MDPI

Article

Classificazione delle Emozioni Musicali tramite Rappresentazioni Tempo-Frequenza e Reti Neurali Convoluzionali

Luca Zerella

Address: luc.zerella@stud.uniroma3.it

Abstract

Questo studio affronta la sfida del Music Emotion Recognition (MER), cercando di tradurre la complessa percezione emotiva della musica in un problema di classificazione automatica. L'obiettivo è stato quello di sviluppare un sistema in grado di riconoscere lo stato emotivo di un brano musicale attraverso l'analisi del segnale audio, sfruttando rappresentazioni tempo–frequenza e modelli di Deep Learning. In particolare, i brani sono stati classificati in quattro categorie fondamentali: Happy, Sad, Angry e Relaxed, ispirate al modello di Russell basato sulle dimensioni di Valence e Arousal. Utilizzando il dataset DEAM (Database for Emotional Analysis of Music), i segnali audio sono stati trasformati in Mel-spectrogrammi, una rappresentazione che consente di evidenziare pattern spettrali rilevanti e compatibili con l'impiego di reti neurali convoluzionali (CNN). Nell'ambito della sperimentazione sono state confrontate una CNN progettata ad hoc e un'architettura basata sul transfer learning tramite VGG16 (Visual Geometry Group 16-layer network), al fine di valutare l'impatto di modelli specializzati rispetto a reti pre-addestrate su domini differenti. I risultati ottenuti mostrano una buona capacità predittiva complessiva, tuttavia, l'analisi delle matrici di confusione evidenzia alcune difficoltà nella distinzione tra emozioni adiacenti nello spazio di Russell, in particolare per stati caratterizzati da livelli di energia simili. Questo comportamento suggerisce una parziale sovrapposizione delle feature acustiche associate a determinate emozioni, confermando la complessità intrinseca del problema. Il flusso di lavoro è stato ulteriormente arricchito da un confronto con un approccio basato sui coefficienti MFCC (Mel-Frequency Cepstral Coefficients), che ha fornito risultati intermedi, e da una validazione del sistema su campioni reali estratti da YouTube. Tali esperimenti dimostrano la solidità della pipeline proposta e la sua capacità di riconoscere pattern emotivi anche al di fuori del dataset di addestramento, pur mettendo in evidenza l'elevata soggettività che caratterizza il dominio musicale. Nel complesso, questo lavoro conferma il potenziale delle tecniche di Deep Learning applicate all'elaborazione del segnale audio per il MER, offrendo una base solida per futuri sviluppi e approfondimenti.

Keywords: Music Emotion Recognition; Reti Neurali Convoluzionali; Mel-spectrogrammi; MFCC; Deep Learning; Modello di Russell; Transfer Learning; Elaborazione del segnale audio.

Received:
Revised:
Accepted:
Published:

1. Introduzione

La musica è spesso definita un linguaggio universale: attraverso ritmo, armonia, melodia e parole è in grado di evocare stati d'animo complessi e profondamente soggettivi [1]. Questa ricchezza espressiva, tuttavia, non è immediatamente accessibile ai sistemi di

calcolo, per i quali un brano musicale si riduce a una sequenza di valori numerici privi di significato semantico. Il campo del MER nasce proprio con l'obiettivo di colmare questo divario, cercando di tradurre il segnale audio grezzo in una rappresentazione capace di riflettere la percezione emotiva umana [2]. Negli ultimi anni, il MER ha conosciuto un notevole sviluppo grazie all'avvento del Deep Learning e all'impiego di rappresentazioni tempo–frequenza, in grado di catturare pattern acustici complessi e altamente informativi [3]. Tuttavia, uno degli ostacoli principali rimane la forte componente soggettiva dell'emozione musicale: ciò che per un ascoltatore può risultare triste, per un altro può evocare sensazioni completamente diverse [4]. Per rendere il problema affrontabile in modo sistematico, questo studio adotta il modello circomplesso di Russell, un approccio ampiamente validato che permette di descrivere le emozioni non tramite etichette arbitrarie, ma attraverso due dimensioni fondamentali [5]. In particolare, la dimensione di Valence esprime il grado di piacevolezza di un'emozione, collocandosi su un asse che va da stati negativi, come tristezza o rabbia, a stati positivi, come felicità e serenità. La dimensione di Arousal, invece, descrive l'intensità energetica dell'emozione, distinguendo stati di calma da condizioni di elevata agitazione o tensione. L'incrocio di questi due assi genera quattro quadranti principali, che in questo lavoro sono stati utilizzati per organizzare il dataset DEAM in altrettante categorie emotive: Happy, Sad, Angry e Relaxed [6]. Dal punto di vista tecnologico, le CNN hanno dimostrato una notevole efficacia anche nell'elaborazione di segnali audio, a condizione che questi vengano trasformati in rappresentazioni adatte [7]. In questo studio, i segnali musicali sono stati convertiti in Mel-spectrogrammi e MFCC utilizzando la libreria Librosa, permettendo di sfruttare le capacità di apprendimento delle CNN su strutture visive che riflettono il contenuto spettrale del suono [8] [9]. Il cuore della ricerca risiede nel confronto tra una CNN progettata specificamente per questo compito e un'architettura di riferimento nel campo della Computer Vision, la VGG16, impiegata secondo il paradigma del transfer learning [10]. Questo confronto permette di indagare se un modello nato per riconoscere oggetti e texture visive possa effettivamente adattarsi a cogliere l'emozione di un brano musicale [11]. Infine, per valutare la robustezza e l'applicabilità del sistema in contesti reali, i modelli sono stati testati anche su brani acquisiti da YouTube, verificandone il comportamento in scenari d'uso quotidiano.

2. Obiettivi

Il fulcro di questo lavoro è la progettazione di un sistema in grado di interpretare autonomamente il contenuto emotivo di un brano musicale. Nel contesto del MER, la sfida principale non risiede soltanto nell'analisi del segnale audio, ma soprattutto nel tentativo di tradurre la natura profondamente soggettiva dell'emozione musicale in una rappresentazione quantitativa e condivisibile [2] [1] [4]. Per affrontare questa complessità, è stato adottato il modello circomplesso di Russell, che consente di mappare le emozioni all'interno di uno spazio bidimensionale definito dalle dimensioni di Valence e Arousal [5]. Per trasformare questa impostazione teorica in un sistema concreto e funzionale, il percorso di ricerca è stato articolato in una serie di fasi fondamentali, ciascuna delle quali ha contribuito in modo significativo alla costruzione della pipeline finale. In una prima fase, è stata condotta un'accurata Exploratory Data Analysis (EDA), indispensabile per comprendere la struttura del dataset e le caratteristiche acustiche dominanti dei segnali audio. Questa analisi preliminare ha permesso di osservare come i campioni si distribuissero nello spazio emotivo definito dal modello di Russell, fornendo indicazioni preziose sulla separabilità delle classi e garantendo una base solida per le successive fasi di classificazione. Successivamente, sono state formalizzate le categorie emozionali, mappando i brani all'interno dello spazio di Valence e Arousal e definendo in modo chiaro i target della classificazione. Questa scelta ha consentito di ridurre l'ambiguità tipica delle etichette emotive, rendendo

