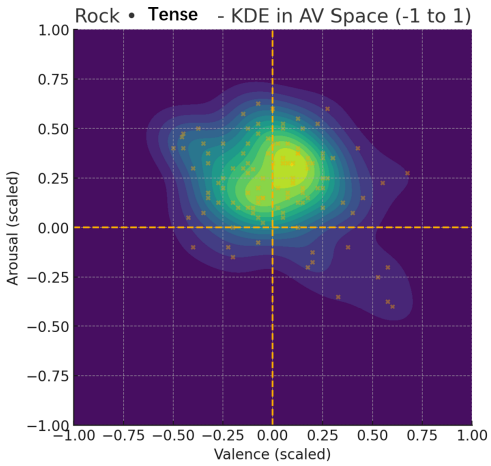

図 2 VA 平面における Rock–Tense のカーネル密度推定（KDE）

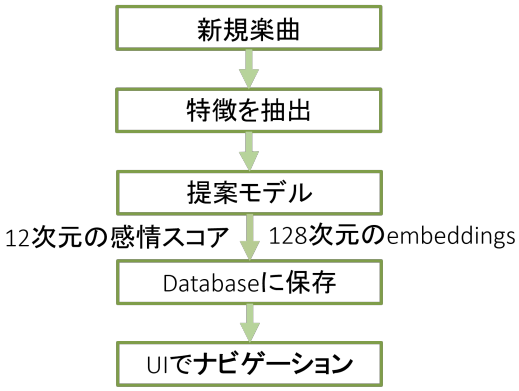

図 3 システムの全体処理フロー

4 提案手法

4.1 システムの全体処理フロー

本システムの処理は (図 3)、楽曲データのプリプロセッシングからユーザーへの提示まで、以下のパイプラインで構成される。

- **オフライン処理：**新規楽曲に対し、メルスペクトログラム、MuQ 埋め込み、および 8 つの低レベル音響特徴量を抽出する。これらの特徴量を提案モデル（QuadFusionNet）に入力し、12 次元の感情ランキングスコアと 128 次元の埋め込みベクトル（Embedding）を推論・算出する。その後、算出結果を楽曲メタデータとともにデータベースへ保存する。
- **オンライン処理：**ユーザーが UI 上で特定のジャンルおよび感情を選択すると、システムはデータベースから予測スコアに基づき該当楽曲をフィルタリングする。さらに、提示された楽曲群の埋め込み空間上で K-means 法によるサブカテゴリ生成を行い、音響特徴量を用いた動的意味解釈

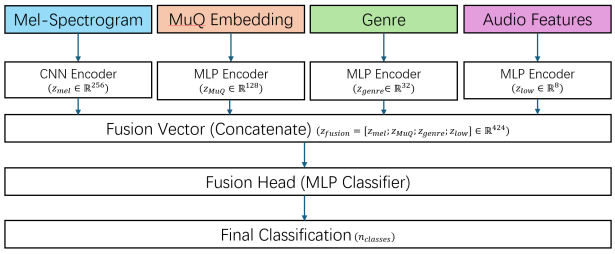

図 4　QuadFusionNet のアーキテクチャ

レイヤーを適用することで、ユーザーの直感的なナビゲーションを支援する。

4.2　QuadFusionNet アーキテクチャ

図 3 に示すように、提案する QuadFusionNet は、4 つの異なるモダリティを統合するために設計されたマルチブランチのレイトフュージョン（late-fusion）アーキテクチャである。モデルのワークフローは、以下の主要な段階で構成される：

