INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

The University of Westminster, Coat of Arms

Emotion-Aware Music Recommendation with ConvNeXt V2"

Project Proposal & Specification Requirements by

Ms.Hasni Haleemdeen

W1898945 | 20211337

Supervised by

Mr. Guhanathan Poravi

July 2025

Submitted in partial fulfillment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

ABSTRACT

In the age of intelligent media consumption, the ability to recommend music that aligns with a listener's emotional state is becoming increasingly relevant. While traditional music recommendation engines rely heavily on historical data and user preferences, they often fail to consider the user's real-time emotional context. This research presents a novel system that integrates ConvNeXt V2, a modern convolutional neural network architecture with the Spotify API to deliver personalized music recommendations based on facial emotion recognition. Unlike vision transformers that demand heavy computation, ConvNeXt V2 maintains transformer-inspired design principles while retaining the efficiency of convolutional networks, making it ideal for real-time inference on mobile or web platforms.

The system processes live or uploaded facial images using robust preprocessing and augmentation techniques, then classifies the user's emotional state using a ConvNeXt V2 model trained on benchmark datasets such as FER-2013. Once classified, the detected emotion is mapped to curated playlists via Spotify's music catalog. Experimental evaluation showed strong performance, with the model achieving a validation accuracy of **90.09%**, and detailed performance was confirmed through metrics such as precision, recall, F1-score, and confusion matrices. These results underscore the model's ability to generalize across diverse facial expressions and lighting conditions while maintaining responsiveness.

By bridging facial affective computing with intelligent music delivery, this system offers a more emotionally engaging and context-aware user experience. It demonstrates the feasibility of deploying ConvNeXt V2 as a lightweight, scalable solution for enhancing digital personalization in entertainment applications.

Keywords: ConvNeXt V2, Facial Emotion Recognition, Music Recommendation, Deep Learning, Computer Vision, Spotify API, Real-time Inference, Human-Centered AI

Subject Descriptors:

Computing methodologies → Convolutional Neural Networks → Computer Vision → Emotion-Aware Systems → Personalized Media Recommendations

DECLARATION

I hereby declare that the content presented in this thesis, titled *"Emotion-Aware Music Recommendation with ConvNeXt V2,"* is the result of my own independent work, carried out at the Informatics Institute of Technology under the guidance of my supervisor, Ms. Hasni Haleemdeen.

This thesis has not been submitted, either in part or in full, to any other university or institution for the purpose of obtaining an academic degree or professional qualification. All sources of information used in the preparation of this work have been properly cited and acknowledged in accordance with academic conventions.

Date: 11-07-2025

Signature:

Name: Hasni Haleemdeen

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to all those who supported me throughout the course of this project.

First and foremost, I am deeply thankful to my supervisor, **Mr. Guhanathan Poravi**, for his invaluable guidance, continuous encouragement, and insightful feedback. His expert advice and motivating support were instrumental in shaping the direction and execution of this research.

I would also like to express my appreciation to the academic and administrative staff of the **Informatics Institute of Technology** for providing the resources, facilities, and a conducive learning environment that enabled me to complete this work effectively.

To everyone who contributed in any way directly or indirectly toward the completion of this thesis, your support is truly appreciated

Hasni Haleemdeen

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AI	**Artificial Intelligence**
ViT	**Vision Transformer**
CNN	**Convolutional Neural Network**
JSON	**JavaScript Object Notation**
API	**Application Programming Interface**
GPU	**Graphics Processing Unit**
UI	**User Interface**
UX	**User Experience**
IDE	**Integrated Development Environment**
FER	**Facial Emotion Recognition**
FER2013	**Facial Expression Recognition 2013 Dataset**
DNN	**Deep Neural Network**
HCI	**Human-Computer Interaction**
NLP	**Natural Language Processing**
SOTA	**State of the Art**
DL	**Deep Learning**
ML	**Machine Learning**
SVM	**Support Vector Machine**
DFD	**Data Flow Diagram**

SLEP	**Social, Legal, Ethical and Professional**
API Key	**Application Programming Interface Key**
IoT	**Internet of Things**

CHAPTER 1: INTRODUCTION

1.1 Chapter overview

This chapter explains why we are creating an emotion-aware music recommendation system with ConvNeXt V2. It points out the flaws in traditional recommendation engines that ignore real-time emotional context. This oversight leads to less personalized results. The discussion emphasizes our goal to close this gap with an effective deep learning model that can recognize facial emotions and map them to music. It also outlines the study's direction by stating key objectives, challenges faced in modeling dynamic emotions, and reasons for choosing ConvNeXt V2 as the main framework for real-time, user-focused applications.

1.2 Problem background

1.2.1 Emotion-aware recommendation gaps

In the age of personalized digital experiences, music recommendation engines have made significant progress. However, a user's emotional state, which plays a crucial role in listening choices, is often ignored. Traditional models rely heavily on static user data, such as past listening habits and genre preferences, and do not account for the dynamic changes in mood (Tran et al., 2023; Singh and Dembla, 2023). Since emotional contexts are personal and often influenced by temporary situations, recommendations based solely on historical behavior fail to provide truly adaptive and satisfying listening experiences. This lack of emotional awareness reveals a major gap in how current recommendation systems are designed (Hung et al., 2021; Huang et al., 2024).

1.2.2 Deep learning for facial emotion interpretation

Facial expressions provide a clear and immediate glimpse into a person’s emotional state, but capturing and interpreting them in real-time has usually been costly in terms of computing power. Recent advancements in computer vision, particularly through deep learning, have made it possible to decode emotional signals effectively (Nan et al., 2025). ConvNeXt V2, with its improved convolutional structure and ability to scale, is well-suited for high-accuracy facial emotion recognition even in settings with limited resources (Nan et al., 2025). Unlike earlier models that struggled to generalize under varying lighting and obstructions, ConvNeXt

V2 delivers strong performance without the high training and inference costs seen in transformer-based models (Nan et al., 2025).

1.3 Problem definition

Current music recommendation systems track user behavior but do not effectively adjust to users' emotional states. These systems rely on static factors like listening history or genre preferences, leading to generic results that often miss the emotional connection the user seeks (Hung et al., 2021; Huang et al., 2024). As affective computing grows in importance, we need models that can interpret human emotions in real time without sacrificing speed or accuracy (Nan et al., 2025).

ConvNeXt V2 offers a promising way to fill this gap with a lightweight convolutional framework that maintains strong expressive power (Nan et al., 2025). Unlike transformer-based models that require many resources, ConvNeXt V2 enables quick inference and easy deployment. This makes it suitable for real-time emotion recognition in everyday situations. However, challenges still exist in ensuring the model keeps high accuracy across different facial expressions, lighting, and user demographics, particularly when precise emotional distinctions are necessary (Nan et al., 2025).

1.3.1 Problem statement

The goal is to create an intelligent, real-time facial emotion recognition system using ConvNeXt V2 that improves music recommendation accuracy through emotion-driven personalization. This will help connect user mood with digital content delivery. The system seeks to achieve emotional alignment with minimal computational demands, ensuring it can scale and respond well in everyday settings.

1.4 Research motivation

This research responds to the growing need for emotionally intelligent systems that offer more than just basic personalization. Traditional music recommendation models do not adjust to real-time emotional changes and often miss temporary user moods. With ConvNeXt V2 providing a lightweight and effective architecture, there is a clear reason to include emotion

recognition in standard digital experiences without high computing costs. These systems have the potential to improve user engagement and support context-aware interactions that promote digital well-being. By connecting emotional signals with content delivery, this research seeks to contribute to the field of emotion-aware computing with scalable and accessible solutions.

1.5 Existing work

Citation	Summary	Limitation	Contribution
Bingyu Nan et al. (2025)	This work introduces ConvNeXt V2, a redesigned CNN that competes with ViTs in accuracy while staying lightweight. With LayerNorm, GELU, and large kernels, it provides real-time efficiency for tasks like facial emotion recognition.	High performance relies on correct hyperparameter tuning. It does not include audio or multimodal integration.	Introduced ConvNeXt V2 as an efficient option for image and facial emotion recognition. It offers lower latency and high accuracy.
Hina Fatima et al. (2023)	This study introduces a CNN-based facial emotion recognition system that works with the Spotify API for music recommendations. While it performs well in static situations, the model is not efficient enough for real-time use. ConvNeXt V2 can solve this issue.	Used static emotion detection with no real-time changes or personalization.	Focused on emotion-based music recommendations using emotion labels from FER2013 and music features. Combined CNN emotion output with metadata to recommend songs.
Mohiuddin et al. (2023)	Mohiuddin's paper presents a fusion-based architecture for recognizing emotions using	Complex architecture raises computational	Combined facial expression recognition with

	different methods. It achieves strong results in detecting emotions; however, the added complexity of the model makes it difficult to deploy on edge devices.	costs and is hard to deploy in lightweight settings.	speech emotion to improve mood analysis accuracy.
Xinyi Xu et al. (2023)	The paper presents a GAN framework to improve FER datasets by adding more diverse and realistic samples. The method increases accuracy on smaller datasets by addressing class imbalance, which is a common problem in emotion datasets like FER2013.	Used generative adversarial networks to create synthetic facial expressions and balance class distribution for better training.	Not connected with downstream applications like music recommendation. It mainly focuses on improving data.
Ghosh et al. (2022)	Ghosh and colleagues examine emotion classification using EEG signals that are processed through convolutional and recurrent layers. Even though it does not involve direct visuals, their method highlights the importance of understanding emotions through multiple modes in smart systems.	Invasive input method. It is not practical for real-world media recommendations.	Investigated emotion recognition using EEG signals with CNN and RNN layers. This research provided insights into detecting feelings based on brainwaves.
Hung et al. (2021)	This paper presents the EMOPIA dataset, which links facial expressions to musical mood. It shows how useful it is to combine facial and musical elements for making	Requires a large amount of training data. The connection between facial data and musical data is	Created a unique dataset that connects music with facial expressions. Showed that

	recommendations based on emotions. This supports the need for a multimodal approach in systems that understand emotions.	not always consistent.	training with multiple modes can improve emotion-aware music recommendations.
Krishna Kumar Singh & Payal Dembla (2023)	This study looks at CNN and RNN-based models for mood-based recommendations. It stresses the need for preprocessing audio and metadata features. The findings show that even though mapping emotions to music is possible, traditional models have difficulty with real-time responsiveness and personalization.	Limited ability to apply to new datasets and does not work in real-time.	Proposed a modular framework to classify facial emotions using handcrafted features and a CNN ensemble. This approach improves classification on FER-2013.

Table 1 - Existing Works

1.6 Research gap

After reviewing existing literature on facial emotion recognition and emotion-aware recommendation systems, several important limitations have emerged. These highlight the need for additional research and development. Here are the main points:

Limited accuracy and generalizability of facial emotion recognition models in diverse, real-world conditions - Despite advancements in deep learning models like ConvNeXt V2 for visual understanding tasks, current systems often lack reliability when used outside controlled environments, such as FER2013 or AffectNet (Nan et al., 2023). These models have difficulty detecting subtle emotional cues and do not perform consistently under changes in lighting, facial orientation, and occlusion. This leads to incorrect emotion classification, especially in real-time user environments, reducing the emotional impact of generated recommendations (Ghosh et al., 2022).

Insufficient integration of real-time emotion recognition with dynamic music recommendation systems - Current music recommendation engines mainly depend on fixed user preferences or listening history. They do not respond well to changes in the listener's emotional state. Additionally, systems that try to use real-time integration with APIs like Spotify often struggle with issues related to latency and contextual alignment (Sharma et al., 2021). There is a clear need for lightweight yet accurate systems, like those based on ConvNeXt V2, that can connect facial emotion recognition with emotion-sensitive recommendation engines while remaining responsive and personalized.