il problema più trattabile dal punto di vista computazionale. Una volta definite le classi, è stata implementata una pipeline di preprocessing dedicata alla pulizia e alla preparazione dei dati. Tale pipeline ha incluso la normalizzazione dei segnali audio e la codifica delle etichette, rendendo il dataset compatibile con le architetture di Deep Learning adottate nello studio. Un passaggio chiave del lavoro è stato rappresentato dall'estrazione delle rappresentazioni tempo–frequenza. L'audio grezzo è stato trasformato in Mel-spectrogrammi e MFCC, due descrittori ampiamente utilizzati che permettono di catturare in modo efficace le caratteristiche timbriche e temporali del segnale musicale, rendendole interpretabili da modelli convoluzionali [9] [8]. Il cuore del sistema è costituito da una CNN progettata ad hoc, sviluppata e addestrata da zero con l'obiettivo di apprendere pattern distintivi direttamente dagli spettrogrammi. Accanto a questo approccio, è stata esplorata anche la strategia di transfer learning, adattando un'architettura pre-addestrata su larga scala come la VGG16 al dominio audio, al fine di valutarne l'efficacia nel riconoscimento delle emozioni musicali. Per estendere l'applicabilità del sistema oltre il contesto sperimentale, il progetto è stato arricchito con un'integrazione dedicata all'analisi di brani acquisiti direttamente da YouTube, consentendo di testare la capacità di generalizzazione dei modelli su contenuti reali e non pre-elaborati. Infine, particolare attenzione è stata dedicata all'interpretabilità dei risultati, attraverso lo sviluppo di visualizzazioni e strumenti diagnostici interattivi. Queste soluzioni hanno facilitato l'analisi degli errori di classificazione e la comprensione delle decisioni prese dalla rete neurale, migliorando al contempo l'usabilità complessiva del sistema. Nel complesso, questo lavoro propone una pipeline completa e replicabile per la classificazione automatica delle emozioni musicali, che combina metodologie consolidate presenti in letteratura con estensioni sperimentali orientate a massimizzare sia l'accuratezza sia l'interpretabilità dei risultati.

3. Materiali e metodi

Il presente capitolo descrive il percorso metodologico seguito per la realizzazione del sistema di MER, illustrando in modo progressivo le scelte e le procedure che hanno permesso di trasformare un segnale sonoro grezzo in una predizione emotiva. L'obiettivo è accompagnare il lettore attraverso le diverse fasi del processo, chiarendo il ruolo di ciascun componente all'interno della pipeline complessiva. In apertura, viene presentato il dataset DEAM, che costituisce il nucleo informativo del progetto. Verranno analizzate le caratteristiche dei brani musicali che lo compongono e descritte le operazioni di preprocessing necessarie per pulire, normalizzare e uniformare i dati audio, rendendoli idonei alle successive fasi di analisi e modellazione [7] [8]. Successivamente, l'attenzione si sposta sulle tecniche adottate per la rappresentazione tempo–frequenza del segnale audio. In questa sezione vengono approfondite le motivazioni che hanno portato alla scelta dei Mel-spectrogrammi e dei MFCC, illustrando come queste rappresentazioni consentano di estrarre informazioni rilevanti dal punto di vista percettivo ed emotivo [9] [1]. Il capitolo prosegue con la descrizione del cuore del sistema, ovvero la CNN utilizzata per il processo di apprendimento automatico. Verranno illustrati i principi alla base dell'architettura adottata e i principali passaggi delle fasi di addestramento e validazione, evidenziando le scelte che hanno guidato lo sviluppo del modello [3] [7]. Infine, il capitolo si conclude con l'analisi dei criteri di valutazione impiegati per misurare le prestazioni del sistema. Attraverso metriche quantitative e strumenti diagnostici, verrà discusso quanto il modello sia effettivamente in grado di cogliere e interpretare le sfumature emotive presenti nella musica [12] [2].

3.1. Dataset e Preprocessing

Il cuore di questo progetto è il dataset DEAM, una risorsa di riferimento per lo studio dell'emozione nella musica [6]. Il dataset mette a disposizione una vasta collezione di brani musicali annotati tramite valori numerici che descrivono il contenuto emotivo lungo i due assi del modello circomplesso di Russell: Valence, che esprime il grado di positività o negatività dell'emozione, e Arousal, che ne rappresenta l'intensità energetica [5]. Per rendere il processo di addestramento più efficace e controllabile, non si è lavorato sui brani nella loro interezza. Come implementato, ogni traccia audio è stata caricata e standardizzata a una frequenza di campionamento pari a 22.050 Hz, un valore comunemente utilizzato nell'elaborazione audio perché garantisce un buon compromesso tra qualità del segnale e complessità computazionale [8]. Successivamente, ciascun brano è stato segmentato in frammenti di 5 secondi. Questa scelta si è rivelata fondamentale per due motivi principali. Da un lato, permette alla rete neurale di concentrarsi su pattern acustici locali, evitando di dover apprendere sequenze troppo lunghe e complesse; dall'altro, aumenta significativamente il numero di campioni disponibili, assicurando che tutti i dati abbiano durata e formato omogenei, condizione essenziale per un addestramento stabile [7]. Una delle fasi più delicate del pre-processing ha riguardato la trasformazione delle annotazioni emotive, originariamente espresse come valori continui di Valence e Arousal, in etichette categoriali. A questo scopo è stata adottata una strategia basata sui percentili (33,33° e 66,66°), suddividendo ciascun asse in tre fasce: Basso, Medio e Alto. In questa fase è stata presa una decisione strategica, documentata anche nel codice: l'eliminazione della fascia intermedia ("Mid" o neutra) per entrambi gli assi. Questa scelta è stata motivata dalla volontà di ridurre l'ambiguità emotiva dei campioni. Le emozioni centrali, infatti, risultano spesso meno definite e possono introdurre confusione nel processo di apprendimento. Concentrandosi esclusivamente sugli stati emotivi più marcati, la CNN è stata messa nelle condizioni di apprendere caratteristiche acustiche più distintive e robuste. Incrociando le fasce rimanenti di Valence e Arousal, sono stati così ottenuti i quattro quadranti emotivi fondamentali: Happy, Sad, Angry e Relaxed. Il passaggio successivo, cruciale per l'applicazione delle CNN, è stato la trasformazione dei segnali audio in rappresentazioni visive. Ogni frammento è stato convertito in un Mel-spectrogramma, successivamente ridimensionato a una matrice 128×128 pixel. Per garantire uniformità dimensionale, sono state applicate operazioni di padding o cropping quando necessario. La scelta della scala Mel non è casuale: essa riproduce la percezione umana delle frequenze sonore, rendendo le rappresentazioni più significative dal punto di vista percettivo ed emotivo [9] [1]. Un aspetto particolarmente critico del preprocessing riguarda la suddivisione del dataset in training set e test set. Poiché ogni brano è stato segmentato in più frammenti, esisteva un rischio concreto di data leakage: se segmenti della stessa canzone fossero finiti in entrambi i set, il modello avrebbe potuto semplicemente riconoscere il brano, invece di apprendere pattern emotivi generalizzabili [11]. Per evitare questo problema, è stata adottata la tecnica del GroupShuffleSplit, raggruppando i campioni in base all'ID della canzone originale. In questo modo, tutti i frammenti di uno stesso brano sono stati assegnati esclusivamente al training o al test set, garantendo una valutazione più realistica delle prestazioni del modello. Infine, è stato affrontato il problema dello sbilanciamento delle classi, dovuto alla diversa frequenza delle emozioni nel dataset. Per prevenire una tendenza del modello a favorire le classi più rappresentate, sono stati introdotti dei pesi di classe utilizzando la funzione compute_class_weight della libreria scikit-learn. Assegnando un peso maggiore alle classi meno frequenti, il modello è stato incentivato a trattare in modo equo tutti i quadranti emotivi. Il dataset finale è stato quindi organizzato in una matrice di input X, contenente i Mel-spectrogrammi, e in un vettore di etichette y. Le etichette sono state codificate tramite LabelEncoder e successivamente trasformate in one-hot encoding, rendendo i

dati pienamente compatibili con lo strato di output softmax della CNN, che restituisce una distribuzione di probabilità sulle quattro classi emotive [3] [7].

