City St George's, University of London
MSc in Data Science

Project Report
2024

Emotion Detection in Images and Music Pairing Through AI

Samantha Georgina Isaac Munoz
Supervised by: Eugenio Alberdi

October 2, 2024

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed:

October 2, 2024

Abstract

The aim of this project is to develop and implement computer vision models capable of identifying emotions in images of facial expressions and landscapes. The final system recommends a musical playlist that corresponds to the detected emotion, with the broader goal of integrating AI into the creative field. A combination of CNNs architectures and CNNs with SVMs algorithms was used, resulting in eight models in total. These included two baseline models and two fine-tuned models, with optimal hyperparameters identified through grid search for each dataset. The fine-tuned CNNs model produced the best results for both datasets, achieving an accuracy of 60% for the face expression dataset, and 55% for the landscapes dataset, indicating that the models did not performed as expected, and there is still potential for improvement. This project also explored emotion detection in non-human subjects, specifically landscapes, contributing to an area of research that remains largely unexplored.

Keywords: Facial Expressions, Landscapes, Emotion Detection, Music Recommendation, Computer Vision

Table of Contents

1. Introduction and Objectives

1.1. Introduction

The right music enhances the emotional appeal of a film, as well as being a powerful guide to emotional responses that increases the effect and memorability of scenes in particular (Hoeckner et al., 2011). This is especially challenging to independent filmmakers and content creators, who work on limited budgets and few resources, making it hard for them to be able to afford professional music composition services. Therefore, they are often compelled to use pre-existing music in their production, which are much less effective at driving an emotional tone in the scene, making the viewer experience less immersive and engaging.

In fact, conventional ways of selecting music take time and require a certain amount of knowledge to make these critical decisions effectively, which can be challenging for filmmakers who may not have the necessary experience. The task usually involves manual search of music libraries, which is not only labour-intensive but also demanding in terms of understanding how various musical elements can convey different emotions (Neumeyer, 2013).

This has created the need for a solution that simplifies the process, allowing high-quality, and emotionally relevant music to be accessed by all creators without a demand on their resources. Sturm (2019) points out that the solution to this problem is currently being formed by integrating machine learning and artificial intelligence to create a model that can analyse visual and audio data for emotional cues and suggest relevant music tracks according to the emotion. It can be trained on large datasets of film scenes with related musical scores, enabling them to learn the complex relationships between visual emotions and musical elements (Ansani et al., 2020).

Machine learning has demonstrated a great potential in the filmmaking industry. AI-driven tools may be used for the required script, scene recognition, or even editing, which proves the flexibility and power of the combination of machine learning with filmmaking to augment the creative process (Gu et al., 2023). The use of these models can reduce both the time and labour costs associated with searching for the right music and speed up the timelines of a project at a lower cost of production, consequently making the filmmaking process more productive and easier to accomplish.

1.2. Motivation and Aim

This project aims to explore the potential of integrating machine learning algorithms into the filmmaking production process, particularly in the selection of music, to achieve significant improvements in both quality and efficiency of the editing phase.

This project aspires to take the first step in providing an innovative solution via machine learning models that are capable of identifying emotions in images, and suggesting a playlist of music that portrays the identified emotion. This accessibility is expected to create far more engaging and immersive visual experiences, empowering, in some way, the creators with the ability to stand up for themselves. The project will also look into the broader outcomes of integrating these technologies into the creative workflow, highlighting how machine learning can transform access to resources and boost creative processes in today's digital era.

1.3. Research Questions

Can a machine learning model specialised in computer vision effectively identify the emotions present in human faces and landscapes? Can it appropriately assign a playlist of music that reflects the emotion detected in the image?

1.4. Objectives

- Find a suitable dataset of human face images that displays different facial expressions, along with landscapes that convey various emotions.
- Develop models that identifies the emotion in images of human expressions and landscapes.
- Evaluate the performance and results obtained from the models using different metrics like precision, accuracy, f1 score, recall, confusion matrix and additional graphs.
- Develop a prototype system that it’s able to suggest a playlist of songs that match the identified emotion.

1.5. Requirements

The following requirements are for the accomplishment of the previous mentioned objectives:

- Search and inform myself about existing studies and cases in which machine learning has been used in the creative media industry. In doing so, analyse what current tools and technologies are in use and/or have been used in the past. And delimit which aspects

of the editing process will be specifically improved by Artificial Intelligence (AI) and machine learning.

- Subsequent implementation of model training techniques, and evaluation and optimisation of model accuracy through iterations.
- Identify and collect data, for the images of faces showing emotions and landscapes, which will meet the requirements, and ensure the quality and consistency of the data. Also provide clear documentation of how the dataset was collected and organised.
- Define metrics to assess the emotional compatibility between the recommended music and the images. Analyse test results to identify areas for improvement in the model. And keep track of possible improvements and optimisations.

1.6. Beneficiaries

This project benefits all those interested in emotion recognition in images and in the emotional expression through music, as it combines both areas. By integrating machine learning techniques to identify emotions in images and pair them with appropriate music, it opens new possibilities for researchers and professionals looking to explore how visual emotions can connect with emotions through music. Additionally, it is particularly useful for filmmakers and content creators who seek a technological solution to streamline the production of visual projects. This technology allows them to optimise the process of selecting music that is emotionally relevant to their scenes, thereby enhancing the narrative and emotional impact of their work.

1.7. Work Plan

The project will be divided into 4 phases and will have set timelines for the completion of each phase:

- **Design Phase (June - July):** complete process of framework development, review literature, selection of methodologies, research, creation and collection of dataset for the images and creation of the playlist.
- **Implementation Phase (July - August):** start of setup for the environment, model development and programming, and model testing.
- **Analysis and Evaluation Phase (August - September):** measurement of performance metrics, refinement and optimisation of the model based on the given results.

- **Reporting Phase (June - September):** process that will be carried out from the beginning of the project by means of a draft, with subsequent revision and review, modifications and final submission.

1.8. Changes of Goals and Methods

During the course of the project, the objectives were revised to make them clearer and more manageable within the given time frame. The scope was also refined to focus on developing simple models, consisting of four versions for each dataset: two models using only Convolutional Neural Networks (CNNs) and two hybrid models that combined CNNs with Support Vector Machines (SVMs), resulting in a total of eight models. Additionally, some models underwent hyperparameter tuning using grid search to optimise performance.

1.9. Report Structure

The structure of this report is as follows:

- **Chapter 1 – Introduction and Objectives:** The objective of the project is presented, as well as the changes of the same during the development of the project.
- **Chapter 2 – Context:** Literature review from projects with similar objectives or areas to this one to set the context and guidance for this project.
- **Chapter 3 – Methods:** Explains the methods and processes of the experiments carried out during the project.
- **Chapter 4 – Results:** The results of the conducted experiments are shown along with an analysis.
- **Chapter 5 – Discussion:** An evaluation and discussion of the results obtained in the previous chapter are provided. It is determined whether the objectives were adequately met.
- **Chapter 6 –Reflections and Conclusions:** A general evaluation of the project is provided. Reflections and conclusions on the results are given, along with the obstacles faced during its development. Opportunities are mentioned, and possible changes for future projects are suggested.
- **Chapter 7 – References:** A list of all the studies and researches mentioned during the process of this project.
- **Chapter 8 – Appendices:** Here are shown all those elements that were not presented directly in the corresponding section but are referenced.

2. Critical Context

The intersection of artificial intelligence and creative industries marks one of the more significant areas of technological development and academic interest, since creatives have always sought new tools to enhance their work, making them quick to embrace technological advancements (Amato et al., 2019). As Mazzone and Elgammal (2019) expresses, emotional analysis includes high-level AI algorithms capable of interpreting and matching the emotional tone of the music being listened to with the corresponding sentiment presented by an image. This methodology not only enhances the aesthetic experience, but also offers new dimensions of interaction and engagement for audiences and for content producers and creators.

According to Hutson (2023) convergence of artificial intelligence and the arts introduces new paths for creativity, broadening the scope and methods of artistic creation. AI technology has the potential to extend the frontiers of human creativity, enhancing both artistic expression and production processes. For example, the automatic synchronization of music with visual imagery, ensuring that the emotional tones are congruent, is transforming various industries by making creative content more immersive and emotionally impactful (Hutson, 2023). As stated by Trattner et al. (2021) this advancement not only redefines artistic practices but also enhances the audience's experience by fostering deeper emotional connections and more engaging narratives.

However, the integration of artificial intelligence with creative practices presents a wealth of opportunities as some big challenges within this dynamic field. The potential of AI to enhance and transform artistic expression is balanced with critical considerations of ethical implications, such as the authenticity of AI-generated content and its impact on employment in the creative sectors. Additionally, there is a growing need to address the technical limitations and potential biases inherent in AI systems, advocating for responsible and transparent AI development. By analysing these factors, the discourse in this area aims to provide valuable insights and guide the ethical evolution of AI within the arts and creative industries (Trattner et al., 2021).

Despite these advancements, the accuracy of emotion detection algorithms poses a significant challenge. As noted by Cai, Li and Li (2023) any current models can misinterpret emotional cues due to the complexity and subtlety of human emotions, leading to possible inaccuracies in the objective of this project, which is pairing music with the emotion of images. In addition,

the training dataset will be extended to include landscape images, adding a different level of complexity and a new challenge for future models and projects.

2.1. Artificial Intelligence Application in Creative Media Production

Artificial intelligence has been an area of immense research in the creative industries. De-Lima-Santos y Ceron (2021) explains that artificial intelligence in media production is at the meeting point of ethical and practical considerations with respect to its potential to automate and enhance the creative process. His work highlights the increasing reliance on artificial intelligence to streamline traditionally human-intensive and time-consuming tasks, such as music selection in film editing. This aligns very well with the project's goal to make the process of associating music with video content more effortless for independent filmmakers, thereby optimising the editing process and making high-quality, emotionally relevant music more accessible.

Artificial intelligence can revolutionise the user experience of music recommendation systems by incorporating emotional recognition into music recommendation engines. Machine learning models enables the analysis of user emotions and the recommendation of music that aligns with their mood at a particular moment (Tran et al., 2023). This is certainly in line with the objectives of this project, which is to link emotionally relevant music with images using similar technology to enhance the emotional result of the visual media. Moreover, the technical aspects of AI in music composition and recommendation are explored by Briot, Hadjeres, and Pachet (2020). They examine various machine learning algorithms used in music generation and their effectiveness in creating emotionally resonant compositions. This technical perspective provides a foundation for selecting appropriate algorithms and techniques for the project's AI model, ensuring that the music recommendations are both relevant and high-quality.

2.2. Emotional Detection with Computer Vision on Human Expressions and Landscapes

In the field of emotion identification, Ekman's (1992) research paper laid the foundation for understanding how emotions can be systematically identified based on facial expressions. It has been applied in the development of effective machine learning models, meaning the emotional cues derived from visual data can be interpreted. These established principles are applied to the use of computer vision to detect emotions in an image, allowing the machine learning model to follow emotional visual cues.

Calvo and D'Mello (2010) explore the broader field of affective machine learning, which involves the study and development of systems that can recognise, interpret, and replicate human emotions. Their work contributes to the idea that the recognition of emotion is highly context-dependent; hence, the training data used by machine learning systems must be comprehensive and inclusive for such systems to be sensitive to and accurately predict responses. The number of diverse and representative human emotions and landscapes in the dataset has a significant impact on the approach to dataset compilation and model training in the project. Being able to use all-inclusive datasets will ensure that the machine learning can adapt to the subtle emotional needs of different visual narrations, in turn adding value to the storytelling of filmmakers, and content creators.

As proposed by Benenti and Meini (2017), landscapes possess the ability to convey dynamic information effectively, with characteristics such as colour, shape, and the arrangement of various elements within the visual field. They also categorize inanimate objects as expressive, capable of evoking emotions such as joy, sadness, liveliness, or melancholy.

The paper by Mehendale (2020) explores the potential of use of Convolutional Neural Networks (CNNs) for facial emotion detection. Specifically, two CNNs models were developed and trained using greyscale images to categorise facial expressions into emotions such as happy, sad, angry, neutral, and fear. Another study made by Canedo and Neves (2019) shows the potential of the CNNs model using feature classification, with the Support Vector Machine (SVMs) model, which is mentioned to help the model to have slightly higher percentage accuracy in emotion classification. These models use several key techniques to enhance accuracy, including batch normalisation and dropout to mitigate overfitting. The best model achieved an accuracy of 80% for detecting four emotions and 72% for five emotions, demonstrating the potential of CNNs in emotion recognition tasks, but having the limitation that adding more emotions to recognize, may affect the accuracy which this project will take into account. By leveraging the good result of the CNNs model shown by previous research papers, this project aims to use two versions of the models to perform a comparison and analysis of the accuracy for the detection and categorisation of emotions in images: simple CNNs model and a model using the CNNs base model but with the implementation of SVMs will be used.

2.3. Contribution to this project

This project draws inspiration from the previously mentioned papers, building upon their ideas to plan and develop the steps and structure for the implementation of the models. One of the major influences on opting to use CNNs models in trying to detect human emotion was Mehendale (2020), who showed that this algorithm had the highest accuracy for both testing and training of the model. And with this, the project will also incorporate a second model also using the same algorithm CNNs which will be able to do the same emotion identification, but for the dataset of landscapes.

This paper will also delve in small quantities into the analysis of the related ethical issues in the authenticity and impact of the use of AI in the creative process. As De-Lima-Santos y Ceron (2021) noted, this is an opportunity to enhance the creativity of emerging talent by providing resources that streamline time-consuming tasks.

For the classification of the human facial expression, five distinct labels will be used: happiness, sadness, anger, fear, and surprise; as referenced by Canedo and Neves (2019). This set of labels provides good balance for reaching a higher accuracy and is also a good starting point in determining whether it is feasible to extend the number of categories in the future. The emotion categories that the project will use in categorizing the emotion found within landscapes are based on the ones proposed by Benenti and Meini, 2017: joy, sadness, liveliness, and melancholy. It goes without saying that these categories are subjective, adding another layer of complexity that will be dealt with during the development of the project.

In this context, the two models—the simple CNNs model and the CNNs model combined with SVMs —will be compared according to criteria like accuracy, sensitivity, and robustness against changes in data. Such a comparison will be useful in showing which model is more appropriate for the project, but more importantly, it gives a foundation that can be used for improvements in the future.

2.4. Contribution from this project

During the search for research papers discussing emotions recognition in landscapes, it became clear that there is a significant lack of studies on this topic. Most existing researches in emotion recognition focuses on facial expressions or other human centred data leaving non-human subjects like landscapes largely unexplored. This project aims to address that gap by taking the

first steps towards methods and models capable of interpreting emotional signals in natural scenes.

