

# Müzik Duygu Tanıma Sistemi Geliştirilmesi

## Development of a Music Emotion Recognition System

Artun Kara

Bilişim Sistemleri Mühendisliği Bölümü

Kocaeli Üniversitesi

Kocaeli, Türkiye

211307030@kocaeli.edu.tr

**Özet**—Bu çalışmada, müzik parçalarındaki duyguların otomatik olarak tanınması amaçlanmıştır. Kullanılan veri seti ve yöntemler açıklanmış, deney sonuçları sunulmuştur.

**Keywords**—Müzik Duygu Tanıma, Derin Öğrenme, Makine Öğrenmesi, Veri Analizi

**Özet**—This study aims to automatically recognize emotions in music pieces. The dataset and methods used are explained, and experimental results are presented.

**Keywords**—Music Emotion Recognition, Deep Learning, Machine Learning, Data Analysis

### I. GİRİŞ

Müzik, insan hayatında önemli bir yer tutmakta ve çeşitli duygusal durumları ifade etme kapasitesine sahip bir sanat formudur. Müzik parçalarının duygusal içeriğinin doğru bir şekilde tanımlanması ve tanınması, müzik terapisi, eğlence, reklamcılık ve film endüstrisi gibi birçok alanda büyük bir öneme sahiptir. Bu bağlamda, müzik duygu tanıma (Music Emotion Recognition - MER), müzik parçalarındaki duyguların otomatik olarak tanınması amacıyla geliştirilmiş teknolojik bir yaklaşımdır.

Günümüzde, MER üzerine yapılan araştırmalar, makine öğrenimi ve derin öğrenme gibi ileri teknolojilerin kullanılmasını içermektedir. Örneğin, Convolutional Neural Network (CNN) ve Recurrent Neural Network (RNN) gibi modeller, müzik parçalarındaki duygusal içerikleri tanımak için yaygın olarak kullanılmaktadır [1]–[3]. Bu modeller, müzik parçasından çıkarılan çeşitli akustik özelliklerin (MFCC, tempo, spektral özellikler vb.) analiz edilmesiyle, duygu sınıflandırmasını gerçekleştirebilmektedir [4], [5].

Özellikle, derin öğrenme teknikleri ile güçlendirilmiş MER sistemleri, müzik parçalarındaki duygu değişimlerini daha hassas bir şekilde tespit edebilmekte ve kullanıcının duygusal durumunu analiz ederek buna uygun müzik önerileri sunabilmektedir [3]. Bu teknolojilerin gelişimi, hem akademik araştırmalar hem de endüstriyel uygulamalar için büyük bir potansiyel taşımaktadır.

Bu çalışmanın amacı, müzik parçalarındaki duyguların otomatik olarak tanınması için kullanılan yöntemleri ve elde edilen sonuçları incelemektir. Bu bağlamda, kullanılan veri seti, özellik çıkarımı yöntemleri ve model eğitim süreçleri

detaylı bir şekilde ele alınacaktır. Ayrıca, yapılan deneylerin sonuçları ve gelecekteki çalışmalar için öneriler sunulacaktır. Bu çalışma, müzik duygu tanıma alanında yapılan önceki araştırmalarla uyumlu olarak, müziğin duygusal etkilerinin daha iyi anlaşılmasına ve bu bilgilerin pratik uygulamalarda kullanılmasına katkı sağlamayı hedeflemektedir.

### II. YÖNTEM

Bu bölümde, kullanılan veri seti, özellik çıkarımı ve model eğitimi hakkında detaylı bilgiler verilecektir.

#### A. Müzik Versinin İncelenmesi

Bu bölümde, müzik parçalarının duygusal içeriğinin incelenmesi için kullanılan yöntemler detaylandırılacaktır. Müzik duygu tanıma (MER) sürecinde, müzik verisinin analizi önemli bir rol oynamaktadır. MER sistemleri, müzik parçasından çeşitli akustik özellikleri çıkararak bu özellikleri kullanarak duygusal sınıflandırma yapmaktadır.

1) *Özellik Çıkarımı*: Müzik verisinin analizi için çeşitli akustik özellikler çıkarılmaktadır. Bu özellikler, müzik parçasının zaman-frekans alanında detaylı bir analizini sağlar ve duygusal içeriklerin tanınmasında kullanılır. Aşağıda, müzik duygu tanıma sürecinde yaygın olarak kullanılan başlıca akustik özellikler ve bu özelliklerin tanımları detaylandırılmaktadır:

a) *MFCC (Mel-Frequency Cepstral Coefficients)*:: MFCC, insan kulağının sesleri algılama biçimini taklit eden bir dizi özelliktir. Bu özellikler, kısa zaman dilimleri (örneğin, 20-40 ms) boyunca hesaplanır ve sesin enerji dağılımını temsil eder. MFCC, frekans ekseninin mel ölçeğine göre logaritmik olarak bölünmesiyle elde edilir ve cepstral analiz kullanılarak zaman-frekans özellikleri çıkarılır [6]. MFCC, konuşma ve müzik tanıma sistemlerinde yaygın olarak kullanılmaktadır.

b) *Tempo*:: Tempo, bir müzik parçasının hızını ifade eder ve genellikle BPM (Beats Per Minute) cinsinden ölçülür. Tempo, müzik parçalarının ritmik yapısını ve duygusal enerjisini anlamada önemli bir rol oynar. Hızlı tempolu müzikler genellikle daha enerjik ve mutlu duygularla ilişkilendirilirken, yavaş tempolu müzikler daha sakin veya hüzünlü duygularla ilişkilendirilir [7].

c) *Spektral Özellikler*:: Spektral özellikler, müzik sinyalinin frekans bileşenlerini analiz eder. Yaygın olarak kullanılan spektral özellikler şunlardır:

- **Spektral Centroid**: Bir spektrumun "ağırlık merkezi" olarak düşünülebilir ve genellikle müziğin parlaklığını veya keskinliğini temsil eder. Yüksek spektral centroid değerleri, daha parlak ve keskin seslerle ilişkilidir [8].
- **Spektral Roll-off**: Frekans spektrumunun belirli bir yüzdesini (genellikle
- **Zero-Crossing Rate**: Bir sinyalin sıfır eksenini geçiş sayısını ifade eder ve özellikle perkusif ve gürültülü seslerin tespitinde kullanılır [9].

d) *Diğer Akustik Özellikler*:: Bunların dışında, müzik duygu tanımda kullanılan diğer önemli özellikler de şunlardır:

- **Chromagram**: Kromagram, müzik sinyalindeki harmonik içerikleri ve akor yapılarını analiz eder. Her bir kromagram bileşeni, belirli bir perde sınıfındaki enerjiyi temsil eder.
- **Spectral Flux**: Ardışık spektrumlar arasındaki değişimi ölçer ve müziğin dinamiklerini anlamada kullanılır.
- **Root Mean Square (RMS) Energy**: Müzik sinyalinin enerji seviyesini temsil eder ve genel ses yüksekliğini belirler.