3.2. Rappresentazioni Tempo-Frequenza

Le rappresentazioni tempo-frequenza costituiscono uno strumento imprescindibile nell'elaborazione dei segnali audio, poiché consentono di descrivere simultaneamente l'evoluzione temporale e il contenuto spettrale di un segnale [9] [1]. In un compito complesso come il MER, l'analisi diretta della forma d'onda risulta insufficiente: l'informazione emotiva non è contenuta nei valori di ampiezza in sé, ma nelle strutture armoniche, timbriche e dinamiche che caratterizzano un brano musicale [2] [4]. Per questo motivo, è necessaria una scomposizione più profonda del segnale nel dominio tempo-frequenza. In questo lavoro l'attenzione si è concentrata su due rappresentazioni fondamentali: il Mel-spectrogram e i MFCC, entrambe ampiamente validate in letteratura per l'analisi audio-emotiva [12][2]. Il Mel-spectrogram è una rappresentazione bidimensionale del segnale audio ottenuta applicando la Short-Time Fourier Transform (STFT) e proiettando le frequenze risultanti sulla scala Mel, una scala percettiva che approssima il funzionamento dell'udito umano. In particolare, l'orecchio è molto più sensibile alle variazioni nelle basse frequenze rispetto a quelle alte, ed è proprio questa caratteristica che rende il Mel-spectrogram particolarmente adatto allo studio delle emozioni musicali [1]. Grazie alla sua struttura visiva, il Mel-spectrogram si presta in modo naturale all'utilizzo con le CNN [7]. In questo progetto è stato utilizzato come input principale, permettendo alla rete di analizzare lo spettrogramma come un'immagine e di individuare pattern complessi di energia, armonia e dinamica che si sviluppano nel tempo [3]. La risoluzione adottata, pari a 128×128 pixel, rappresenta un compromesso efficace tra ricchezza informativa e compatibilità con l'architettura convoluzionale, evitando al contempo un eccessivo carico computazionale [7]. Accanto ai Mel-spectrogrammi, è stata condotta una sperimentazione parallela utilizzando i MFCC, estratti in questo caso nella configurazione a 40 coefficienti. Gli MFCC forniscono una descrizione più compatta del segnale, ottenuta applicando la Trasformata Discreta del Coseno (DCT) al logaritmo delle energie delle bande Mel [9]. Questa rappresentazione, pur comportando una compressione delle informazioni spettrali e temporali, è in grado di catturare efficacemente le caratteristiche timbriche globali del suono [12][2]. Nel contesto sperimentale, i 40 MFCC sono stati utilizzati come input per una CNN dedicata, consentendo di valutare il comportamento del modello anche su feature più compatte e meno ridondanti [11]. I risultati hanno mostrato che, sebbene gli MFCC permettano prestazioni soddisfacenti e una maggiore efficienza computazionale, la perdita di dettaglio rispetto ai Mel-spectrogrammi limita la capacità del modello di cogliere le sfumature emotive più sottili, soprattutto in presenza di emozioni acusticamente simili [4]. Per questo motivo, nel corso della sperimentazione, è stato dato maggiore rilievo ai Mel-spectrogrammi come rappresentazione principale, mentre gli MFCC sono stati utilizzati come termine di confronto per valutare l'impatto della compressione delle feature sulle prestazioni della rete. Nel complesso, l'adozione di queste rappresentazioni ha permesso di trasformare il problema del riconoscimento emotivo musicale in un vero e proprio problema di Computer Vision. Invece di fornire alla macchina una definizione esplicita di concetti astratti come felicità o tristezza, le sono stati presentati dei veri e propri scatti visivi del suono, lasciando alla rete neurale il compito di apprendere autonomamente quali configurazioni di frequenza e tempo corrispondessero ai quattro quadranti emotivi del modello di Russell [5] [2].

3.3. Architettura della Rete Neurale

Per la classificazione delle emozioni musicali a partire dai Mel-spectrogrammi, è stata progettata e implementata una CNN profonda. Questa scelta nasce dalla naturale affinità delle CNN con dati bidimensionali strutturati, che nel nostro caso assumono la forma di immagini tempo-frequenza. I Mel-spectrogrammi, infatti, possono essere interpretati come vere e proprie rappresentazioni visive del suono, rendendo le CNN particolarmente efficaci nell'estrarne pattern informativi [3]. L'architettura adottata segue un approccio sequenziale, articolato in tre blocchi principali di estrazione delle caratteristiche, seguiti da una fase finale di classificazione. Il modello riceve in input matrici di dimensione (128, 256, 1), dove 128 rappresenta il numero di bande Mel (risoluzione in frequenza) e 256 i passi temporali, corrispondenti ai segmenti audio di 5 secondi utilizzati nella fase di preprocessing. La rete è strutturata in una gerarchia di tre blocchi convoluzionali, progettati per incrementare progressivamente la complessità delle informazioni apprese:

- Primo blocco: utilizza 32 filtri convoluzionali di dimensione 3×3, con il compito di individuare le caratteristiche locali di base dello spettrogramma, come variazioni di energia e pattern spettrali semplici. L'operazione di convoluzione è seguita da una funzione di attivazione ReLU e da uno strato di Batch Normalization, che contribuisce a stabilizzare il processo di apprendimento e a velocizzare la convergenza [3]. Infine, uno strato di MaxPooling (2×2) riduce la dimensionalità spaziale, mantenendo solo le attivazioni più rilevanti.
- Secondo blocco: aumenta il numero di filtri a 64, permettendo alla rete di combinare le feature elementari in pattern più complessi [3]. La struttura del blocco ricalca quella del precedente, garantendo continuità nel processo di estrazione e riducendo il rischio di problemi legati al gradiente.
- Terzo blocco: a questo livello di profondità, la CNN è in grado di catturare rappresentazioni ad alto livello, come strutture ritmiche più articolate e sfumature timbriche complesse. Queste informazioni risultano fondamentali per distinguere emozioni che occupano regioni adiacenti nello spazio di Valence e Arousal [3] e che presentano caratteristiche acustiche parzialmente sovrapposte.

Terminata la fase di estrazione delle caratteristiche spaziali, il modello entra nella fase decisionale. Le mappe di attivazione vengono prima trasformate in un vettore monodimensionale tramite uno strato di Flatten, rendendole compatibili con i livelli completamente connessi. Questo vettore viene poi elaborato da uno strato Dense composto da 128 neuroni, che ha il compito di interpretare e combinare le feature apprese nei livelli precedenti. Anche in questo caso viene utilizzata l'attivazione ReLU, per introdurre non linearità nel processo decisionale. Per contrastare il rischio di overfitting, è stato inserito uno strato di Dropout con una probabilità del cinquanta percento. Questa scelta si è rivelata particolarmente importante considerando la natura soggettiva delle annotazioni emotive del dataset DEAM [6]. Il dropout costringe la rete a non affidarsi eccessivamente a specifiche attivazioni, favorendo l'apprendimento di rappresentazioni più robuste e generalizzabili. L'ultimo livello del modello è l'output layer, composto da 4 neuroni con funzione di attivazione Softmax. Questo strato restituisce una distribuzione di probabilità sulle quattro classi emotive considerate (Happy, Angry, Sad e Relaxed), permettendo di identificare la classe dominante in modo probabilistico. Il modello è stato infine compilato seguendo criteri di stabilità e precisione. L'ottimizzatore Adam è stato configurato con un learning rate ridotto pari a 0.0001, una scelta che consente una discesa del gradiente più graduale e controllata [3], evitando convergenze premature verso minimi locali sub-ottimali. Come funzione di Loss è stata utilizzata la Categorical Cross-Entropy, standard per problemi di classificazione multiclasse con etichette in one-hot encoding [3], in grado di penalizzare in modo efficace le discrepanze tra le probabilità predette dal modello e le etichette reali del dataset.