By providing a framework for understanding how visual elements in nature evoke emotions, this project lays the groundwork for further exploration and could inspire more sophisticated models and techniques that push the boundaries of emotion recognition beyond traditional human subjects. Ultimately this work paves the way for a broader understanding of how emotions are conveyed through non-human visual stimuli offering fresh perspectives, for both academic research and practical applications across various fields.

3. Methods

This chapter outlines the entire process for the developing and integration of the models. It begins with an explanation of how the underlying algorithms work, providing insights into their design and functionality. Following this, the chapter covers the selection and preparation of the data, including steps such as cleaning and transformation, which were used to optimise the data for the training of the model.

Finally, there is an explanation of the process for the implementation of each model, focusing on the technical and practical considerations. This includes software, libraries used, the steps followed, the challenges encountered during implementation, and the solutions developed to overcome these challenges, all aimed at integrating the functionalities of emotion prediction and playlist recommendation.

3.1. Algorithmic Foundations

The following section provides a brief explanation of the algorithms used in the implementation of the models, along with the justification for their selection in the task of emotion classification in images.

3.1.1. Convolutional Neural Networks

The following section is adapted from the paper by O'Shea and Nash (2015). CNNs shares similarities to the structure of traditional Artificial Neural Networks (ANNs), an example of this is that both include a perceptive score function that takes inputs and gives outputs and usually is expressed in terms of weights. They also have loss functions for classification and neurons that optimize through learning. But the CNNs algorithms are more inclined for pattern recognition in images due to their specialised architecture.

This algorithm was chosen since they are designed to directly encode features from images making them more efficient for tasks that relate to image data, this is done by reducing the number of parameters needed, which in turn decreases the computational load. A typical architecture of this model will include three types of layers: convolutional layers, pooling layers, and fully connected layers.

Note: Section 3 was occasionally written in first person to clarify my contribution in the development of the model, logic, and code, as suggested by my supervisor.

3.1.2. Support Vector Machine

Based on the paper by ElSayed et al. (2021), one of the most popular forms of supervised machine learning methods applied to classification and regression problems is the SVMs algorithm, this is because it finds the best separating hyperplane by mapping data into a high-dimensional feature space using the kernel trick. And SVMs, although designed to solve binary classification, can be used in solving multiclass classification problems.

3.1.1. Combination of Convolutional Neural Networks with Support Vector Machine

It has been shown by Ahlawat and Choudhary (2020) that incorporating a SVMs at the final layer of a CNNs-based system ensures higher accuracy in classification, particularly for problems where generalization error on unseen data is to be minimized. A SVMs is strong in its ability to construct robust decision boundaries even when applied in high-dimensional spaces, hence quite well-suited for operating with the output obtained from a feature extraction process like CNNs.

This combination not only makes the model more accurate since it brings out the best of both CNNs and SVMs techniques, but also makes it more robust to noise and variability in the input data, hence offering a comprehensive solution for complex classification tasks (Ahlawat and Choudhary, 2020). For these reasons, it was decided to implement this within the CNNs algorithm to compare whether incorporating the SVMs technique would enhance the model's performance and results.

3.2. Data

For the data used in this project, it was decided to use two datasets: one consisting of facial expressions and the other of landscapes. This choice was made since human faces and landscapes are two fundamental components often featured in visual storytelling both of which play a crucial role in conveying emotions and setting the tone of a scene. By focusing on these categories, the project aims to provide valuable tools for those in the creative industry, particularly in enhancing their ability to match visuals elements with music, thereby strengthening the emotional impact and overall narrative. The combination of facial expressions and landscapes offers a broad range of emotional cues that can be used to create a more immersive and emotionally resonant experience for audiences.

3.2.1. Facial expression dataset

For the facial expression dataset, the open source dataset given by Kaggle for the Challenges in Representation Learning: Facial Expression Recognition Challenge competition will be used. This decision was made since this dataset features high-quality images that are representative, diverse, and available under usage conditions suitable for this project: furthermore, they are open source and accessible to the public. It has a total of 35,257 images separated into two sets, the training set consists of 28,709 samples, and the test set has 7,178 in total. The images have a 48 x 48 pixel grayscale format, with the focus on the faces. It has a total of 7 emotions: surprise, anger, happiness, sad, neutral, disgust, and fear; but for this project only 5 emotions will be used.

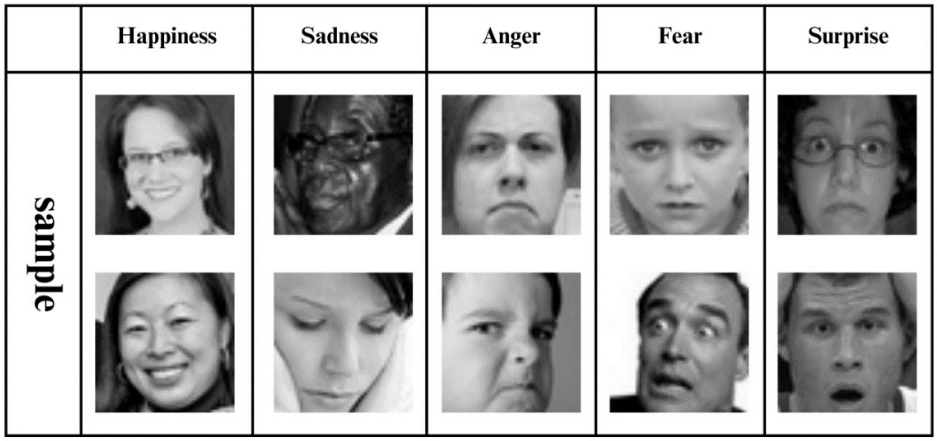


Figure 1. Samples of the images for the facial expression dataset.

3.2.2. Landscape dataset

For the landscape dataset, a collection of Public Domain images from Kaggle was obtained, which includes images of mountains, deserts, the sea, beaches, islands, and cities. This decision was made since the images are of high-quality, are varied, and representative, as well as they are open source to the public. The process of classification was conducted manually, which implies potential subjectivism. However, it should be noted that this does not affect the functionality of the model, as this section of identifying emotions in landscapes is experimental and aims to determine whether the AI can interpret and identify emotions in landscapes. The classification will be as follows: joy, sadness, liveliness, or melancholy.

I developed a framework for classifying landscape emotions based on visual characteristics, inspired by principles from environmental psychology and art theory. For instance:

- Joy: landscapes with clear skies, giving a feeling of tranquillity, with few visual elements and evoking happiness.

- Sadness: dark landscapes, with few elements, with a feeling of desolation.
- Liveliness: landscapes that show activity, life, and dynamism. With bright colours and clear skies, with various elements.
- Melancholy: landscapes with scenes with a sense of longing, with old elements, sunsets, or with buildings.

Figure 2. Sample of the images of the landscape dataset.

3.3. Work Environment Preparation

The programming process was carried out entirely in the Google Colab environment, using the Python programming language. It was decided to use Google Colab since it facilitates its use, has access to cloud computing resources, such as the Graphics Processing Unit (GPU), and access to libraries is simple and straightforward. For the creation and implementation of the deep learning models, we used the libraries of: TensorFlow and its API Keras for building and training neural networks more efficiently; Numpy for matrix manipulation; Seaborn and Matplotlib for data and results visualisation, facilitating data exploration and graphical interpretation of model performance; and finally scikit-learn for data processing tasks, model evaluation, and metrics generation.

3.4. Implementation and Pre-processing of the Data for Face Expressions

I created a function named *load_images_and_labels* to load the images and their corresponding labels into the directory in which they are located. This directory contains two subfolders, one for the training data named 'Training', and another for testing data 'Testing'. Both of them have respective subfolders for each class name, which are 'Angry', 'Fear', 'Surprise', 'Neutral', and 'Happy', as described in the previous section. The function first retrieves the name of each sub-

folder to use as the label for the images. Then, it loads the images and pairs them with their corresponding labels. Each image is loaded in greyscale mode using *cv2.imread()*, and both the image and its label are stored in a list for both the test and training datasets. I also created a function to check the image labels to see if this was done correctly, these images can be found in the Appendix 8.1.

Subsequently, the labels were converted to a numerical format by creating a dictionary, where a numerical index was assigned to each unique label. Using Kera's *to_categorical()* function, the numerical labels were transformed into binary vectors with a length equal to the number of classes. This process, known as one-hot encoding, is commonly used in classification problems. Finally, the images were converted into NumPy arrays, and the pixel values were normalised to a range of 0 to 1 to enhance the efficiency of the model training.

For the adaptation of the format of the images, the following pre-processing steps were performed by me. First, the reshaping of the images was made. Reshaping is the process of making an image into a specific size; this helps to work uniformly with the dimensions for the models while processing the data. For the images in the facial expression dataset, a channel dimension was added, which is necessary for the CNNs models to perform properly. The format given is as follows: number of images, height, width, channels.

Another important step was data augmentation, where various transformations such as rotation, shifting, zooming, and horizontal flipping were applied. These transformations generate new versions of the training images to enhance the model's ability to generalise.

3.5. Implementation of the CNNs Model for Facial Expressions

For the definition, compilation and training of the CNNs model for the classification of facial emotion images, I made three versions of the model, a baseline, a model that served for the hyper parameters tuning with the implementation of a grid search method, and finally the final version that uses the best combination determined by the grid search. An important note is that I used as reference the project by Skillcate (2022) for the structure of the CNN models throughout the project.

For the baseline model, I first defined the input layer to accept 48 x 48 pixel greyscale images, followed by blocks of convolutional layers. The first block has 16 filters, and the second one

has 32 filters, both with the Rectified Linear (ReLU) activation functions and a *same* padding to maintain the size of the images. Each convolutional layer is followed by a MaxPooling layer that reduces the spatial dimensions of the extracted features. After the convolutional layers, the features are flattened to connect to a fully connected dense layer with 64 neurons and ReLU activation. Lastly, an output layer with *softmax* activation produces a probability for each class for the classification of each emotion. The model is compiled using the *adam* optimiser and crossentropy categorical loss, and trained for 20 epochs with a batch size of 32. The decision to use the *adam* optimiser and categorical cross-entropy loss was based on their effectiveness, robustness, and ability to efficiently handle complex data classification tasks, as highlighted by Kingma and Ba (2015) in their research paper. A table outlining the model's structure is available in Appendix 8.2.

In the second version of the model, I implemented a grid search technique to find the optimal hyperparameter configuration. As Belete and Huchaiah (2021) explain, grid search generates all possible combinations of these hyperparameters and train the model with each combination. During this process, validation accuracy is monitored to identify and save the hyperparameter combination that achieves the best model performance. Ultimately, the code prints out the best accuracy achieved and the associated hyperparameters.

This was achieved by building the model within a function called *build_model*, which allows for the specification of various activation functions, optimisers, dropout rates, and L2 regularisation rates as hyperparameters. One significant modification from the baseline model was to increase the number of convolutional layers to four. This adjustment was necessary since the baseline model exhibited a high level of overfitting. By adding more convolutional layers, the model gains complexity, enabling it to learn more useful and general features from the images.

I began the architecture with an input layer configured identically to that of the baseline model, followed by four convolutional layers that include ReLU activation to extract features. Each convolutional layer is followed by a dropout layer for regularisation and a Max Pooling layer to reduce dimensionality. As the model progresses through these layers, the number of filters in the convolutional layers increases to 32, 64, 128, and 256 respectively, allowing the model to capture more complex features. After these convolutional layers, the output is flattened and fed into a fully connected dense layer with 128 units and ReLU activation, followed by an

additional dropout layer. The final output layer utilises *softmax* activation to generate probabilities for each class, facilitating multi-class classification. To prevent the overfitting from the baseline model, this second version incorporates L2 regularisation and dropout techniques. It is trained using the *adam* optimiser and categorical cross-entropy loss function as with the baseline.

Finally, for the last version of the model, the best combination of hyperparamenters obtained through the mentioned grid search was implemented. While the overall architecture remains similar to the previous model used for the hyperparameter search, specific adjustments were made to enhance the performance. Notably, the activation function was set to *leaky_relu*, the optimiser remained as before with *adam*, the dropout rate was adjusted to 0.3, and an L2 regularisation rate of 0.0001 was applied. These changes were implemented to maximise the model's ability to generalise and improve accuracy in images classifications tasks. Appendix 8.3 contains a table depicting the model's structure.

3.6. Implementation of the CNNs with SVMs Model for Facial Expressions

For the second set of models, the CNNs framework for facial emotion image classification was extended by including the SVMs algorithm. As with the previous set of models, I created three versions: a baseline, a model used for hyperparameter tuning by implementing a grid search method, and the final version with the best combination of hyperparameters determined by the grid search.

I kept the structure for the first baseline model simple, with two convolutional and MaxPooling layers followed by a fully connected layer. The first convolutional layer applies 16 filters of size 3 x 3, using the ReLU activation function, followed by a MaxPooling layer with a filter size of 2 x 2. The second convolutional layer applies 32 filters of 3 x 3, also using ReLU, followed by another MaxPooling layer. After this, the resulting feature maps are flattened into a vector using a flatten layer, which is then feed to a dense layer with 64 units and ReLU activation. Finally, the output is generated through a *softmax* layer to classify the five classes. The model still uses the *adam* optimizer, with categorical cross-entropy loss for the training.

Once the CNNs is trained, it is used as a feature extractor; the flattened output is fed into a SVMs model, to make the final classification instead of the original *softmax* layer. The extracted features are normalised using a StandardScaler which adjusts the data to have a mean

of 0 and a variance of 1, and the SVMs is trained with a linear kernel and a regularisation parameter C (which controls the penalty for misclassification), set to 1. A table outlining the model's structure is available in Appendix 8.4.

A grid search was implemented for the second version of the model to find the optimal parameters for the SVMs classifier. For the structure, first I defined a function named *build_feature_extractor*, containing the CNNs architecture that acts as a feature extractor, processing input images through four blocks of convolutional and pooling layers. Each convolutional layer applies filters to capture patterns and textures in the images. Following each convolutional layer, ReLU activation functions are added, along with L2 regularisation to prevent overfitting and dropout to enhance the model's generalisation ability. MaxPooling layers are then applied to reduce the spatial dimensions of the extracted features, consolidating the most relevant information. At the end of the CNNs, the resulting features are flattened into a 128 dimensional vector that encapsulates the abstract features of each image.

These features are then normalised using a StandardScaler, which serves as the final classifier. Then, to determine the best SVMs parameters, the grid search is employed, tuning hyperparameters such as the kernel type, linear or RBF; the C parameter and *gamma* (which controls the influence of each training point in the RBF kernel).