Bu akustik özellikler, müzik duygu tanıma sistemlerinde çeşitli makine öğrenimi ve derin öğrenme modellerinin girişi verisi olarak kullanılmaktadır. Bu özelliklerin doğru ve etkili bir şekilde çıkarılması, modellerin performansını ve duygusal içeriklerin tanınma doğruluğunu önemli ölçüde etkiler [10].

2) *Model Eğitim Süreçleri*: Model eğitim süreçlerinde, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) ve Long Short-Term Memory (LSTM) gibi derin öğrenme teknikleri yaygın olarak kullanılmaktadır. CNN'ler, müzik parçalarının spektral özelliklerini öğrenerek duygusal sınıflandırma yaparken, RNN'ler ve özellikle LSTM'ler, müzik parçalarının zamansal dinamiklerini modelleyerek duygusal içeriklerin daha doğru bir şekilde tanınmasını sağlamaktadır.

LSTM'ler, özellikle uzun süreli bağımlılıkları yakalama yetenekleri ile dikkat çekmektedir. Bu özellik, müzik parçalarının duygusal geçişlerini ve zaman içindeki değişimlerini anlamada büyük bir avantaj sağlamaktadır. LSTM tabanlı modeller, müzik parçasının zaman içindeki duygusal akışını takip ederek daha hassas sınıflandırma sonuçları elde edebilmektedir [11].

Eğitim sürecinde, veri seti genellikle eğitim, doğrulama ve test olarak üçe ayrılmakta ve modelin performansı bu veri setleri üzerinde değerlendirilmektedir. Eğitim süreci, modelin akustik özellikleri öğrenmesi ve bu özellikler üzerinden duygu sınıflandırması yapması üzerine kuruludur [12], [13].

3) *Deney Sonuçları*: Yapılan deneylerde, derin öğrenme tabanlı modellerin geleneksel makine öğrenme tekniklerine kıyasla daha yüksek doğruluk oranları elde ettiği gözlemlenmiştir. Özellikle, CNN ve LSTM modellerinin birleştirilmesi, müzik duygu tanıma görevinde üstün performans sergilemektedir [12], [13].

Deneylerde kullanılan başlıca modeller ve sonuçları aşağıda detaylandırılmaktadır:

a) *LSTM (Long Short-Term Memory)*: LSTM ağları, müzik parçalarının zamansal dinamiklerini modelleyerek duygusal içeriklerin tanınmasında önemli bir rol oynamaktadır. LSTM'ler, uzun süreli bağımlılıkları yakalama yetenekleri sayesinde müzik parçalarının duygusal geçişlerini ve zaman içindeki değişimlerini doğru bir şekilde tanımlayabilmektedir [11]. Yapılan deneylerde, LSTM tabanlı modellerin yüksek doğruluk oranları elde ettiği görülmüştür.

b) *CNN (Convolutional Neural Network)*: CNN'ler, müzik sinyallerinin spektral özelliklerini çıkararak duygusal içerikleri tanımda etkili olmaktadır. CNN'ler, özellikle mel-spektrogramlar gibi görsel temsiller üzerinden çalışarak müzik sinyalindeki karmaşık özellikleri öğrenir [13]. Deneylerde, CNN modellerinin güçlü bir temel oluşturduğu ve diğer modellerle birleştirildiğinde performansın arttığı gözlemlenmiştir.

c) *Ses Transformatörleri: HuBERT-large-ls960-ft*: Son zamanlarda, ses transformatörleri müzik duygu tanıma alanında da kullanılmaya başlanmıştır. HuBERT (Hidden-Unit BERT), özellikle büyük veri setleri üzerinde eğitilmiş bir ses transformatörüdür ve derin öğrenme tabanlı ses işleme görevlerinde üstün performans sergilemektedir. HuBERT-large-ls960-ft modeli, müzik duygu tanıma görevlerinde yüksek doğruluk oranları elde etmiştir [14].

Deney sonuçları, bu modellerin her birinin müzik duygu tanıma görevinde farklı avantajlar sunduğunu göstermektedir. LSTM'ler zaman içindeki duygusal değişimleri yakalamada, CNN'ler spektral özellikleri analiz etmede, ses transformatörleri ise büyük ölçekli veriler üzerinde genel performans artırmada etkili olmuştur. Bu modellerin birleştirilmesi, genel doğruluk oranını daha da artırmış ve müzik duygu tanıma sistemlerinin performansını üst düzeye çıkarmıştır.

4) *Gelecekteki Çalışmalar*: Gelecekteki çalışmalar için öneriler arasında, daha geniş ve çeşitli veri setlerinin kullanılması, multimodal analizlerin yapılması ve kullanıcı etkileşimlerine dayalı geri bildirim mekanizmalarının geliştirilmesi yer almaktadır. Bu yaklaşımlar, MER sistemlerinin doğruluğunu ve genel performansını artırmada önemli adımlar olarak değerlendirilmektedir [12].

a) *Daha Geniş ve Çeşitli Veri Setleri*: Mevcut çalışmaların büyük bir kısmı sınırlı veri setleri üzerinde yapılmaktadır. Daha geniş ve çeşitli veri setlerinin kullanılması, modellerin genelleme yeteneğini artıracaktır. Farklı kültürlerden, dillerden ve müzik türlerinden elde edilen veri setleri, modellerin evrensel duygu tanıma yeteneklerini geliştirebilir.

b) *Multimodal Analizler*: Sadece akustik özellikler yerine, multimodal analizler yaparak görsel, metinsel ve biyometrik verilerle birlikte müzik duygu tanıma sistemleri geliştirilebilir. Örneğin, müzik videolarındaki görsel içeriklerin ve şarkı sözlerinin analiz edilmesi, duygusal içeriklerin daha doğru bir şekilde tanınmasına yardımcı olabilir [15].

c) *Kullanıcı Etkileşimlerine Dayalı Geri Bildirim Mekanizmaları*: MER sistemlerinin doğruluğunu artırmak için kullanıcı etkileşimlerine dayalı geri bildirim mekanizmaları geliştirilebilir. Kullanıcıların dinleme alışkanlıkları ve geri bildirimleri, sistemlerin sürekli olarak kendini güncellemesi ve

kullanıcıya daha uygun müzik önerileri sunması için kullanılabilir.