3.4. Fasi di Addestramento

L'addestramento della CNN è stato condotto seguendo un protocollo strutturato in quattro fasi principali: partizionamento del dataset, configurazione dei parametri, addestramento e valutazione finale. Questo approccio ha permesso di controllare in modo sistematico ogni passaggio del processo, garantendo risultati affidabili e riproducibili. In una prima fase, il dataset è stato suddiviso in due sottoinsiemi indipendenti: l'80% dei dati è stato utilizzato per l'addestramento, mentre il restante 20% è stato riservato al test. Una delle principali criticità emerse in questa fase riguarda la distribuzione non uniforme delle classi emotive all'interno del dataset DEAM. Alcune emozioni risultano infatti più rappresentate di altre, con il rischio che il modello impari a favorire le classi più frequenti a discapito di quelle meno presenti. Per mitigare questo problema, durante l'addestramento è stata adottata la tecnica dei class weights, integrata direttamente nella funzione di training. Assegnando un peso maggiore alle classi meno rappresentate, la CNN è stata costretta a dedicare un'attenzione bilanciata a tutti i quadranti emotivi (Happy, Angry, Sad e Relaxed) [5]. Questa strategia si è rivelata fondamentale per migliorare la capacità del modello di riconoscere correttamente anche le emozioni più rare, evitando una classificazione sbilanciata. L'addestramento è stato eseguito per 20 epoche, con un batch size di 32, valori che rappresentano un buon compromesso tra stabilità dell'apprendimento e tempi computazionali. Per monitorare l'evoluzione del modello durante il training, è stato utilizzato un validation set, pari al 20% dei dati di addestramento. Attraverso la variabile history, è stato possibile osservare l'andamento dell'accuratezza e della funzione di perdita sia sui dati di training sia su quelli di validazione, epoca dopo epoca. Questo monitoraggio continuo ha svolto un ruolo cruciale nella prevenzione dell'overfitting, consentendo di verificare che la rete non si limitasse a memorizzare i campioni di addestramento, ma stesse effettivamente apprendendo caratteristiche generali del segnale audio, come il timbro, il ritmo e la distribuzione energetica nel tempo. Al termine delle 20 epoche, il modello è stato sottoposto a una valutazione finale sul test set, costituito esclusivamente da dati mai visti durante l'addestramento. Oltre all'accuratezza complessiva, particolare attenzione è stata dedicata all'analisi della matrice di confusione, uno strumento fondamentale per comprendere nel dettaglio il comportamento del classificatore. L'osservazione della confusion matrix ha permesso di analizzare come le predizioni del modello si distribuiscono rispetto alle classi reali, evidenziando sia i punti di forza sia le aree di maggiore incertezza. In particolare, l'analisi ha mostrato che il modello tende a confondere più facilmente emozioni caratterizzate da livelli di energia simili nello spazio di Russell, come la rabbia e la felicità, o stati emotivi più sfumati come Relaxed. Nel complesso, i risultati confermano la capacità della CNN di cogliere molte delle sfumature acustiche necessarie per il riconoscimento delle emozioni musicali [2]. Tuttavia, l'analisi degli errori mette in luce alcune criticità residue, soprattutto nella distinzione tra le classi Relaxed e Angry, suggerendo la necessità di strategie future più raffinate per separare emozioni con profili acustici parzialmente sovrapposti.

4. Estensioni Sperimentali

Per mettere davvero alla prova la robustezza del sistema e valutarne l'efficacia in contesti realistici, il progetto è stato progressivamente arricchito con una serie di estensioni che vanno oltre il semplice addestramento di una CNN standard. Una delle prime implementazioni ha riguardato la gestione del salvataggio e del ripristino del modello (model.save), una scelta fondamentale per garantire la replicabilità degli esperimenti e consentire l'utilizzo della rete senza dover ripetere ogni volta il lungo processo di training. In parallelo, è stata condotta un'analisi comparativa tra diverse rappresentazioni del segnale audio. Accanto ai Mel-spectrogrammi, sono stati estratti anche i MFCC, con l'obiettivo

di capire quale di queste tecniche fosse in grado di fornire una sintesi più efficace delle informazioni rilevanti per il riconoscimento delle emozioni musicali. Un ulteriore passo in avanti è stato l'esplorazione del Transfer Learning tramite il modello pre-addestrato VGG16. Nonostante sia un'architettura nata per la visione artificiale, VGG16 è stata adattata all'analisi degli spettrogrammi audio, trattati come immagini [10]. Questo ha permesso di verificare se la conoscenza visiva appresa su grandi dataset potesse essere sfruttata anche per individuare pattern musicali complessi, soprattutto in presenza di un dataset audio di dimensioni limitate [11]. Un vero punto di svolta in termini di interattività è stato raggiunto con l'integrazione dell'API di YouTube. Grazie a questa estensione, il sistema non è più vincolato all'uso di file audio locali, ma può attingere direttamente all'enorme catalogo musicale della piattaforma, ampliando notevolmente le possibilità di utilizzo e sperimentazione. Per rendere i risultati più chiari e intuitivi, sono stati introdotti strumenti di ascolto diretto e visualizzazioni grafiche avanzate. In particolare, la waveform (forma d'onda) riveste un ruolo centrale: mostrando l'andamento dell'ampiezza del segnale nel tempo, consente di individuare visivamente attacchi, pause e variazioni dinamiche. Questo crea un collegamento immediato tra ciò che l'utente ascolta e i pattern che la rete neurale sta analizzando e interpretando. Il lavoro si è infine concentrato su un'analisi approfondita dei casi di errata classificazione, come la confusione emersa tra le classi Relaxed e Angry. L'obiettivo non è stato semplicemente quantificare gli errori, ma comprenderne le cause, studiando in dettaglio i campioni che hanno tratto in inganno la rete. Questa analisi ha permesso di individuare possibili direzioni di miglioramento future, come l'aumento della risoluzione temporale o l'introduzione di nuove caratteristiche legate al ritmo, che verranno approfondite nella sezione conclusiva.