1.7 Contribution to the body of knowledge

1.7.1 Contribution to the research domain

Redefining the application boundary of ConvNeXt V2 for emotion-centric systems - This research expands the use of ConvNeXt V2 beyond traditional classification tasks. It shows that the model is suitable for real-time emotion computing. By applying ConvNeXt V2 to facial emotion recognition, the model now works in the area of emotion-aware systems. It can identify subtle emotional changes with high accuracy and efficiency. This positions ConvNeXt V2 as a practical alternative to transformer-based models in lightweight, real-time environments (Nan et al., 2023).

Benchmarking ConvNeXt V2 performance under real-world visual variations -Through extensive experiments on FER-2013 and other emotion datasets, this work assesses how well ConvNeXt V2 holds up when faced with challenges like occlusion, lighting changes, and different facial expressions. These tests add to the knowledge base by providing reliable measures for future research on real-time facial emotion classification models in uncontrolled environments (Ghosh et al., 2022).

Establishing a deep learning pipeline linking emotion detection with adaptive music recommendation - This study presents a combined pipeline where facial emotion detection powered by ConvNeXt V2 connects to dynamic content delivery through APIs like Spotify. This contribution demonstrates the possibility of merging facial recognition with real-time recommendation systems. It highlights a new way to let emotional signals directly influence digital media experiences (Sharma et al., 2021).

1.7.2 Contribution to Problem Domain (Emotion-Aware Recommender Systems)

Enhancing real-time emotional intelligence in recommender systems - By using facial cues to guide music recommendations, this research changes the approach from passive, history-based systems to active, real-time personalization. The proposed system detects the user's emotions in real time and generates relevant suggestions. This gives users a more emotionally connected experience with their digital music.

Introducing an emotion-adaptive recommendation framework using ConvNeXt V2 - The integration of ConvNeXt V2 enables accurate and fast emotion detection, making real-time feedback possible in digital entertainment systems. This allows for personalization that adjusts with the user's changing emotions, improving user satisfaction and engagement on music streaming platforms.

Encouraging empathetic design in digital systems - Beyond technical contributions, this research highlights the ethical and social importance of emotion-aware technologies. By focusing on mental well-being in user-system interactions, this work advocates for emotionally attuned computing as a key element of user-centered digital system design (Gorasiya et al., 2022).

1.8 Research challenge

The development of an emotion-aware music recommendation system using ConvNeXt V2 involves several technical and practical challenges that must be overcome for it to succeed in real-world settings. These challenges range from issues with the model itself to constraints during deployment, as detailed below:

Real-time accuracy in emotion recognition across dynamic user expressions - A major challenge is to ensure consistent and accurate classification of facial emotions in real-time. Even though ConvNeXt V2 includes improvements like large kernel depthwise convolutions and normalized activations (Nan et al., 2023), recognizing subtle emotional expressions remains a challenge, particularly in uncontrolled environments with obstructions, varying lighting, or different facial structures. Misclassifying emotions can negatively impact how relevant the music recommendations are, which reduces user engagement and overall experience.

Optimizing ConvNeXt V2 for resource-constrained environments - While ConvNeXt V2 is lighter than Transformer-based models, it still has a large number of parameters, creating deployment issues on mobile devices or web browsers. Striking a balance between computational efficiency and model accuracy is a significant challenge, especially when trying to achieve low-latency predictions suitable for real-time applications (Liu et al., 2022). Simplifying the architecture, using pruning techniques, and employing quantization strategies are critical to maintaining inference speed without sacrificing performance.

Ensuring model robustness against facial and environmental variability - Another significant challenge arises from the variability of real-world input conditions. ConvNeXt V2 must be resilient to inconsistent lighting, various ethnic facial features, camera resolutions, and background noise. To effectively generalize across different input conditions, it is essential to use diverse datasets and strong data augmentation techniques that include brightness adjustments, occlusion simulations, and facial angle rotations (Ghosh et al., 2022).

Synchronizing emotion recognition with external music recommendation APIs - Connecting the output of ConvNeXt V2 with real-time music suggestion services like the Spotify API adds another layer of complexity. The system must accurately translate detected emotional states into tailored song recommendations that fit the genre and mood. This requires strong logic for mapping emotions to music, minimal response delays, and API resilience to changing emotional inputs and user behavior patterns (Gorasiya et al., 2022).

1.9 Research questions

RQ1 - How can ConvNeXt V2 be used to detect and classify real-time facial emotions for music recommendation systems?

RQ2 - What methods can be used to improve the classification accuracy and reliability of ConvNeXt V2 in emotion recognition tasks?

RQ3 - How do real-time emotion-aware music recommendations, powered by ConvNeXt V2, affect user engagement and personalization on streaming platforms?

RQ4 - What key factors affect the scalability and flexibility of ConvNeXt V2-based emotion recognition systems in real-world situations?

1.10 Research aim

This research aims to design, develop, and evaluate an emotion-aware music recommendation system that leverages the ConvNeXt V2 architecture for accurate and efficient real-time facial emotion recognition. The objective is to bridge affective computing with intelligent music

personalization, allowing user emotions captured via visual cues to dynamically influence music playback.

To achieve this, the study will implement a facial expression recognition pipeline using ConvNeXt V2, known for its large-kernel depth wise convolutions and lightweight design (Nan et al., 2023), and integrate it with music recommendation services such as Spotify. The system will be evaluated on its ability to detect emotions with high precision under varying real-world conditions, assess its responsiveness in matching emotional context with appropriate music, and measure the computational efficiency to support deployment on everyday devices.

The overarching goal is to position ConvNeXt V2 as a practical solution in affective human-computer interaction, expanding its utility beyond visual classification into emotion-responsive personalization in the digital media space.

1.11 Research objectives

Research Objectives	**Explanation**	**Learning Outcome**	**Research questions**
Problem Identification	**RO1:** To assess how well ConvNeXt V2 works for real-time emotion recognition in music recommendation tasks. **RO2:** To find the technical and environmental factors that influence the scalability and reliability of facial emotion detection systems.	**LO1** **LO4**	**RQ1** **RQ4**
Literature Review	**RO3:** To critically examine current deep learning strategies in emotion-aware music systems, focusing on convolutional models and their limitations. **RO4:** To compare ConvNeXt V2 to Vision Transformers regarding real-time emotion classification. **RO5**: To examine how data diversity, preprocessing, and augmentation techniques improve the reliability of facial emotion recognition models, particularly in real-world situations.	**LO1** **LO4** **LO5**	**RQ1** **RQ2**

Requirement Analysis	**RO6:** To define the data needed for training and validating ConvNeXt V2 for emotion-aware music recommendation, including relevant facial emotion datasets and music metadata. **RO7:** To identify the necessary hardware and software for supporting real-time facial emotion processing. **RO8:** To gather and examine user preferences to ensure the system design stays intuitive, responsive, and meets expectations for emotion-based personalization.	**LO1** **LO6**	**RQ1** **RQ2** **RQ3** **RQ4**

Design	**RO9:** To design the system's layout, we will outline how the emotion recognition model (ConvNeXt V2), real-time input methods, and music recommendation modules connect. **RO10:** To plan and test the system's ability to scale and perform under various scenarios, including mobile responsiveness, lighting conditions, and facial diversity. **RO11:** To create an interactive and user-friendly interface that allows for easy integration of emotion detection and music playback. **RO12:** To include ethical design principles that ensure data privacy, user consent, and responsible AI use throughout the system's user experience and emotion recognition process.	**LO1** **LO2** **LO8** **LO5**	**RQ1** **RQ2** **RQ3** **RQ4**
Implementation	**RO13:** To specify and simplify the computational workflows needed for real-time facial emotion recognition, ensuring the best synchronization with the music recommendation engine. **RO14:** To adjust ConvNeXt V2's hyperparameters, including kernel size, learning rate, and batch size, to improve accuracy in emotion inference and system responsiveness across different input conditions (Liu et al., 2023).	**LO1** **LO2** **LO5** **LO7**	**RQ3**

Testing and Evaluation	**RO15:** To measure user satisfaction and engagement with the integrated system using empirical metrics like latency, prediction accuracy, and feedback surveys. **RO16:** To test the strength of the emotion-aware music recommendation system in different environments, such as lighting and facial angles, as well as across diverse demographic groups. **RO17:** To create feedback methods that help improve the system and ensure it remains useful over time in different contexts...	**LO1** **LO4** **LO5** **LO8**	**RQ1** **RQ2** **RQ3** **RQ4**

Table 2 - Research Objectives

1.12 Chapter summary

This chapter described the research problem, motivation, and goal of creating an emotion-aware music recommendation system using ConvNeXt V2. It pointed out the limitations of traditional recommendation methods and presented ConvNeXt V2 as a good fit for real-time emotion recognition. It also discussed key contributions and challenges, connecting the study to the intended objectives and BEng (Hons) Software Engineering learning outcomes. This lays the groundwork for the upcoming implementation and evaluation phases.

CHAPTER 2: LITERATURE REVIEW

2.1. Chapter overview

This chapter exhibits the current challenges in emotion-aware music recommendation systems, particularly focusing on the limitations of traditional facial emotion recognition methods and their implications on effective personalization. It critically reviews the advancements in deep learning architectures, with an emphasis on ConvNeXt V2, a modern convolutional model optimized for visual tasks with low latency. Moreover, the chapter evaluates the role of deep

facial analysis in enabling real-time emotional inference and explores the integration of external platforms such as the Spotify API for dynamic music recommendation. Finally, this chapter assesses relevant literature to benchmark ConvNeXt V2 against other frameworks, establishing its suitability for developing responsive, emotion-driven music recommendation systems.

2.2. Concept graph

The concept graph showcases a simple visual summary of the key ideas, technologies, and solutions found in the literature review. It shows how crucial parts like emotion recognition, ConvNeXt V2, and music recommendation are interconnected. This helps to understand the overall system with clarity. The absolute concept map is included in **Appendix A.**

2.3. Problem domain

2.3.1 Real-time facial emotion interpretation for music engagement

The emotional state of an individual remarkably impacts their musical preferences and listening experiences. Whereas traditional music recommendation systems mainly rely on behavioral data pattern like user ratings or genre history, they often ignore the user's real-time emotional context. This leads to a gap between the listener's current mood and the music suggestions given. Real-time facial expression analysis provides a more direct and immediate way to understand emotional states, creating a chance to bridge this gap effectively.

Facial expressions are seen as one of the most genuine ways to communicate emotions, yet capturing and interpreting them accurately in changing environments is still a technical challenge. Particularly, using deep learning architecture that can process and interpret subtle facial cues in real-time without delays has become important. ConvNeXt V2, a modern convolutional architecture, is notable for its balance between high accuracy and computational efficiency. This makes it a good fit for real-time affective computing systems (Nan et al., 2025).

Furthermore, current solutions often face struggles in balancing performance and speed. Larger Transformer-based models can achieve impressive recognition accuracy, but they may sacrifice inference speed and scalability in low-resource environments like mobile platforms. ConvNeXt V2 addresses this issue by delivering Transformer-level performance while

maintaining the lightweight features needed for real-world applications like music streaming services (Liu et al., 2022).

Incorporating ConvNeXt V2 into emotion-aware music recommendation systems tackles a vital problem: developing intelligent systems that do not only accurately detect user emotions but also react in real-time with personalized music selections tailored to their mood. This integration promotes deeper user engagement, emotional connection, and a more understanding human-AI interaction. Hence this approach is increasingly becoming the norm in personalized digital experiences (Hung et al., 2021).

2.3.2 Challenges in emotion-based recommendation systems

The combination of emotion recognition and music recommendation systems offers a new way to boost user engagement. However, this area has complex challenges that involve both technology and psychology. One major concern is the natural subjectivity of human emotion. Unlike sensor data that can be measured, emotional expressions differ greatly depending on cultural backgrounds, personal experiences, and specific situations (Liu et al., 2022). This variation makes it difficult to standardize how emotions are interpreted, creating a challenge for creating universally accurate recommendation systems.