For the final version of the model, the best combination of hyperparameters obtained from the previous grid search was applied. While the overall architecture remained consistent with the earlier feature extraction model, specific adjustments were made to improve classification accuracy. The optimised SVMs classifier was configured with an RBF kernel, C=10, and gamma='scale', based on the best results from the grid search. This configuration enables the SVMs to effectively handle non-linear decision boundaries. Appendix 8.5 contains a table depicting the model's structure.

3.7. Implementation and Pre-processing of the Data for Landscapes

Similar to the integration process for the Facial Expressions image dataset, I developed the *load_images_and_labels* function to load images and their associated labels in a similar way. The directory contains a single folder with four classes: 'Joy', 'Melancholy', 'Sadness', and 'Liveliness'. An example of these images can be found in Appendix 8.6.

The images were initially read in BGR format and converted to RGB for correct visualisation and processing. Each image was resized to 128 x 128 pixels to ensure uniformity in the model's input, and normalised by dividing the pixel values by 255 to scale them between 0 and 1. To assign classes to each image I created a dictionary named *label_map*, mapping the name of each class which is taken from the folder name to a unique numerical value. Each image was then assigned a corresponding numerical label based on its class. These numerical labels were transformed into a one-hot encoding format using the function *to_categorical*, converting the labels into binary vectors. In this way, the images were efficiently prepared for use in the neural network with their respective encoded classes. Finally, a data generator was applied to perform augmentation, similarly to the Face Expression dataset.

3.8. Implementation of the CNNs Model for the Landscapes

For the CNNs model for the classification of the landscapes images, I made three versions, a baseline, a model that served for the hyper parameters tuning with the implementation of a grid search method, and finally the final version which uses the best combination determined by the grid search.

The baseline model begins with an input layer that accepts 128 x 128 images with 3 RGB colour channels. This is followed by three sequential convolutional blocks. The first block uses 32 filters with a 3 x 3 kernel size, followed by a MaxPooling operation to reduce the spatial dimensions. The second convolutional block increases the number of filters to 64 and again applies MaxPooling. The third block is similar to the previous ones but further increases the filters to 128. After these blocks, the output is flattened and passed to a dense layer with 128 neurons activated by ReLU. To prevent overfitting, dropout is applied with a rate of 0.5 before the output layer. The output layer consists of 4 neurons, corresponding to the total number of classes, and uses *softmax* activation to generate classification probabilities. Like previous models, the *adam* optimiser is used, along with the *categorical_crossentropy* loss function, and accuracy is measured during training. A table showing the model's structure is available in Appendix 8.7.

The second model was built in a *build_model* function which was made from scratch, allowing for the specification of various hyperparameters such as activation functions, optimisers, dropout rates, and L2 regularisation rates. Compared to the baseline model, one significant modification was increasing the number of convolutional layers to four. This adjustment was

made because the baseline model exhibited signs of overfitting, where it performed well on the training data, but poorly on the validation data.

The architecture starts with an input layer identical to that of the baseline model, followed by four convolutional layers, each designed to extract increasingly complex features from the images. These convolutional layers use different numbers of filters, 64, 128, 256, and 512, which allow the model to capture detailed hierarchical patterns. Each convolutional layer is followed by a dropout layer for regularisation and a MaxPooling layer to reduce the dimensionality. Unlike the baseline model, which only used three convolutional layers, this deeper structure helps the model learn more nuanced and generalised patterns. After passing through these convolutional blocks, the output is flattened and fed into a fully connected dense layer with 256 units and an activation function like ReLU or Leaky ReLU. This dense layer is followed by an additional dropout layer to prevent overfitting.

A key change from the baseline is the use of L2 regularisation applied to each convolutional layer, combined with dropout, which further combats overfitting. In terms of training, the *adam* optimiser was used alongside categorical cross-entropy as the loss function, consistent with the baseline. However, this grid search model included experimentation with two optimisers, *adam* and RMSprop; activation functions, and regularisation rates, leading to a labour-intensive process due to the large number of hyperparameter combinations. Despite these improvements, the model often struggled with performance during the search, and tuning it for better results required significant time and computational resources.

For the last model, I initially attempted to develop a more complex model inspired by the hyperparameter search conducted through grid search. This deeper model incorporated four convolutional blocks with an increasing number of filters of 64, 128, 256, and 512; along with advanced regularisation techniques such as Leaky ReLU, L2 regularisation, and Batch Normalisation in each convolutional layer. Additionally, I applied dropout after each convolutional block to prevent overfitting. Despite these techniques, the model did not perform well in terms of accuracy and failed to classify the images correctly across the different classes. As the training progressed, signs of either overfitting or undertraining were observed, as the validation accuracy remained low despite improvements in training accuracy.

Due to these unsatisfactory results, the model architecture was simplified, opting for a structure closer to the original baseline model, which had proven more robust in generalising. The new model consisted of only three convolutional blocks with a reduced number of filters, 32, 64, and 128; and utilised ReLU as the activation function. The hyperparameters identified as optimal during the grid search were still used for this simple model, such as the *adam* optimiser, a dropout rate of 0.4, and L2 regularisation, but with a more manageable structure that was less prone to overfitting.

This shift towards a simpler model proved successful, as it significantly improved validation accuracy, outperforming the more complex model. This demonstrates that, simpler and well-tuned models can be more effective than complex architectures, particularly when the dataset size is limited or the problem's complexity does not require multiple layers for optimal performance. The combination of a simple structure with optimised hyperparameters allowed for a better balance between generalisation and performance as in this case. Appendix 8.8 contains a table depicting the model's structure.

3.9. Implementation of the CNNs with SVMs Model for Landscapes

As with the facial emotion classification models, the CNNs model was extended by including the SVMs algorithm. The same way as before, I created three versions: a baseline, a model used for hyperparameter tuning by implementing a grid search method, and finally, the optimal version, utilizing the best combination of hyperparameters identified by the grid search.

The baseline model begins with an input layer that accepts 128 x 128 images with three RGB colour channels. Next, two successive convolutional layers apply 3 x 3 filters, with 16 and 32 filters respectively, using the ReLU activation function. After each convolutional layer, I used a MaxPooling2D layer with a pool size of 2 x 2 to reduce the spatial dimensions of the extracted features, retaining important information while lowering the computational cost. Following these layers, the output is flattened into a 1D with a flatten layer, converting the 2D representation into a one-dimensional vector. Then is passed through a fully connected dense layer with 64 units and ReLU activation, followed by a *softmax* output layer.

The CNNs model is trained using the *adam* optimiser and the categorical crossentropy loss function. Once trained, the flattened layer is used as a feature extractor. The features extracted from the training and test images are normalised using StandardScaler and then used to train a

SVMs model with a linear kernel and a regularisation hyperparameter of C=1. Finally, the SVMs is trained on these features and used to predict the classes of the test images. Appendix 8.9 contains a table depicting the model's structure.

I implemented a grid search for the second version of the model to find the optimal parameters for the SVMs classifier. The process began by defining a function named *build_feature_extractor*, which contains the CNNs architecture acting as a feature extractor. It processes the input images through four blocks of convolutional layers with increasing filter sizes of 64, 128, 256, and 512; and pooling layers. After each of these layers, I defined Batch Normalization, and ReLU. Additionally, L2 regularization is included to prevent overfitting, while dropout layers with a dropout rate of 0.5, help enhance the model's generalization ability. MaxPooling layers are also used to reduce the spatial dimensions of the extracted features, ensuring that only the most relevant information is retained. The output from the CNNs model is then flattened into a 256-dimensional vector, representing the abstracted features of each image. Then they are normalized using StandardScaler before being passed into the SVMs classifier. With the help of grid search, the search for the best hyperparameters was made by tuning parameters such as the kernel type, linear, RBF, or poly; the C parameter, the gamma parameter, and the degree for the polynomial kernel.

In the last model, the CNNs architecture is similar to the one used during the grid search, since it gave good results. The model uses the hyperparameters RBF kernel, C=1, degree=2, and gamma = 'auto'. The CNNs processes the input images through four convolutional blocks, with increasing filter sizes of 64, 128, 256, and 512; followed by Batch Normalization, ReLU activations, L2 regularization, and dropout. Then MaxPooling layers are applied after each convolutional layer. The features are then flattened into a 256 dimensional vector, which is then normalized using StandardScaler. A table outlining this model's structure is available in the Appendix 8.10.

3.10. Music Playlist Recommendation

For music playlist recommendations based on the identified emotion, I created two versions of the code, one for each respective dataset. While both versions work in essentially the same way, they differ in the dataset they load and the method used to process the corresponding image, as each dataset requires a different processing approach.

I first established a dictionary called *emotion_to_playlist*, which associates the emotion classes from both datasets with a URL of a Spotify playlist. Then I created two functions, *run_random_image_test* and *run_random_landscape_test*. These functions randomly load an image from the respective dataset, either in colour or black and white, resize the image to the dimensions expected by the model, and, if necessary, convert the colour format from BGR to RGB (for landscape images). After this, the image is normalised by dividing the pixel values by 255.

Once pre-processed, the image is passed through the corresponding model for emotion prediction. The model returns a probability for each emotion, and the most likely emotion is selected using *np.argmax*. Finally, the function returns the identified emotion label showing the image used, along with the recommended playlist.

3.11. Evaluation Metrics

To evaluate the effectiveness and functionality of the models, a confusion matrix was generated for all models across both datasets. As Beauxis-Aussalet and Hardman (n.d.) explain, the confusion matrix helps evaluate errors in classification problems by providing a summarised view of how well the model has identified the correct classes and whether there is a bias towards any particular class that could be causing misclassification. Also, a classification report is generated for all models, displaying the precision, recall, F1-score, and accuracy for each model. Below is an explanation of the metrics in the classification report:

- Precision: Measures how many of the predicted positive classifications are truly positive.

$$\text{Precision} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) \quad + \quad Fa \quad Positives\ (FP)}$$

- Recall: Measures how many actual positive classifications were correctly identified by the model.

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) \quad + \quad False\ Negatives\ (FN)}$$

- F1-Score: A metric that combines precision and recall, with more weight given to recall.

$$\text{F1} = \frac{(1+2^2) \quad * \quad (Precision \quad * \quad Recall)}{(2^2 \quad * \quad Precision) \quad + \quad Recall}$$

- Support: Indicates the distribution of the classes.

Additionally, a graph is generated to display two key components of the training process, accuracy and loss for both the training and validation sets across the epochs.

4. Results

In this section, the results of the eight models developed are presented: four using the face expressions dataset and other four using the landscapes dataset. Evaluating these results is essential for assessing the performance of each model, particularly in terms of their ability to generalise, minimise overfitting, and accurately classify the images.

4.1. Models Performance

The models were evaluated using classification metrics, such as accuracy, precision, recall, F1-score, and support, across both the face expressions and landscapes datasets. Additionally, confusion matrices were generated to illustrate the performance of each model and highlight any misclassifications. This review intends to provide an objective overview of the results examining more closely how each model performed at classifying the emotions, and potential errors. The in-depth analysis and interpretation of these results will be covered in the next chapter. The following table present the accuracy percentage for each model across both datasets. In both cases, the CNNs models with optimised hyperparameters achieved the best results, and on the contrary, lowest performance was noted for the CNNs with SVMs models with fine-tuned hyperparameters.

Face Expressions

	Accuracy
CNNs Baseline	50%
CNNs Fine Tuned	60%
CNNs & SVMs Baseline	45%
CNNs & SVMs Fine Tuned	40%

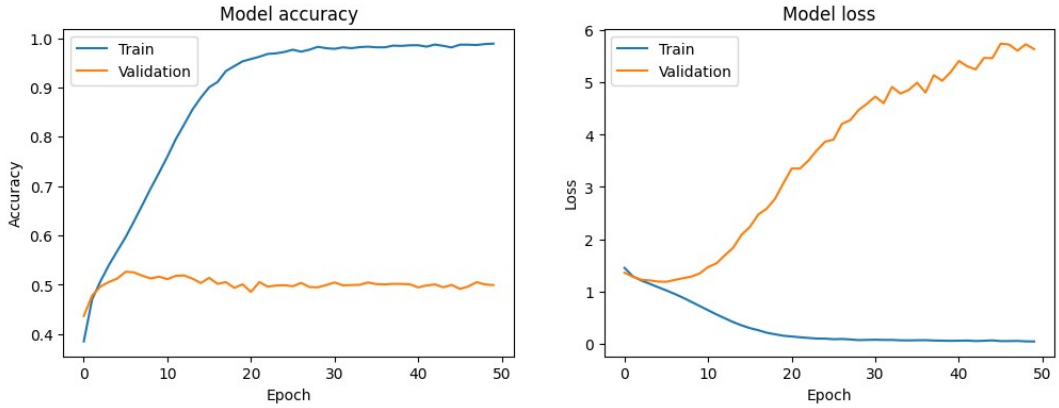
Landscapes

	Accuracy
CNNs Baseline	53%
CNNs Fine Tuned	55%
CNNs & SVMs Baseline	52%
CNNs & SVMs Fine Tuned	48%

Figure 3. Percentages of the accuracy for all the models for both datasets.

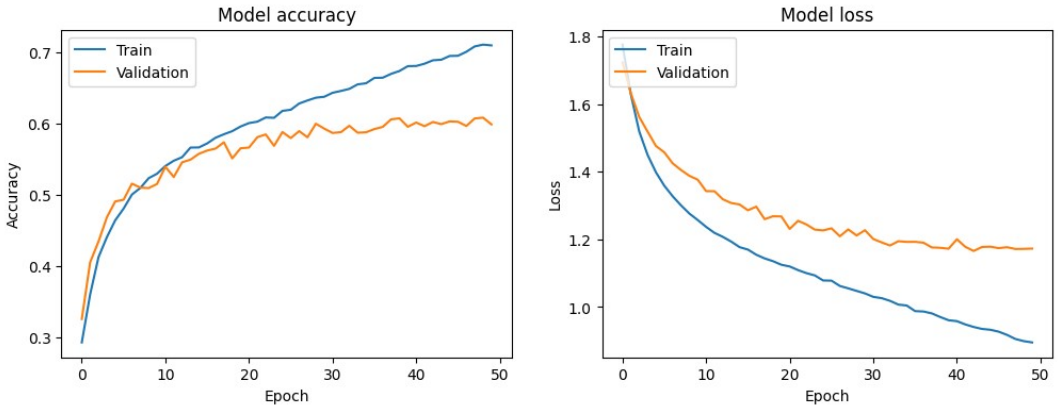
Graphs were also developed to show the behaviour and performance of each model. These include two types of results: accuracy and loss per epoch for the CNNs models, and precision, recall, and F1-score per class for the CNNs models with SVMs.