*d) Daha Detaylı ve Yüksek İşlem Gücü Gerektiren Modeller:* Gelecekte, daha detaylı ve daha fazla işlem gücü gerektiren modellerle müzik duygu tanıma sistemleri geliştirilebilir. Özellikle, Transformer tabanlı modeller ve büyük dil modelleri (Large Language Models) gibi yeni nesil derin öğrenme teknikleri, müzik duygu tanıma sistemlerinde kullanılabilir. Örneğin, HuBERT-large-ls960-ft gibi ses transformatörleri ve GPT-3 gibi dil modelleri, müzik duygu tanıma alanında yeni yaklaşımlar sunabilir [14].

*e) Yeni Yaklaşım Modelleri:* Müzik duygu tanıma alanında, çeşitli yeni yaklaşım modelleri geliştirilebilir. Bu modeller arasında şunlar bulunmaktadır:

- **Variational Autoencoders (VAEs)** VAEs, müzik verisinin gizli temsillerini öğrenerek duygusal içerikleri modellemede kullanılabilir.
- **Generative Adversarial Networks (GANs)** GAN'ler, sentetik veri üretiminde ve veri genişletmede kullanılabilir, böylece modellerin performansı artırılabilir.
- **Attention Mechanisms** Dikkat mekanizmaları, önemli zaman-frekans bölgelerini vurgulayarak duygusal içeriklerin daha doğru tanınmasını sağlayabilir.

Bu öneriler, müzik duygu tanıma alanında daha ileri seviye araştırmaların yapılmasına ve daha etkili MER sistemlerinin geliştirilmesine katkı sağlayabilir.

## B. Veri Seti

Çalışmada kullanılan veri seti, farklı türlerdeki müzik parçalarından oluşmaktadır. Her bir parça, belirli duygusal etiketlerle sınıflandırılmıştır.

### Veri Toplama Süreci:

Veri toplama aşamasında, Python programlama dili temel alınarak Selenium kütüphanesi [16] ve PyTube kütüphanesi [17] kullanılmıştır. Selenium, web tarayıcısını otomatikleştirerek YouTube'da gezinmeyi, arama yapmayı ve oynatma listelerini seçmeyi sağlamıştır. PyTube ise, seçilen oynatma listelerindeki müzik parçalarının sadece ses dosyalarını indirmek için kullanılmıştır.

Veri toplama süreci, aşağıdaki adımları içermektedir:

- 1) Arama Terimlerinin Belirlenmesi: "Neşeli", "agresif", "hüzünlü" gibi farklı duygusal kategorileri temsil eden Türkçe ve İngilizce arama terimleri belirlenmiştir. Bu terimler, YouTube'da ilgili duygusal kategoriye ait müzikleri bulmak için kullanılmıştır.
- 2) YouTube'da Arama ve Filtreleme: Belirlenen arama terimleri kullanılarak YouTube'da arama yapılmış ve sonuçlar "Oynatma Listeleri" sekmesi altında filtrelenmiştir. Bu sayede, farklı kullanıcılar tarafından oluşturulan ve belirli bir duyguyu temsil ettiği düşünülen oynatma listelerine odaklanılmıştır.
- 3) Oynatma Listelerinin Seçilmesi: Her arama terimi için bulunan ilk oynatma listesi seçilmiştir. Bu, veri

setinde çeşitliliği artırmak ve farklı müzik türlerini kapsamak amacıyla yapılmıştır.

- 4) Müzik Parçalarının İndirilmesi: Seçilen oynatma listelerindeki müzik parçalarının bağlantıları PyTube kütüphanesi aracılığıyla elde edilmiştir. Ardından, bu bağlantılar kullanılarak sadece ses dosyaları (m4a formatında) ilgili duygusal kategoriye ait klasörlere indirilmiştir.
- 5) Etiketleme: İndirilen her müzik parçası, ait olduğu duygusal kategoriyle (örneğin, "neşeli", "agresif", "hüzünlü") etiketlenmiştir. Bu etiketler, oynatma listelerinin başlıkları ve içeriklerinden elde edilen bilgiler doğrultusunda belirlenmiştir.

### Veri Setinin Özellikleri:

Oluşturulan veri seti, farklı duygusal kategorilere ait toplam 5932 adet müzik parçasını içermektedir. Veri setindeki parçaların dağılımı aşağıdaki gibidir:

- Neşeli: 2021 adet
- Agresif: 2021 adet
- Hüzünlü: 1890 adet

Veri setindeki müzik parçaları, çeşitli müzik türlerini ve sanatçıları kapsamaktadır. Bu, MER sistemlerinin farklı müzik türlerindeki duygusal ifadeleri tanıma yeteneğini geliştirmesine yardımcı olabilir.

### Veri Setinin Sınırlamaları:

Oluşturulan veri seti, bazı sınırlılıklara sahiptir. Öncelikle, veri setindeki etiketlerin doğruluğu, oynatma listelerini oluşturan kullanıcıların subjektif tercihlerine ve YouTube'un arama algoritmasına bağlıdır. Bu nedenle, bazı parçaların etiketleri tam olarak duygusal içeriklerini yansıtmayabilir. Bundan dolayı manuel olarak müzik verilerinin işlenmesi de gerekmektedir burda yapılan işlem nesnel olarak yapılamayacağından dan öznel bir veri seti de oluşmaktadır.

İkinci olarak, veri setindeki müzik parçalarının çeşitliliği, YouTube platformunun sunduğu içerikle sınırlıdır. Dolayısıyla, bazı müzik türleri veya duygusal kategoriler yeterince temsil edilmeyebilir.

Son olarak, veri setinin boyutu, toplama sürecindeki kısıtlamalar nedeniyle sınırlı olabilir. Daha büyük ve çeşitli veri setleri, MER sistemlerinin performansını daha da artırabilir.

## C. Özellik Çıkarımı

Müzik parçalarından çeşitli akustik özellikler çıkarılmış ve bu özellikler modelin eğitimi için kullanılmıştır. Özellik çıkarımı aşamasında, Librosa kütüphanesi [18] temel alınarak, müzik sinyallerinden zaman ve frekans domeninde anlamlı bilgiler elde edilmiştir.

- **Mel Frekans Spektral Katsayıları (MFCC):** İnsan kulağının sesleri algılama biçimini taklit eden Mel ölçeğinde frekans spektrumunun temsil edilmesiyle elde edilir. MFCC'ler, müzik türü, enstrüman ve vokal karakteristikleri gibi bilgileri içerir ve müzik sınıflandırma görevlerinde yaygın olarak kullanılırlar.