4.2.1　並列特徴エンコーディング

モデルは、4 つの入力モダリティを、それぞれ専用のエンコーダブランチで処理する。

- メルスペクトログラム は CNN エンコーダ [4] に入力され、256 次元の特徴ベクトル $z_{mel} \in \mathbb{R}^{256}$ を抽出する。
- MuQ 埋め込みは MLP エンコーダによって処理され、128 次元のベクトル $z_{MuQ} \in \mathbb{R}^{128}$ を生成する。
- ジャンル情報 は、別の MLP エンコーダによって 32 次元のベクトル $z_{genre} \in \mathbb{R}^{32}$ に変換される。
- 最後に、手作業で設計された 低レベルの音声特徴 が 4 番目の MLP エンコーダを通過し、8 次元のベクトル $z_{low} \in \mathbb{R}^{8}$ を生成する

4.2.2　特徴融合

4 つのブランチからの潜在ベクトルは、単一の包括的な表現を形成するために連結される。この融合ベクトル z_{fusion} は次のように作成される：

$$z_{fusion} = [z_{mel}; z_{MuQ}; z_{genre}; z_{low}] \in \mathbb{R}^{424}$$

結果として 424 次元のベクトル $z_{fusion} \in \mathbb{R}^{424}$ となる。

4.2.3　分類

統合された融合ベクトルは、 MLP ベースの分類器である融合ヘッドに入力される。このヘッドは、融合された特徴を $n_{classes}$ 個の感情カテゴリにわたる最終的な確率分布にマッピングし、最終分類を生成する。

4.3　低レベル音声特徴

人間の感情知覚に関連する音響属性 [10] を直接捉えるために、8 つの低レベル特徴が Essentia [2] を用いて選択・抽出された。各グループの詳細は以下の通りである：

Tense:　median_dE_dt, mean_dissonance . エネルギーの変化率（ダイナミックな推進力の代理）と感覚的な不協和音は、より高い緊張または不快感と関連している。

Happy:　p90_centroid, high_band_ratio. 高いスペクトルセントロイドのパーセンタイル値と高周波エネルギーの比率の高さは、音色の明るさを反映し、楽器の音色におけるエネルギッシュ/ポジティブな感情と相関している。

Sad:　low_band_ratio, energy_std. 支配的な低周波成分と狭いラウドネスのダイナミックレンジは、低い覚醒度と演奏における悲しみのパターンと一致している。

Relax:　p10_centroid, median_flux. 低いスペクトルセントロイド（暗い音色）と減少したスペクトルフラックスは、より滑らかで変化の少ないスペクトルを示し、低い覚醒度と一致している。

4.4　4 ブランチエンコーダ

CNN エンコーダ（メルスペクトログラム用）： 2D メルスペクトログラムは、チャネル深度を 64 から 128、そして 256 へと段階的に増加させる 3 つの畳み込みブロックのスタックによって処理される。各ブロックは、3*3 の畳み込み層 （Conv2d）、バッチ正規化 （BN） 、 ReLU （正規化線形ユニット）活性化関数、2*2 のマックスプーリング層、およびドロップアウト（p=0.25）で構成される。その後、Adaptive Average Pooling 層が出力をフラット化し、最終的な特徴ベクトル $z_{mel} \in \mathbb{R}^{256}$ を生成する。

MLP エンコーダ（ベクトル入力用）： 3 つの 1D ベクトル入力については、構造的に類似した多層パーセプトロン（ MLP ）エンコーダが、それらを専用の特徴空間に射影するために使用される。各エンコーダは、通常、全結合（ FC ）層、 BN 、 ReLU 、 およ びドロップアウトで構成される。各ブランチの具体的な入力および出力次元は以下の通りである：

- MuQ 埋め込みブランチ ： 1024 次元の MuQ ベクトルが、特徴ベクトル $z_{MuQ} \in \mathbb{R}^{128}$ にマッピングされる。
- ジャンルブランチ ： 入力ジャンルベクトル $z_{genre} \in \mathbb{R}^{32}$ が、特徴ベクトルにマッピングされる。
- 低レベル特徴ブランチ ： 8 次元の標準化ベクトル x_{low} が処理され、特徴ベクトル $z_{low} \in \mathbb{R}^{8}$ が生成される。