Baseline CNNs for the Face Expressions Dataset



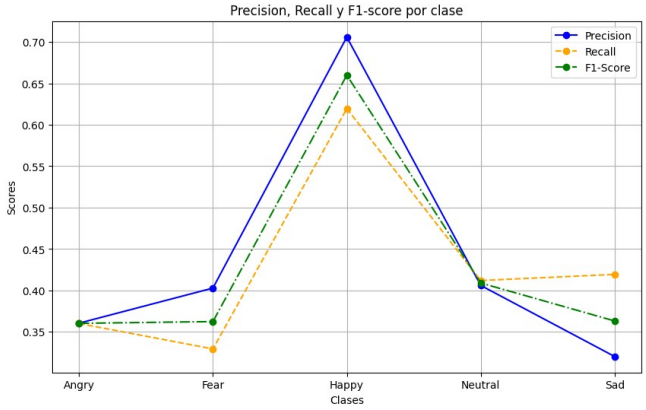
The training data reached a high accuracy of almost a 100%. In contrast, the test data stabilises around 50%, failing to generalise well to unseen data, also meaning that the model is suffering from overfitting. The test set loss increases while the training loss decreases to nearly zero, supporting the overfitting hypothesis. The model is excessively fitted to the training data and does not perform well on data it has not seen before.

Fine-tuned CNNs for the Face Expressions Dataset



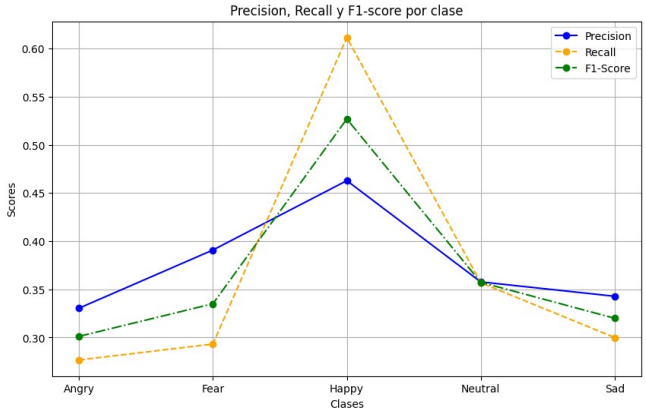
For this model, both the training and testing sets have shown an improvement in their accuracy. The test accuracy is much more stable and closely aligned to the training accuracy, suggesting that the model has refined its ability to generalise. In terms of loss, the test set loss remains higher than the training loss but it has a decreasing trend, meaning that this model fits the data better than the baseline model.

Baseline CNNs with SVMs Model for the Face Expressions Dataset



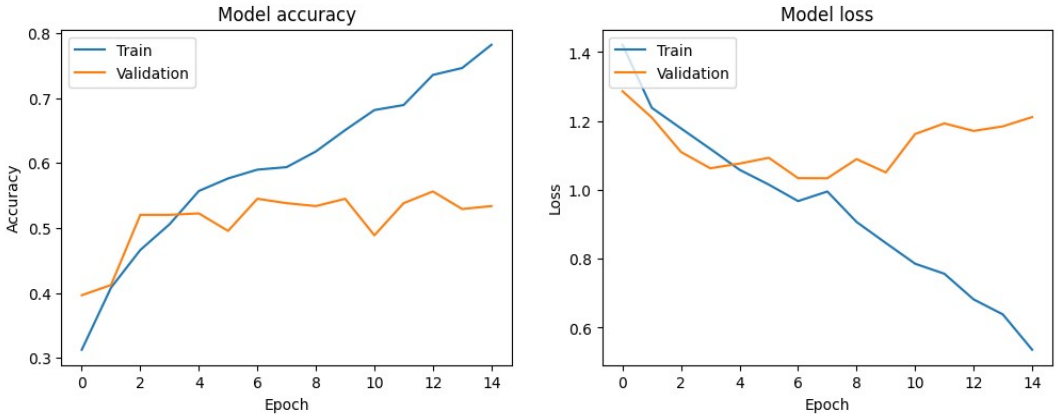
The “Happy” class shows the highest scores across all three metrics, with precision being around 0.7, recall of 0.65, and a F1-score to 0.65, indicating that the model is most efficient in identifying this emotion in particular. In the case of the other emotions, the scores are considerably lower. The “Angry” and “Sad” classes give the lowest in all metrics, all of which are around 0.4. The “Fear” class has slightly better scores, but still remains low, and the “Neutral” class shows slightly improved recall but also has a low precision.

Fine-tuned CNNs & SVMs for the Face Expressions Dataset



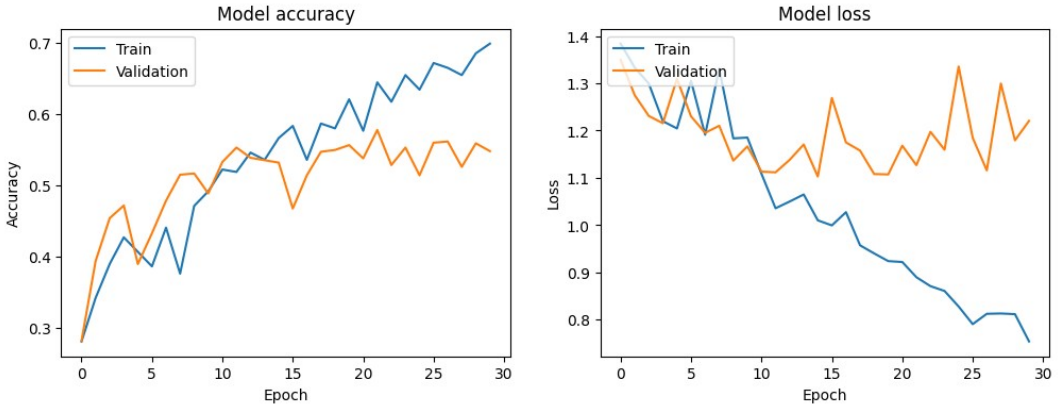
All classes show a slightly increase. The "Happy" class remains the best classified emotion with a recall of over 0.6 and both precision and F1-score around 0.5. The "Fear" emotion shows moderate results, having a recall close to 0.35, and precision and F1-score closer to 0.4. Lastly, the "Angry" class has the lowest recall, less than 0.3, with a slightly higher precision and F1-scores.

Baseline CNNs for the Landscape Dataset



The training set raises gradually, reaching approximately to 0.8 by the end of the training period. However, the testing accuracy vary but keeps stable at around 0.50 after an initial rise in the early epochs. The growing gap between the training and test accuracy suggests that the model performs well on the training data but struggles to generalise to the test set. In terms of the loss gradually reduces, reaching around 0.60, however, the testing loss remains higher, oscillates and stabilizes around 1.2. This indicates that the model may be overfitting.

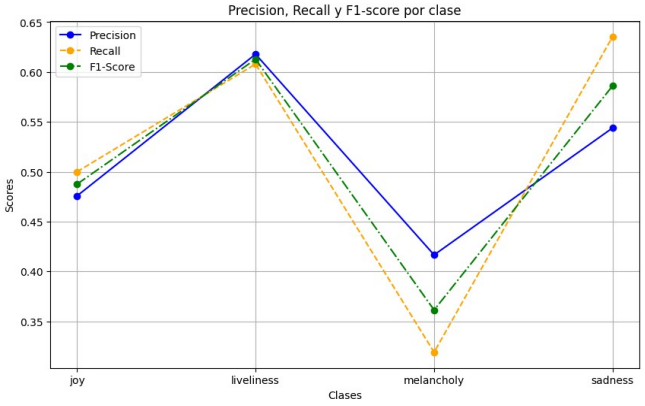
Fine-tuned CNNs for the Landscape Dataset



The training accuracy shows a consistent increase, reaching approximately 0.7 by the end of the training process. The test accuracy follows a similar pattern, with even more oscillations, stabilising around 0.55 after initial increases. Despite the oscillations, the test accuracy remains close to the training accuracy, indicating that the model is generalising better compared to previous iterations. The training loss steadily decreases, reaching around 0.8, while the test loss

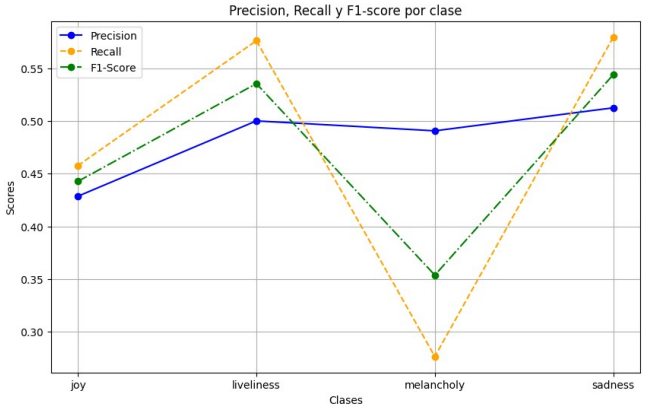
oscillates throughout but follows a decreasing trend overall, ending around 1.2. Although the test loss is higher than the training loss, the gap between them is less pronounced than in the previous model, suggesting that overfitting may be less of an issue in this case.

Baseline CNNs with SVMs Model for the Landscape Dataset



The "Liveliness" class outperforms all the other metrics, with precision, recall, and F1-score at approximately 0.60. The "Joy" class follows closely behind with the three metrics around 0.50 to 0.55. On the contrary, the "Melancholy" class shows the lowest values, with all three metrics dropping to approximately 0.40, indicating that the model struggles to classify this emotion. The class "Sadness" shows an improvement in the recall value, reaching around 0.65, although it's precision and F1-score are lower, at around 0.55. This indicates that the model performs better at identifying the "Sadness" class, but risks to have a higher rate of false positives in this category.

Fine-tuned CNNs & SVMs for the Landscape Dataset

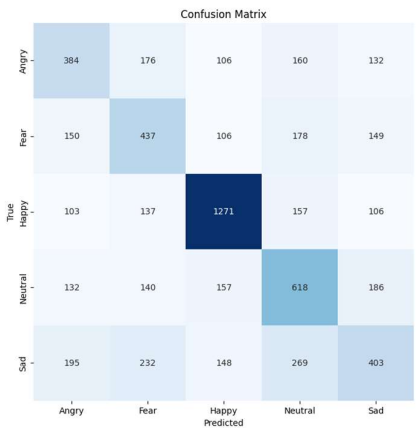


The "Liveliness" class has the highest recall value at approximately 0.57, while the precision and F1-score are both around 0.50. The "Joy" class shows slightly lower but balanced performance across all metrics close to 0.45 to 0.50. The "Melancholy" class has the lowest recall rate, dropping to around 0.30, while precision and F1-score values remain around 0.45, indicating that the model has a poor performance to recall instances of this class correctly. The "Sadness" class, on the other hand, has a high recall of approximately 0.60, while its precision and F1-score are around 0.50, suggesting that the model is better at identifying "Sadness" but may produce a higher number of false positives in this category.

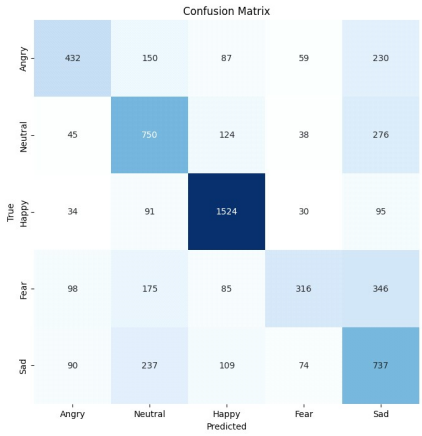
Additionally, confusions matrix were made for the four models of each dataset to provide a detailed breakdown of the classification performance by showing the number of correct and incorrect predictions for each class or emotion. These allow for a more granular evaluation of how well each of the four models distinguishes between different classes and help identify specific areas where the models may be misclassifying certain classes or emotions. First the matrices of the models for the face expressions dataset will be shown, following by the results for the landscape dataset.

CNNs for the Face Expressions Dataset

Baseline

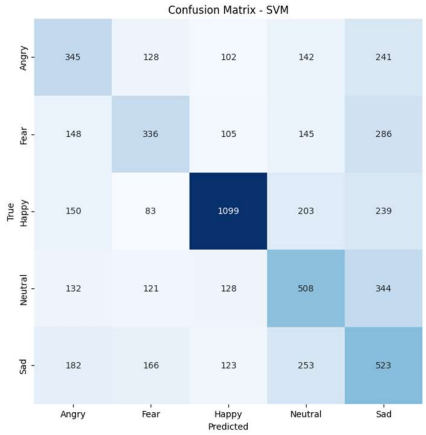


Fine-tuned

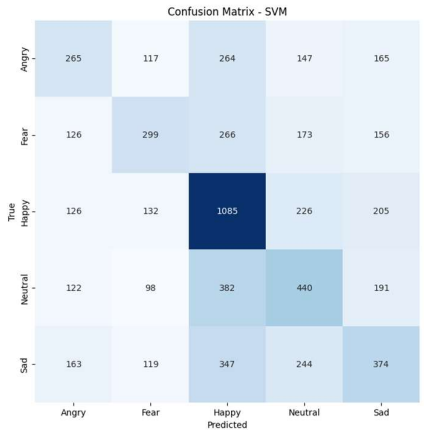


CNNs & SVMs for the Face Expressions Dataset

Baseline



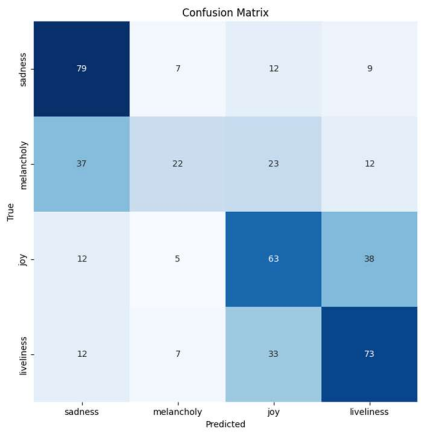
Fine-tuned



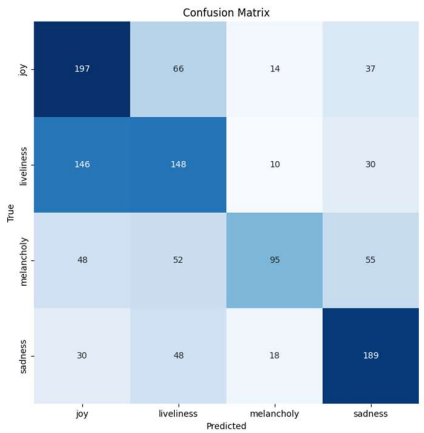
The baseline CNNs model struggled to differentiate between negative emotions like "Angry", "Fear", and "Sad", with numerous misclassifications between these classes, while "Happy" was the most accurately recognised class. With the fine-tuned CNNs model showed an overall improvement in the classification for the "Neutral" and “Sad” classes, but showing a decrease for the emotion “Fear”. The baseline model with CNNs and SVMs performed better in the classification of "Happy" and "Neutral", though still misclassifying between "Angry" and "Fear". Lastly, the fine-tuned model with CNNs and SVMs showed a significant decrease in misclassifications across most classes, except for negative emotions such as "Angry", "Fear", and "Sad" which continued to pose challenges for accurate classification.