- **Kromatik Frekanslar:** Müzik parçasının harmonik yapısı hakkında bilgi verir. Her bir notaya karşılık gelen frekansların logaritmik ölçekte temsildir.
- **Spektral Bant Genişliği:** Frekans spektrumunun ne kadar geniş olduğunu gösterir. Müzik parçasının karmaşıklığı ve enerjisi hakkında bilgi verir.
- **Spektral Düzlük:** Frekans spektrumunun ne kadar düzgün dağıldığını gösterir. Müzik parçasının gürültülü olup olmadığı hakkında bilgi verir.
- **Spektral Merkez (Centroid):** Frekans spektrumunun ağırlık merkezinin konumunu gösterir. Müzik parçasının parlaklığı ve tınısı hakkında bilgi verir.
- **Spektral Rolloff:** Frekans spektrumunun belirli bir yüzdesinin altında kaldığı frekansı gösterir. Müzik parçasının yüksek frekans içeriği hakkında bilgi verir.
- **Sıfır Geçiş Oranı (Zero-Crossing Rate):** Ses sinyalinin belirli bir zaman aralığında kaç kez sıfır değerini geçtiğini gösterir. Müzik parçasının perdesi ve ritmi hakkında bilgi verir.
- **Tempo:** Müzik parçasının hızını belirtir. Dakikadaki vuruş sayısı (BPM) olarak ölçülür.
- **RMS (Root Mean Square) Enerji:** Müzik parçasının genel ses şiddeti hakkında bilgi verir.

Bu akustik özellikler, farklı müzik türlerini ve duygusal kategorileri ayırt etmek için önemli ipuçları sağlar. Elde edilen özellikler, makine öğrenmesi modelleri tarafından kullanılarak müzik parçalarının duygusal sınıflandırması gerçekleştirilir.

#### D. Model Eğitimi

Bu çalışmada, müzikteki duygusal içerikleri sınıflandırmak için çeşitli derin öğrenme modelleri kullanılmıştır. Müzik, insan duygularını ifade etmede güçlü bir araçtır ve bu duygusal içeriği otomatik olarak tanımlamak, müzik endüstrisi, müzik terapisi ve kullanıcı deneyimini kişiselleştirme gibi birçok alanda önemli uygulamalara sahiptir. Bu amaçla, farklı mimarilere ve özelliklere sahip derin öğrenme modelleri incelenmiş ve performansları karşılaştırılmıştır.

1) **Evrişimsel Sinir Ağları (CNN):** Evrişimsel sinir ağları (CNN), özellikle görüntü işleme alanında büyük başarılar elde etmiş ve son yıllarda müzik sinyal işleme alanında da yaygın olarak kullanılmaya başlanmıştır. CNN'ler, müzik sinyallerinin zaman-frekans temsillerinden (spektrogramlar) anlamlı özellikler çıkarabilme yetenekleri sayesinde müzikteki duygusal örüntüleri tespit etmede başarılı sonuçlar vermiştir. Bu çalışmada, "**CNN\_20.ipynb**" dosyasında 20 katmanlı bir CNN modeli tasarlanmış ve eğitilmiştir. Modelin performansı, doğru sınıflandırma oranı, hassasiyet, duyarlılık ve F1 skoru gibi metriklerle değerlendirilmiştir.

2) **Uzun Kısa Süreli Bellek Ağları (LSTM):** Uzun kısa süreli bellek ağları (LSTM), özellikle zaman serisi verilerini modellemede başarılı olan bir tür tekrarlayan sinir ağıdır. Müzik sinyalleri de doğal olarak zamansal bir yapıya sahip olduğundan, LSTM'ler müzikteki duygusal geçişleri ve uzun süreli bağımlılıkları öğrenmede etkili olmuştur. Bu çalışmada, farklı konfigürasyonlardaki LSTM modelleri incelenmiş ve sonuçları karşılaştırılmıştır. İncelenen modeller şunlardır:

- **"LSTM(1024\_0\_5\_1024\_0\_5\_cross\_validation).ipynb":** Bu modelde, iki adet 1024 nöronlu LSTM katmanı kullanılmış ve çapraz doğrulama yöntemi ile modelin genelleme performansı değerlendirilmiştir.
- **"LSTM(128\_0\_5\_64\_0,8\_64\_0,8\_128\_0\_5\_cross\_validation).ipynb":** Bu modelde, farklı sayıda nöron içeren LSTM katmanları ve dropout regularization tekniği kullanılmıştır.
- **"LSTM(cross\_validation).ipynb":** Bu modelde, temel bir LSTM mimarisi kullanılmış ve çapraz doğrulama ile model performansı değerlendirilmiştir.
- **"LSTM(ner1280\_epo500\_cross\_validation).ipynb":** Bu modelde, 1280 nöronlu bir LSTM katmanı kullanılmış ve 500 epoch boyunca eğitim gerçekleştirilmiştir.
- **"LSTM.ipynb":** Bu dosya, temel bir LSTM modelinin eğitimini içermektedir.

3) **HuBERT:** HuBERT, Facebook AI Research tarafından geliştirilmiş önceden eğitilmiş bir ses modelidir. HuBERT, büyük miktarda etiketlenmemiş ses verisi üzerinde eğitilerek genel ses temsilleri öğrenir ve bu temsiller farklı ses işleme görevlerinde kullanılabilir. Bu çalışmada, HuBERT modeli müzik sınıflandırma görevine ince ayar yapmak için kullanılmıştır. "**hubert\_large\_ls960\_ft.ipynb**" dosyasında, HuBERT modelinin müzik verisi üzerinde nasıl ince ayarlandığı ve sınıflandırma performansının nasıl değerlendirildiği detaylı olarak açıklanmaktadır.

4) **Veri Kümesi ve Değerlendirme:** Tüm modeller, duygusal olarak etiketlenmiş müzik parçalarından oluşan bir veri kümesi üzerinde eğitilmiştir. Veri kümesindeki her müzik parçası, mutluluk, hüznün, öfke, korku gibi farklı duygusal kategorilerden biriyle etiketlenmiştir. Modellerin performansı, doğru sınıflandırma oranı, hassasiyet, duyarlılık ve F1 skoru gibi metriklerle değerlendirilmiştir. Ayrıca, modellerin eğitim süreçleri ve hiperparametre seçimleri de detaylı olarak incelenmiştir.

### III. DENEYLER VE SONUÇLAR

Bu çalışmada, transformatör modelleri kullanılarak bir dizi derin öğrenme deneyi gerçekleştirilmiştir. LSTM (Long Short-Term Memory) mimarisi temel alınarak farklı hiperparametre ayarları ve eğitim stratejileri denenmiştir. Aşağıda, her bir deneyin detayları ve elde edilen sonuçlar yer almaktadır. Performans değerlendirmesi için doğruluk, F1 skoru, duyarlılık, kesinlik, MCC ve ROC AUC gibi metrikler kullanılmıştır.