4.5　特徴融合とランキング学習ヘッド

融合されたベクトル z_{fusion} は、2 つの隠れ層を持つ融合ヘッド（ MLP 分類器）に入力される：

- 最初の隠れ層は、次元を 424 から 128 に削減し、 BN、ReLU 、およびドロップアウトを使用する。
- 2 番目の隠れ層は、次元を 128 から 64 に削減し、同様に BN、ReLU、 およびドロップアウトを使用する。
- 出力層は、64 から $n_{classes}$ への線形マッピングを行い、各感情タグに対する関連度スコアを出力する。

モデルの学習には、ListNet[3] アルゴリズムを採用した。nDCG のようなランキング指標は直接微分が困難であるた

め、ListNet では予測スコアと正解ラベルをそれぞれ確率分布（Top-k 確率）として捉える。この 2 つの分布間のクロスエントロピー誤差を損失関数として最小化することで、リスト全体の順序を最適化し、結果として nDCG の最大化を図る。

4.6 埋め込み表現に基づくサブカテゴリ生成

提案モデル（QuadFusionNet）によって学習された埋め込み空間は、単なる分類のための平面だけでなく、楽曲間の聴感的な類似性を距離として保持する空間を形成する。本研究では、この特性を利用し、タグ検索の限界を克服するための「サブカテゴリ」生成手法を提案する。具体的なシステムの手順は以下の通りである。まず、ユーザーが指定したジャンル・感情条件に対し、モデルが出力する予測スコアに基づいて候補楽曲群をフィルタリングする。次に、これら候補楽曲の埋め込みベクトル集合に対し、K-means 法を用いて K 個のクラスタに分割する。この処理により生成された各クラスタは、巨視的なタグは同一でありながら、微視的な音響特徴が異なる「サブカテゴリ（Sub-category）」として定義される。最終的に、システムは各サブカテゴリの重心に位置する楽曲を代表曲として提示することで、ユーザーに「言葉にできないスタイルの違い」を直感的に選択させることを可能にする。

4.7 埋め込み空間の動的意味解釈

次元削減によって得られる埋め込み空間（PCA 軸）は、数学的に抽象的であり、ユーザーが楽曲配置の理由を直感的に理解することは困難である。この課題を解決するため、現在の視点における空間軸に対し、物理的な音響特徴の意味を動的にマッピングする「動的意味解釈レイヤー」を導入した。

具体的には、Essentia[2] を用いて抽出した音響特徴量（ダイナミクス、リズム、音色、複雑性など）と、現在の投影軸との間でピアソン相関係数 r をリアルタイムに計算する。システムは最も相関の高い特徴量をその軸の意味として特定し、例えば X 軸が「音色」と相関する場合、「Simple $\longleftrightarrow$ Complex」のような極性を持った言語ラベルを軸上に自動的に可視化する。これにより、ブラックボックスになりがちな特徴空間に対し、解釈性を与えることを目的とする。

5 実験設定

5.1 実験計画と手順

提案する QuadFusionNet の有効性を評価するため、以下の二段階の実験アプローチを採用する 。

1. **モデルの有効性検証：**第 1 段階では、公開ベンチマークデータセット（DEAM）を用いた 4 感情分類タスクを実施する 。ベースライン手法 [9] との比較分析およびアブレーションスタディを通じて、4 ブランチによる特徴融合の有効性を評価し、最適なモデル構成を特定する 。
2. **複雑なタスクへの適用：**第 2 段階では、検証されたモデルをより複雑なタスクへと拡張する。具体的には、12 種類の感情タグを持つ ME-14k データセットに対し、マルチラベルランキング学習を適用する。

5.2 データセット

本研究では 2 つのデータセットを利用する。標準化された比較のため、公開ベンチマークである DEAM [1] を 4 クラスの感情分類タスクに使用する。より複雑なタスクでの性能を評価するため、新たに構築した ME-14k を採用する。これには、12 クラスの感情タスクのための約 14,000 トラックが含まれている。ME-14k データセットは、訓練用（80%）と検証用（20%）に分割された。