CNNs for the Landscape Dataset

Baseline

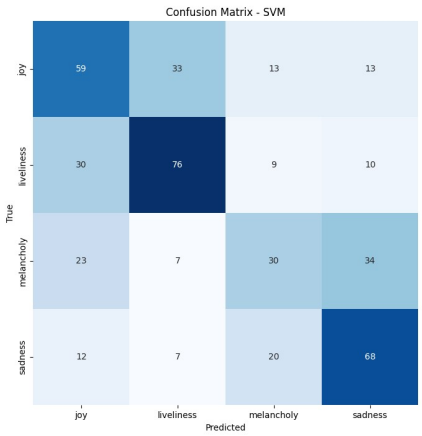


Fine-tuned

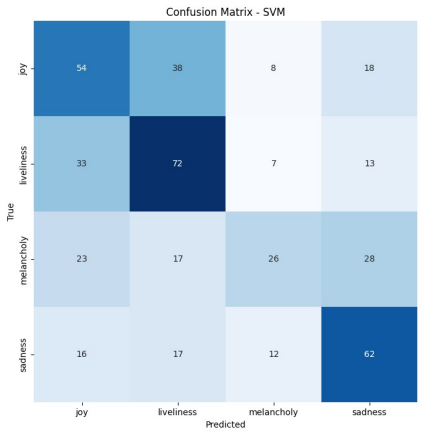


CNNs & SVMs for the Landscape Dataset

Baseline



Fine-tuned



The baseline CNNs model shows a significant number of misclassifications, particularly between "Melancholy" and "Sadness", as well as the "Liveliness" and "Joy" classes. This suggests that the model has difficulty distinguishing emotions related to the landscape elements. On the other hand, the fine-tuned CNNs model shows a slight improvement, especially in the "Joy" class, which is correctly classified in a higher number of cases, although there are still some misclassifications between "Liveliness" and "Joy". When incorporating SVMs into the CNNs baseline model, the overall accuracy improves for certain classes such as "Liveliness", but the confusion between "Melancholy" and "Sadness" remains, indicating that these emotions are harder for the model to differentiate. Finally, the fine-tuned CNNs model with SVMs demonstrates improved performance with fewer misclassifications across all classes, particularly in distinguishing between "Sadness" and "Melancholy".

4.2. Results of music playlist suggestion

As previously mentioned, two relatively simple systems were developed for the music playlist recommendation system, each using the model that performed best for its respective dataset. For the facial expressions dataset, the fine-tuned CNN model was used. The system selects a random image from the dataset, predicts the emotion detected, and suggests an appropriate playlist based on the prediction. This ensures that the music recommendation aligns with the emotional tone of the image. The system performed lower than expected, achieving 60% accuracy for facial expression images and 50% accuracy for landscape images. For each randomly selected image, the correct and predicted labels are displayed, followed by the

corresponding playlist recommendation. Examples of these image predictions and their corresponding playlist suggestions can be found in Appendix 8.11.

Face Expression Playlist Results

#Try	Correct	Predicted
1	Angry	Fear
2	Neutral	Sad
3	Sad	Sad
4	Neutral	Angry
5	Sad	Sad
6	Happy	Happy

Landscape Playlist Results

#Try	Correct	Predicted
1	Liveliness	Liveliness
2	Joy	Liveliness
3	Sadness	Liveliness
4	Sadness	Sadness
5	Joy	Joy
6	Melancholy	Sadness

Although an accuracy of 55% and 60% might suggest a relatively low performance, this results still fall within or close to the average percentage of accuracy of similar projects, particularly for the facial expression dataset, as mentioned by Guo et al. (2023) . However, these results are suboptimal for practical purposes, and this indicates there is a lot of room for improvement.

4.3. Key Findings

For the face expressions dataset, as mentioned in the methods chapter, four models were created: two models used only the CNNs algorithm, and two combined CNNs with SVMs. Among these, the model that produced the best results was the CNNs with the optimal hyperparameter combination found through grid search, achieving an accuracy of 60%. The other models performed below 50%, showing signs of overfitting, and made incorrect classifications. In general, the baseline CNNs model showed difficulty in distinguishing between negative emotions like "Angry", "Fear", and "Sad", while "Happy" was classified with higher accuracy. The fine-tuned CNNs model improved the classification of "Neutral" and "Sad". The baseline CNNs with SVMs model improved the classification of "Happy" and "Neutral", though misclassifications between "Angry" and "Fear" remained. The fine-tuned

CNNs with SVMs model reduced overall misclassifications but continued to face challenges with negative emotions.

Similarly, for the landscapes dataset, four models were developed with the same structure: two models using CNNs and two implementing SVMs to the CNNs. The scores resulted in the models for this dataset were generally lower, with the CNNs model using the best hyperparameter combination performing the best, achieving an accuracy of 55%. Notably, this model did not show signs of overfitting. In general, the baseline CNNs model showed significant misclassifications between "Melancholy" and "Sadness", as well as between "Liveliness" and "Joy". The fine-tuned CNNs model showed improvement, particularly in the classification of "Joy". The baseline CNNs with SVMs model demonstrated improved accuracy for "Liveliness", but misclassifications between "Melancholy" and "Sadness" persisted. The fine-tuned CNNs with SVMs model showed a reduction in misclassifications across all classes, particularly in distinguishing between "Sadness" and "Melancholy".

Lastly, the music playlist recommendation system performed as expected, providing an emotion prediction and recommending the corresponding playlist based on the detected emotion on randomly selected images. The best-performing models for both datasets were used in this system, resulting in an accuracy of 60% for facial expressions and 55% for landscapes.

5. Discussion

This chapter will evaluate and discuss if the project's objectives were fulfilled and reflect on the strengths and weaknesses of the developed models.

5.1. Assessing the Accomplishment of Project Goals

The main objective of this project was to develop machine learning models capable of identifying emotions in images of both human facial expressions and landscapes, and to use these emotional cues to recommend music playlist according to the identified emotion. Although the models did not show high performance or accuracy, they still demonstrate that there is potential for the development and improvement in the identification of emotions for both datasets, achieving 60% accuracy for facial expressions and 55% for landscapes. While these results may not be as expected, it still shows the accomplishment of the goals mentioned in Section 1.

Both datasets were effective for training and testing the models, as they provided the necessary variety of emotions on their corresponding area. In particular, the facial expressions dataset, with its diversity and size, offered enough variation to support the identification of multiple emotions in human faces. However, there is room for improvement for both, especially in the landscape dataset, where the limited size may have affected its overall effectiveness, and since it was created manually, this process may have introduced bias in the classification. Expanding the size of both datasets could enhance their potential for more accurate emotion classification in future developments.

Lastly, the playlist recommendation system also proved to be functional. By using the models with the best performance metrics for both dataset, the system was able to recommend a playlist aligned with the emotion identified in the image. This demonstrates the future potential for integrating emotion detection in images with music recommendation, marking the beginning of a concept that could expand and benefit anyone interested in blending machine learning with creative fields.

5.2. Performance in Relation to Theoretical and Applied Work

When considered in the context of existing theoretical and applied work, as discussed in Chapter 2, the models developed in this project demonstrate both strengths and limitations that reflect the current state of the field. Research by Ekman (1992) on the recognition of basic

human emotions provided the foundational principles for emotion classification from facial expressions, and this framework was instrumental in guiding the development of the facial emotion recognition models used in this project. Ekman's research focused on distinct emotions like happiness, sadness, fear, neutral, and anger, which are also the core emotional categories used in the facial expression dataset. By adopting methodologies similar to Ekman's for identifying basic emotions, this project achieved favourable results. However, it is important to note that the accuracy percentages were lower compared to those reported in Ekman's original findings.

The work of Benenti and Meini (2017) on how landscapes convey emotional cues also played a significant role in shaping the landscape emotion classification model. Their study demonstrated how elements like colour and composition in landscapes can evoke emotions such as melancholy, joy, and sadness. Building on their research, this project applied machine learning techniques to classify these emotions in landscapes, an area that remains relatively underexplored. However, the models encountered difficulties in distinguishing between similar emotions, such as "Liveliness" and "Joy" or "Sadness" and "Melancholy." These challenges align with the findings of Benenti and Meini, and Cai et al (2023), who also reported difficulties in recognising subtle or overlapping emotional cues in visual media. This suggests that further improvements in model architecture or dataset design are needed. Despite these challenges, this project contributes to the growing body of work by taking initial steps towards automating emotional analysis in landscapes, while also highlighting the limitations of current models in handling this complex task.

Regarding the model architecture, the decision to use CNNs as the primary model for both facial expression and landscape emotion classification aligns with the approach commonly used in many state-of-the-art emotion recognition systems, as noted by Mehendale (2020). The CNNs models in this project had the highest accuracy for both dataset, with 60% accuracy for facial expressions and 55% for landscapes. While this barely falls in the lower end of the average results for facial expression recognition projects, it isn´t a good performance and it could have a lot of improvement. As for the landscape dataset, the models were more exploratory in nature, and because of this, there isn´t a way to compare fully if the final results are good or bad, but when using the average accuracy for facial recognition models as reference, the performance falls below that. As Guo et al. (2023) explains, a typical accuracy range for similar models is between 60% and 70%, while more advanced and robust

architectures tend to reach between 70% and 90%. This disparity could be attributed to differences in dataset size, model complexity, or pre-processing techniques. Future improvements to this project could include more complex architectures, such as deeper CNNs, or the introduction of additional regularisation techniques to prevent overfitting.

The decision to experiment with SVMs in combination with CNNs was informed by the research of Ahlawat and Choudhary (2020), which suggested that hybrid models can sometimes improve emotion classification performance. However, in this project, the hybrid CNN-SVM models did not consistently outperform the CNN-only models. This result mirrors findings in the literature, particularly in complex datasets, where SVMs may struggle with multi-class classification tasks, as noted by ElSayed et al. (2021). These findings suggest that while the hybrid approach has theoretical potential, in practice, CNNs may be better suited for tasks involving complex, multi-class emotion detection. This is particularly true for the landscape dataset, where the visual features are more abstract and less easily separable by SVMs, which may explain the observed performance gap.

In summary, this project demonstrates the potential of CNNs for emotion recognition tasks, particularly in multi-class scenarios like facial expressions and landscapes, while highlighting the limitations of hybrid CNNs with SVMs models. Particularly, by focusing on interpreting emotional cues in landscapes, this project lays a foundation for further research, addressing the gap in non-human emotion recognition studies. The challenges of recognising emotions in landscapes arise from the abstract and diverse nature of visual cues, but the results shown in this study suggest that with more sophisticated models and expanded datasets, significant progress can be made in the future. This opens opportunities for practical applications in fields such as filmmaking, gaming, virtual environments, and more, where understanding the emotional tone of natural settings could be important.

5.3. Confidence in Results and Scope

The confidence in the results comes from the careful approach used in optimising and regularising the models, which helped ensure they were as accurate and generalisable as possible, given the limitations of the datasets. The 60% accuracy achieved in facial expression classification and 55% accuracy in landscape emotion recognition, while not perfect, demonstrate that the models were capable of performing the task they were designed for. These results are consistent with similar studies in the field and suggest that the models were

effectively capturing the essential emotional cues in the data. Similarly, the music recommendation system performed as expected, particularly with facial expression images, demonstrating that it is indeed possible to combine machine learning with emotion recognition and music recommendation.

However, it is important to acknowledge the scope of these results. The limited number of emotion categories and the relatively small dataset size for landscapes mean that the models ability to generalise to more complex or subtle emotional states is constrained. The scope of the models is largely confined to the specific emotions and dataset used in this project, and the results may not fully translate to other contexts or datasets where emotional cues might be more diverse or less clearly defined. That said, this project makes a contribution by demonstrating that machine learning models, particularly CNNs, can effectively automate emotion detection not only in facial expressions but also in landscapes, an area which hasn't been fully explored. This ability to extend emotion recognition beyond human subjects showcases the potential of CNNs to handle non-human visual stimuli, providing a valuable foundation for future research and applications in fields such as filmmaking, creative media, and virtual environments.

6. Reflections and Conclusions

This section will discuss reflections and conclusions gained during the development of this project. From an analysis of the project's objectives, methods and implementation of the models, to future work and lessons learned by the author.

6.1. Review of the Project Objectives and Achievements

As previously mentioned, the main objective of this project is to explore the possibility of using machine learning models to identify emotions in images, focusing on both human facial expressions and landscapes, and then pairing these emotions with relevant playlist music. In this regard, the project achieved the expected goals. The models demonstrated that it is possible to extract emotional cues from both types of visual content, however, they still need to be refined and improved, which highlights the potential of CNNs for these tasks. Nonetheless, there were limitations that impacted performance, particularly in the landscape emotion recognition.

6.2. Future Work

Throughout the making of this project, several important lessons were learn regarding the methods and implementation of the models. One of the key insights is the importance of dataset quality and variety. The relatively small and limited dataset for landscape images constrained the models ability to generalise and accurately classify more subtle emotional states. Expanding the dataset to include a wider variety of landscapes and emotions would improve the model’s robustness and performance. Additionally, since the labelling of the landscape images was done manually, this may have impacted the results due to the potential bias in the classification process. A more systematic approach to categorising landscape emotions, possibly by incorporating insights from psychology, could lead to more consistent and accurate classifications.

In terms of model architectures, it became clear that the development of more advanced and robust models could have led to better results, particularly in cases where emotional cues are subtle. Implementing deeper models or architectures like transformers or Recurrent Neural Networks (RNNs) could capture finer details and hierarchical features, improving the ability to distinguish between nuanced emotions. For example, the subtle differences between a "neutral" expression and a slightly "happy" or "sad" expression could be better detected by these improvements. Combining CNNs with RNNs could allow for the temporal sequencing

of facial movements, which is crucial in identifying transitions between emotional states. Also, further experimentation with different model architectures could enhance performance in emotion recognition tasks. Testing transformers or refining the hybrid models, like the ones done in this research of CNNs combined with SVMs, would most definitely give higher results. Additionally, experimenting with different hyperparameter combinations could optimise the existing models, allowing them to better capture the emotional details in both facial expressions and landscapes.