4.1. Integrazione con l'API di YouTube e analisi di casi studio

L'integrazione con la piattaforma YouTube ha rappresentato un vero salto di qualità per il progetto, rendendo possibile testare il modello su brani reali, al di fuori dei tradizionali dataset statici. Grazie all'utilizzo delle YouTube Data API v3, il sistema è in grado di cercare automaticamente i brani desiderati e recuperarne i principali metadati. Una volta estratto l'audio, il segnale viene normalizzato e trasformato in un Mel-spectrogramma, assicurando la piena coerenza con il formato dei dati utilizzati durante la fase di addestramento della rete neurale. Per verificare concretamente l'efficacia del sistema, sono stati analizzati tre brani iconici, selezionati per le loro marcate differenze in termini di dinamica, timbro ed espressività emotiva:

- "Happy" di Pharrell Williams: il modello ha classificato correttamente il brano all'interno del quadrante Happy. La CNN ha saputo riconoscere la brillantezza timbrica e la regolarità del pattern ritmico, associando questi elementi a un'elevata probabilità di uno stato emotivo positivo ed energico [5] [2].
- "Someone Like You" di Adele: utilizzato come esempio per il quadrante Sad, questo brano ha messo in evidenza la sensibilità dei filtri convoluzionali all'assenza di percussioni e alla predominanza di frequenze medie morbide. Il modello ha individuato con successo la bassa energia complessiva e il carattere malinconico dell'arrangiamento [5] [2].
- "Thunderstruck" degli AC/DC: questo caso si è rivelato il più interessante dal punto di vista analitico. Nonostante l'elevata energia percepita dal punto di vista umano, il modello ha classificato il brano come Relaxed. Questo risultato apparentemente paradossale suggerisce che il celebre riff iniziale, pur essendo veloce, presenta una struttura armonica estremamente pulita e ripetitiva, che la rete neurale ha interpretato come un pattern stabile e rassicurante, associandolo quindi a una bassa tensione emotiva.

L'analisi di questi casi studio consente una validazione non solo quantitativa, ma anche qualitativa del sistema. Questo approccio mette in luce come, nonostante l'elevata efficacia della CNN, la percezione emotiva della musica sia fortemente legata a sottili variazioni temporali, che rappresentano ancora oggi una delle principali sfide per i futuri sviluppi del sistema.

4.2. Confronto tra Mel-Spectrogram e MFCC

Nel corso della sperimentazione è emerso con chiarezza quanto la scelta della rappresentazione del segnale audio sia un fattore determinante per le prestazioni complessive del modello [2]. Per questo motivo è stato condotto un confronto sistematico tra le due tecniche più utilizzate nel campo del MER: il Mel-spectrogramma e i MFCC. L'obiettivo era capire quale delle due consentisse alla rete neurale di estrarre in modo più efficace le informazioni necessarie a rappresentare la complessità emotiva dei brani musicali. Il Mel-spectrogramma fornisce una rappresentazione spettro-temporale molto ricca, in cui l'intensità delle diverse frequenze viene tracciata nel tempo e proiettata sulla scala Mel, pensata per imitare la percezione non lineare dell'orecchio umano [2]. In questo lavoro, l'utilizzo di 128 bande Mel ha permesso di ottenere una vera e propria "mappa termica" del suono, capace di descrivere in modo dettagliato l'evoluzione energetica del segnale. Questa forma di rappresentazione si integra in modo naturale con le architetture CNN [7], poiché trasforma il problema dell'analisi audio in un compito assimilabile alla visione artificiale. La rete non lavora più su semplici sequenze numeriche, ma su immagini ricche di texture acustiche, dalle quali può apprendere pattern spaziali complessi legati al ritmo, al timbro e alla dinamica del brano. Le MFCC, al contrario, offrono una descrizione del segnale molto più compatta. Ottenute applicando la DCT alle bande Mel, queste feature si concentrano principalmente sull'inviluppo spettrale, sintetizzando i tratti più caratteristici del timbro [9]. Nei test condotti con 40 coefficienti, questa scelta ha comportato una drastica riduzione della dimensionalità dei dati e, di conseguenza, del carico computazionale. L'analisi diagnostica ha però evidenziato come questa maggiore efficienza abbia un costo in termini di contenuto informativo. La forte compressione del segnale tende infatti a smussare o perdere quelle variazioni temporali e dinamiche che risultano cruciali per la corretta interpretazione delle emozioni musicali. I risultati sperimentali hanno quindi confermato il Mel-spectrogramma come la soluzione più adatta quando l'obiettivo principale è massimizzare la precisione del sistema. La sua capacità di preservare la struttura profonda del brano consente alla CNN di operare su una base informativa completa e più rappresentativa della realtà musicale. Le MFCC restano comunque una valida alternativa in contesti con risorse hardware limitate o per compiti di classificazione meno fini, come la distinzione tra generi musicali molto diversi tra loro. Tuttavia, nel contesto del MER, si sono dimostrate meno efficaci nel catturare le sfumature emotive più sottili. In definitiva, la superiorità del modello basato su Mel-spectrogrammi conferma la scelta di privilegiare la ricchezza del dato rispetto alla velocità di calcolo, garantendo una maggiore robustezza e affidabilità del classificatore finale.

4.3. Transfer Learning con VGG16

Accanto alla CNN addestrata end-to-end, è stato esplorato anche un approccio basato sul transfer learning, utilizzando la rete VGG16 pre-addestrata sul dataset ImageNet [10]. L'obiettivo di questo esperimento era capire fino a che punto un modello nato per la visione artificiale potesse adattarsi all'analisi del dominio audio, rappresentato tramite descrittori tempo-frequenza come gli spettrogrammi. Poiché VGG16 richiede input tridimensionali in formato RGB, le rappresentazioni audio mono-canale sono state opportunamente adattate per rispettare le specifiche dell'architettura, cercando al contempo di preservare il contenuto

spettrale originale. Il modello pre-addestrato è stato quindi utilizzato come estrattore di feature, mantenendo congelati i livelli convoluzionali e addestrando soltanto il classificatore finale, progettato specificamente per il numero di classi emozionali considerate. Questa strategia permette di sfruttare la profondità e la stabilità di una rete già consolidata, con il vantaggio di ridurre sia il rischio di overfitting sia i tempi complessivi di addestramento [11]. Dal punto di vista qualitativo, il modello basato su VGG mostra un apprendimento coerente e stabile, ma complessivamente inferiore rispetto alla CNN addestrata end-to-end sullo stesso dataset. In particolare, mentre la CNN personalizzata riesce a cogliere pattern temporali e dinamiche spettrali strettamente legate alle emozioni musicali, VGG16 tende ad affidarsi a caratteristiche più generiche, risultando meno sensibile alle micro-variazioni che distinguono classi emozionali simili. In conclusione, l'esperimento con VGG16 rappresenta un utile termine di paragone dal punto di vista metodologico, mettendo in evidenza i limiti del transfer learning da ImageNet quando manca un adattamento profondo dei livelli convoluzionali. Queste osservazioni rafforzano la scelta di privilegiare architetture CNN progettate ad hoc per l'elaborazione di segnali audio, soprattutto quando l'obiettivo è una discriminazione fine e accurata di stati emozionali complessi.

4.4. Visualizzazione e Ascolto dei Dati

Per migliorare l'interpretabilità del sistema e consentire una valutazione qualitativa dei risultati, sono state implementate funzionalità di visualizzazione interattiva e di ascolto diretto dei campioni attraverso la funzione personalizzata show_and_play, basata sulla libreria librosa [8]. Questo strumento trasforma l'analisi da un processo puramente numerico a un'esperienza realmente multimodale, in cui i dati vengono affiancati da un riscontro visivo e uditivo immediato. Il sistema opera su tre livelli strettamente integrati. Il primo livello riguarda l'analisi temporale (waveform): utilizzando la funzione librosa.display.waveshow, viene generata la rappresentazione dell'ampiezza del segnale nel dominio del tempo [8], consentendo di cogliere rapidamente la dinamica del brano e mettendo in evidenza attacchi, transitori e pause che giocano un ruolo chiave nella percezione emotiva. A questa si affianca l'analisi spettrale tramite il Mel-spectrogramma, in cui il segnale audio viene convertito mediante librosa.feature.melspectrogram [8] e trasformato in scala logaritmica per migliorarne la leggibilità. La visualizzazione ottenuta con specshow, utilizzando la mappa di colori viridis, permette di osservare la distribuzione dell'energia nel tempo e nelle frequenze, fornendo alla CNN la base informativa fondamentale per la classificazione [7]. Infine, l'integrazione di un player audio interattivo, basato sul modulo IPython.display.Audio, permette di riprodurre il campione nei primi cinque secondi di durata, esattamente come viene processato dalla rete. L'integrazione di queste funzionalità ha reso possibili verifiche rapide e intuitive tra le caratteristiche visive dei campioni e il loro contenuto sonoro. Un approccio di questo tipo non solo facilita l'individuazione di eventuali anomalie o artefatti nel segnale, ma consente anche una validazione più profonda e consapevole del modello. In questo modo, la predizione algoritmica non rimane un semplice valore numerico, ma si trasforma in un giudizio critico supportato da evidenze visive e uditive.