5.3 特徴抽出プロトコル

一貫性を確保し、意味のある比較を可能にするため、研究全体を通じて統一された特徴抽出プロトコルが採用された。このプロトコルの構成は、後続のタスクで使用される ME-14k の主要なソースである Melon Playlist Dataset [5] の既存の特徴によって決定された。 DEAM ベンチマークでのモデル評価が直接比較可能で転移可能であることを保証するため、すべての DEAM 音声のメルスペクトログラムは、 Melon Playlist Dataset のパラメータを正確に再現するものを使用して抽出された。特徴抽出は、 Essentia[2] ライブラリ（v2.1b5.dev677）を使用し、以下の設定で行われた：サンプルレート 16 kHz 、フレームサイズ 512、ホップサイズ 256、 Hann ウィンドウ、および 48 メルフィルタバンク。

5.4 比較モデルとベースライン

DEAM の 4 クラスタスクにおける評価は、提案モデルとそのアブレーションバリアントを、発表済みのベースラインと比較することに焦点を当てる。

ベースライン ： Medina[9] らの研究で最も性能の高かったモデル。このモデルは、 MLP 分類器と SMOTETomek データバランシング戦略を組み合わせ、 Macro-F1 スコア 0.50 を達成したと報告されている。

QuadFusionNet ： 提案する 4 ブランチモデルで、4 つすべてのモダリティを統合する。

Model V1 ： 提案モデルのアブレーションバリアントで、MuQ 、ジャンル、および低レベル特徴のみを使用する。このバリアントは、メルスペクトログラムから学習された深層音響特徴の寄与を評価するために設計され ている。

Model V2 ： メルスペクトログラム、ジャンル、および低レベル特徴のみを使用するアブレーションバリアント。このバリアントは、 MuQ 埋め込みによって提供される事前学習済みの意味情報の寄与を評価することを目的としている。

Model V3 ： メルスペクトログラム、MuQ 埋め込み、および低レベル特徴のみを使用するアブレーションバリアント。このバリアントは、 ジャンルの感情コンテキストの寄与を評価することを目的としている。

この段階の主な目標は、後続のアプリケーションに最も効果的なモデル構成を特定することである。

5.5 実装詳細とハイパーパラメータ

すべてのモデルは PyTorch を使用して実装され、GPU で訓練された。各実験は、データ分割のための異なるランダムシードで 3 回繰り返され、結果は平均化された。共通の訓練設定。 ベースラインの研究 [9] の方法論に従い、データセットは層化アプローチを用いて訓練用（80%）とテスト用（20%）に分割され、訓練データの 20% は検証用に留保された。すべてのモデルは、バッチサイズ 32 で最大 50 エポック訓練された。最大ノルム 1.0 の勾配クリッピングが適用された。過学習を緩和するため、検証損失に関してペイシェンス 8 エポックの早期停止戦略が使用された。モデル固有の設定。 最適なパフォーマンスを確保するため、主要なハイパーパラメータと正則化戦略は、各モデル構成ごとに個別に調整された：

QuadFusionNet および Model V2, V3 の場合：

- オプティマイザ： AdamW (lr=3e-4, wd=1e-4).
- LR スケジューラ： ReduceLROnPlateau (factor=0.2, patience=3).
- 損失関数： クラス重み付きクロスエントロピー。 Model V2 は追加で 0.1 のラベルスムージングを使用した。
- ドロップアウト (p): CNN ブロックは 0.25; MuQ 、ジャンル、音声特徴ブランチは 0.5; 融合ヘッドの 2 つの層は 0.4 と 0.3。

Model V1 の場合：

- オプティマイザ： Adam (lr=1e-3, wd=1e-5).
- LR スケジューラ： OneCycleLR .
- 損失関数： クラス重み付きクロスエントロピー。
- ドロップアウト (p): MuQ 、ジャンル、音声特徴ブランチは 0.5; 融合ヘッドの層は 0.4 と 0.3。