Lastly, to evaluate the effectiveness of the emotion classification and music recommendation system developed in this project, a user study could be implemented as a future work. Participants would be asked to classify the emotions in the images presented to them and assess whether the selected music appropriately matches the emotional tone of each image. This human evaluation would provide valuable feedback on how well the system performs in real-world applications and offer insights for further refinements.

6.3. Contributions of the Project

One of the key contributions of this project is its exploration of emotion recognition in landscapes, an area that has received little attention in previous researches. The project's results show that CNNs are capable of identifying emotions in landscapes, but with some limitations. By demonstrating the application of machine learning models to non-human subjects, the project opens new possibilities for practical applications of this. Additionally, the project contributes to the experimentation in the broader field of emotion recognition in visual media and production.

6.4. Reflections and Lessons Learned

This project was a satisfying experience, as I have always been interested and intrigued by computer vision. Having the opportunity to give life to an idea of a prototype system that could be implemented in the creative field, particularly in music and video production, was especially exciting given my background in these areas. Despite using CNNs models, which are known to be almost a baseline to work for image recognition, this project motivated me to experiment further and take on more ambitious and complex projects, including developing full systems that could be integrated into editing or animation software. It also helped me expand my curiosity on of how machine learning and AI can support creatives in overcoming the challenges they could often face during the process of creative production.

7. References

Ahlawat, S. and Choudhary, A. (2020). Hybrid CNNs-SVMs Classifier for Handwritten Digit Recognition. *Procedia Computer Science*, 167, pp.2554–2560. doi: https://doi.org/10.1016/j.procs.2020.03.309.

Amato, G., Behrmann, M., Bimbot, F., Caramiaux, B., Falchi, F., Garcia, A., Geurts, J., Gibert, J., Gravier, G., Holken, H., Koenitz, H., Lefebvre, S., Liutkus, A., Lotte, F., Perkis, A., Redondo, R., Turrin, E., Vieville, T. and Vincent, E. (2019). AI in the media and creative industries. *AI in the media and creative industries*. doi: https://doi.org/10.48550/arxiv.1905.04175.

Ansani, A. et al. (2020) 'How Soundtracks Shape What We See: Analyzing the Influence of Music on Visual Scenes Through Self-Assessment, Eye Tracking, and Pupillometry', *Frontiers In Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.02242.

Beauxis-Aussalet, E. and Hardman, L. (n.d.). Visualization of Confusion Matrix for Non-Expert Users. [online] CWI - Information Access Group. Available at: https://vis.cs.ucdavis.edu/vis2014papers/VIS_Conference/infovis/posters/beauxis-aussalet.pdf [Accessed 16 Sep. 2024].

Belete, D.M. and Huchaiah, M.D. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 44(9), pp.1–12. doi: https://doi.org/10.1080/1206212x.2021.1974663.

Benenti, M. and Meini, C. (2017). The Recognition of Emotions in Music and Landscapes: Extending Contour Theory. *Philosophia*, 46(3), pp.647–664. doi: https://doi.org/10.1007/s11406-017-9874-4.

Briot, J.-P., Hadjeres, G. y Pachet, F.-D. (2020) 'Deep Learning Techniques for Music Generation', *Computational synthesis and creative systems*. https://doi.org/10.1007/978-3-319-70163-9.

Cai, Y., Li, X. and Li, J. (2023). Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors*, 23(5), p.2455. doi: https://doi.org/10.3390/s23052455.

Calvo, R.A. y D'Mello, S. (2010) 'Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications', *IEEE Transactions On Affective Computing*, 1(1), pp. 18-37. https://doi.org/10.1109/t-affc.2010.1.

Canedo, D. and Neves, A.J.R. (2019). Facial Expression Recognition Using Computer Vision: A Systematic Review. Applied Sciences, 9(21), p.4678. doi: https://doi.org/10.3390/app9214678.

De-Lima-Santos, M.-F. y Ceron, W. (2021) 'Artificial Intelligence in News Media: Current Perceptions and Future Outlook', *Journalism And Media*, 3(1), pp. 13-26. https://doi.org/10.3390/journalmedia3010002.

Dumitru, Goodfellow, I., Cukierski, W. and Bengio, Y. (2013). Challenges in Representation Learning: Facial Expression Recognition Challenge | Kaggle. [online] Kaggle.com. Available at:

http://www.kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge/overview/ [Accessed 6 Aug. 2024].

Ekman, P. (1992) 'An argument for basic emotions', *Cognition And Emotion*, 6(3-4), pp. 169-200. https://doi.org/10.1080/02699939208411068.
ElSayed, M.S., Le-Khac, N.-A., Albahar, M.A. and Jurcut, A. (2021). A novel hybrid model for intrusion detection systems in SDNs based on CNNs and a new regularization technique. *Journal of Network and Computer Applications*, 191, p.103160. doi: https://doi.org/10.1016/j.jnca.2021.103160.

Gu, R., Li, H., Su, C. and Wu, W., (2023). Innovative Digital Storytelling with AIGC: Exploration and Discussion of Recent Advances. *arXiv preprint* arXiv:2309.14329 (Accessed: 16 May 2024).
Guo, X., Zhang, Y., Lu, S. and Lu, Z. (2023). Facial expression recognition: a review. Multimedia Tools and Applications, 83. doi: https://doi.org/10.1007/s11042-023-15982-x.

Hoeckner, B. et al. (2011) 'Film music influences how viewers relate to movie characters.', *Psychology Of Aesthetics, Creativity, And The Arts*, 5(2), pp. 146-153. https://doi.org/10.1037/a0021544.

Hutson, J. (2023). *AI and the Creative Process: Part One*. [online] JSTOR Daily. Available at: https://daily.jstor.org/ai-and-the-creative-process-part-one/ [Accessed 8 Jul. 2024].

Kingma, D. and Ba, J. (2015). adam: A Method for Stochastic Optimization. In: Computer Science. [online] 3rd International Conference for Learning Representations. doi: https://doi.org/10.48550/arXiv.1412.6980.

Mazzone, M. and Elgammal, A. (2019). Art, Creativity, and the Potential of Artificial Intelligence. Arts, [online] 8(1), pp.1–9. doi:https://doi.org/10.3390/arts8010026.

Mehendale, N. (2020) 'Facial emotion recognition using convolutional neural networks (FERC)', *SN Applied Sciences/SN Applied Sciences*, 2(3). https://doi.org/10.1007/s42452-020-2234-1.

Neumeyer, D. (2013b) The Oxford Handbook of Film Music Studies, *Oxford University Press eBooks*. https://doi.org/10.1093/oxfordhb/9780195328493.001.0001.

O'Shea, K. and Nash, R. (2015). An Introduction to Convolutional Neural Networks. doi: https://doi.org/1511.08458v2.

Skillcate (2022). emotion-detection-cnn-model. [online] GitHub. Available at: https://github.com/skillcate/emotion-detection-cnn-model/tree/main [Accessed 15 Sep. 2024].

Sturm, B.L. et al. (2019) 'Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis', *Arts*, 8(3), p. 115. https://doi.org/10.3390/arts8030115.

Tran, H. et al. (2023) 'Emotion-Aware music recommendation', *Proceedings Of The ... AAAI Conference On Artificial Intelligence*, 37(13), pp. 16087-16095. https://doi.org/10.1609/aaai.v37i13.26911.

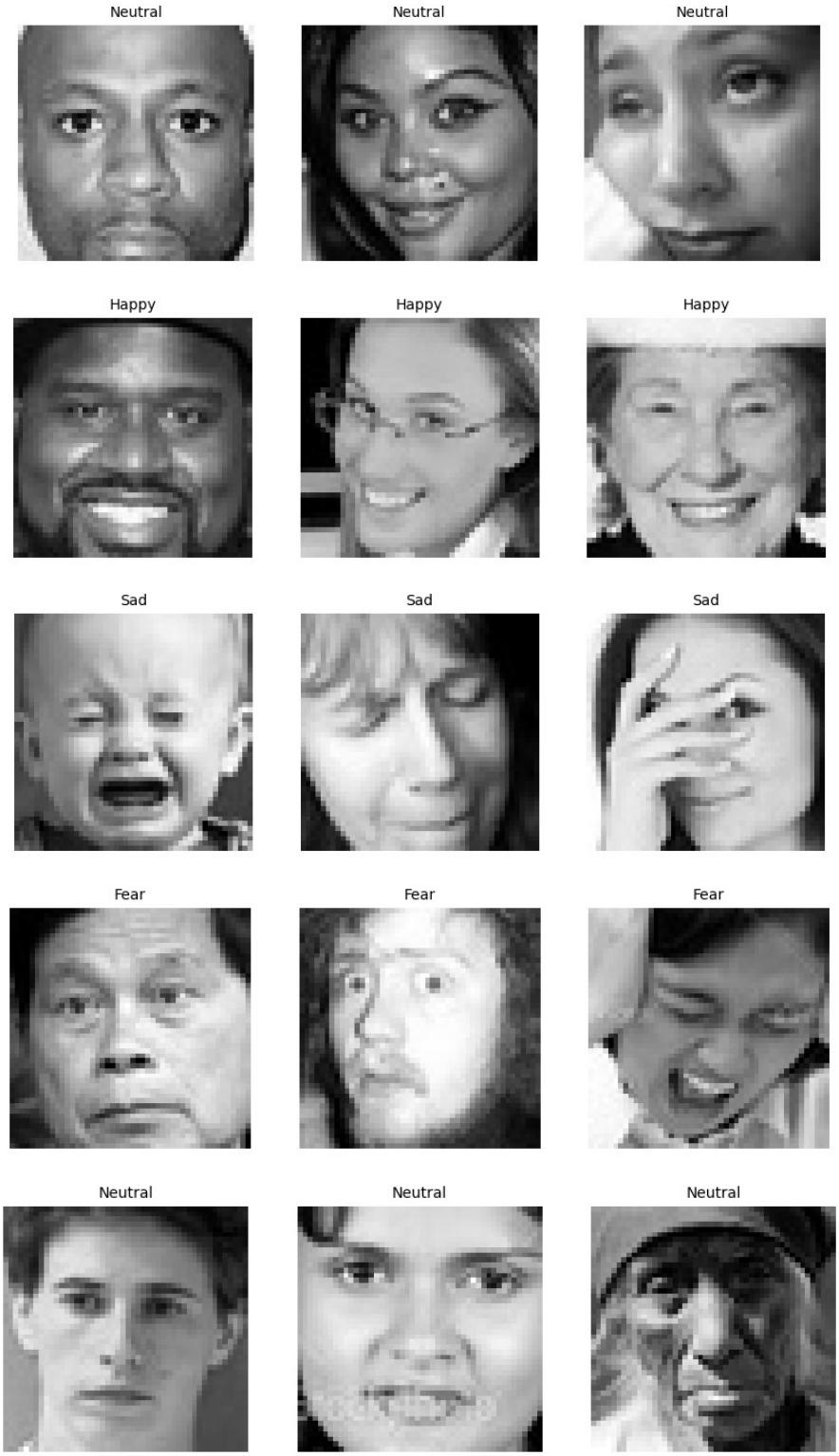
Trattner, C., Jannach, D., Motta, E., Costera Meijer, I., Diakopoulos, N., Elahi, M., Opdahl, A.L., Tessem, B., Borch, N., Fjeld, M., Øvrelid, L., De Smedt, K. and Moe, H. (2021). Responsible media

technology and AI: challenges and research directions. *AI and Ethics*, [online] 2, pp.585–594. doi: https://doi.org/10.1007/s43681-021-00126-4.

Truong, N.B.T., Venkatesh, S. y Dorai, C. (2003) 'Scene extraction in motion pictures', *IEEE Transactions On Circuits And Systems For Video Technology*, 13(1), pp. 5-15. https://doi.org/10.1109/tcsvt.2002.808084.

8. Appendix

8.1. Examples of the Classification and Processing of Facial Expression Images



8.2. CNNs Baseline Model Structure for Facial Expressions

Layer (type)	Output Shape	Param #
Input_1 (InputLayer)	(None, 48, 48, 1)	0
conv2d (Conv2D)	(None, 48, 48, 16)	160
max_pooling2d (MaxPooling2D)	(None, 24, 24, 16)	0
conv2d_1 (Conv2D)	(None, 24, 24, 32)	4640
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 32)	0
flatten (Flatten)	(None, 4608)	0
dense(Dense)	(None, 64)	294976
dense_1 (Dense)	(None, 5)	325

8.3. CNNs Best Hyperparameters Model Structure for Face Expressions

Layer (type)	Output Shape	Param #
Input_layer_2 (InputLayer)	(None, 48, 48, 1)	0
conv2d_8 (Conv2D)	(None, 48, 48, 32)	320
leaky_re_lu_4 (LeakyReLU)	(None, 48, 48, 32)	0
dropout_10 (dropout)	(None, 48, 48, 32)	0
max_pooling2d_8 (MaxPooling2D)	(None, 24, 24, 32)	0
conv2d_9 (Conv2D)	(None, 24, 24, 64)	18,496
leaky_re_lu_5 (LeakyReLU)	(None, 24, 24, 64)	0
dropout_11 (dropout)	(None, 24, 24, 64)	0
max_pooling2d_9 (MaxPooling2D)	(None, 12, 12, 64)	0
conv2d_10 (Conv2D)	(None, 12, 12, 128)	73,856
leaky_re_lu_6 (LeakyReLU)	(None, 12, 12, 128)	0
dropout_12 (dropout)	(None, 12, 12, 128)	0
max_pooling2d_10 (MaxPooling2D)	(None, 6, 6, 128)	0
conv2d_11 (Conv2D)	(None, 6, 6, 256)	295,168
leaky_re_lu_7 (LeakyReLU)	(None, 6, 6, 256)	0
dropout_13 (dropout)	(None, 6, 6, 256)	0
max_pooling2d_11 (MaxPooling2D)	(None, 3, 3, 256)	0
flatten_2 (Flatten)	(None, 2304)	0
dense_4 (Dense)	(None, 128)	295,040
dropout_14 (dropout)	(None, 128)	0
dense_5 (Dense)	(None, 5)	645