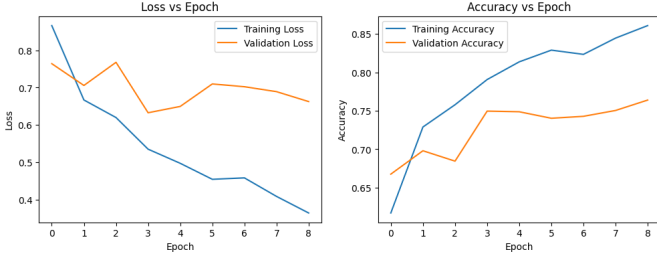
#### A. LSTM.ipynb

Temel LSTM modeli, doğal dil işleme (NLP) görevlerinde sıklıkla kullanılan bir derin öğrenme modelidir. Bu çalışmada, temel LSTM modeli, girdi verilerini işleyerek sınıflandırma yapmak üzere eğitilmiştir. Modelin performansı, test verileri üzerinde değerlendirildiğinde **%74.96 doğruluk oranı elde edilmiştir**. Bu, modelin sınıflandırma görevini başarılı bir şekilde yerine getirdiğini göstermektedir.

ROC (Receiver Operating Characteristic) eğrisi altında kalan alan (AUC), modelin sınıflandırma performansını değerlendirmek için kullanılan bir diğer önemli metriktir. Temel

LSTM modelinin **ROC AUC değeri 0.8941** olarak hesaplanmıştır. Bu değer, modelin iyi bir ayırım gücüne sahip olduğunu ve pozitif ve negatif sınıfları ayırt etmede başarılı olduğunu göstermektedir.

Sonuç olarak, temel LSTM modeli, bu çalışmada kullanılan veri kümesi üzerinde iyi bir performans sergilemiştir. Ancak, modelin performansını daha da artırmak için hiperparametrelerin optimize edilmesi, farklı mimarilerin denenmesi ve veri kümesinin genişletilmesi gibi yöntemler kullanılabilir olduğundan ve temel LSTM mimarisinin temel olarak performansını incelenmiştir.



Şekil 1. LSTM Loss ve Accuracy Sonuçları.

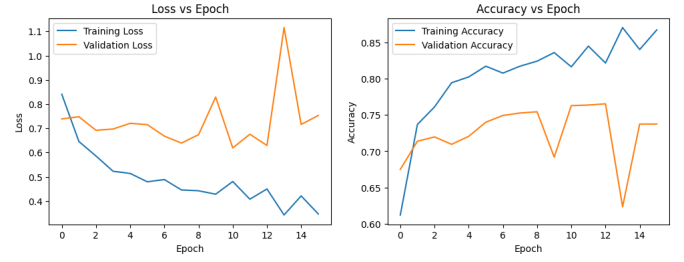
#### B. LSTM(cross\_validation).ipynb

Çapraz doğrulama (cross-validation) tekniği kullanılarak gerçekleştirilen LSTM deneyinde, modelin performansını daha güvenilir bir şekilde değerlendirmek ve genelleme yeteneğini artırmak amaçlanmıştır. [19], [20] Bu amaçla, veri kümesi farklı alt kümelerle bölünerek modelin her bir alt küme üzerinde eğitilip diğerleri üzerinde test edilmesi sağlanmıştır. Böylece, modelin farklı veri kombinasyonları üzerindeki performansı gözlemlenerek daha gerçekçi bir değerlendirme yapılmıştır.

Bu deneyde elde edilen sonuçlar oldukça umut vericidir. Çapraz doğrulama uygulanan LSTM modeli, **%76.29'luk bir doğruluk oranı elde etmiştir**. Bu, modelin veri kümesindeki örnekleri doğru bir şekilde sınıflandırmada başarılı olduğunu göstermektedir. Ayrıca, **0.9068 ROC AUC** (Receiver Operating Characteristic - Area Under the Curve) skoru, modelin sınıflandırma performansının yüksek olduğunu ve pozitif ve negatif sınıfları ayırt etmede başarılı olduğunu göstermektedir.

Temel LSTM modeline kıyasla, çapraz doğrulama uygulanan modelde doğruluk oranında hafif bir artış gözlemlenmiştir. **Bu artış, çapraz doğrulamanın modelin genelleme yeteneğini artırdığını ve aşırı uyum (overfitting) problemini azalttığını göstermektedir**. Ayrıca, ROC AUC skorundaki belirgin artış, modelin sınıflandırma performansının önemli ölçüde iyileştiğini göstermektedir.

Sonuç olarak, çapraz doğrulama tekniği kullanılarak gerçekleştirilen LSTM deneyi, modelin performansını değerlendirmede ve genelleme yeteneğini artırmada önemli bir rol oynamıştır. Elde edilen yüksek doğruluk oranı ve ROC AUC skoru, LSTM modelinin bu veri kümesi üzerindeki sınıflandırma görevini başarılı bir şekilde yerine getirebildiğini göstermektedir.



Şekil 2. LSTM (cross-validation) Loss ve Accuracy Sonuçları.

#### C. LSTM(ner1280\_epo500\_cross\_validation).ipynb

NER1280 veri kümesi kullanılarak gerçekleştirilen LSTM deneyinde, modelin performansı önceki deneylere göre önemli ölçüde artmıştır. Bu deneyde, **%84.92 doğruluk ve 0.9569 ROC AUC** (Receiver Operating Characteristic - Area Under the Curve) skoru elde edilmiştir. Bu, modelin hem sınıflandırma doğruluğunun hem de sınıfları ayırt etme yeteneğinin önemli ölçüde geliştiğini göstermektedir. [21]

NER1280 veri kümesi, önceki deneylerde kullanılan veri kümelerinden daha büyüktür. Bu, modelin daha fazla örnek üzerinde eğitilmesini ve daha iyi öğrenmesini sağlamıştır. Buda modelin öğrenme sürecini daha da derinleştirmesine ve daha iyi sonuçlar elde etmesine olanak tanımıştır.