5.6 評価指標

nDCG@3 において、ユーザーは提示されたリストの上位数件に注目する。そのため、全体の正解率よりも「上位にいかに適切な感情が含まれているか」が重要となる。本研究では、ランキング品質を評価する指標として nDCG@3 (Normalized Discounted Cumulative Gain at rank 3) を採用する。 nDCG は、正解アイテムが上位に出現するほど高い値となり、順位の重要性を考慮した実用的な指標である。

5.7 可視化およびサブカテゴリ生成の設定

提案モデルが学習した埋め込み空間の構造特性を定性的に評価するため、ME-14k データセットの全楽曲に対して推論を行い、埋め込みベクトルを抽出した。本実験の分析対象としては、詳細なスタイルの多様性が期待される「Jazz（ジャンル）」かつ「Romantic（感情タグ）」の条件を選択した。なお、特定のスタイルを持つ楽曲群を正確に抽出して空間構造を検証するため、候補のフィルタリングには正解ラベルを使用した。これは、分類誤差によるノイズを排除し、モデルが獲得した内部表現の質を直接的に検証するために不可欠である。クラスタリングには scikit-learn ライブラリの K-means 法を適用した。最適なクラスタ数 K を決定するため、本実験では K を 2 から 6 の範囲で探索し、生成されたクラスタの重心間の距離が最大となる K を採用した。この基準は、聴感上の特徴が最も明確に分離される粒度を自動的に特定するために設けられたものである。

表 1　ME-14k の 12 感情カテゴリにわたる分布

感情	カウント	割合 (%)
Relaxation	5459	37.38
Excitement	4792	32.81
Healing	3921	26.85
Romantic	3558	24.36
Happiness	3115	21.33
Nostalgia	3041	20.82
Fresh	2479	16.97
Quiet	2037	13.95
Touching	1928	13.20
Loneliness	1537	10.52
Sadness	1239	8.48
Missing	817	5.59

6 結果と分析

6.1 ME-14k の統計

ME-14k における感情カテゴリの分布を表 1 に示す。本データセットは全 14,606 トラックに対し、延べ 33,923 個のタグが付与されており、1 曲あたり平均約 2.32 個のタグを持つ密なマルチラベル構造となっている。分布を分析すると、クラス不均衡（ロングテール分布）が確認できる。例えば、ポジティブな感情である「 Relaxation 」は 5459 曲（全体の約 37.4%）、「 Excitement 」は 4792 曲（約 32.8%）と支配的であるのに対し、ネガティブな感情である「 Sadness 」（1239 曲、8.5%）や「 Missing 」（817 曲、5.6%）は少数派に留まっている。この不均衡は、従来の単純な分類モデルにおいては多数派クラスへの過学習を引き起こす課題となる。しかし、本研究が採用するランキング学習（ ListNet ）では、単にタグの有無を予測するのではなく、タグ間の相対的な重要度（順序）を学習対象とする。したがって、たとえ出現頻度の低い「 Missing 」のようなタグであっても、その楽曲にとって主要な感情であれば、リストの上位にランク付けすることが求められる。このデータ特性は、順位の妥当性を評価する nDCG 指標の採用と、分布全体の順序を最適化する ListNet の導入を動機付けるものである。

6.2 DEAM 4 クラスベンチマークにおけるモデル性能

表 2 に DEAM ベンチマークにおける性能比較を示す。QuadFusionNet は Macro-F1=0.542 を達成し、Baseline [9]（0.500）より高い値を示した。これは、複数モダリティを統合する設計が、本設定において一定の有効性を持つ可能性を示唆する。アブレーションの結果、以下の傾向が観察された。