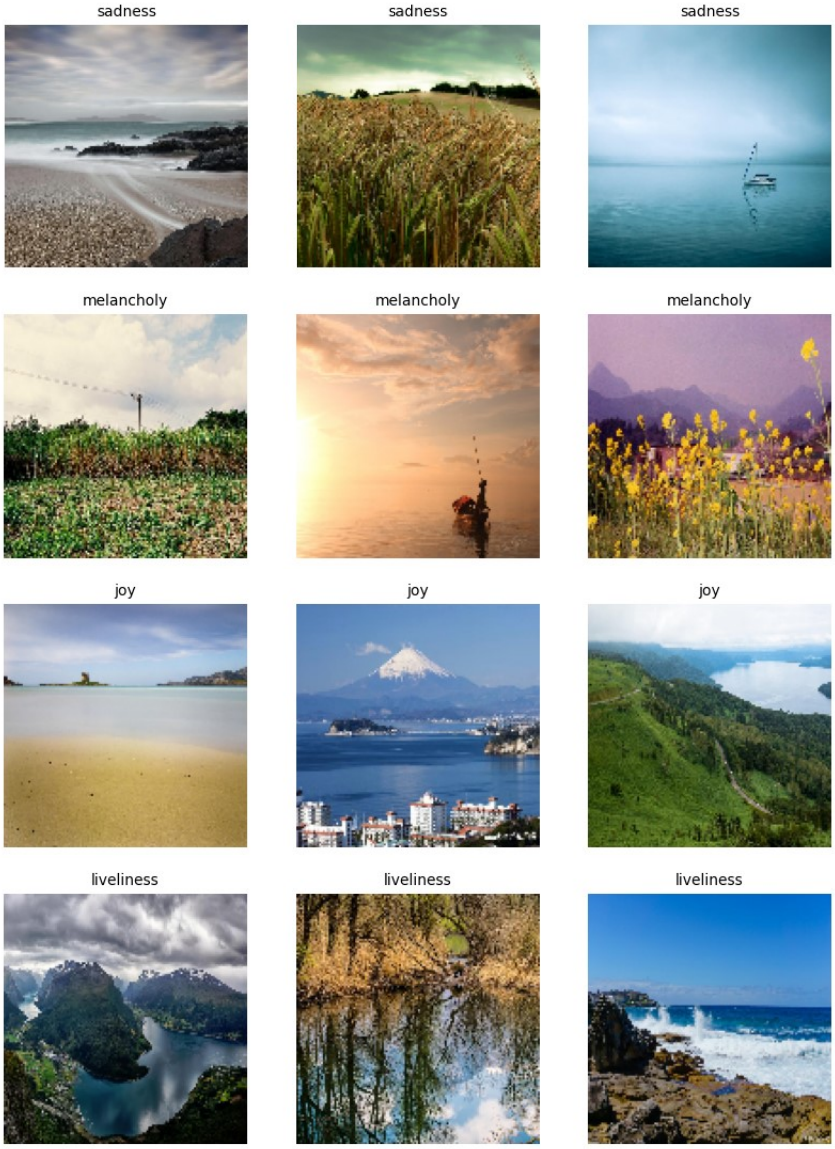
8.4. CNNs with SVMs Baseline Model Structure for Face Expressions

Layer (type)	Output Shape	Param #
Input_layer_2 (InputLayer)	(None, 48, 48, 1)	0
conv2d (Conv2D)	(None, 48, 48, 16)	160
max_pooling2d (MaxPooling2D)	(None, 24, 24, 16)	0
conv2d_1 (Conv2D)	(None, 24, 24, 32)	4,640
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 32)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 64)	294,976
dense_1 (Dense)	(None, 5)	325

8.5. CNNs with SVMs Best Hyperparameters Model Structure for Face Expressions

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 48, 48, 1)	0
conv2d_4 (Conv2D)	(None, 48, 48, 32)	320
batch_normalization (BatchNormalization)	(None, 48, 48, 32)	128
leaky_re_lu_4 (LeakyReLU)	(None, 48, 48, 32)	0
dropout_5 (dropout)	(None, 48, 48, 32)	0
max_pooling2d_4 (MaxPooling2D)	(None, 24, 24, 32)	0
conv2d_5 (Conv2D)	(None, 24, 24, 64)	18,496
batch_normalization_1 (BatchNormalization)	(None, 24, 24, 64)	256
leaky_re_lu_5 (LeakyReLU)	(None, 24, 24, 64)	0
dropout_6 (dropout)	(None, 24, 24, 64)	0
max_pooling2d_5 (MaxPooling2D)	(None, 12, 12, 64)	0
conv2d_6 (Conv2D)	(None, 12, 12, 128)	73,856
batch_normalization_2 (BatchNormalization)	(None, 12, 12, 128)	512
leaky_re_lu_6 (LeakyReLU)	(None, 12, 12, 128)	0
dropout_7 (dropout)	(None, 12, 12, 128)	0
max_pooling2d_6 (MaxPooling2D)	(None, 6, 6, 128)	0
conv2d_7 (Conv2D)	(None, 6, 6, 256)	295,168
batch_normalization_3 (BatchNormalization)	(None, 6, 6, 256)	1,024
leaky_re_lu_7 (LeakyReLU)	(None, 6, 6, 256)	0
dropout_8 (dropout)	(None, 6, 6, 256)	0
max_pooling2d_7 (MaxPooling2D)	(None, 3, 3, 256)	0
conv2d_8 (Conv2D)	(None, 3, 3, 512)	1,180,160
batch_normalization_4 (BatchNormalization)	(None, 3, 3, 512)	2,048
leaky_re_lu_8 (LeakyReLU)	(None, 3, 3, 512)	0
dropout_9 (dropout)	(None, 3, 3, 512)	0
max_pooling2d_8 (MaxPooling2D)	(None, 1, 1, 512)	0
flatten_1 (Flatten)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131,328
dropout_10 (dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32,896
dropout_11 (dropout)	(None, 128)	0
dense_4 (Dense)	(None, 5)	645

8.6. Examples of the Classification and Processing of Landscapes Images



8.7. CNNs Baseline Model Structure for Landscapes

Layer (type)	Output Shape	Param #
Input_layer (InputLayer)	(None, 128, 128, 3)	0
conv2d (Conv2D)	(None, 128, 128, 32)	896
max_pooling2d (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_1 (Conv2D)	(None, 64, 64, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 32, 32, 64)	0
conv2d_2 (Conv2D)	(None, 32, 32, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 16, 16, 128)	0
flatten (Flatten)	(None, 32768)	0
dense (Dense)	(None, 128)	4,194,432
dropout (dropout)	(None, 128)	0
dense_1 (Dense)	(None, 4)	516

8.8. CNNs Best Hyperparameters Model Structure for Landscapes

Layer (type)	Output Shape	Param #
Input_layer_5 (InputLayer)	(None, 128, 128, 3)	0
conv2d_17 (Conv2D)	(None, 128, 128, 32)	896
max_pooling2d_17 (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_1_18 (Conv2D)	(None, 64, 64, 64)	18,496
max_pooling2d_18 (MaxPooling2D)	(None, 32, 32, 64)	0
conv2d_19(Conv2D)	(None, 32, 32, 128)	73,856
max_pooling2d_19 (MaxPooling2D)	(None, 16, 16, 128)	0
flatten (Flatten)	(None, 32768)	0
dense_10 (Dense)	(None, 128)	4,194,432
dropout_20 (dropout)	(None, 128)	0
dense_11 (Dense)	(None, 4)	516

8.9. CNNs with SVMs Baseline Model Structure for Landscapes

Layer (type)	Output Shape	Param #
Input_layer (InputLayer)	(None, 128, 128, 3)	0
conv2d (Conv2D)	(None, 128, 128, 16)	448
max_pooling2d (MaxPooling2D)	(None, 64, 64, 16)	0
conv2d_1 (Conv2D)	(None, 64, 64, 32)	4,640
max_pooling2d_1 (MaxPooling2D)	(None, 32, 32, 32)	0
flatten (Flatten)	(None, 32768)	0
dense (Dense)	(None, 64)	2,097,216
Dense_1 (Dense)	(None, 4)	260

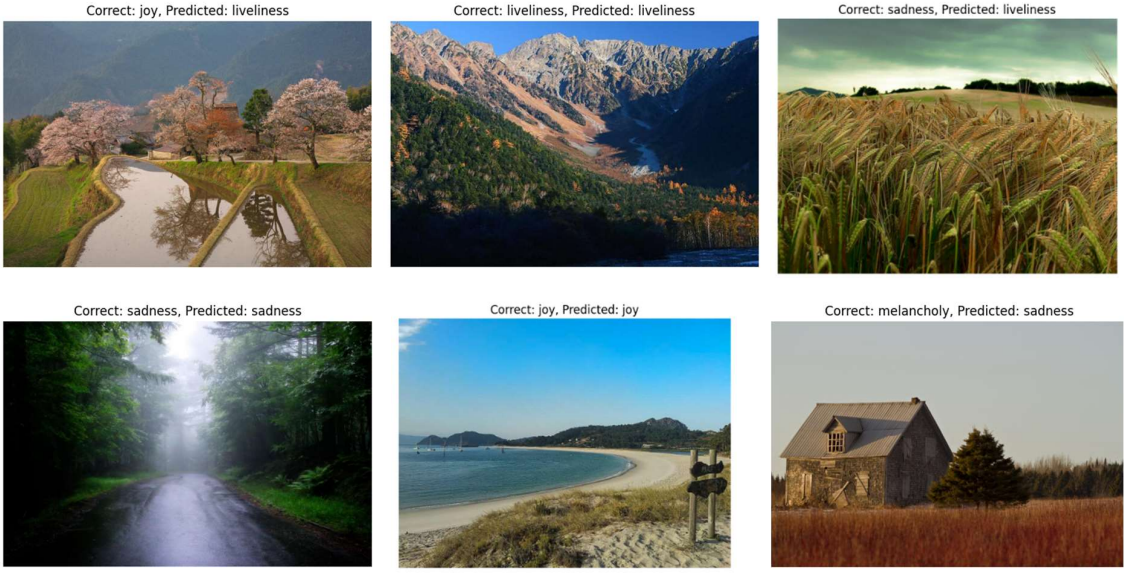
8.10. CNNs with SVMs Best Hyperparameters Model Structure for Landscapes

Layer (type)	Output Shape	Param #
Input_layer_1 (InputLayer)	(None, 128, 128, 3)	0
conv2d_4 (Conv2D)	(None, 128, 128, 64)	1,792
batch_normalization_4 (BatchNormalization)	(None, 128, 128, 64)	256
activation_4 (Activation)	(None, 128, 128, 64)	0
max_pooling2d_4 (MaxPooling2D)	(None, 64, 64, 64)	0
dropout_4 (dropout)	(None, 64, 64, 64)	0
conv2d_5 (Conv2D)	(None, 64, 64, 128)	73,856
batch_normalization_5 (BatchNormalization)	(None, 64, 64, 128)	512
activation_5 (Activation)	(None, 64, 64, 128)	0
max_pooling2d_5 (MaxPooling2D)	(None, 64, 64, 128)	0
dropout_5 (dropout)	(None, 32, 32, 128)	0
conv2d_6 (Conv2D)	(None, 32, 32, 128)	295,168
batch_normalization_6 (BatchNormalization)	(None, 32, 32, 256)	1,024
activation_6 (Activation)	(None, 32, 32, 256)	0
max_pooling2d_6 (MaxPooling2D)	(None, 16, 16, 256)	0
dropout_6 (dropout)	(None, 16, 16, 256)	0
conv2d_7 (Conv2D)	(None, 16, 16, 512)	1,180,160
batch_normalization_7 (BatchNormalization)	(None, 16, 16, 512)	2,048
activation_7 (Activation)	(None, 16, 16, 512)	0
max_pooling2d_7 (MaxPooling2D)	(None, 8, 8, 512)	0
dropout_7 (dropout)	(None, 8, 8, 512)	0
flatten_1(Flatten)	(None, 32768)	0
dense_1 (Dense)	(None, 256)	8,388,864

8.11. Playlist Results Images

Face Expression Dataset

Landscape Dataset



8.12. Link to Full Code

Here is the link to the full code in which is located in Google Drive.

https://www.google.com/url?q=https%3A%2F%2Fdrive.google.com%2Fdrive%2Ffolders%2F1_NAstIXbev-V0BiBxFbmukq4gDJ7pt4a%3Fusp%3Dsharing

8.13. Proposal

Project Proposal
M.Sc. in Data
Science

School of Science and Technology at City University of London

Identifying Emotions and Paring No Copyright Music to Photographs: An AI Approach to Enhance Creative Media Production

Author: Samantha Georgina Isaac Munoz
Supervisor: Jialie Shen

1. Introduction

1.1. Problem statement

The right music enhances the emotional appeal and memorability of a film, as well as being a powerful guide to emotional responses that increases the effect and memorability of scenes in particular (Hoeckner et al., 2011). This is especially relevant to independent filmmakers and content creators, who work on limited budgets and few resources, making it hard for them to be able to afford professional music composition services. Therefore, they are often compelled to use pre-existing music in their production, which are much less effective at driving an emotional tone in the scene, making the viewer experience less immersive and engaging.

In fact, the absence of such effective and reliable ways to identify and recommend the most emotionally suitable music worsens these problems. Conventional ways of selecting music take time and require a certain amount of knowledge to make these critical decisions effectively, which can be challenging for novice filmmakers who may not have the necessary experience. The task usually involves manual search of music libraries, which is not only labor-intensive but also demanding in terms of understanding how various musical elements can convey different emotions (Neumeyer, 2013b).

This has created the need for a solution that simplifies the process, allowing high-quality, emotionally relevant music to be accessed by all creators without a demand on their resources. The solution to this problem is currently being formed by integrating machine learning and artificial intelligence to create a model that can analyse visual and audio data for emotional cues and suggest relevant music tracks according to the emotion (Sturm *et al.*, 2019). It can be trained on large datasets of film scenes with related musical scores, enabling them to learn the complex relationships between visual emotions and musical elements (Ansani et al., 2020).

Machine learning has demonstrated a great potential in the filmmaking industry. AI-driven tools may be used for the required script, scene recognition, or even editing, which proves the flexibility and power of the combination of machine learning with filmmaking to augment the creative process (Gu et al., 2023). The use of these models can reduce both the time and labour costs associated with searching for the right music and hasten the timelines of a project at a lower cost of production, consequently making the filmmaking process more productive and easier to accomplish.

1.2. Aim

The aim of this project is to explore the potential of using machine learning with the filmmaking industry, and the potential benefits it can bring to small or independent filmmakers and content creators by optimizing the editing process by giving access to high-quality, emotionally relevant music to produce more engaging and immersive visual content.