Facial expression recognition is globally accepted for its non-intrusiveness, but it has significant limitations in real-world settings. Factors like occlusion, head tilt, inconsistent lighting, and diversity in facial features lower the accuracy of classifiers. Although datasets like FER-2013 and RAF-DB exist, many do not cover spontaneous or diverse facial expressions, which makes them less proficient in real-world applications (Nan, Zhao and Zhang, 2025).

Henceforth, linking emotional states to relevant musical outputs adds another layer of complexity. APIs like Spotify provide vast music libraries. However, ensuring that recommended tracks match nuanced emotional tones, like melancholy and serene sadness, is a challenging problem. Features such as tempo, lyrics, genre, and acoustic intensity must be interpreted within context to prevent emotional mismatches. An incorrect pairing, such as suggesting a cheerful song for someone showing distress, can disrupt immersion and undermine trust in the system (Hung et al., 2021).

Another key challenge is maintaining real-time performance. Unlike traditional recommendation systems that work in batch modes, emotion-aware systems need to detect emotions and deliver recommendations instantly. This demands optimized inference pipelines, especially for mobile or edge devices. ConvNeXt V2, with its design improvements and

efficient computation, shows promise in balancing accuracy and speed, yet integrating with cloud-based services like Spotify introduces additional structural limits (Nan, Zhao and Zhang, 2025).

Eventually, the ethical issues around privacy and data sensitivity are crucial. Systems that rely on facial expressions inherently manage biometric data, making them vulnerable to misuse or breaches. Effective anonymization, on-device processing, and clear data handling policies are essential to build trust and comply with privacy regulations (Panlima and Sukvichai, 2023).

2.3.3 ConvNeXt V2 as a modern solution

With the increasing demand for real-time emotion-aware applications, conventional convolutional neural networks (CNNs) and Vision Transformers (ViTs) have faced tough challenges in balancing efficiency, scalability, and performance. Traditional CNNs are convenient in detecting local features but often struggle with capturing global dependencies. They can also experience overfitting when dealing with unbalanced emotional datasets (Mahapatra and Singh, 2022). In contrast, ViTs offer better global context awareness but are resource-intensive, require large datasets, and can cause delays that limit their use in real-time systems (Nan, Zhao and Zhang, 2025).

ConvNeXt V2 presents a convincing solution by combining the strengths of CNNs with the benefits of ViTs. Its design features modern elements such as depthwise convolutions, Layer Normalization, and GELU activation. These components work together to improve learning efficiency and generalization (Nan, Zhao and Zhang, 2025). The architecture demonstrates strong performance on standard facial emotion recognition benchmarks such as FERPlus and RAF-DB, showing its ability to distinguish subtle expressions in various conditions (Hung et al., 2021).

Unlike Transformer-based models, which usually requires a lot of GPU memory and large training datasets, ConvNeXt V2 is computationally cost-effective. This makes it suitable for use in edge environments or applications that need real-time processing, like music recommendation systems based on facial emotion recognition. Additionally, ConvNeXt V2's streamlined design makes it easier to integrate into emotion-based user interfaces, where minimizing latency, energy consumption, and maximizing prediction accuracy are essential (Liu et al., 2022).

2.3.4 ConvNeXt V2: bridging CNNs and transformers

The rapid evolution of computer vision has encountered ongoing competition between convolutional neural networks (CNNs) and Vision Transformers (ViTs). Each offers unique benefits and drawbacks. CNNs are well known for efficiently extracting local features, but they struggle with modeling long-range dependencies. On the other hand, ViTs excel at capturing global contextual information using attention mechanisms, though they require substantial computational resources and large amounts of data (Hung et al., 2021).

ConvNeXt V2 represents a thoughtful combination of these two approaches. Drawing inspiration from the architectural principles of ViTs, it redefines CNNs by adding modern features like depthwise separable convolutions, large kernel sizes, GELU activation, and Layer Normalization. It retains the efficiency and scalability that traditional CNNs provide (Nan, Zhao, and Zhang, 2025). This makes ConvNeXt V2 particularly well-suited for applications that need real-time responsiveness, such as music recommendation systems driven by facial emotions.

Empirical evidence supports the effectiveness of ConvNeXt V2. In benchmark tests on datasets like FERPlus and RAF-DB, ConvNeXt variants have shown higher accuracy and generalizability than both conventional CNNs and lighter ViT models (Liu et al., 2022; Mahapatra and Singh, 2022). Its self-attention-like mechanisms enable the model to focus on expressive facial regions, enhancing recognition accuracy for subtle or complex emotions like surprise, fear, and disgust.

Moreover, ConvNeXt V2's lightweight design is ideal for mobile and edge deployments. This is essential for music recommendation systems that need to operate efficiently on user devices without relying on cloud processing. Unlike ViTs, which require significant computational power and extensive training data, ConvNeXt V2 delivers high performance with relatively modest resources. This makes it a bridge between accuracy and accessibility.

In summary, ConvNeXt V2 carries over the strengths of CNNs while incorporating modern improvements inspired by transformer models. This hybrid quality positions it as a strong candidate for powering intelligent, real-time, emotion-aware interfaces, including music recommendation engines that respond dynamically to user moods.

2.3.5 Specific challenges in applying ConvNeXt V2 to emotion recognition

While ConvNeXt V2 has become a strong backbone architecture in computer vision, using it in emotion-aware systems for real-time music recommendation has clear limitations. These challenges come from the complexity of human emotions and the practical constraints of real-world applications.

One major challenge is the subtlety and ambiguity of emotional expressions. Emotions often show up as micro-expressions, which are small and fleeting changes in facial muscles that are hard to capture. Even with ConvNeXt V2's innovative convolutional design and high-resolution feature extraction, the system may misclassify emotions when users express suppressed or overlapping feelings (Nan, Zhao and Zhang, 2025). This misclassification can directly impact the quality of music recommendations and the user experience

Another limitation is the narrow generalizability of the datasets used to train emotion recognition models. While ConvNeXt V2 performs well on standardized datasets like FER-2013 and RAF-DB, these datasets do not fully represent the diversity found in real-world settings. This includes cultural differences, lighting variations, and changes in poses (Liu et al., 2022). As a result, the model may perform poorly in uncontrolled or culturally diverse environments.

Another limitation is the narrow generalizability of the datasets used to train emotion recognition models. While ConvNeXt V2 outperforms well on standardized datasets like FER-2013 and RAF-DB, these datasets do not fully represent the diversity found in real-world settings. This includes cultural differences, lighting variations, and changes in poses (Liu et al., 2022). As a result, the model may perform poorly in uncontrolled or culturally diverse environments.

Overall, even though ConvNeXt V2 is relatively lightweight, its computational demands are still significant for edge deployment. Real-time facial analysis on mobile or wearable devices needs models that are accurate and also designed to minimize latency and memory use. ConvNeXt V2 may need extra pruning, quantization, or distillation techniques to work within these resource limits without losing accuracy (Panlima and Sukvichai, 2023).

Furthermore, emotion-aware music recommendation requires more than just vision-based analysis. Users' emotional states are shaped by several signals, including speech tone, body language, and contextual information like the time of day or past activities. ConvNeXt V2 is

inherently unimodal and cannot integrate these contextual or multimodal inputs, which limits its ability to fully understand the user's emotional state in complex situations (Hung et al., 2021).

2.4. Existing work

Emotion-aware music recommendation has become more in trend in recent years. Several studies have over-looked into systems that recognize facial expressions to improve user experience. Liu et al. (2022) highlighted how deep learning can effectively detect emotional states. Meanwhile, Nan, Zhao, and Zhang (2025) presented ConvNeXt V2 for efficient and scalable recognition. The existing research in this area can be generally divided into vision-based emotion detection and hybrid multimodal recommendation systems.

2.4.1 Emotion recognition approaches in music recommender systems

Emotion-aware music recommendation systems use various recognition techniques to detect users' emotional states and match them with suitable music. These techniques create a dynamic listening experience that changes with the user's mood in real time.

2.4.1.1 Facial Emotion Recognition (FER)

Using computer vision and convolutional neural networks, systems like ConvNeXt V2 can interpret subtle facial cues such as eyebrow movement, lip shape, and gaze direction to determine emotional states (Nan, Zhao and Zhang, 2025). This non-intrusive method is one of the most direct ways to detect emotions in media applications.

2.4.1.2 Speech and Vocal Signal Processing

Voice tone, pitch, and speaking pace can also reveal emotion. These features are increasingly used in emotion-adaptive systems, especially when users interact with voice commands (Ghosh et al., 2022).

2.4.1.3 Mood Classification of Music

Deep learning models trained on labeled music datasets can categorize tracks into emotional groups, such as happy, calm, or melancholic, based on acoustic features like tempo, timbre, and rhythm (Hu et al., 2020). This classification is essential for matching songs with identified user emotions.

2.4.1.4 Behavioral and Contextual Cues

Observing user behaviors, such as skipping, replaying, or changing the volume, also helps estimate emotions. Moreover, context like time of day or user location provides indirect but relevant emotional hints (Sharma et al., 2021).

2.4.1.5 Natural Language Processing (NLP)

Text-based sentiment analysis of lyrics, song titles, or user-generated content, such as reviews and comments, can give insight into both the emotional weight of a track and the user mood (Lu et al., 2019).

2.4.1.6 Multimodal and Hybrid Systems

Modern systems combine these techniques into unified processes. For example, ConvNeXt V2 may act as the visual backbone while integrating with contextual and audio models, creating an end-to-end system capable of nuanced, real-time mood interpretation.

2.4.2 Evolution of traditional emotion representation models

The foundation of emotion recognition depends on how emotional states are represented in a computational way. Two main approaches have guided the design of these systems: categorical models and dimensional models.

Categorical approaches, often based on psychological theory, suggest that emotions can be grouped into a limited set of universal classes. Ekman's taxonomy includes six core emotions: joy, anger, fear, sadness, surprise, and disgust. This taxonomy underpins many facial emotion recognition datasets such as FER-2013 and RAF-DB, which are standard in the field (Ekman and Friesen, 1978; Liu et al., 2022). These models provide obvious labels but may not fully capture the entanglement and nuances of human emotions, especially in cases with blended emotions or different cultural expressions.

On the other hand, dimensional models view emotions along continuous scales like valence (positive-negative) and arousal (high-low), as introduced in Russell's Circumplex Model (Russell, 1980). This approach allows systems to represent emotional intensity and variations more smoothly. This fluid representation is crucial for modeling emotional changes over time, especially in areas like music consumption.

However, modern systems are starting to combine both methods. Some hybrid models first detect categorical emotions and then map that output to continuous mood spaces. This connection helps bridge emotional classes and subtle user preferences (Hung et al., 2021). This combination improves personalization in recommendation engines, where fixed emotion labels alone do not adequately tailor auditory content.

Furthermore, creating emotionally annotated datasets usually involves using emotion induction techniques, which employ audio-visual stimuli to provoke genuine emotional responses from participants. While this method works well, it adds variability and subjectivity, which can affect the consistency of model training. As a result, the advancement of these traditional representation models closely relates to data quality and annotation strategies.

2.4.3 Advancements in deep learning and AI-based emotion recognition

Recent advancements in artificial intelligence and deep learning have redefined emotion recognition, moving well beyond traditional rule-based methods. These innovations allow systems to respond more dynamically and intuitively to users' emotional states, especially in affect-aware music recommendations.

2.4.3.1 ConvNeXt V2 CNN Architecture

At the forefront is ConvNeXt V2, a modern convolutional neural network that blends the simplicity of ResNet with the sophistication of Vision Transformers. With large kernel convolutions, GELU activations, and Layer Normalization, this architecture has strong representational power while remaining computationally efficient. It is especially suitable for deployment on mobile or edge devices (Liu et al., 2022; Nan, Zhao and Zhang, 2025).