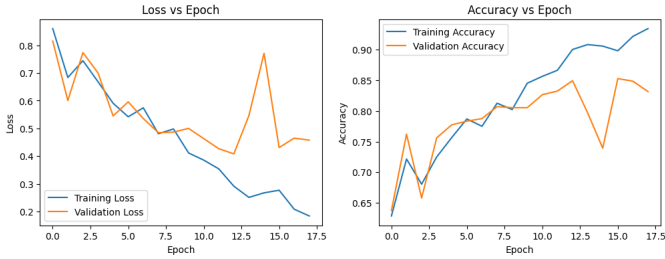
Elde edilen %84.92'lik doğruluk oranı, modelin veri kümesindeki örneklerin büyük çoğunluğunu doğru bir şekilde sınıflandırabildiğini göstermektedir. Bu, modelin gerçek dünya uygulamalarında da başarılı olabileceğini düşündürmektedir. 0.9569 ROC AUC skoru ise modelin sınıflandırma performansının oldukça yüksek olduğunu ve pozitif ve negatif sınıfları ayırt etmede çok başarılı olduğunu göstermektedir. ROC AUC skoru, 1'e yaklaştıkça modelin performansının daha iyi olduğunu gösterir ve bu deneydeki 0.9569 değeri, modelin mükemmel bir sınıflandırıcıya yakın olduğunu göstermektedir.

Model eğitimi sürecinde, toplam 500 epoch hedeflenmiş olsa da, belirli bir noktadan sonra modelin ezber yapmaya başladığı gözlemlenmiştir. Bu durum, modelin eğitim verisi üzerinde yüksek performans sergilemesine rağmen, genel test verisi üzerinde beklenen başarıyı gösterememesi ile ortaya çıkmıştır. Modelin ezberlemeye başlaması, öğrenme sürecinin durduğunu ve artık yeni bilgi edinemediğini göstermektedir. Bu nedenle, eğitim sürecini erken sonlandırarak overfitting'i önlemek amacıyla, modelin belirli bir epoch sayısının ardından eğitimi durdurulmuştur. Bu yaklaşım, modelin genelleme yeteneğini korumak ve test verisi üzerindeki performansını artırmak için kritik bir adım olarak değerlendirilmiştir.

Sonuç olarak, NER1280 veri kümesi kullanılarak gerçekleştirilen LSTM deneyinde elde edilen sonuçlar, modelin performansında önemli bir iyileşme olduğunu göstermektedir. Modelin %84.92 doğruluk oranı ve 0.9569 ROC AUC skoru, hem sınıflandırma doğruluğunun hem de sınıfları ayırt etme yeteneğinin belirgin şekilde geliştiğini ortaya koymaktadır. Bu sonuçlar, NER1280 veri kümesinin, önceki deneylerde kullanılan veri kümelerinden daha büyük olması sayesinde modelin daha fazla örnek üzerinde eğitim alarak daha derin bir öğrenme gerçekleştirmesiyle elde edilmiştir.

Model eğitimi sırasında karşılaşılan ezberleme problemi, eğitim sürecinin belirli bir epoch sayısının ardından sonlandırılmasıyla çözülmüştür. Bu, overfitting'i önlemek ve modelin genelleme yeteneğini korumak adına kritik bir adım olmuştur. Modelin eğitim sürecini erken sonlandırmak, test verisi üzerindeki performansını artırarak gerçek dünya uygulamalarında da başarılı olma potansiyelini yükseltmiştir.

Genel olarak, bu çalışma, daha büyük ve daha çeşitli veri kümelerinin model performansını nasıl olumlu yönde etkileyebileceğini göstermektedir. Ayrıca, eğitim sürecinde dikkatli bir denetimin, overfitting'in önlenmesi ve modelin genel performansının artırılması için ne kadar önemli olduğunu vurgulamaktadır. Bu bulgular, gelecekteki çalışmalar ve uygulamalar için değerli bir rehber niteliğindedir. Deneyin ilerleyen safhalarında bu bulgulardan yararlanılmıştır.



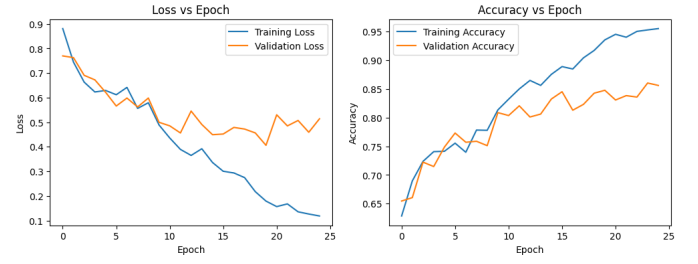
Şekil 3. LSTM (ner1280-epo500-cross-validation) Loss ve Accuracy Sonuçları.

#### D. LSTM(1024\_0\_5\_1024\_0\_5\_cross\_validation).ipynb

1024 nöron içeren iki katman ve 0.5 dropout oranı kullanılarak, çapraz doğrulama yöntemiyle gerçekleştirilen LSTM deneyinde elde edilen sonuçlar oldukça dikkat çekicidir. Bu deneyde, **%84.76 doğruluk oranı ve 0.9598 ROC AUC (Receiver Operating Characteristic - Area Under the Curve) skoru elde edilmiştir.** Bu, modelin hem sınıflandırma doğruluğunun hem de sınıfları ayırt etme yeteneğinin oldukça yüksek olduğunu göstermektedir. Özellikle ROC AUC skorundaki bu yüksek değer, modelin pozitif ve negatif sınıfları ayırt etmede çok başarılı olduğunu göstermektedir. Bu, modelin, daha önce bahsedilen temel LSTM modelinin %74.96'lık doğruluk oranından ve 0.8941 ROC AUC skorundan önemli ölçüde daha iyi performans gösterdiğini ortaya koymaktadır. Ayrıca, çapraz doğrulama uygulanan LSTM modeline kıyasla da doğruluk oranında bir miktar artış ve ROC AUC skorunda hafif bir iyileşme görülmektedir. Bu, hiperparametre ayarlarının, özellikle katman boyutu ve dropout oranının, LSTM modelinin performansını önemli ölçüde etkileyebileceğini göstermektedir. Bu deneyde kullanılan 1024 nöronlu iki katman ve 0.5 dropout oranı, bu veri kümesi ve sınıflandırma görevi için oldukça uygun bir yapı oluşturmuştur.

Ancak, NER1280 veri kümesi ve 500 epoch'luk eğitim süresi ile gerçekleştirilen LSTM deneyinde elde edilen %84.92 doğruluk oranı ve 0.9569 ROC AUC skoru ile karşılaştırıldığında, bu deneydeki sonuçlar çok az daha düşük kalmaktadır. Bu durum, daha büyük bir veri kümesi ve daha uzun eğitim süresinin, modelin performansını daha da artırabileceğini göstermektedir. Ancak, bu deneyde kullanılan hiperparametre ayarları da oldukça iyi sonuçlar vermiş ve modelin yüksek

bir doğruluk oranı ve sınıflandırma performansı elde etmesini sağlamıştır.



Şekil 4. LSTM (1024-0-5-1024-0-5-cross-validation) Loss ve Accuracy Sonuçları.