表 2　DEAM 4 クラスタスクにおける全モデルの F1 スコア比較

Baseline [9]	0.720	0.310	0.620	0.340	0.500
QuadFusionNet	0.739	**0.429**	0.642	0.356	**0.542**
Model V1 (w/o Mel)	**0.761**	0.413	0.611	**0.376**	0.540
Model V2 (w/o MuQ)	0.655	0.390	**0.645**	0.240	0.485
Model V3 (w/o Genre)	0.697	0.443	0.608	0.327	0.518

- **MuQ の寄与：** Model V2（w/o MuQ）では Macro-F1 が 0.485 まで低下し、特に *Relaxed* での低下が大きい。MuQ による表現が、本タスクで有用な情報を含む可能性がある。
- **Genre の寄与：** Model V3（w/o Genre）は 0.518 であり、フルモデルより低い。DEAM の 4 象限分類ではジャンル情報の寄与は大きくない可能性があるが、補助的に働くケースがあると考えられる。
- **Mel-spectrogram の寄与：** Model V1（w/o Mel）は 0.540 とフルモデルに近い。MuQ が多くの音響・文脈情報を含む場合、Mel-CNN の寄与が一部重複する可能性がある。一方で、フルモデルが最良であることから、組み合わせにより追加的な利得が得られる可能性も残る。

全体として、 QuadFusionNet が感情カテゴリ全体で最もバランスの取れたパフォーマンスを示したため、 ME-14k データセットでの 12 クラス感情分類タスクの最終構成として選択された。

6.3　ME-14k におけるランキング性能

本節では、ME-14k を用いたマルチラベルランキングの評価結果を示す。本タスクと同一設定の先行比較が限定的であるため、音楽類似検索で用いられる metric learning 系手法を参考実装し、ベースラインとして比較した。

6.4　比較対象：Triplet Network baseline

Won ら [13] は、音楽タグ付け・類似検索において Triplet Network の有効性を報告している。本研究では、同手法に倣い、CNN エンコーダを用いてアンカー、ポジティブ、ネガティブの 3 つ組から Triplet Loss で学習し、推論時は埋め込み間のコサイン類似度でランキングを生成した。評価には nDCG@3 を用いた。

表 3　ME-14k におけるランキング性能（nDCG@3）

Method	**nDCG@3**
Random baseline	0.242
Baseline [13]	0.412
QuadFusionNet (ours)	**0.426**

6.5　結果と考察

表 3 より、提案手法は nDCG@3=0.426 を示し、Triplet baseline（0.412）およびランダム基準（0.242）より高い値であった。

この差の要因として、以下の点が考えられる。

1. **Metric learning の制約：** Triplet Loss は「近い／遠い」の関係学習に適している一方、複数タグの優先順位（強弱）やタグ間の共起構造を直接目的関数に含めにくい。例えば、Happiness と Excitement は別タグであるため、サンプリング次第では負例として扱われる可能性がある。
2. **Listwise 学習の利点：** 提案手法では、正解リストの順位構造を直接最適化する Listwise 損失（ListMLE）を用いる。これにより、上位タグと下位タグの相対関係を目的関数に取り込みやすく、nDCG@3 の向上につながった可能性がある。

6.6　サブカテゴリの聴感的妥当性の検証

提案モデルが学習した埋め込み空間の構造を解明するため、“Jazz × Romantic” 条件下の楽曲群に対して、提案する動的意味解釈機能（Dynamic Axis Interpretation）を適用した。

(1) 軸の意味的マッピング

解析の結果、第 1 主成分（X 軸）に対して「Timbre Variation（音色の変化）」および「High-Frequency（高域エネルギー）」との間に相関（$r \approx 0.64$）が検出された。これに基づき、システムは X 軸の左方向を「Complex / Crisp（複雑・鮮明）」、右方向を「Simple / Muffled（単純・暗め）」とする意味的ラベルを自動生成した。また、第 2 成分（Y 軸）には「Note Density（音符密度）」との相関（$r \approx 0.51$）が見られ、上下方向が「Sparse（疎）」対「Dense（密）」の構造を持つことが示唆された（図 5 参照）。