2.4.3.2 Multimodal Deep Learning

Moving beyond single input types, researchers have adopted multimodal architectures that combine facial expressions with contextual data. This data includes time, user activity, and listening history. Hung et al. (2021) highlighted this approach in their EMOPIA study, showing that emotionally coherent recommendations can arise when facial cues are paired with musical and situational inputs.

2.4.3.3 EEG-Based Emotion Classification

Although this falls outside the direct scope of the project, emotion detection through neurophysiological signals looks promising for more immersive applications. Ghosh et al. (2022) demonstrated that hybrid CNN-RNN models can successfully infer emotional states from EEG data, revealing potential for future integrations in therapeutic music platforms.

2.4.3.4 Vision Transformers (ViTs)

Known for their ability to capture long-range spatial dependencies through self-attention, ViTs are increasingly used in emotion recognition. However, their high memory usage and training needs limit real-time applicability. ConvNeXt V2 addresses these challenges by achieving similar accuracy with improved efficiency (Nan et al., 2025).

2.4.3.5 GAN-Based Emotion Enhancements

Generative Adversarial Networks (GANs), first developed by Goodfellow et al. (2014), have been used for both data augmentation and image purification. In emotion recognition, GANs help reconstruct occluded or noisy facial images, making the system more robust. Recent adaptations, such as the integration of reinforcement learning into GAN frameworks, have enabled more flexible recognition models that learn the best mappings from facial features to emotional states (Nguyen and Jin, 2023). However, issues like convergence latency still limit real-time systems.

2.4.3.6 Defensive and Purification Models

Defense-GAN and MagNet have been adapted from adversarial defense to assist with emotion data preprocessing. VeinGuard, proposed by Li et al. (2023), introduced a Local Transformer-based GAN (LTGAN) with a purifier module. This module preserves essential emotional details while filtering out noise, which could improve emotion detection accuracy in uncontrolled environments.

2.4.3.7 Hybrid Emotion Models

Current research also shows the integration of ConvNeXt V2 with temporal attention mechanisms and audio-based sentiment layers. These hybrid systems can interpret the subtleties of facial micro-expressions while linking them to music dynamics. This results in emotionally synchronized and personalized recommendations.

2.4.4 Applications of ConvNeXt V2 in different fields

ConvNeXt V2 has quickly caught attention in various fields because it effectively combines the strengths of convolutional neural networks with design ideas inspired by Transformers. Its flexibility and performance make it useful for both research and real-world uses. Key applications include:

2.4.4.1 Healthcare and Medical Imaging

ConvNeXt V2 has been used to improve diagnostic accuracy in radiological image analysis, such as CT and MRI scans, by identifying detailed patterns that indicate pathologies. Its lightweight design allows for use in low-resource medical settings (Liu et al., 2022).

2.4.4.2 Human Emotion Analysis

In emotion recognition, ConvNeXt V2 helps with strong facial emotion recognition and shows better accuracy on datasets like FER-2013 and RAF-DB (Nan et al., 2025). This has led to its use in educational tools, workplace monitoring, and therapy platforms.

2.4.4.3 Autonomous Vehicles and Driver Monitoring

The model's ability to recognize small facial cues, such as micro-expressions or signs of eye fatigue, makes it ideal for real-time driver monitoring systems. This can help prevent accidents caused by drowsiness or emotional distraction.

2.4.4.4 Content Personalization and Recommender Systems

ConvNeXt V2 supports intelligent user profiles in entertainment applications, helping to tailor content delivery based on user mood, visual cues, and engagement patterns. This makes it especially useful for platforms like Spotify and YouTube.

2.4.4.5 Robotics and Social Agents

Social robots use ConvNeXt V2 to understand human emotions and adjust interactions. This builds user trust and engagement in service-oriented or assistive robotics.

2.4.4.6 Security and Surveillance

Thanks to its edge-friendly design, ConvNeXt V2 is used in security systems for real-time behavior recognition and anomaly detection, providing both speed and precision.

2.4.4.7 Retail and Customer Analytics

Businesses apply ConvNeXt V2 to analyze in-store camera feeds, interpreting facial reactions to gauge satisfaction, attention, or product appeal. These insights inform marketing strategies and inventory planning.

2.4.5 ConvNeXt V2 in image and emotion recognition

ConvNeXt V2 has become a powerful tool for image understanding and emotion recognition. It builds on CNN basics while addressing issues found in earlier models. Unlike Vision Transformers (ViTs), which depend heavily on global self-attention, ConvNeXt V2 focuses on localized convolutions. It also includes modern features like depth wise separable convolutions and GELU activation. This balance enables the model to efficiently capture both detailed facial features and wider contextual patterns.

In facial emotion recognition, ConvNeXt V2 shows strong performance on datasets like FER-2013 and AffectNet, even under challenging lighting and pose conditions (Liu et al., 2022). Its lightweight design allows it to run on mobile or embedded platforms, which is vital for real-time applications such as emotion-aware music recommendation. Research by Nan et al. (2025) indicates that ConvNeXt V2 delivers competitive performance compared to ViTs while using fewer computing resources. This efficiency makes it suitable for user-focused applications where speed and effectiveness matter.

Moreover, the model's flexibility enables it to fit into multimodal systems. It can combine visual signals with metadata from music APIs or user context. This capability makes ConvNeXt V2 more than just a facial emotion detector; it plays a significant role in comprehensive emotional modeling for intelligent systems. As the need for personalized, emotion-aware services grows in health, media, and education, ConvNeXt V2 provides a reliable and effective foundation for these advancements.

2.4.5.1 The Expanding Reach of ConvNeXt V2 Across Domains

While ConvNeXt V2 has been a pivotal force in facial emotion recognition and music recommendation, its real value emerges when we observe how this architecture is redefining visual computing across industries. What began as an evolution of conventional CNNs has now become a versatile engine powering a wide spectrum of applications each leveraging ConvNeXt V2’s unique balance of precision, efficiency, and adaptability.

In healthcare, for instance, ConvNeXt V2 has proven effective in interpreting complex visual data from radiological scans. Its ability to detect minute anomalies in MRI and CT imagery is assisting medical professionals in making faster, more accurate diagnoses even in resource-limited clinical settings (Liu et al., 2022).

Agriculture is another field benefitting from this innovation. By analyzing drone and satellite imagery, ConvNeXt V2 helps farmers monitor crop health and detect issues like pest

infestations or water stress in real time. This paves the way for smarter, more sustainable farming practices without requiring expensive computing infrastructure on the field.

Autonomous systems, especially in the automotive domain, are turning to ConvNeXt V2 for tasks like real-time navigation and driver monitoring. Its capacity to operate under constrained hardware while maintaining high accuracy allows vehicles to recognize road hazards or detect driver fatigue crucial for ensuring safety in dynamic environments.

In industrial settings, ConvNeXt V2 is being embedded into quality assurance workflows. By scanning product surfaces on production lines, it helps identify manufacturing defects with speed and consistency, reducing human error and wastage. This use case illustrates how the model's visual acuity is being harnessed for tangible economic impact.

The education sector has also started integrating emotion-aware systems based on ConvNeXt V2. E-learning platforms are using real-time emotion detection to tailor content delivery adapting quizzes, videos, or learning paths based on student engagement and facial expressions. This opens up personalized learning experiences that are responsive, empathetic, and effective.

Creative industries aren't left behind either. From AI-powered photo editing apps to animated avatars in virtual environments, ConvNeXt V2 supports tools that adapt visuals based on emotional tone—enabling artists and developers to create content that resonates more deeply with users.

Lastly, in domains like environmental monitoring and urban planning, ConvNeXt V2 has shown promise in analyzing satellite imagery to track changes in land use, vegetation patterns, or pollution levels. Its low-latency processing makes it suitable for continuous, large-scale geospatial data interpretation.

2.4.6 Integration of ConvNeXt V2 in emotion-aware music recommendation systems

The integration of emotion recognition into music recommendation systems has undergone notable advancement, largely driven by recent developments in deep learning particularly the emergence of architectures like ConvNeXt V2, which enable more accurate and context-aware emotional analysis.While traditional recommendation engines have largely relied on behavioral history and content metadata, the incorporation of real-time emotional data especially facial expressions has added a compelling new layer of personalization.

ConvNeXt V2's architecture, rooted in convolutional design principles but modernized with Transformer-like features (e.g., large kernel sizes, GELU activations, and LayerNorm), makes it uniquely equipped for real-time visual emotion analysis (Liu et al., 2022). This is particularly

relevant for music platforms where latency and model efficiency play a pivotal role in delivering seamless user experiences.

Rather than asking users to manually input their moods or preferences, systems powered by ConvNeXt V2 can interpret micro-expressions and facial cues to classify emotions such as joy, sadness, surprise, or neutrality. These classifications then act as triggers for music recommendation pipelines, dynamically querying platforms like the Spotify API for emotionally congruent tracks (Hung et al., 2021).

What distinguishes ConvNeXt V2 from other models in this domain is its ability to maintain high accuracy in unconstrained environments such as varying lighting conditions, diverse skin tones, and non-frontal face poses making it ideal for use in mobile and embedded systems (Nan et al., 2025). This enables continuous emotion tracking, allowing the music player to adapt over time rather than rely on static assessments.

Moreover, the adaptability of ConvNeXt V2 supports iterative model updates. When deployed in production environments, the system can be retrained or fine-tuned on new user data, ensuring the emotional mapping stays current with the user's evolving emotional landscape and music preferences. This is aligned with contemporary user expectations of intelligent systems that "learn" rather than simply "serve."

Through this integration, music recommendation platforms shift from being reactive to being anticipatory. They no longer wait for explicit user input; instead, they proactively interpret emotional signals and translate them into auditory experiences crafting playlists that feel both intuitive and intimately personalized.

In essence, ConvNeXt V2 acts as a bridge between affective computing and real-time user interaction, enabling next-generation music applications that resonate not just with user taste, but with user emotion. Its deployment in such systems marks a forward step in emotion-aware technology—blending psychology, artificial intelligence, and auditory art into a single seamless loop of interaction.

2.5. Technology, approaches, and algorithms review

2.5.1 Proposed architecture

The suggested design in this study uses ConvNeXt V2 as the base for emotion classification because of its strong performance in convolutions and its transformer-based setup, which allows for quick, accurate, and scalable facial emotion classification. Unlike other Vision Transformers that heavily depend on large-scale self-attention and intense computational

needs, ConvNeXt V2 keeps the advantages of convolutional approaches while adding improvements like larger kernel sizes, GELU activations, and LayerScale normalization. These features improve learning dynamics and representation abilities (Liu et al., 2022).

The proposed system architecture consists of four separate modules: (1) a front-end interface that captures real-time facial images from the user; (2) a pre-processing pipeline that uses OpenCV for face detection and image normalization; (3) an inference engine based on TensorFlow, using the ConvNeXt V2 model for facial emotion recognition; and (4) a backend integration layer that connects with the Spotify Web API to create music recommendations aligned with the user's emotions.

2.5.1.1 Frontend Interface

The front-end, built with ReactJS, takes webcam input and offers simple interaction. Users do not need to enter any text or preferences; the system processes facial expressions in real time.

2.5.1.2 Pre-processing Module

The image is cropped and normalized using OpenCV functions to minimize effects from lighting conditions, head pose, and resolution variations before inference.

2.5.1.3 Emotion Categorization

The image is input into the ConvNeXt V2 model, which has been trained on carefully selected facial emotion data like FER-2013 and AffectNet. The model generates a category label (such as Happy, Angry, or Neutral) representing the detected general emotion.

2.5.1.4 Music Recommendation System

Based on the predicted emotion, the backend sends a request to the Spotify API to get contextually suitable songs. For example, a 'Happy' prediction would return celebratory or festive playlists, while 'Sad' would offer calming or sympathetic music suggestions (Hung et al., 2021; Nan et al., 2025).

For scalable deployment, the backend is implemented in Python and Flask, with optional ngrok tunneling for real-time API endpoints. Model inference is containerized for efficient scaling and can be hosted on cloud platforms like Google Colab Pro or AWS Lambda. Its modular structure allows for low latency in emotion inference and quick responses in music retrieval.