4.5. Analisi degli Errori

L'analisi degli errori rappresenta una fase cruciale per valutare la reale efficacia del sistema. Capire non solo quanto il modello sbaglia, ma soprattutto dove e perché lo fa, fornisce indicazioni preziose per futuri miglioramenti, sia a livello di architettura della CNN sia in relazione alla qualità e alla distribuzione del dataset utilizzato. All'interno di questo progetto, l'analisi è stata condotta combinando tre livelli di verifica complementari: Il primo livello riguarda l'analisi visiva comparativa, in cui i Mel-spectrogrammi dei campioni

classificati in modo errato sono stati confrontati con quelli correttamente riconosciuti. Tale confronto ha evidenziato come pattern visivi estremamente simili possano trarre in inganno la rete, specialmente quando le distinzioni tra le classi emozionali risultano sottili. A questo si aggiunge lo studio della confusion matrix, rivelatosi fondamentale per individuare sovrapposizioni sistematiche tra le classi; in particolare, è emersa una confusione ricorrente tra gli stati emotivi Relaxed e Sad, spesso dovuta alla condivisione di componenti spettrali e dinamiche molto affini. L'analisi ha inoltre evidenziato come le classi rappresentate da un numero ridotto di campioni siano più esposte a errori di classificazione. Questo risultato suggerisce che sviluppi futuri dovrebbero concentrarsi su strategie di data augmentation o su un riequilibrio del dataset [11], al fine di migliorare la sensibilità e la robustezza del modello. In definitiva, questa fase ha confermato l'importanza di affiancare alle metriche quantitative tradizionali, come accuratezza e loss, un'analisi interpretativa più approfondita. Solo attraverso questo approccio è possibile garantire una valutazione solida e affidabile del sistema sviluppato, andando oltre il semplice valore numerico delle prestazioni.

5. Risultati

In questo capitolo vengono presentati e discussi i risultati sperimentali ottenuti dall'addestramento e dalla valutazione del modello di classificazione basato su reti neurali convoluzionali. L'analisi delle prestazioni non si limita a una semplice lettura dei valori numerici, ma integra metriche quantitative standard con strumenti diagnostici e visualizzazioni interpretative, offrendo una valutazione approfondita delle reali capacità del sistema di MER sviluppato. La validazione del modello è stata condotta esclusivamente sul test set, composto da campioni audio mai utilizzati nelle fasi di addestramento o di validazione. Questa scelta, coerente con la struttura del codice implementato, garantisce l'imparzialità dei risultati e consente di valutare in modo affidabile la capacità di generalizzazione della CNN su dati completamente non visti. Le predizioni sul test set sono state ottenute tramite inferenza diretta del modello addestrato, seguita dalla conversione delle probabilità in etichette discrete attraverso l'operazione di argmax. La valutazione delle prestazioni si basa su due elementi fondamentali, in primo luogo, sono state analizzate le metriche di classificazione, con particolare riferimento all'accuratezza complessiva (accuracy) e all'andamento della funzione di perdita (categorical cross-entropy), monitorati durante le epoche di addestramento. Queste metriche consentono di valutare la convergenza del modello e di individuare eventuali fenomeni di overfitting, fornendo una misura globale dell'efficacia del classificatore.

Parallelamente, la matrice di confusione è stata utilizzata come strumento diagnostico per analizzare in dettaglio il comportamento del modello sulle singole classi emozionali [3]. Questo approccio permette di individuare pattern di errore ricorrenti e di osservare quali emozioni tendano a essere maggiormente confuse tra loro, mettendo in luce possibili sovrapposizioni acustiche o limiti intrinseci della rappresentazione adottata.

Le sottosezioni successive approfondiscono ciascuno di questi aspetti, con particolare attenzione ai risultati ottenuti utilizzando diverse rappresentazioni tempo-frequenza degli input audio. Viene inoltre analizzato l'impatto delle principali strategie di ottimizzazione e regolarizzazione adottate nel codice, tra cui la scelta dell'ottimizzatore, l'impiego del dropout e la configurazione dei parametri di addestramento, evidenziandone il ruolo nel determinare le prestazioni finali del modello.

5.1. Accuratezza e Metriche di Valutazione

Per valutare in modo completo l'efficacia dei modelli sviluppati nel compito di MER, è stata adottata una strategia di valutazione multidimensionale. L'obiettivo non è stato soltanto misurare quante volte il modello fornisce una predizione corretta, ma comprendere

la natura degli errori e la solidità delle predizioni quando il sistema viene applicato a dati mai visti, ovvero al test set. Le metriche principali, calcolate tramite la libreria scikit-learn, includono:

- Accuracy (accuratezza): fornisce una misura globale della percentuale di predizioni corrette sul totale dei campioni e rappresenta una prima indicazione della capacità di generalizzazione del modello.
- Precision e Recall: particolarmente rilevanti in presenza di dataset sbilanciati. La precision misura la capacità del modello di limitare i falsi positivi, mentre la recall indica quanto efficacemente il modello riesca a individuare tutti i campioni appartenenti a una determinata classe.
- F1-score: essendo la media armonica tra precision e recall, l'F1-score è stato utilizzato come indicatore sintetico della qualità della classificazione per ciascuna emozione, permettendo un confronto più equilibrato tra le classi.
- Support: indica il numero reale di campioni presenti nel test set per ciascuna classe ed è fondamentale per contestualizzare la rilevanza statistica dei risultati ottenuti.

Il fulcro della sperimentazione è stato il confronto tra tre approcci differenti: una CNN progettata ad hoc e addestrata end-to-end su Mel-spectrogrammi, un modello basato su coefficienti MFCC e un'architettura VGG16 pre-addestrata, utilizzata secondo il paradigma del transfer learning. Dall'analisi quantitativa dei risultati emergono alcune considerazioni di particolare rilievo. In primo luogo, si osserva la superiorità del modello custom addestrato su Mel-spectrogrammi, il quale ha raggiunto un'accuratezza finale del 74%. Questo risultato dimostra che un'architettura relativamente compatta, se addestrata specificamente sulle caratteristiche tempo–frequenza del segnale audio, è in grado di catturare in modo efficace le sfumature emozionali della musica, confermando l'elevata capacità informativa di questa rappresentazione nel contesto del MER.

Parallelamente, le prestazioni del modello basato su MFCC hanno fatto registrare un'accuratezza pari al 68% con una loss finale di 1.17. Sebbene tale valore sia inferiore rispetto all'approccio basato sui Mel-spectrogrammi, il risultato evidenzia come gli MFCC rappresentino una descrizione compatta ed efficiente del contenuto spettrale. Essi sono in grado di catturare informazioni emotive rilevanti, pur mostrando una minore sensibilità alle dinamiche temporali più complesse rispetto agli spettrogrammi completi.