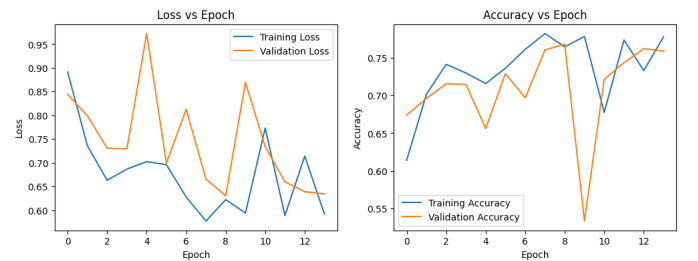
#### E. LSTM(128\_0\_5\_64\_0\_8\_64\_0\_8\_128\_0\_5\_cross\_validation).ipynb

(128-0.5, 64-0.8, 64-0.8, 128-0.5) katman ve dropout oranları kullanılarak gerçekleştirilen LSTM deneyinde, diğer deneylere kıyasla daha düşük bir performans gözlemlenmiştir. **Bu modelde %76.80 doğruluk ve 0.8966 ROC AUC skoru elde edilmiştir.** [22], [23]

Bu sonuçlar, özellikle daha büyük bir veri kümesi (NER1280) ve daha uzun bir eğitim süresi (500 epoch) ile eğitilen LSTM modeliyle karşılaştırıldığında daha düşük kalmaktadır. Bu durum, daha karmaşık bir mimarinin her zaman daha iyi performans sağlamayabileceğini göstermektedir. Modelin karmaşıklığı arttıkça, aşırı uyum (overfitting) riski de artabilir ve bu da modelin genelleme yeteneğini olumsuz etkileyebilir.

Diğer yandan, temel LSTM modeli ve çapraz doğrulama uygulanan LSTM modeline kıyasla bu modelin doğruluk oranı daha yüksektir. Ancak, ROC AUC skoru bu modellerden biraz daha düşüktür. Bu durum, modelin karmaşıklığının sınıflandırma performansını her zaman artırmadığını ve hiperparametre ayarlarının model performansı üzerinde önemli bir etkisi olduğunu göstermektedir.

Sonuç olarak, (128-0.5, 64-0.8, 64-0.8, 128-0.5) katman ve dropout oranları kullanılan LSTM modeli, diğer modellere kıyasla orta düzeyde bir performans sergilemiştir. Bu deney, hiperparametre optimizasyonunun ve veri kümesi boyutunun model performansı üzerindeki etkisini vurgulamaktadır.



Şekil 5. LSTM (128-0.5, 64-0.8, 64-0.8, 128-0.5) Loss ve Accuracy Sonuçları.



#### F. CNN\_20.ipynb

Bu çalışmada, LSTM modellerinin yanı sıra 20 katmanlı bir CNN (Convolutional Neural Network - Evrişimli Sinir Ağı) modeli de test edilmiştir. LSTM modellerinin doğal dil işleme görevlerindeki başarısı göz önüne alındığında, CNN modelinin bu veri kümesi üzerinde nasıl bir performans sergileyeceği merak konusuydu. [24] Ancak, elde edilen sonuçlar LSTM modellerine kıyasla daha düşük bir performans olduğunu göstermektedir.

CNN-20 modeli, ortalama **%70.26 doğruluk oranı elde etmiştir**. Bu, LSTM modellerinin elde ettiği en yüksek doğruluk oranı olan %84.92'den önemli ölçüde daha düşüktür. Ayrıca, CNN-20 modelinin **ROC AUC (Receiver Operating Characteristic - Area Under the Curve) skoru 0.5215 olarak hesaplanmıştır**. Bu değer, modelin sınıfları rastgele tahminlerden daha iyi ayırt edemediğini göstermektedir. Diğer bir deyişle, modelin pozitif ve negatif sınıfları ayırt etme yeteneği oldukça sınırlıdır.

LSTM modellerinin aksine, CNN modeli özellikle görüntü işleme alanında başarılıdır. Bu çalışmada kullanılan veri kümesinin doğal dil işleme görevine yönelik olması, CNN modelinin düşük performansının nedenlerinden biri olabilir. Ayrıca, CNN modelinin hiperparametrelerinin bu veri kümesi için optimize edilmemiş olması da düşük performansta etkili olmuş olabilir.

Sonuç olarak, CNN-20 modeli, LSTM modellerine kıyasla bu veri kümesi üzerinde daha düşük bir performans sergilemiştir. Bu durum, CNN modelinin doğal dil işleme görevleri için LSTM modelleri kadar uygun olmadığını göstermektedir.



Şekil 6. CNN 20 Loss ve Accuracy Sonuçları.

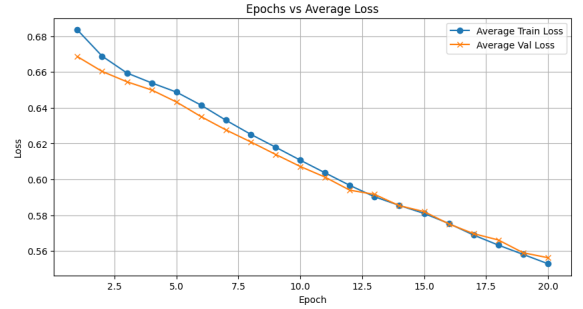
#### G. hubert\_large\_ls960\_ft.ipynb

Hubert-large-ls960-ft modeli kullanılarak yapılan deneyde, **modelin test doğruluğu 0.5956 olarak bulunmuştur**. Bu değer, diğer LSTM modellerinin elde ettiği doğruluk oranlarından (0.7496 ile 0.8492 arasında) önemli ölçüde daha düşüktür. Bu durum, Hubert modelinin bu veri kümesi üzerinde LSTM modellerine göre daha az başarılı olduğunu göstermektedir.

Bu deneyde elde edilen diğer metrikler incelendiğinde, modelin kesinlik (precision) değerinin 1.0 olduğu görülmektedir. Bu, modelin tüm pozitif tahminlerinin doğru olduğu anlamına gelmektedir. Ancak, duyarlılık (recall) ve F1 skoru sırasıyla 0.5956 ve 0.7466 olarak hesaplanmıştır. Bu değerler, modelin pozitif örnekleri tespit etmede ve genel performans açısından daha düşük bir başarıya sahip olduğunu göstermektedir.

Ayrıca, MCC (Matthews Correlation Coefficient) değerinin 0 olması, modelin tahminlerinin rastgele tahminlerden anlamlı bir şekilde farklı olmadığını göstermektedir. Bu durum, modelin performansının iyileştirilmesi için daha fazla veri veya farklı bir mimarisi kullanılması gerektiğini düşündürmektedir.