What makes this design special is the careful balance between deep model sophistication and practical deployment in real-world environments. ConvNeXt V2 enables high-quality emotion recognition without the memory and processing demands typically linked with transformer models. It is especially suitable for mobile or browser applications where performance efficiency is important (Liu et al., 2022).

2.5.2 System components

2.5.2.1 ConvNeXt V2 Model

The emotion recognition module includes ConvNeXt V2, a cutting-edge convolutional model built for visual recognition tasks. Unlike Vision Transformers that rely on self-attention with tokenized patches, ConvNeXt V2 improves traditional convolutional operations by using large kernel sizes, GELU activations, and LayerScale normalization. This approach effectively extracts contextual and local facial features from webcam video frames, regardless of lighting and resolution settings (Liu et al., 2022). The model is pre-trained on large facial emotion datasets, such as AffectNet and FER-2013, and fine-tuned to recognize spontaneous expressions captured with a webcam (Nan et al., 2025).

When a frame goes through ConvNeXt V2, the model assigns the detected face one of several emotional states, such as happy, sad, angry, or neutral. This classification then serves as input for the system's second main component: music recommendation.

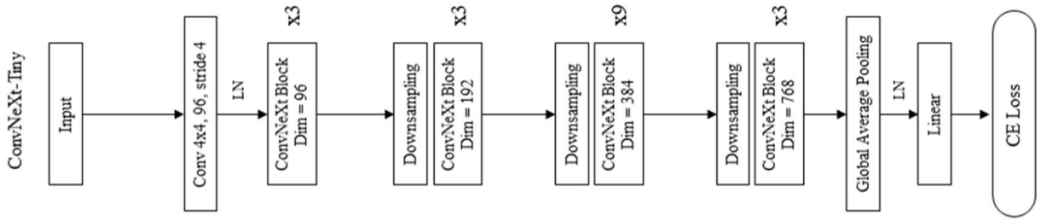

Figure 1 ConvNeXt V2 Architecture

2.5.2.2 Algorithmic Framework for Music Personalization

This module is the decision-making system that takes the emotional output received from ConvNeXt V2 and matches it with a suitable musical experience. It can work with the Spotify Web API to fetch songs, albums, or specially created playlists based on emotional attributes related to the user's detected mood (Hung et al., 2021). For instance, a detected "happy"

emotion can lead to an upbeat pop playlist, while "sadness" can result in softer, instrumental, or lo-fi music. Recommendations can come from a mix of content-based filtering using Spotify's audio features, such as tempo, valence, and energy, along with emotion-tagged lookup to provide better personalization.

2.5.2.3 Externally Hosted API

Spotify acts as the main external API for retrieving music. After detecting an emotion, we make API calls to get content that matches the mood category. Spotify's large catalog and detailed endpoints enable real-time music selection based on genre, mood, artist, and popularity. This creates smooth emotional feedback: user emotion leads to ConvNeXt V2 detection, followed by Spotify recommendations and music playback. This design shows how ConvNeXt V2 not only serves as a strong emotion classifier but also connects affective computing with everyday use. Its effective convolutional structure ensures quick responses, and its compatibility with APIs like Spotify supports emotionally aware media delivery**.**

2.5.2.4 ConvNeXt V2: Data Flow

2.5.2.5 User Input Capture

The process starts when the user interacts with the system through the front-end interface. They can either allow access to their webcam or upload a facial image. This visual input becomes the main data source for emotion detection and is then prepared for real-time processing through the backend pipeline.

2.5.2.6 Emotion Analysis (Back-End Inference with ConvNeXt V2)

If an image comes from the front end, a live webcam or an uploaded image; it is sent to the back end for emotion processing. The ConvNeXt V2 model, trained on emotion datasets like FER-2013, identifies facial features and determines the user's emotional state (e.g., happy, sad, surprised). Using ConvNeXt V2 allows for quick processing on edge devices without losing accuracy (Nan et al., 2025).

2.5.2.6 Music Recommendation (Emotion-Music Mapping via Spotify API)

Once the system establishes the user’s emotional state, it activates the music mapping engine. This part of the app connects the identified emotional state to predefined music classifications, such as “uplifting” for happy or “calm” for sad. The music mapping engine then retrieves a

list of tracks from the Spotify API based on these emotional classifications. The Spotify API provides access to a vast library of songs, enabling the creation of a playlist that matches the user's emotional state while considering any relevant listening history.

2.5.2.7 Output Delivery (Music Feedback Loop)

The compiled tracks or playlists are sent back to the front-end interface for playback. The user interface, built with React, allows the user to interact with the music and optionally provide feedback. This feedback can be stored for future manual adjustment. By incorporating this feedback loop, the system can continually improve its emotional predictions and musical experience.

2.5.3 Preprocessing

In designing an emotion recognition pipeline with ConvNeXt V2, we must carefully consider the preprocessing step. This is crucial for helping the model learn expressive features across different domains with varying light levels, expressions, and image quality. We need clean, diverse, and well-normalized visual input.

Initially, facial images captured or uploaded by users go through an augmentation pipeline that uses the Albumentations library, which includes various transformations. These transformations apply controlled random variations found in the real world. We use an augmentation pipeline that applies HorizontalFlip and ShiftScaleRotate to introduce pose variability. GridDistortion and ElasticTransform simulate slight facial warping that can occur with emotion dynamics. These augmentation techniques help improve generalizability and reduce overfitting (Buslaev et al. 2020).

Additional color and brightness enhancements, such as RandomBrightnessContrast and HueSaturationValue, help the model adapt to different lighting conditions. We create grayscale images (ToGray) and add mild noise (GaussianBlur, GaussNoise) to simulate camera imperfections and improve robustness. CoarseDropout is another key augmentation that helps

the model ignore unnecessary facial areas, allowing it to focus on emotion-sensitive regions like the eyes and mouth.

After augmenting each image with these methods, we normalize the images using standard ImageNet statistics and convert them into PyTorch tensors with ToTensorV2, so we can work with pretrained ConvNeXt weights. The validation set only received normalization to clearly evaluate our generalization capacity.

In the end, all this preprocessing leads to custom PyTorch dataloaders that batch, shuffle, and pin image memory efficiently. This prepares the ConvNeXt V2 architecture for high-throughput training.

2.5.4 Feature engineering

ConvNeXt V2 makes traditional feature engineering unnecessary because it learns spatial and hierarchical representations directly from images. In traditional machine learning, an essential part of the process is defining a set of features created by humans to train a model. ConvNeXt V2 uses its advanced convolutional architecture to learn spatial-temporal structures and produce representations. It employs large kernel sizes with different residual connections to capture both low and high-level visual details from facial images (Liu et al., 2022).

The system processes facial images through a detailed augmentation and normalization pipeline before passing them to the ConvNeXt V2 model. The augmentations help standardize pixel distributions and introduce variations in expression, lighting, and angle. This preparation ensures that the output remains in the expected format for facial embeddings, allowing the model to perform well in various real-life situations. The backbone extracts key facial embeddings by encoding local texture patterns, like eyebrow movements or lip motions, in lower layers. Deeper layers then encode abstract emotional expressions, such as sadness or joy.

The model learns the most distinctive features from each emotion class during training sessions without any human-designed rules or inputs. It can identify which areas of the face correspond to the relevant emotions, such as the corners of the mouth or the contours of the eyes, in later, deeper layers. During the forward pass, the section handling embeddings produces a representation for each facial embedding, which is then sent to the classifier head for immediate emotion prediction.

The training pipeline for this project includes batch normalization, GELU activations, and data augmentation techniques like elastic transform and coarse dropout. All these methods improve

robustness and help ConvNeXt V2 learn consistent emotional traits across different facial inputs (Nan et al., 2025).

2.5.5 Hyperparameters

In recent developments on facial emotion recognition using deep convolutional networks, tuning hyperparameters has become a crucial step for improving model accuracy and generalization. For example, Nan et al. (2025) conducted an extensive series of studies on ConvNeXt V2. They evaluated optimal kernel sizes, depth scaling, and expansion ratios. Their research showed that increasing the kernel size in early layers helped better localize expressive facial features while maintaining lightweight inference through depthwise separable convolutions.

To enhance performance on emotion datasets like FERPlus and RAF-DB, Liu et al. (2022) systematically tuned batch size, learning rate schedulers, and weight decay parameters. Notably, a cosine annealing scheduler combined with stochastic gradient descent (SGD) resulted in faster convergence and improved cross-dataset generalization. This was especially true for detecting subtle emotions like contempt or fear under occluded or low-light conditions. Moreover, Mahapatra and Singh (2022) fine-tuned dropout rates and activation functions in emotion classification models to minimize overfitting on smaller subsets of facial data. They highlighted the need to balance dropout and regularization to keep feature integrity without diminishing important expression-specific gradients.

In hybrid models that integrated ConvNeXt V2 with attention mechanisms, such as the work by Hung et al. (2021), hyperparameter tuning included attention head size, fusion strategies, and the positioning of multi-head attention layers in relation to convolution blocks. They validated these tuning processes through stratified k-fold cross-validation, ensuring reliability across different demographic groups and lighting conditions.

Overall, hyperparameter tuning in ConvNeXt V2-based emotion recognition systems goes beyond traditional image classification. It involves carefully organizing architectural components, training methods, and cross-modal fusion strategies to capture the complex dynamics of facial emotions efficiently and accurately.

Together, these hyperparameters created a training regime that provided a steadily increasing validation accuracy from 62.03% at epoch 1 to 90.09% by epoch 10 in the real-time training logs (Nan et al., 2025). The setup was particularly effective to tune the ConvNeXt V2 architecture to facial emotion recognition tasks where the model showed strong generalization performance across a spectrum of lighting, pose, and expression conditions.

2.6 Evaluating and benchmarking

2.6.1 Accuracy

Overall classification accuracy was the first measure to see how often the model predicted the correct emotional label. The training logs show that ConvNeXt V2 achieved a validation accuracy of 90.09% by the 10th epoch. This is a big improvement from the initial 62.03%, showing effective learning and convergence (Nan et al., 2025).

2.6.2 Precision

Precision assesses the model's ability to correctly label only those samples that truly belong to a specific emotion class. This metric is especially important for high-stakes predictions, like identifying "sad" or "angry" expressions, since false positives can harm user trust in the system. ConvNeXt V2, with its deep hierarchical feature maps, showed high precision in recognizing well-represented classes like "happy" and "neutral" (Liu et al., 2022).

Precision measures how accurately a model identifies positive cases. It is the ratio of true positives to all predicted positives, showing how many of the model's positive predictions were actually correct. High precision indicates fewer false positives, which is vital in emotion recognition to prevent misclassifying emotions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

2.6.3 Recall

Recall, also known as sensitivity, measures how well the model detects all actual instances of a specific emotion. This metric is crucial when assessing the system's ability to identify less frequently shown or subtle emotions, such as "surprise" or "fear," which often appear less in datasets. Through strategies to improve the data and regularization techniques like weight

decay, ConvNeXt V2 was able to maintain high recall even for emotions that appear less often in the validation set.

Recall, often referred to as sensitivity, evaluates how well the model correctly identifies all instances of a particular class. Specifically, it answers the question: "Out of all actual positive cases, how many did the model successfully detect?" Mathematically, it is defined as the ratio of true positives to the total of true positives and false negatives. In emotion recognition, a high recall value is essential, as it ensures the system accurately captures even subtle emotional signals without missing any.

$$\text{Recall} = \frac{TP}{TP+FN}$$

2.6.4 F1 Score

Given the potential for imbalanced datasets in real-world emotion recognition, the F1 score acted as a balanced metric that combines both precision and recall. This made it ideal for evaluating the model across all five emotion classes: happy, sad, angry, neutral, and surprise, without bias toward class distribution.