(2) クラスタの聴取分析による検証

この自動解析ラベルと、実際の楽曲の聴感印象が一致するかを検証するため、K-means 法（$K = 2$）によって分割されたクラスタの代表曲を試聴分析した。結果、システムが提示した軸の意味は、著者が行った以下の聴取分析の結果と合致することが確認された。

- **Cluster 1（左側領域・紫）：**システムは「Complex / Crisp（複雑・鮮明）」と判定した。実際の聴取において、このクラスタは「High Energy / Up-tempo」な楽曲群であり、明確なドラムのビートや、サックス等のリード楽器による明るい音色（Bright Timbre）が多く含まれていることが確認された。これは、高域成分が多く音色が複雑であるというシステムの解析結果と一致する。
- **Cluster 2（右側領域・黄）：**システムは「Simple / Muffled（暗め・単純）」と判定した。実際の聴取において、このクラスタは「Low Energy / Slow-tempo」であり、ゆったりとした旋律やマイナー調（短調）の雰囲気が特徴であった。また、ドラムが含まれないか、ブラシ奏法のような静かなリズムが中心であった。これは、高域が抑えられ（Muffled）、構成がシンプルであるというシステムの解析結果と一致する。

以上の結果は、提案モデルが明示的なタグ情報（Jazz, Romantic）には含まれない「楽器の奏法」や「音色の明暗」と

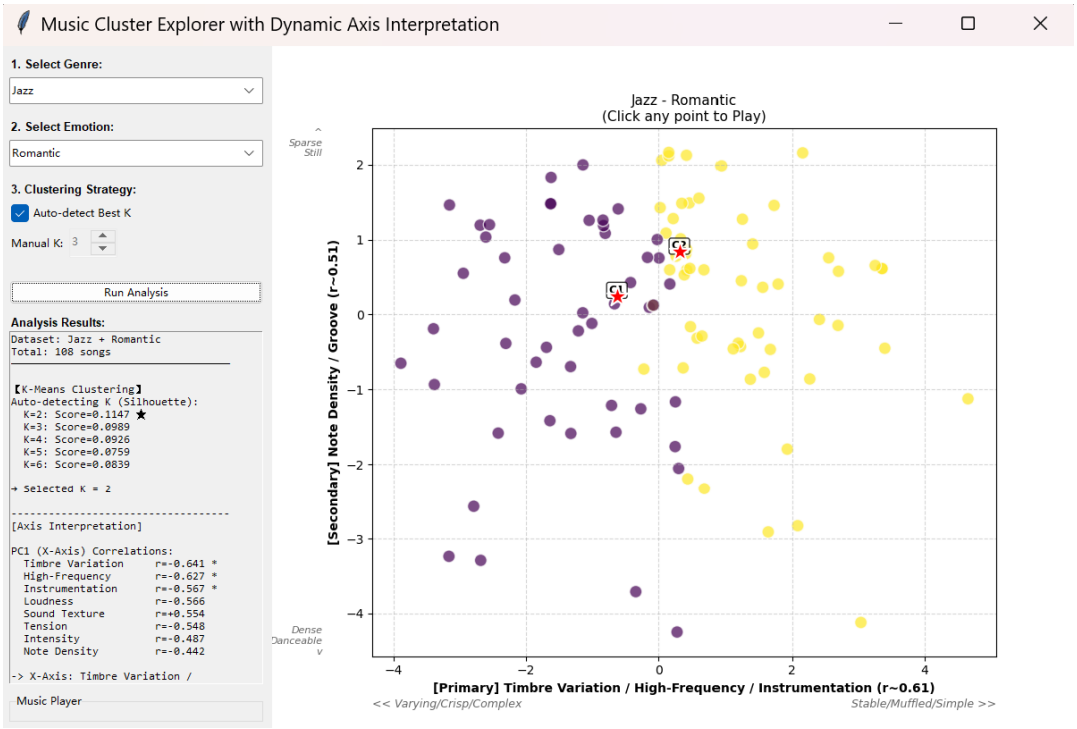

図 5　Jazz × Romantic におけるサブカテゴリ生成の可視化例 ($K = 2$)