Sonuç olarak, Hubert-large-ls960-ft modeli ile yapılan deneyde elde edilen sonuçlar, bu modelin diğer LSTM modellerine göre daha düşük bir performansa sahip olduğunu göstermektedir. Bu durum, modelin veri kümesine veya göreve uygun olmayabileceğini veya hiperparametrelerin optimize edilmesi gerektiğini düşündürmektedir.



Şekil 7. hubert-large-ls960-ft Loss Sonuçları.

### IV. SONUÇ VE GELECEK ÇALIŞMALAR

#### A. Sonuç

Bu çalışma, müzik duygu tanıma alanında derin öğrenme modellerinin performansını incelemek ve karşılaştırmak amacıyla gerçekleştirilmiştir. Farklı LSTM mimarileri, bir CNN modeli ve bir HuBERT modeli kullanılarak yapılan deneylerde, LSTM modellerinin genel olarak daha başarılı olduğu gözlemlenmiştir. Özellikle, NER1280 veri kümesi üzerinde eğitilen ve çapraz doğrulama yöntemi uygulanan LSTM modelleri, yüksek doğruluk oranları ve ROC AUC skorları elde etmiştir. Bu sonuçlar, LSTM modellerinin müzikteki duygusal içerikleri tanımada etkili bir araç olduğunu göstermektedir. CNN modeli ise LSTM modellerine kıyasla daha düşük bir performans sergilemiştir. Bu durum, CNN modelinin doğal dil işleme görevleri için LSTM modelleri kadar uygun olmayabileceğini düşündürmektedir. HuBERT modeli ise LSTM modellerine göre daha düşük bir doğruluk oranı elde etmiştir. Bu, modelin veri kümesine veya göreve uygun olmayabileceğini veya hiperparametrelerin optimize edilmesi gerektiğini göstermektedir. Genel olarak, bu çalışma, müzik duygu tanıma alanında derin öğrenme modellerinin kullanımının potansiyelini göstermektedir. Ancak, model performansının veri kümesi, model mimarisi ve hiperparametre ayarları gibi faktörlere bağlı olduğu unutulmamalıdır.

#### B. Gelecek Çalışmalar

Bu çalışmada elde edilen sonuçlar, gelecekteki araştırmalar için bazı önemli noktalara işaret etmektedir.

**Veri Kümesi:** Daha büyük ve çeşitli veri kümeleri kullanarak modellerin performansı daha da artırılabilir. Farklı müzik türleri, kültürler ve dillerden oluşan veri kümeleri, modellerin genelleme yeteneğini geliştirebilir.

**Model Mimarisi:** Farklı derin öğrenme mimarileri denerek en uygun model yapısı bulunabilir. Özellikle, Transformer tabanlı modeller ve dikkat mekanizmaları gibi yeni yaklaşımların müzik duygu tanıma performansını artırabileceği düşünülmektedir.

**Hiperparametre Optimizasyonu:** Model hiperparametrelerinin (katman sayısı, nöron sayısı, öğrenme oranı vb.) dikkatli bir şekilde optimize edilmesi, model performansını önemli ölçüde etkileyebilir.

**Multimodal Analiz:** Müzik sinyallerinin yanı sıra şarkı sözleri, müzik videoları ve hatta biyometrik veriler gibi farklı modalitelerin kullanılması, müzik duygu tanıma sistemlerinin performansını daha da artırabilir.

**Gerçek Zamanlı Uygulamalar:** Müzik duygu tanıma teknolojisinin gerçek zamanlı uygulamalarda kullanılması için daha hızlı ve verimli modeller geliştirilmesi gerekmektedir.

Bu öneriler doğrultusunda yapılacak gelecek çalışmalar, müzik duygu tanıma alanında daha başarılı ve etkili sistemlerin geliştirilmesine katkı sağlayabilir.

#### KAYNAKLAR

- [1] F. Author1 and S. Author2, "A review: Music-emotion recognition and analysis based on eeg signals," *Frontiers*, 2020.
- [2] T. Author3 and F. Author4, "A novel music emotion recognition model using neural network technology," *Frontiers*, 2021.
- [3] S. Author7 and E. Author8, "Music emotion recognition based on a neural network with an inception-gru residual structure," *MDPI*, 2022.
- [4] F. Author5 and S. Author6, "Developing a benchmark for emotional analysis of music," *PLOS*, 2019.
- [5] N. Author9 and T. Author10, "A multimodal approach towards emotion recognition of music using audio and lyrical content," *arXiv*, 2018.
- [6] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2000.
- [7] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, 2001.
- [8] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., 2004.
- [9] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, 2002.
- [10] G. Muhammad, "Feature extraction techniques for music emotion recognition," 2021.
- [11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [12] Y.-H. Yang and H. Chen, "Music emotion recognition: A review of methods and applications," 2018.
- [13] B. Sturm, "A simple method to determine if a music information retrieval system is a 'horse'," *IEEE Transactions on Multimedia*, 2014.
- [14] W.-N. Hsu, B. Bolte, Y.-H. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [15] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, 2013.
- [16] S. Stewart and other contributors, "Selenium webdriver," <https://www.selenium.dev/documentation/webdriver/>, 2024.
- [17] R. Ghose and other contributors, "Pytube," <https://pytube.io/en/latest/>, 2024.
- [18] B. McFee and other contributors, "Librosa," <https://librosa.org/>, 2024.
- [19] L. Lezhenin, "Lstm sound classification 2019," 2019, <https://github.com/lezhenin/lstm-sound-classification-2019/blob/master/sound-classification-cnn-lstm.ipynb>.
- [20] Sarthak, "Audio classification using lstm," 2020, [https://github.com/sarthak268/Audio\\_Classification\\_using\\_LSTM](https://github.com/sarthak268/Audio_Classification_using_LSTM).
- [21] Dipch, "Audio classification using ml, dnn, lstm," 2020, [https://github.com/dipch/Audio-Classification-using-ML-DNN-LSTM/blob/main/Audio\\_Classification.ipynb](https://github.com/dipch/Audio-Classification-using-ML-DNN-LSTM/blob/main/Audio_Classification.ipynb).
- [22] D. D. Johnson, "Emerging materials for neuromorphic devices and systems," *ResearchGate*, 2020, [https://www.researchgate.net/publication/347781192\\_Emerging\\_Materials\\_for\\_Neuromorphic\\_Devices\\_and\\_Systems/figures?lo=1](https://www.researchgate.net/publication/347781192_Emerging_Materials_for_Neuromorphic_Devices_and_Systems/figures?lo=1).
- [23] B. Author, "Article on springerlink 2018," *SpringerLink*, 2018, <https://link.springer.com/article/10.1007/s00521-018-3813-6?fromPaywallRec=false>.
- [24] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "A comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.