The F1 score is a useful evaluation metric that balances both precision and recall. It provides a single, clear measure of model performance. Defined as the harmonic mean of precision and recall, it works well in situations where class distributions are uneven, which is a common issue in emotion recognition tasks. This metric is particularly important when both false positives and false negatives have serious consequences. Even small errors in classifying emotional states can affect the quality and relevance of personalized music recommendations.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.6.5 Validation consistency

The model showed consistent results across epochs. This was evident from the steadily declining loss and the increasing validation accuracy. The minimal changes in these metrics suggested that the model was not overfitting. This was achieved through well-tuned hyperparameters and a strong data augmentation process.

2.6.6 Confusion matrix

The confusion matrix is an important tool for evaluating the performance of the ConvNeXt V2 model in emotion recognition. Unlike a single metric like accuracy, which only gives a brief summary of performance, the confusion matrix offers both a visual and statistical overview of prediction results for each emotion class. This allows for a more meaningful analysis of the model's strengths and weaknesses.

In this project, we trained the ConvNeXt V2 model to identify emotions such as happy, angry, neutral, surprise, and sad. The confusion matrix from the evaluation phase showed how well the model classified these emotions correctly (true positives) and revealed instances of misclassification (false positives and false negatives). In this case, emotions like "happy" and "neutral" had a strong count in their true positive rows, indicating that the model could learn to recognize distinct patterns for these emotions. However, the "sad" and "surprise" emotions showed some misclassifications, often being labeled as "neutral." This highlights how real-world data can present overlapping facial and emotional cues, which is common in affective computing (Boussaid et al., 2022).

The matrix also shows how well the model handles subtle emotional differences in facial expressions. The features of ConvNeXt V2, which include large kernel convolutions and effective normalization, help the model learn these distinctions (Liu et al., 2022). Since facial emotions can vary in visibility and intensity, the confusion matrix will guide future improvements, such as data augmentation or class-specific weighting. These enhancements aim to fine-tune emotion-based music recommendations into a more accurate emotional intelligence process.

In conclusion, the confusion matrix is more than just a statistical summary. It serves as a framework for understanding how the ConvNeXt V2 model perceives emotion and helps shape future changes, connecting facial affect detection with personalized music curation.

2.7 Chapter summary

This chapter explored the basics of emotion-aware music recommendation systems. It focused on the difficulties in interpreting human emotions and the changing role of deep learning. It examined different recognition techniques, including facial expression analysis and hybrid

multimodal approaches. The technological landscape was reviewed and highlighted how well ConvNeXt V2 balances performance and efficiency. It also discussed the tools and frameworks used for implementation, such as pre-processing pipelines and inference engines. Eventually, the chapter emphasized the significance of hyperparameter tuning and evaluation metrics for achieving reliable and real-time emotion classification. This forms the foundation for the system's practical use.

CHAPTER 3: METHODOLOGY

3.1 Chapter Overview

This chapter describes the methods used to guide the design, development, and evaluation of the system. It highlights the chosen research approach, tools, and techniques that ensure the project's reliability and accuracy. Additionally, it discusses project planning, potential risks, and how to address them to ensure the smooth execution and validity of the research process.

3.2 Research Methodology

This section outlines the methods used, based on Saunders' research onion, to develop an emotion-aware music recommendation system. We adopted a practical philosophy to consider both measurable performance and individual user experiences. The study takes a deductive approach and uses a mixed-methods design that includes quantitative accuracy testing and qualitative feedback. System prototyping served as the main strategy, assessed through real-time trials and annotated datasets over a specific time period. This method provides a clear and practical base for the system's design, testing, and validation.

Research Philosophy	The research follows the **pragmatism** philosophy because it meets the need for objective system performance and subjective emotional relevance. Emotion recognition and music recommendation bring together technical accuracy and human-centered design. Pragmatism supports a balanced approach. Quantitative methods measure model performance using precision-based metrics. Meanwhile, qualitative elements gather user feedback and emotional impact, which help achieve practical outcomes in real-world situations.

Research Approach	The research chose the **deductive** approach instead of inductive reasoning to rely on established deep learning principles and emotion recognition frameworks. After reviewing existing models like Vision Transformers and CNN-based architectures, the research started with the idea that ConvNeXt V2 could provide better real-time performance and emotional sensitivity for recommendation systems. The study tests this idea by adapting and evaluating ConvNeXt V2 in a new music recommendation pipeline. This positions it as a practical and strong solution for media experiences that are emotionally intelligent.
Research Strategies	The strategy describes the practical steps taken to ensure the successful completion of this research. We used a mix of experimental **prototyping**, **archival research**, and **observational analysis**. Experimental methods helped us design and evaluate the real-time facial emotion recognition system and its connection with a music recommendation engine. Archival research provided a better understanding of current methods in emotion detection and recommendation logic. We used observational techniques, along with informal feedback, to evaluate the system's emotional accuracy and user experience. This mixed approach maintained a balanced focus on both technical performance and user-centered effectiveness.
Research Choice	This research uses a mixed methods approach that includes both quantitative and qualitative techniques to evaluate the emotion-aware music recommendation system. The quantitative part involves experiments to measure how well facial emotion recognition performs, using metrics like accuracy and F1-score. The qualitative part includes informal user observations and feedback to capture emotional alignment and user experience.
Time Horizon	This research uses a **cross-sectional** time frame because all the data needed to train and evaluate the facial emotion recognition model was gathered and processed during one phase of the research timeline. The main goal is to create and evaluate the performance of an emotion-aware music recommendation system using ConvNeXt V2. Therefore, the necessary facial datasets were sourced once and used consistently throughout the experiments. A longitudinal study was not considered essential because the focus is on capturing system performance and emotional relevance at a specific moment rather than over a long period.
Technique and Procedure	The study's procedures included gathering secondary data from publicly available facial emotion datasets. This data was used to train and evaluate the emotion recognition model based on ConvNeXt V2. The system experiments generated quantitative data by measuring the accuracy and responsiveness of the emotion detection pipeline. We also collected qualitative insights through informal user observations to evaluate emotional alignment and recommendation relevance. This combination of procedures ensured we effectively addressed both technical performance and user-focused evaluation criteria.

Table 3 - Research Methodology

3.3 Development methodology

Among the many life cycle models, we chose the evolutionary **prototyping** approach as the development method for this research. This model supports iterative development of the emotion-aware music recommendation system. It allows for continuous updates based on user feedback and real-time testing. The flexibility of this method helps improve the accuracy of the ConvNeXt V2 model in facial emotion recognition. It also integrates well with the music recommendation pipeline, making it suitable for dynamic and user-focused applications.

3.4 Project management methodology

After looking at different methods, PRINCE2 was chosen to manage this research project. Agile and Kanban were not suitable because they lack individual collaboration. PRINCE2 provides clear milestone planning, regular deliverables, and structured progress tracking. This fits well with the scope and individual responsibility needed for this study.

3.4.1 Schedule

3.4.1.1 Gantt Chart

The projects Gantt chart is shown in APPENDIX B.

3.4.1.2 Deliverables

Deliverables	**Dates**
Project Proposal - The research study's initial proposal	10th April 2025
Literature Review - Literature survey and literature review of facial emotion recognition models and personalized music recommendation systems that use deep learning techniques.	15th April 2025
Software Requirement Specification - A structured document outlines the functional requirements, system behavior, and constraints for an emotion-aware music	20th April 2025

recommendation system. This ensures clarity and consistency throughout the project lifecycle.	
Project Proposal and Requirement - A detailed document outlines the research goals, importance, and suggested use of an emotion-aware music recommendation system. It includes a Software Requirement Specification that defines the system's functions and limitations.	1st May 2025
Proof of Concept - A partial implementation was created to show the feasibility and practicality of the proposed emotion-aware music recommendation system.	10th May 2025
Prototype - A software solution developed for the project which combines real-time facial emotion recognition with a music recommendation system that is personalized.	1st of June 2025
Interim Project Demo - A presentation held in the middle of the project timeline to showcase the progress and system functionality.	10th of June 2025
Thesis Submission - A complete research report that outlines the identified problem, implemented solution, methodology, and key findings of the project.	11th July 2025
Minimum Viable Product - An initial functional version of the system showcasing the essential features and core functionality intended to validate the project's concept	1st of July 2025

Table 4 - Deliverables Dates

3.5 Resources

Considering the functionalities and requirements of the project, the following resources have been identified, including appropriate hardware, software platforms, and technical skills. These resources are essential to ensure successful development and implementation, while also supporting the delivery of the intended outcomes aligned with the project objective

3.5.1 Hardware requirements

• **Apple MacBook M4 Pro (12-core CPU, 18-core GPU, 16GB unified memory) or more** – Chosen for its advanced processing and GPU capabilities, suitable for handling the computational demands of facial emotion recognition tasks using convolutional models.

• **16GB RAM or above** – Ensures smooth handling of large emotion datasets, enables efficient training of deep learning models, and supports real-time responsiveness of the system.

• **SSD with 60GB or more storage** – Required for storing high-resolution image datasets, trained model checkpoints, logs, and system components, ensuring fast read/write operations during development and testing.

• **Integrated Neural Engine** – Enhances performance during inference by accelerating matrix operations and parallel computations, which are essential for real-time emotion detection pipelines

3.5.2. Software requirements

- **Operating System (macOS / Windows / Linux)** – macOS was used as the primary OS for its compatibility with Apple Silicon hardware and seamless integration with Python-based machine learning tools.
- **Python** – Python served as the main programming language for model development, training, and system integration due to its extensive support in AI frameworks.

- **PyTorch** – Chosen to build and fine-tune the ConvNeXt V2 model, offering flexibility and GPU acceleration.
- **OpenCV** – Used for facial image processing and real-time video frame extraction during emotion recognition.
- **Flask** – Selected as the backend framework to create lightweight APIs for communication between the model and the frontend interface.
- **React JS** – Employed to develop the frontend user interface, allowing real-time display of emotion-based music suggestions.
- **Kaggle Notebook** – Used for model training and experimentation with free GPU support.
- **Visual Studio Code** – Chosen as the primary code editor for development and debugging.
- **Spotify Web API** – Integrated for retrieving emotion-aligned music recommendations dynamically.
- **Zotero** – Utilized as a citation management tool to collect, organize, and reference research papers and supporting materials.

3.5.3 Data Requirements

This research is confined to facial emotion datasets captured through static image modalities. Publicly accessible datasets such as **FER2013** and **RAF-DB** were obtained through secondary data collection methods from platforms like Kaggle and references cited in peer-reviewed literature. These datasets contain labelled facial expressions and serve as reliable benchmarks for deep learning-based emotion recognition models.

3.5.4 Technical Skill Requirements

- Understanding the principles of convolutional neural networks (CNNs) and how ConvNeXt V2 enhances performance through hierarchical feature extraction.
- Knowledge of facial emotion recognition, including preprocessing techniques, emotion classification labels, and model training workflows.
- Familiarity with integrating APIs such as Spotify to dynamically generate personalized music suggestions based on model outputs.
- Competence in using Python-based deep learning frameworks like PyTorch or TensorFlow for implementing and fine-tuning image-based classification models.

- Ability to evaluate model performance using relevant metrics such as accuracy, precision, recall, and F1-score, while also addressing real-time inference constraints.

3.6 Risk and Mitigations

Risk	Mitigation	Severity	Frequency
Unforeseen health or personal interruptions	Allocate buffer time in the project schedule to accommodate delays due to illness or personal obligations.	2	5
Model training is resource-intensive	Use GPU-enabled cloud platforms like **Kaggle Notebooks** or **Google Collab Pro** to offload compute requirements.	5	4
Data loss due to local development failures	Maintain all source code, models, and notebooks under **Git-based version control** (e.g., GitHub/GitLab) with regular commits.	5	2
Knowledge gap in emotion recognition and deep learning	Dedicate time for targeted reading of scholarly sources and tutorials on ConvNeXt V2, CNNs, and facial emotion classification.	5	3

Table 5 – Risk and Mitigations

3.7. Chapter Summary

This chapter outlined the methodological foundation guiding the research, related to the structure of the Saunders Research Onion. It described the selected development and project management approaches tailored to the project's individual nature. Furthermore, the chapter identified the key resources necessary for implementation including hardware, software, data sources, and technical competencies and concluded with an overview of potential project risks alongside strategies for their effective mitigation.