1.3. Objectives

- Explore the potential of machine learning technologies to improve the filmmaking process, focusing on benefits for small and independent filmmakers or content creators.
- Create a model to automatize the editing workflow by automatically pairing emotionally appealing, copyright-free music with audiovisual projects.
- Develop and compile a dataset that contains copyright-free music.
- Validate the effectiveness of the model through testing with several photographs and comparing the emotional compatibility of the music recommendations the model will give.

1.4. Research Question

Can a machine learning model specialized in computer vision effectively identify the emotions present in human faces and landscapes, and appropriately assign music that reflects the emotion detected in the image?

1.5. Outputs

The project will provide a detailed framework for the development of a machine learning model that identifies emotions in photographs and automatically linking them with emotion-relevant copyright-free music. This will also include technical specifications, including an explanation of the algorithms used, and an accompanying guideline of the implementation. In addition, a full working prototype of the model will also be provided.

1.6. Beneficiaries

The main beneficiaries will be independent filmmakers and content creators who want to concentrate on developing and producing stories without having to be concerned about collecting and choosing copyright-free music for their projects. Furthermore, this project has the potential to be a starting point for those interested in developing a more sophisticated model with the same objective of simplifying the editing process of any audiovisual project.

1.7. Scope

- To avoid copyright-related conflicts, the project will not use frames from scenes from Hollywood movies or animated series. Instead, a dataset including landscape photographs and human faces will be used. This selection will not affect the main

objective of the project. As mentioned by Truong, Venkatesh y Dorai (2003), a scene is defined as a section of a motion picture that is unified in time and space, composed of a series of shots from different angles. As a result, the utilization of a collection of static images are consistent to this definition and preserves the authenticity of the investigation.

- Also, for the copyright-free music dataset, a manual compilation will be carried out, in which each song will be carefully selected and categorised according to the emotions assigned by the authors and the descriptions provided in their metadata. This approach ensures that the dataset is consistent and also avoids subjective interpretations of the music that will be utilised.

- Given the short time frame available for its development, the outcome of this project will be a prototype.

2. Critical Context

2.1. Introduction

The intersection of artificial intelligence and creative media production marks one of the more significant areas of technological development and academic interest. Based within a critical review of relevant literature, this project, with a focus on pairing copyright-free music with photographs based on emotional analysis, contextualises the research question and informs the methodological choices, with elements from a broad range of scientific and technical sources.

2.2. Artificial Intelligence Application in Creative Media Production

Artificial intelligence has been an area of immense research in the creative industries. De-Lima-Santos y Ceron (2021) explains that artificial intelligence in media production is at the meeting point of ethical and practical considerations with respect to its potential to automate and enhance the creative process. His work highlights the increasing reliance on artificial intelligence to streamline traditionally human-intensive and time-consuming tasks, such as music selection in film editing. This aligns very well with the project's goal to make the process of associating music with video content more effortless for independent filmmakers, thereby optimizing the editing process and making high-quality, emotionally relevant music more accessible.

Artificial intelligence can revolutionize the user experience of music recommendation systems by incorporating emotional recognition into music recommendation engines. Machine learning models enables the analysis of user emotions and the recommendation of music that aligns with their mood at a particular moment (Tran et al., 2023). This is certainly in line with the objective of the project, which is to link emotionally relevant music with photographs using similar technology to enhance the emotional result of the visual media. To the independent filmmaker, it is the provision of a high-tech tool that significantly helps in integrating it in a delicate fashion and leading the music with the visual for a better fit of the two, thus achieving emotional more resonance in the final product.

Moreover, the technical aspects of AI in music composition and recommendation are explored by Briot, Hadjeres, and Pachet (2020). They examine various machine learning

algorithms used in music generation and their effectiveness in creating emotionally resonant compositions. This technical perspective provides a foundation for selecting appropriate algorithms and techniques for the project's AI model, ensuring that the music recommendations are both relevant and high-quality. By integrating such advanced algorithms, the project can offer filmmakers music options that elevate the emotional depth and quality of their visual content.

2.3. Emotional Detection with Computer Vision

In the field of emotion identification, Ekman's (1992) research paper laid the foundation for understanding how emotions can be systematically identified based on facial expressions. It has been applied in the development of effective machine learning models, meaning the emotional cues derived from visual data can be interpreted. These established principles are applied to the use of computer vision to detect emotions in a photograph, allowing the machine learning model to follow emotional visual cues. By leveraging these principles, the project aims to create more engaging and immersive visual content for small or independent filmmakers.

Calvo and D'Mello (2010) explore the broader field of affective machine learning, which involves the study and development of systems that can recognise, interpret, and replicate human emotions. Their work contributes to the idea that the recognition of emotion is highly context-dependent; hence, the training data used by machine learning systems must be comprehensive and inclusive for such systems to be sensitive to and accurately predict responses. The number of diverse and representative human emotions and landscapes in the dataset has a significant impact on the approach to dataset compilation and model training in the project. Being able to use all-inclusive datasets will ensure that the machine learning can adapt to the subtle emotional needs of different visual narrations, in turn adding value to the storytelling of filmmakers, and content creators.

The paper by Mehendale (2020) explores the potential of use of Convolutional Neural Networks (CNNs) for facial emotion detection. Specifically, two CNN models were developed and trained using greyscale images to categorize facial expressions into emotions such as happy, sad, angry, neutral, and fear. These models use several key techniques to enhance accuracy, including batch normalization and dropout to mitigate overfitting. The best model achieved an accuracy of 80% for detecting four emotions and 72% for five emotions, demonstrating the potential of CNNs in emotion recognition tasks. By leveraging CNNs, this project can develop a model that understands and categorizes the emotional tone of visual content, allowing for the automatic selection of music that matches the emotional context of scenes. This can enhance the immersive experience for viewers and simplifies the editing process for independent filmmakers and content creators, making advanced machine learning techniques accessible and beneficial to the filmmaking industry.

2.4. Conclusion

These sources collectively provide a comprehensive context for the research question, illustrating the current capabilities and future potential of machine learning in media production. They highlight the importance of accurate emotion recognition and the role of machine learning in automating creative processes. This project builds on these insights, aiming to develop a model that can autonomously pair copyright-free music with photographs based on emotional analysis. The methodological choices, from dataset compilation to

algorithm selection, are informed by the reviewed literature, ensuring that the project is both innovative and grounded in established research.

3. Approaches: Methods & Tools for Design, Analysis & Evaluation

3.1. Design

The design phase will follow a systematic and structured methodical process. An in depth literature review will be conducted to identify the most promising and proven techniques used in similar projects related to emotion detection through photographs of human faces and sceneries. Subsequently, the development of a framework for guidance on how to successfully complete the objectives of the project and adequately answer the research question will be made. This part of the project is totally theoretical, therefore, no specific tools are needed at this stage.

3.2. Implementation

During the model implementation phase, code will be developed using my personal laptop as the primary platform for development. The model will be implemented using Python on Google Colab since it provides access to superior computational resources, ensuring that the processing demands of the model are properly within the hardware limitations of the personal device. Libraries from Python will include NumPy, Pandas, TensorFlow, and PyTorch, which have strong functionalities to perform data manipulation, machine learning, and deep learning processes. The implementation will take into close interest the effectiveness and performance of the system for proper responsiveness and accuracy under all computational loads.

3.3. Analysis and Evaluation

I will perform a series of test to measure various performance metrics such as accuracy of emotion detection, relevance of the music selected, and computational efficiency. Key test scenarios will include:

- Testing the emotion recognition on a range of images to ensure the model accurately identifies the underlying emotions.
- Validating the music matching algorithm by comparing the emotional tone of the selected music with the emotions identified in the images.
- Performance testing to guarantee that the model operates efficiently within the computing constraints typically found in film production settings.

3.4. Ethical Issues

The datasets used in this study will be designed to meet research ethics standards or will be already publicly available and cited as corresponded. In accordance with these standards, proper ethical measures will be strictly adhered to throughout the project, despite the fact it does not directly involve human subjects, to ensure compliance with privacy and data protection laws. External resources, such as datasets, algorithms, and code snippets, will be properly referenced to avoid plagiarism.

4. Risk

Risk	Probability	Impact	Contingency Plan
The datasets may not be varied or large enough to train the model effectively.	Medium	High	Expand the search for data collection sources, including collaborations with photographers and musicians.
The model may not be able to accurately identify emotions due to limitations in training or the complexity of emotional analysis.	Medium	Medium	Implement additional phases of model validation and tuning.
The project might face time limitations that prevent achieving all the objectives.	Medium	High	A workflow plan will be established that will be closely followed, with no deviations, focusing exclusively on the main objectives that must be met.
Programming codes, datasets, model results, and documentation, could be lost due to hardware failures, software errors, or malware attacks on the personal computer used for project development.	Low	High	Establish an automatic backup system (or manual) that saves copies of all important project files at regular intervals.
Not understanding libraries or theories that are crucial for the development of the project.	Medium	Medium	Asking my supervisor for advice and researching about the topic.
The available computer equipment does not have the necessary capacity or power to efficiently run the machine learning models.	Medium	Medium	Working on optimizing the code and models to ensure they are as efficient as possible.

Figure 1. Risks Table

5. Work Plan

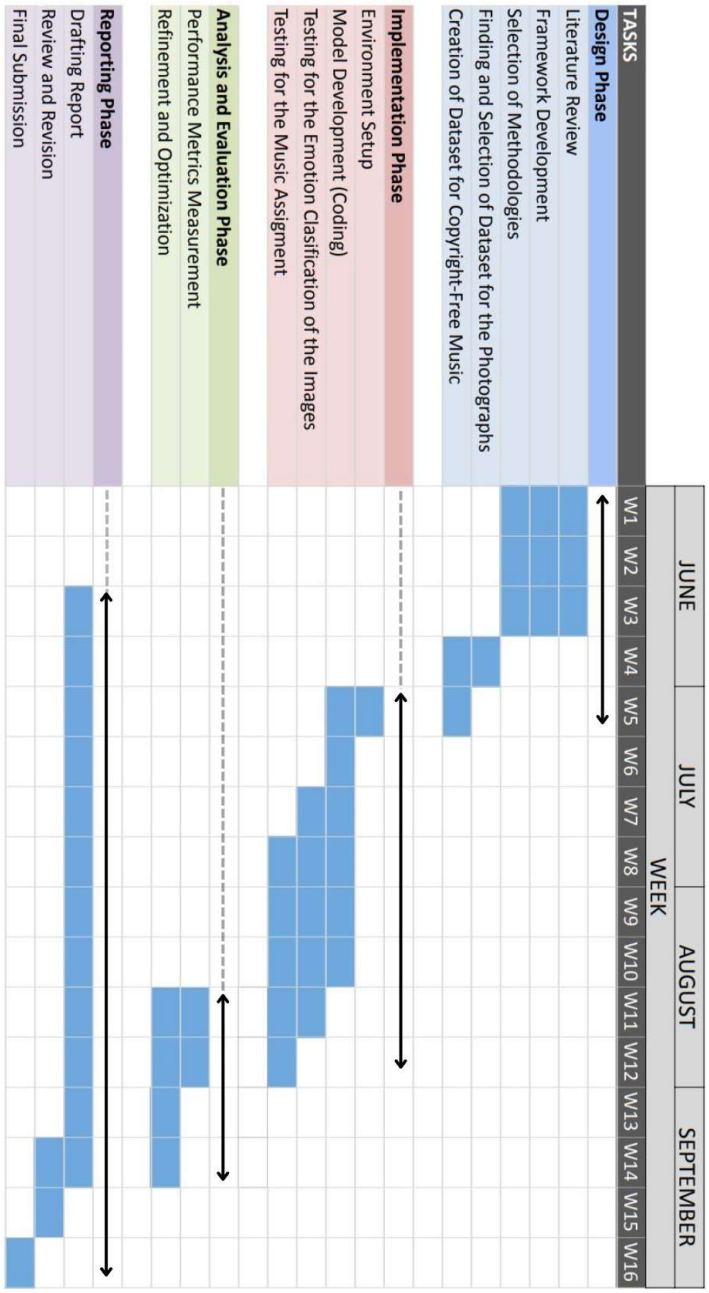


Figure 2. Working Plan

6. Ethics Review

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**

2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**

7. References

Hoeckner, B. et al. (2011) 'Film music influences how viewers relate to movie characters.', *Psychology Of Aesthetics, Creativity, And The Arts*, 5(2), pp. 146-153. https://doi.org/10.1037/a0021544.

Neumeyer, D. (2013b) The Oxford Handbook of Film Music Studies, *Oxford University Press eBooks*. https://doi.org/10.1093/oxfordhb/9780195328493.001.0001.

Sturm, B.L. et al. (2019) 'Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis', *Arts*, 8(3), p. 115. https://doi.org/10.3390/arts8030115.

Ansani, A. et al. (2020) 'How Soundtracks Shape What We See: Analyzing the Influence of Music on Visual Scenes Through Self-Assessment, Eye Tracking, and Pupillometry', *Frontiers In Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.02242.

Gu, R., Li, H., Su, C. and Wu, W., (2023). Innovative Digital Storytelling with AIGC: Exploration and Discussion of Recent Advances. *arXiv preprint* arXiv:2309.14329 (Accessed: 16 May 2024).

Truong, N.B.T., Venkatesh, S. y Dorai, C. (2003) 'Scene extraction in motion pictures', *IEEE Transactions On Circuits And Systems For Video Technology*, 13(1), pp. 5-15. https://doi.org/10.1109/tcsvt.2002.808084.

De-Lima-Santos, M.-F. y Ceron, W. (2021) 'Artificial Intelligence in News Media: Current Perceptions and Future Outlook', *Journalism And Media*, 3(1), pp. 13-26. https://doi.org/10.3390/journalmedia3010002.

Ekman, P. (1992) 'An argument for basic emotions', *Cognition And Emotion*, 6(3-4), pp. 169-200. https://doi.org/10.1080/02699939208411068.

Calvo, R.A. y D'Mello, S. (2010) 'Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications', *IEEE Transactions On Affective Computing*, 1(1), pp. 18-37. https://doi.org/10.1109/t-affc.2010.1.

Tran, H. et al. (2023) 'Emotion-Aware music recommendation', *Proceedings Of The ... AAAI Conference On Artificial Intelligence*, 37(13), pp. 16087-16095. https://doi.org/10.1609/aaai.v37i13.26911.

Mehendale, N. (2020) 'Facial emotion recognition using convolutional neural networks (FERC)', *SN Applied Sciences/SN Applied Sciences*, 2(3). https://doi.org/10.1007/s42452-020-2234-1.

Briot, J.-P., Hadjeres, G. y Pachet, F.-D. (2020) 'Deep Learning Techniques for Music Generation', *Computational synthesis and creative systems*. https://doi.org/10.1007/978-3-319-70163-9.