Infine, l'analisi ha messo in luce i limiti del transfer learning con l'architettura VGG16. Nonostante la profondità della rete e i pesi pre-addestrati su ImageNet [6], il modello si è attestato su un'accuratezza del 61%. Il divario di circa 13% rispetto al modello custom evidenzia la cosiddetta "distanza di dominio": i filtri ottimizzati per riconoscere bordi e texture di immagini naturali faticano a interpretare correttamente le relazioni armoniche e i transienti tipici degli spettrogrammi audio senza un processo di fine-tuning più profondo.

Un'ulteriore analisi basata su F1-score e pattern di confusione ha permesso di valutare in modo più fine l'incertezza e la robustezza dei modelli. La CNN custom mostra un comportamento complessivamente bilanciato tra le classi, mentre il modello MFCC evidenzia leggere flessioni per emozioni caratterizzate da pattern meno marcati. La VGG16, invece, presenta una tendenza più pronunciata alla confusione sistematica tra classi acusticamente simili, come Relaxed e Sad. Questo conferma che la maggiore sensibilità della CNN personalizzata alle micro-variazioni temporali e spettrali consente una discriminazione più accurata di stati emotivi con livelli di arousal simili. In conclusione, sebbene la VGG16 rappresenti un'architettura di riferimento nel panorama del Deep Learning e il modello MFCC offra un buon compromesso tra semplicità e prestazioni, la CNN progettata specificamente in questo studio e addestrata su Mel-spectrogrammi si è dimostrata la soluzione più efficace e affidabile per gli obiettivi di classificazione musicale prefissati. Questi risultati

confermano la validità di un approccio orientato al dominio audio e costituiscono una solida base per futuri sviluppi nel campo del MER.

5.2. Confusion Matrix

La matrice di confusione rappresenta uno strumento fondamentale per analizzare in profondità le prestazioni di un classificatore multiclasse, poiché consente di visualizzare in modo chiaro sia le predizioni corrette sia gli errori commessi per ciascuna classe. Ogni cella (i,j)(i, j)(i,j) indica il numero di campioni appartenenti alla classe reale i che il modello ha classificato come classe j. Nel presente studio, la matrice di confusione è stata calcolata tramite la funzione confusion_matrix della libreria scikit-learn ed è stata utilizzata come principale strumento diagnostico per interpretare il comportamento dei modelli rispetto alle classi emotive considerate (Angry, Happy, Relaxed, Sad). Nel caso della CNN addestrata end-to-end su Mel-spectrogrammi, l'osservazione della diagonale principale evidenzia una buona capacità di classificazione per alcune emozioni specifiche. In particolare, le classi Happy e Sad mostrano un numero elevato di predizioni corrette, rispettivamente pari a 266 e 344 campioni. Questo risultato suggerisce che il modello riesce ad apprendere in modo efficace le caratteristiche distintive associate a questi stati emotivi, probabilmente grazie alla presenza di pattern acustici ben definiti e a una maggiore rappresentazione di tali classi all'interno del dataset. Al contrario, le classi Angry e Relaxed presentano una minore concentrazione di valori lungo la diagonale, indicando una maggiore difficoltà del modello nel riconoscerle correttamente. In particolare, Angry viene frequentemente confusa con Happy e Sad, mentre Relaxed mostra errori distribuiti prevalentemente verso le stesse classi. Queste confusioni suggeriscono la presenza di caratteristiche acustico-spettrali parzialmente sovrapposte, per le quali la distinzione automatica richiede la capacità di cogliere micro-variazioni temporali e dinamiche più sottili. Un comportamento simile, ma più marcato, emerge dall'analisi della matrice di confusione del modello basato su VGG16. Anche in questo caso le classi Happy e Sad risultano le meglio riconosciute, ma la distribuzione degli errori al di fuori della diagonale appare più pronunciata rispetto alla CNN personalizzata. In particolare, il modello VGG tende a confondere più frequentemente Angry e Relaxed con le classi dominanti, evidenziando una minore capacità di separazione tra emozioni acusticamente affini. Questo risultato è coerente con l'utilizzo di un modello pre-addestrato su immagini naturali, le cui feature risultano meno sensibili alle specificità temporali e spettrali del segnale audio. L'analisi della matrice di confusione del modello basato su MFCC mostra invece un comportamento intermedio. Pur mantenendo una struttura simile a quella osservata con i Mel-spectrogrammi, si nota una riduzione complessiva dei valori sulla diagonale principale. In particolare, il modello fatica maggiormente a distinguere tra Sad e Relaxed: poiché le MFCC comprimono l'informazione spettrale concentrandosi sull'inviluppo timbrico [9], vengono perse quelle "texture" temporali e dinamiche che aiutano a differenziare un brano malinconico da uno semplicemente rilassante. Inoltre, gli errori risultano più distribuiti tra le classi, indicando che una rappresentazione basata su soli 40 coefficienti fornisce alla rete meno informazioni discriminanti rispetto alle 128 bande Mel. Nel complesso, il confronto tra le matrici di confusione conferma il miglior comportamento generale della CNN addestrata end-to-end, che mostra una maggiore concentrazione di predizioni corrette lungo la diagonale e una distribuzione degli errori più contenuta. Il modello VGG, pur garantendo risultati coerenti e stabili, evidenzia invece limiti strutturali nella discriminazione fine degli stati emotivi, soprattutto in presenza di classi con elevata similarità acustica. L'analisi congiunta delle matrici di confusione si rivela quindi uno strumento essenziale non solo per la valutazione quantitativa delle prestazioni, ma anche per comprendere in modo interpretativo le differenze di comportamento tra architetture progettate specificamente per il dominio audio e modelli basati sul transfer

learning. Le criticità emerse suggeriscono che futuri sviluppi potrebbero beneficiare di strategie di data augmentation mirata, di un riequilibrio delle classi o dell'adozione di architetture ibride in grado di modellare in modo più efficace le dinamiche temporali del segnale musicale.