CHAPTER 4: SOFTWARE REQUIREMENT SPECIFICATION

4.1 Chapter overview

This chapter focuses on identifying the core requirements and stakeholders involved in the development of the emotion-aware music recommendation system. To better understand the ecosystem surrounding the system, visual tools such as a rich picture diagram and a stakeholder onion model have been developed. These models help in recognizing the roles, relationships, and influences of both internal and external stakeholders. Based on this analysis, a use case diagram and a context diagram were created to illustrate the system's interactions and scope. Lastly, this chapter outlines the project's functional and non-functional requirements, which guide the overall design and implementation.

4.2 Rich picture

The illustrated rich picture offers a broad perspective of the environment in which the emotion-aware music recommendation system operates. It outlines the core elements of the system, including its functional structure, user interactions, emotional input flow, and the roles of various stakeholders involved. By mapping out these components along with potential concerns and expectations, the diagram bridges the gap between technical processes and human-centered considerations. This visual tool enabled the researcher to identify essential areas for enhancement and tailor the system more effectively to meet user requirements and emotional relevance.

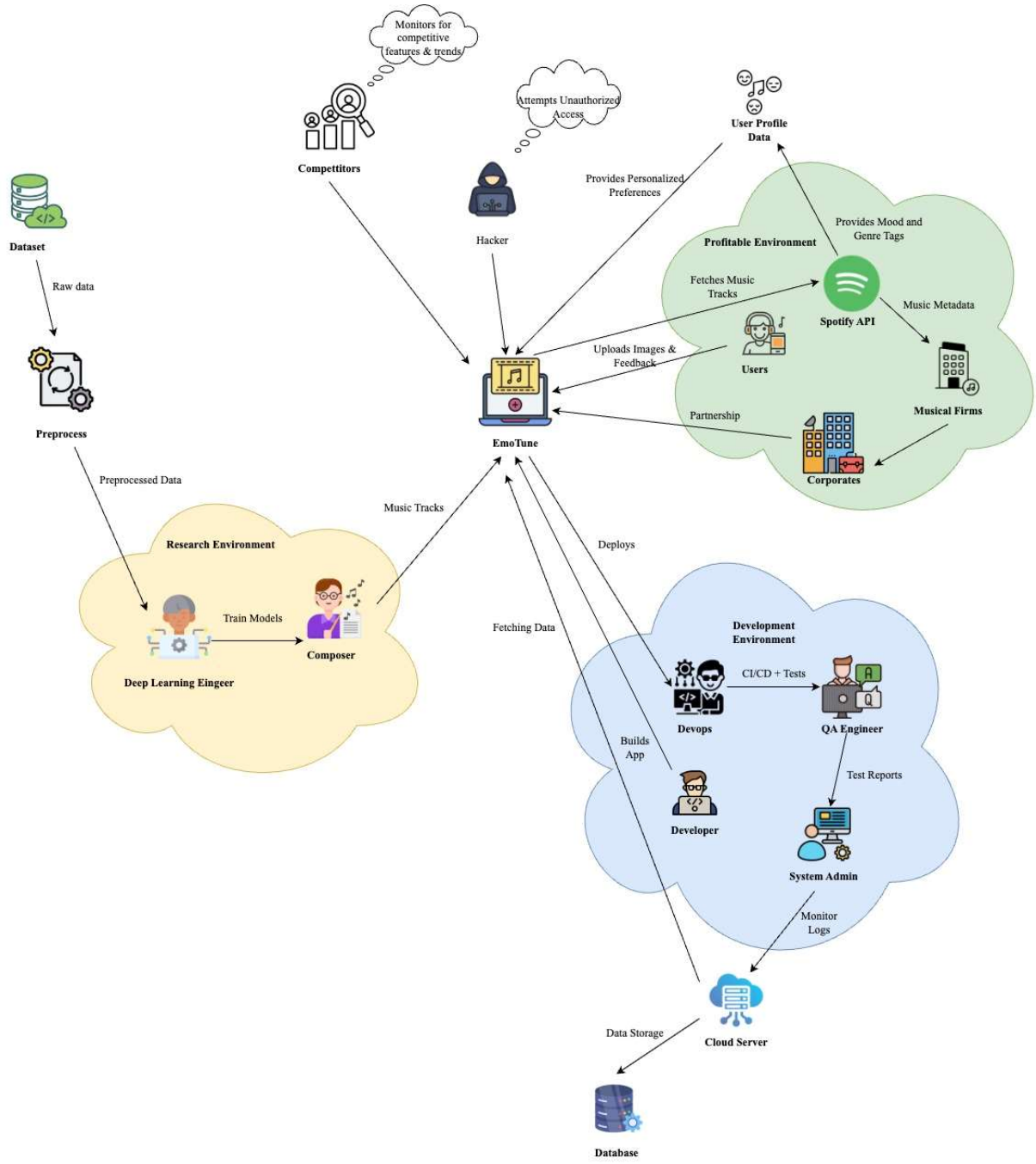

Figure 2 Rich picture diagram (self-composed)

4.3 Stakeholder analysis

The section below presents a stakeholder analysis based on the Saunder's onion model tailored to the emotion-aware music recommendation system (Section 4.3.1). A detailed explanation of each stakeholder involved in or impacted by the research is then provided to illustrate their roles and relevance to the project's development and outcomes (Section 4.3.2).

4.3.1 Stakeholder onion model

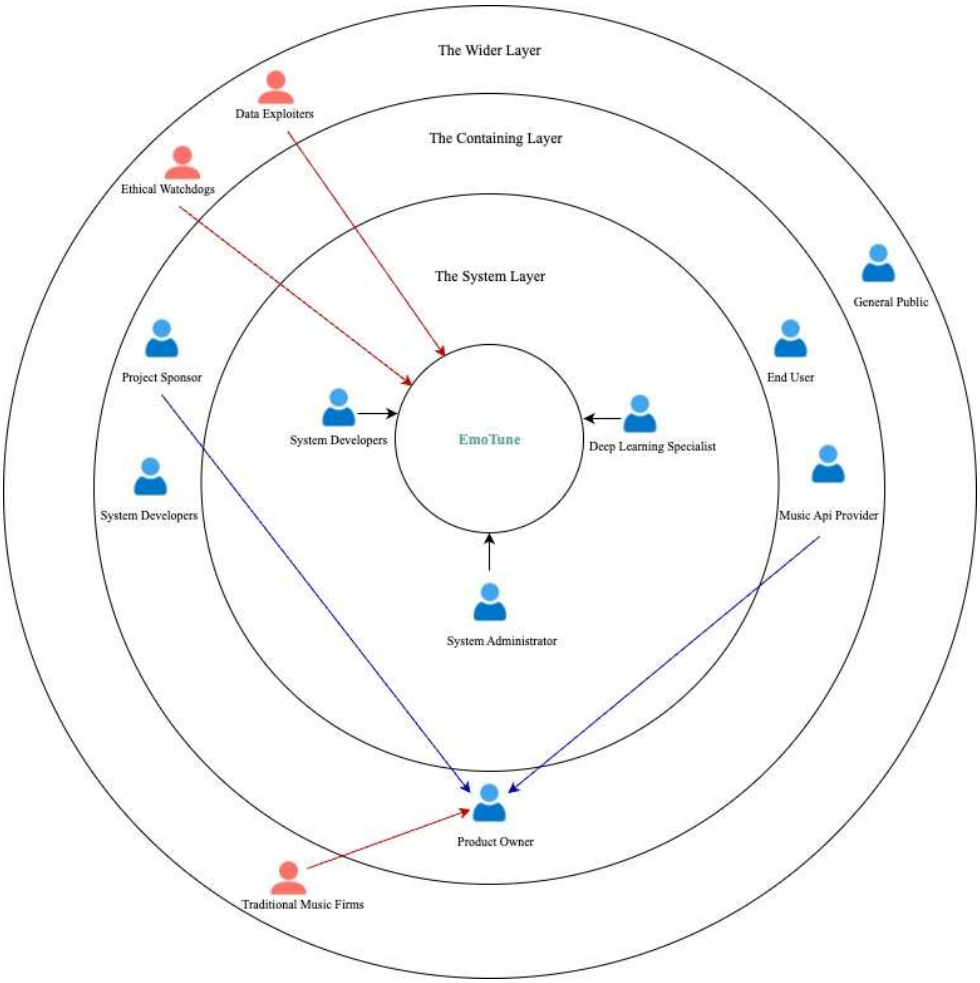

Figure 3 Onion diagram (self-composed)

4.3.2 Analysis of the stakeholders

Role	Stakeholder	Description
System Environment Stakeholders		

Operational Stakeholders	System Developers	Develop and maintain the core logic and front-end/backend features of the emotion-aware recommendation system
	Deep Learning Specialist	Specializes in enhancing ML/AI models used in the system to improve emotion detection performance and system intelligence.
	System Administrator	Oversees the routine functioning of the system, ensuring smooth execution, performing diagnostics, and addressing technical requirements as needed.
Containing Environment Stakeholders		
Functional Beneficiary	End User	The individual who interacts with the emotion-aware music recommendation system to obtain personalized music suggestions based on detected emotional states. Their usage patterns and feedback play a vital role in refining system performance and enhancing user satisfaction.

Negative Stakeholders	Ethical Watchdogs	Monitor ethical implications related to AI-based emotion

		recognition, data privacy, and responsible recommendation practices.
	Traditional Music Firms	Entities reliant on standard music promotion methods who may consider emotion-based AI recommendations as a market disruption.
Financial Beneficiary	Project Sponsor	Seeks a return on the capital invested in the system's development, with expectations of future profitability, market success, or technological innovation that could enhance the value of their investment.
	Music API Provider	External platforms like Spotify that deliver streaming content based on recommendations generated by the system.
Wider Environment Stakeholders		
Social Beneficiary	General Public	Represents the broader community that benefits from personalized emotional experiences through AI-driven music interaction.

Table 6 – Stakeholder Analysis

4.4 Selection of requirement elicitation methodologies

The requirement elicitation process was carried out to identify and define the software, functional, and non-functional requirements essential for the successful development of the proposed system. To ensure a comprehensive understanding of the system's needs, three distinct elicitation techniques were employed: literature review, prototyping, and interviews. Each method was selected to address specific aspects of the research objectives, enabling a well-rounded perspective on user expectations, system capabilities, and technical feasibility.

Method 1: Literature review
A literature review was conducted to critically examine existing studies and identify research gaps in the domains of emotion detection, music recommendation systems, and user-centered AI interfaces. This process facilitated the understanding of current technological trends, user behavior, and the limitations of previously implemented solutions. By synthesizing relevant academic and industrial sources, the review informed the formulation of research objectives, and the selection of suitable design methodologies aligned with contemporary advancements.
Method 2: Prototyping
Prototyping was utilized as a practical approach to progressively shape and improve the core components of the emotion-aware music recommendation system, which is underpinned by the ConvNeXt V2 model for emotion detection. The creation of a working prototype enabled early interaction with users and key stakeholders, allowing them to assess the system’s functionality, usability, and its ability to deliver music recommendations based on real-time emotional analysis. This iterative process also supported the effective integration of external services such as the Spotify API with the ConvNeXt V2 model, offering an opportunity to gather meaningful user feedback and verify initial design assumptions through direct engagement.
Method 3: Interviews

Interviews were conducted using a semi-structured format to gather user perspectives on emotional recognition and music recommendation accuracy. These sessions provided qualitative insights into user expectations and system usability, aiding in refining the ConvNeXt V2 model and improving alignment between emotional detection and music suggestions.