いった微細な音響構造を潜在的に獲得していることを示す。さらに、提案 UI がその構造を適切な言語表現へと翻訳できており、ユーザーの探索を有効に支援しうることが実証された。

7　結論と今後の課題

本研究では、音楽感情認識 (MER) におけるデータの非商用依存や、単一ラベル分類による表現力の限界に取り組んだ。再現可能な商用コーパス ME-14k を構築し、ListNet を用いたマルチラベルランキング学習を適用した結果、nDCG@3 スコア 0.426 を達成し、ランダムベースラインを上回る精度を確認した。

さらに、ブラックボックスになりがちな埋め込み空間に対し、軸解性を与える動的軸解釈」機能を提案・実装した。「Jazz & Romantic」条件下のケーススタディにおいて、本システムは空間内の楽曲配置が「音色の変化」や「高域エネルギー」に基づいていることを自動的に特定し、「Crisp (鮮明)」対「Muffled (暗め)」といった直感的な言語ラベルとして可視化することに成功した。聴取実験の結果、これらのラベルは人間の知覚と一致しており、提案手法がタグ情報以上の音響構造を獲得・言語化できていることが実証された。

7.1　今後の課題

本研究の今後の主要な課題の一つは、現行システムを「検索」中心のインタラクションから、「推薦」へと発展させることである。現在のプロトタイプは、ユーザーがジャンルや感情タグを明示的に選択して候補集合を絞り込む設計であり、インタラクションの性質が検索に近い。これを推薦システムとして拡張するため、今後はユーザーの行動ログやフィードバックを取り込み、パーソナライズを行う機構の導入を検討する。具体的には、埋め込み空間上でユーザーが特定楽曲に対して「いいね」等のポジティブ反応を示した場合、その楽曲近傍に位置する未視聴楽曲を候補として優先的に提示する方策を実装する。これにより、ユーザーが言語化していない音響的嗜好を、探索過程のフィードバックから段階的に推定し、提示内容へ反映することを目指す。今後は、この機能拡張とユーザー評価を通じて、不完全なモデル推定をユーザーのフィードバックで補完できるか、ならびに探索過程を組み込んだ推薦が発見性や満足度に与える影響を検討する。

参考文献

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.

[2] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. 2013.

[3] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to

listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

[5] Andres Ferraro, Yuntae Kim, Soohyeon Lee, Biho Kim, Namjun Jo, Semi Lim, Suyon Lim, Jungtaek Jang, Sehwan Kim, Xavier Serra, et al. Melon playlist dataset: A public dataset for audio-based playlist generation and music tagging. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540. IEEE, 2021.

[6] Christine Hosey, Lara Vujović, Brian St. Thomas, Jean Garcia-Gathright, and Jennifer Thom. Just give me what i want: How people use and evaluate music search. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[7] Qiong Hu, Masrah Azrifah Azmi Murad, and Qi Li. Advancing music emotion recognition: large-scale dataset construction and evaluator impact analysis. *Multimedia Systems*, 31(2):1–16, 2025.

[8] Jaeyong Kang and Dorien Herremans. Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges. *IEEE Transactions on Affective Computing*, 2025.

[9] Yesid Ospitia Medina, José Ramón Beltrán, and Sandra Baldassarri. Emotional classification of music using neural networks with the mediaeval dataset. *Personal and Ubiquitous Computing*, 26(4):1237–1249, 2022.

[10] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 14(1):68–88, 2020.

[11] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[12] Philip Tovstogan, Xavier Serra, and Dmitry Bogdanov. Visualization of deep audio embeddings for music exploration and rediscovery. 2022.

[13] Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra. Multimodal metric learning for tag-based music retrieval. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 591–595. IEEE, 2021.

[14] Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025.