6. Discussione

I risultati sperimentali ottenuti in questo studio offrono un quadro chiaro e articolato delle potenzialità e dei limiti delle diverse architetture analizzate nel contesto del MER . L'analisi combinata delle metriche quantitative, delle matrici di confusione e del confronto tra modelli addestrati end-to-end e approcci basati su transfer learning ha permesso di trarre considerazioni significative, sia dal punto di vista metodologico sia applicativo. Il modello CNN addestrato direttamente su Mel-spectrogrammi si è dimostrato complessivamente il più efficace, raggiungendo un livello di accuratezza nettamente superiore rispetto al modello VGG pre-addestrato. Questo risultato sottolinea l'importanza di progettare architetture e pipeline di addestramento specificamente orientate al dominio audio, capaci di sfruttare in modo mirato la struttura tempo–frequenza del segnale musicale. La CNN personalizzata riesce infatti a catturare pattern spettrali e dinamiche temporali rilevanti per la discriminazione emotiva, mostrando anche una buona capacità di generalizzazione sul test set. L'analisi dettagliata delle metriche di classificazione rafforza questa osservazione: la CNN presenta valori di precisione, recall e F1-score più equilibrati tra le diverse classi, indicando un apprendimento più robusto e meno influenzato dallo sbilanciamento del dataset. In particolare, emozioni caratterizzate da pattern acustici ben definiti, come Happy e Sad, vengono riconosciute con elevata affidabilità, mentre le principali difficoltà emergono per classi acusticamente più ambigue. Il modello VGG, utilizzato come termine di confronto, evidenzia invece prestazioni inferiori e una maggiore variabilità delle metriche tra le classi. Sebbene si tratti di un'architettura profonda e consolidata nel campo della visione artificiale, i risultati suggeriscono che le feature apprese su immagini naturali non siano facilmente trasferibili al dominio audio senza un fine-tuning profondo dei livelli convoluzionali [11]. Questo limite si traduce in una minore capacità di discriminare emozioni caratterizzate da differenze sottili e da micro-variazioni temporali, elementi fondamentali nel riconoscimento emotivo musicale [2]. Le matrici di confusione forniscono ulteriori elementi interpretativi a supporto di queste conclusioni. Nel caso della CNN end-to-end, la concentrazione dei valori lungo la diagonale principale indica un comportamento complessivamente affidabile del classificatore, con errori che si manifestano soprattutto tra classi semanticamente e acusticamente affini, come Relaxed e Sad [5]. Tali confusioni appaiono coerenti con la natura continua e soggettiva delle emozioni musicali, che raramente presentano confini netti [1]. Nel modello VGG, al contrario, le confusioni risultano più diffuse e marcate, in particolare a svantaggio delle classi meno rappresentate o meno caratterizzate da pattern distintivi. Questo comportamento suggerisce una maggiore dipendenza dalle classi dominanti del dataset e una sensibilità ridotta alle specificità del segnale audio, confermando i limiti di un trasferimento diretto da un dominio eterogeneo come quello delle immagini naturali [10]. Nel complesso, i risultati indicano che, nel contesto del MER, l'addestramento end-to-end di modelli convoluzionali su rappresentazioni audio dedicate risulta più efficace rispetto all'utilizzo di modelli pre-addestrati su immagini, almeno in assenza di strategie avanzate di adattamento del dominio. Le criticità emerse suggeriscono che sviluppi futuri potrebbero beneficiare dell'integrazione di tecniche di data augmentation specifiche per l'audio, di metodi di riequilibrio delle classi [11] e dell'adozione di architetture ibride in grado di modellare in modo più esplicito le dipendenze temporali, come CNN-LSTM o modelli basati su Transformer [3]. In conclusione, questo studio conferma che la scelta dell'architettura e della rappresentazione del segnale riveste un ruolo centrale nel riconosci-

mento automatico delle emozioni musicali [2]. L'approccio proposto costituisce una base solida per ulteriori approfondimenti, offrendo spunti concreti sia per il miglioramento delle prestazioni sia per l'esplorazione di soluzioni più avanzate nel campo del MER.

7. Conclusioni

In questo lavoro è stato sviluppato e valutato un sistema di MER basato su reti neurali convoluzionali, con l'obiettivo di riconoscere stati emotivi a partire da segnali audio, rappresentati attraverso diversi descrittori tempo–frequenza. La sperimentazione ha preso in esame tre approcci distinti: una CNN progettata ad hoc e addestrata end-to-end su Mel-spectrogrammi, un modello basato sui coefficienti MFCC e un'architettura VGG16 pre-addestrata su ImageNet, utilizzata secondo il paradigma del transfer learning. I risultati sperimentali mostrano in modo chiaro che la CNN personalizzata basata su Mel-spectrogrammi offre le prestazioni complessivamente migliori in termini di accuratezza, precisione, richiamo e F1-score. Questo modello si è dimostrato particolarmente efficace nel catturare pattern spettrali e dinamiche temporali fondamentali per la discriminazione delle emozioni musicali. L'analisi delle matrici di confusione ha ulteriormente confermato questa superiorità, evidenziando una maggiore concentrazione di predizioni corrette lungo la diagonale principale e una distribuzione degli errori più contenuta, principalmente limitata a classi acusticamente affini. Il modello basato su MFCC ha ottenuto risultati intermedi, confermando l'efficacia di questi coefficienti come rappresentazione compatta del contenuto spettrale del segnale audio. Pur mostrando prestazioni inferiori rispetto all'approccio basato su Mel-spectrogrammi, il modello MFCC è riuscito comunque a catturare informazioni emozionali rilevanti, rappresentando una valida alternativa in contesti in cui siano richieste una maggiore semplicità computazionale o risorse di addestramento più contenute. Il modello VGG16, pur costituendo un interessante termine di confronto, ha evidenziato limiti significativi nell'adattamento al dominio audio. In particolare, la minore sensibilità alle differenze più sottili tra le classi emozionali e la maggiore influenza della distribuzione del dataset suggeriscono che le feature apprese su immagini naturali non siano pienamente trasferibili al contesto del MER senza un processo di fine-tuning profondo. Questo risultato sottolinea l'importanza di progettare architetture specificamente dedicate al dominio audio, soprattutto quando l'obiettivo è una discriminazione fine di stati emotivi complessi. Le osservazioni emerse aprono diverse prospettive per sviluppi futuri, tra cui l'integrazione di tecniche di data augmentation specifiche per il segnale audio [11], l'adozione di strategie di riequilibrio delle classi per ridurre l'impatto di dataset sbilanciati [6] e lo sviluppo di architetture ibride, come CNN-LSTM [3] o modelli basati su Transformer, in grado di modellare in modo più efficace le dipendenze temporali e dinamiche del segnale musicale. In sintesi, questo studio conferma che un approccio basato su CNN addestrate direttamente sul dominio audio, in particolare mediante l'uso dei Mel-spectrogrammi, rappresenta la strategia più efficace per il MER [2]. I risultati ottenuti costituiscono una base solida per futuri miglioramenti e per l'applicazione di queste tecniche in sistemi avanzati di analisi automatica delle emozioni musicali.

Abbreviazioni

MER	Music Emotion Recognition
DEAM	Database for Emotional Analysis of Music
EDA	Exploratory Data Analysis
CNN	Convolutional Neural Network
VGG16	Visual Geometry Group 16-layer network
MFCC	Mel-Frequency Cepstral Coefficients

https://doi.org/10.3390/computers1010000

STFT	Short-Time Fourier Transform
DCT	Discrete Cosine Transform

References

1. Meyer, L.B. *Emotion and Meaning in Music*; University of Chicago Press: Chicago, IL, USA, 1956; pp. 1–50.
2. Yang, Y.-H.; Chen, H.H. *Music Emotion Recognition*; CRC Press: Boca Raton, FL, USA, 2012; pp. 1–250.
3. Choi, K.; Fazekas, G.; Sandler, M.; Cho, K. Convolutional Recurrent Neural Networks for Music Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2017**, pp. 2392–2396.
4. Juslin, J.; Laukka, P. Expression, perception, and induction of musical emotions. *Psychological Bulletin* **2004**, *131*, 217–252.
5. Russell, J.A. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* **1980**, *39*, 1161–1178.
6. Aljanaki, A.; Yang, Y.-H.; Soleymani, M. DEAM: MediaEval Database for Emotional Analysis in Music. In *Proceedings of the MediaEval Workshop* **2017**, pp. 1–6.
7. Dieleman, S.; Schrauwen, B. End-to-end learning for music audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2014**, pp. 6964–6968.
8. McFee, B. et al. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference* **2015**, pp. 18–25.
9. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1980**, *28*, 357–366.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)* **2015**, pp. 1–14.
11. Pons, A.; Serra, J.; Serra, X. Training neural audio classifiers with limited data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2019**, pp. 696–700.
12. Yang, Y.-H.; Lin, Y.-C.; Su, Y.-F.; Chen, H.H. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **2008**, *16*, 448–457.

https://doi.org/10.3390/computers1010000