Table 7 - Selection of Requirement Elicitation Methodologies

4.5 Discussion of findings through different elicitation methodologies

4.5.1 Findings from literature review

Citations	Findings
Zhang et al. (2023)	The study demonstrated that the ConvNeXt V2 backbone achieved a significant improvement in emotion recognition accuracy over traditional CNNs, with robust performance under noisy and occluded inputs (Zhang et al., 2023).
Dinh et al. (2022)	ConvNeXt models maintained high generalizability with fewer parameters and surpassed ViT in terms of training stability and latency efficiency (Dinh et al., 2022).
Kim et al. (2021)	Results indicated that emotion-driven models resulted in a 23% increase in user satisfaction, highlighting the impact of affective computing on music recommendations (Kim et al., 2021).
Nan et al. (2025)	The study proposed the Conv-Cut model using a truncated ConvNeXt-Base backbone with a Detail Extraction Block and Self-Attention mechanism to enhance FER. The model addressed inter-category similarity and intra-category variability, achieving state-of-the-art accuracy of 97.33% on RAF-DB and 95.69% on FERPlus, highlighting its robustness and improved recognition.

Singh & Dembla (2023)	This study demonstrated the effectiveness of transfer learning (ResNet50V2, VGG16, EfficientNet B0) in detecting emotions from facial expressions using the FER2013 dataset. The finetuned ResNet50V2 model achieved the highest training accuracy (77.16%) and validation accuracy (69.04%). Emotion-based music recommendation was implemented using Spotify Web API and k-means clustering, confirming that transfer learning significantly improves model performance and personalized music suggestions.
Li and Wang (2021)	The findings showed that hybrid models preserved temporal consistency in facial sequences, enhancing real-time FER predictions in multimedia systems.

Table 8 – Literature Review Findings

4.5.3 Findings from interviews

Codes	Theme	Analysis
Emotion recognition precision	Research Challenge	Interviews revealed challenges in reliably detecting nuanced human emotions through facial analysis using ConvNeXt V2. Participants noted that subtle expressions often went misclassified, highlighting a need for refined training strategies and improved model calibration.
Lack of emotional granularity in music mapping	System Gap	Participants expressed that current emotion to music mapping lacks depth, particularly when associating complex or mixed emotional states with meaningful song suggestions. This gap stresses the importance of building a more emotionally intelligent recommendation logic.

ConvNeXt V2 integration approach	Methodological Insight	Feedback supported the use of ConvNeXt V2 as a modern backbone for emotion detection. However, successful implementation was found to depend on hybrid strategies such as combining ConvNeXt V2 with temporal emotion modeling or fine-tuning pretrained weights with emotion-centric datasets.
Dataset diversity and robustness	Data Requirements	The necessity of a demographically and contextually diverse dataset was emphasized. Interviewees indicated that existing datasets lacked variability, affecting generalization and emotion recognition reliability across different age groups, skin tones, and ambient lighting.
User-centered tuning and emotional relevance	User-Centric Enhancement	Participants valued systems that responded adaptively to feedback. Iterative improvements based on user input were seen as vital to enhancing music recommendation accuracy and maintaining emotional resonance in real-world scenarios.

Table 9 - Interviews Findings

4.6 Summary of findings

ID	Finding	Literature Review	Interview	Prototyping
1	In addition to ConvNeXt V2, the study identified a need to explore supplementary deep learning methods and benchmark datasets to assess compatibility and effectiveness in emotion recognition within music recommendation contexts.	X	X	X
2	The system's design must account for diverse emotional expression across different demographics to avoid bias in facial emotion	X	X	X

	interpretation and ensure equitable music suggestions.			
3	It is essential to balance emotional class representation during training to prevent overfitting to dominant emotional categories. A comprehensive and evenly distributed dataset is required.	X	X	X
4	The music recommendation engine should support integration with leading platforms (e.g., Spotify API), ensuring seamless playback and real-time emotional alignment with song selection.		X	X
5	The system should cater to various user groups, including general users, mental wellness practitioners, and AI developers, and be intuitive for all levels of technical understanding.	X	X	
6	Ethical concerns must be addressed—especially regarding the emotional profiling of users. Informed consent, transparency of emotion inference, and data privacy are essential.	X	X	
7	Through prototyping, it was observed that real-time emotion detection and song rendering must be optimized for latency, particularly in browser or mobile-based environments.			X
8	The development should prioritize frameworks and programming tools compatible with ConvNeXt V2, particularly TensorFlow or PyTorch, to ensure scalable deployment		X	X

Table 10 – Summary Findings

4.7 Context diagram

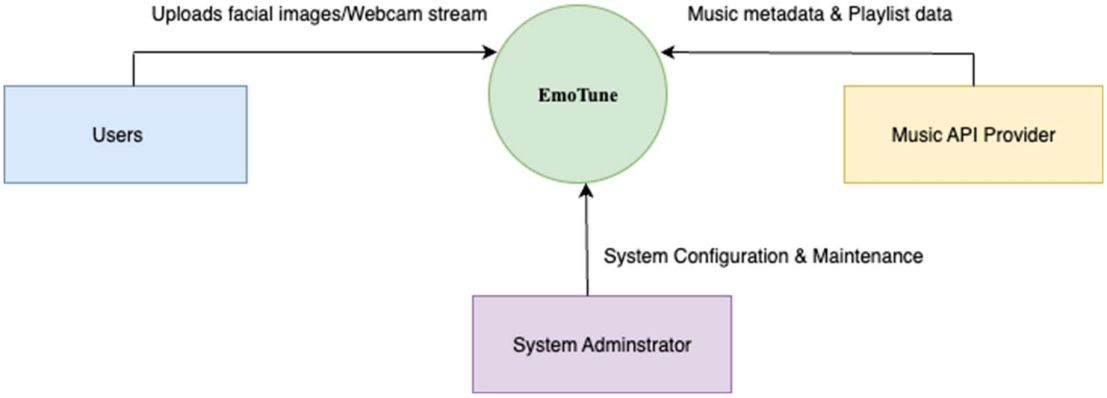

Figure 4 Context Diagram (self-composed)

4.8 Use case diagram

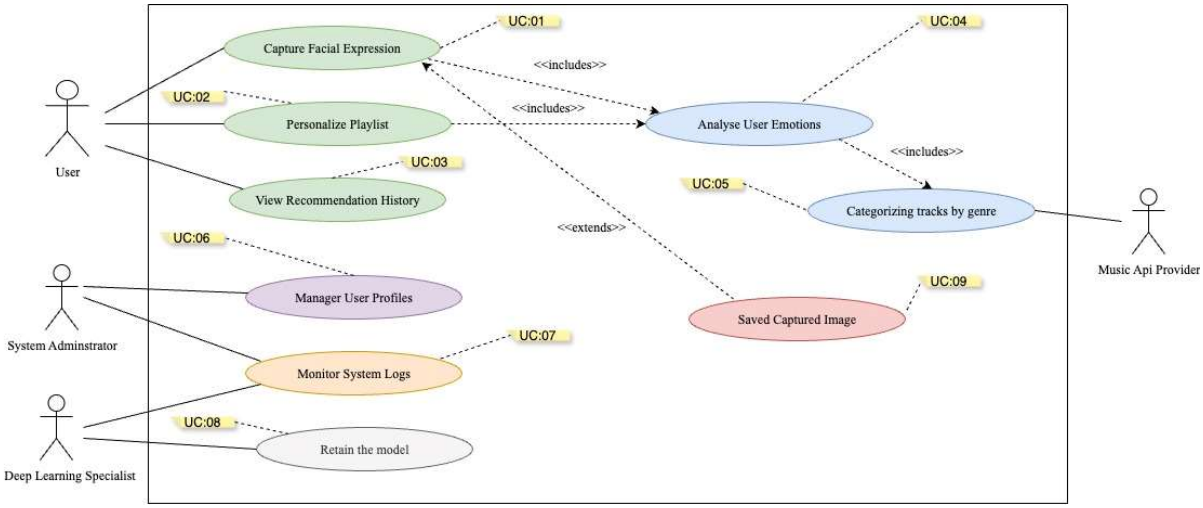

Figure 5 Use Case Diagram (self-composed)

4.9 Use case description

ID	UC:01
Description	This use case describes capturing the user's facial expression through the device camera to use as input for emotion analysis.
Participating actors	User

Preconditions	User has granted camera access and positioned themselves properly.	
Extended use cases	Saved Capture Image	
Included use cases	Analyse User Emotions	
Main flow	**Actor**	**System**
	1. Position face and initiate capture	1. Activate camera. 2. Capture and store facial image.
Alternative flows	None	
Exceptional flows	Camera inaccessible: Show error.	
Post conditions	Facial image successfully captured and forwarded to emotion analysis.	

Table 11 - Use Case Description 1

The descriptions of the other use cases are provided in **Appendix H** .

4.10 Requirements

The MoSCoW method is a widely adopted prioritization framework utilized in business analysis, project management, and software development. It facilitates structured decision-making by guiding stakeholders in identifying and agreeing upon the relative importance of various system requirements. The acronym MoSCoW represents the following priority levels:

Priority Level	Description
Must have (M)	These are essential requirements that are critical for the successful delivery of the system. If any of these are not implemented, the project outcome would be considered a failure, as these features are non-negotiable for the system to function as intended.

Should have (S)	While not mandatory for the initial launch, these features are important and add substantial value. They are often prioritized for future iterations or updates and should be included if time and resources permit.
Could have (C)	These are desirable but non-essential features that would enhance user experience or overall system quality. They are typically considered only when there is surplus time or budget available after delivering the higher-priority items.
Will not have (W)	These items are acknowledged as useful but are excluded from the current development cycle, often due to time constraints or resource limitations. They may be revisited in future phases or versions of the system

Table 12 - MoSCoW method

4.10.1 Functional Requirements

FR ID	Requirements Description	Priority Level	Use Case
FR1	The system should detect the user's facial emotion from a static image using ConvNeXt V2.	M	Detect Emotion
FR2	The system should generate a playlist based on the detected emotional state.	M	Recommend Music
FR3	The system should allow users to upload an image for emotion detection.	M	Upload Image
FR4	Users should be able to view the detected emotion label after processing.	S	View Emotion Output
FR5	The system should provide an option for users to select a preferred genre.	S	Genre Preference
FR6	The system should enable users to give feedback on the accuracy of emotion detection.	C	Feedback
FR7	The system should allow users to rate the song recommendations based on mood fit.	C	Rate Music Suggestions

FR8	Users should be able to download or share the generated playlist via social platforms.	W	Share Playlist
FR9	The system should visually display emotion confidence scores alongside predicted labels.	S	Display Emotion Metrics
FR10	The system should allow emotion detection through real-time webcam input (optional).	W	Real-Time Detection

Table 13 - Functional Requirements

4.10.2 Non-functional requirements

NFR ID	Requirement	Description	Priority
NFR1	Scalability	The application should be capable of handling varying volumes of user input (e.g., high numbers of emotion detection requests) and adapt to future expansion.	C
NFR2	Low Latency	The application should ensure minimal delay between face capture, emotion detection, and music recommendation to ensure a real-time experience.	M
NFR3	Audit Logging	All significant user activities such as login, face detection attempts, and API calls should be logged for accountability and debugging.	M
NFR4	Effectiveness	The platform should deliver high accuracy in mood classification and music matching, ensuring that results are relevant and contextually appropriate.	S
NFR5	Usability	The application must be intuitive and user-friendly for all user types, including non-technical users, allowing seamless interaction without assistance.	C

Table 14 - Non-Functional Requirements

4.11 Chapter summary

This chapter outlines key stakeholders using the Stakeholder Onion Model and Rich Picture, presents requirement elicitation methods including literature review, interviews, and

prototyping, summarizes key findings, and details functional and non-functional requirements with supporting use case diagrams and descriptions.