National College of Ireland

Advancing Safety in Vehicles with AI-Driven Emotion Recognition

MSc Research Project
Artificial Intelligence

Sonali Subhash Jadhav
Student ID: x21236071

School of Computing
National College of Ireland

Supervisor: Prof. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sonali Subhash Jadhav
Student ID:	X21236071
Programme:	MSc. Artificial Intelligence **Year:** 2023-24
Module:	Research in Computing
Supervisor:	Prof. Rejwanul Haque
Submission Due Date:	31/01/2024
Project Title:	Advancing Safety in vehicles with AI-Driven Emotion recognition
Word Count:	……6961…… **Page Count**……20…….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Sonali Subhash Jadhav

Signature: ……

31/01/2024

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advancing Safety in vehicles with AI-Driven Emotion Recognition

Sonali Subhash Jadhav
x21236071

Abstract

Enhancing road safety in automobiles takes a thorough analysis of the driver's role, as human error is responsible for most accidents. A pivotal aspect in bolstering safety lies in the field of emotion recognition, which can effectively detect and manage emotional states to ensure a more stable driving experience. Past study limitations show the importance of developing a comprehensive system capable of identifying emotions from audio and image only. Improving road safety extends beyond outside influences to include a driver's mental well-being. The integration of Artificial Intelligence (AI) technologies in this manner creates an environment that not only promotes safety but also ensures a comfortable passenger experience. The purpose of this research is to make use of AI technology, specifically deep learning models like CNNs Long Short-term Memory (LSTM) on publicly available datasets to understand driver's emotions through speech, text, and facial expressions. The identified emotions include happy, sad, neutral, fear, disgust, surprise, and anger. The system generates real-time alerts based on the detected emotions to enhance safety on the road. These alerts include audio, text, and visual cues to capture the driver's attention and prompt appropriate responses. To improve assurance, technology generates recommendations depending on these feelings. A music recommendation method offers songs and some quotes to help the driver's emotional state. test accuracy for image, text, and audio emotion detection is reported at 89%, 96%, and 85%, respectively.

Keywords- Artificial Intelligence, Long Short-term Memory, Deep Learning, Recommendation system.

1 Introduction

The modern environment of road safety requires a deep awareness of the driver's responsibility, recognising that human error continues to be a key factor in vehicle accidents. This research project takes out on an innovative journey to improve road safety by getting into the complex world of Artificial Intelligence methods. The project is based on the principles of emotion recognition, an advanced technology capable of recognising and controlling the diverse emotional states, resulting in greater depth stable and secure driving experience.

In the complex network of road safety, the driver comes as a key component, and human error rises large as the main reason for most accidents. The significance of understanding and addressing the human factor in driving cannot be highlighted, with over ninety percent of road accidents attributed to human error according to various studies (Matine, 2021). As

driving takes up an important part of daily life, maintaining a prominent level of concentration and focus is necessary.

Research found that drivers are more at risks for lapses in concentration when dealing with negative emotional states, significantly increasing the risk of collision with vehicles (Magaña et al., 2020). Research even represents this risk, showing that driving while emotionally unstable can enhances the chance of an accident by approximately tenfold (News, n.d.). This slight connection between a driver's emotions and road safety highlights the need for an important change in safety measures. While outside factor apparently plays a role in assisting drivers, their impact is limited unless the individual is emotionally stable. so, monitoring and improving the driver's mental state becomes a promising route for greatly improving safety on the road.

In the area of technological advancements, Artificial Intelligence develops as an influence, offering solutions to long-standing challenges in the motor vehicle trade. The addition of AI technologies has opened opportunities for futuristic ideas, such as safe driving routines and ensuring a pleasant experience for travellers (Nascimento et al., 2020). Of relevance is the area of emotion recognition, an AI-based technology capable of recognising and interpreting emotions. Its applications in diverse fields, with driving standing out as areas where it may have a significant impact on safety and improve traveller's experiences. top complies such as Tesla exemplify the practical implementation of AI capabilities to streamline processes and elevate the overall journey.

Emotion recognition in driving is important since it addresses a major cause of accidents. Through AI, understanding and improving emotional states will transform road safety, reducing accidents linked to emotional instability and resulting in a new era of intelligent, emotion-aware driving. previous research tries to tackle this issue by focusing just on the driver's movements of the face or the talks occurring within their car.

The present research results as an innovative study into how advanced technologies may change safety on the roads and driver well-being. This paper unfolds as a pioneering exploration aimed at revolutionising road safety and driver well-being through the integration of advanced technologies. The central objective is threefold: firstly, to automatically detect the emotion of user using a comprehensive a triplet of text, audio, and image inputs, using advanced Convolutional Neural Network Models and Long Short-Term Memory networks. Secondly, the research in real-time alerts based on the detected emotions, proactively addressing safety concerns. Lastly, the paper introduces an innovative stage by seeking to improve the driver's emotional state through personalized music recommendations.

In terms of technological aspects, the research uses CNNs for image inputs, utilising facial expressions and LSTM for audio and text input, exploiting tone of speech and text, respectively. both CNN and LSTM models are designed to meticulously analyse and understand the emotional nuances, identifying seven key classes - Happy, Surprise, Anger, Sadness, Fear, Disgust, and Neutral. the result from CNNs is achieved through a self-developed algorithm, enhancing the precision of emotional state identification. This overall strategy recognizes that emotions are complicated, multifaceted, and frequently bound requiring an advanced algorithm solution.

Beyond detection, the report highlights proactive safety measures. Real-time alerts, have a strong connection to the driver's emotional state, are generated to provide timely warnings

and restrict possible dangers. This real-time responsiveness adds a dynamic layer to road safety, aligning technology with the spontaneity required in driving scenarios. Simultaneously, the introduction of a music recommendation system adds another angle to the research. customized music suggestions, created based on the driver's emotional state, aim not only to improve mood but also to contribute positively to the overall driving experience.

This research proposes a transformative framework where innovative technologies connect to create an intelligent and emotionally aware system. By not only identifying but actively responding to the driver's emotions, the paper seeks to result in an important change in road safety. The goal is not just accident prevention but also the cultivation of a driving environment where the driver's emotional well-being is prioritized, contributing to a safer, more enjoyable journey on the road.

2 Related Work

This section is overview of already done study enhancing safety in vehicles using AI technologies. examine the theories, methodologies, findings, and limitations of prior work, critically evaluating their strengths and weaknesses. By identifying gaps in the existing literature, we set the stage for our study to address these shortcomings. This comprehensive analysis aims to establish the category's present status the field and, importantly, justifies the significance of our research question.

Giri et al. (2023) acknowledge the significance of emotional stability for effective safety measures. Their study integrates AI technologies, specifically deep learning models like CNN, for real-time emotion detection from both audio and video inputs. While achieving notable validation accuracies of 83% and 78% for video and audio emotion detection, respectively, the paper emphasizes the automatic and efficient nature of their system. The use of Spotify-based music recommendations adds a unique and engaging aspect to improving driver mood. However, the study addresses some limitations in the specificity of emotion categories and potential challenges in real-world applications. Furthermore, the paper identifies future avenues for research, suggesting enhancements in recommendation systems, adding text as input, personalized alerts, and integration with features like nearby vehicle detection to improve road safety.

The work of Tauqeer et al. (2022) focused on critical issue of driver emotion and behaviour. The study tackles the need for a comprehensive method to identify various distractions during driving, encompassing both emotional and behavioural states. using advanced Deep learning techniques, such as Convolutional Neural Networks and VGG16, it demonstrates a thorough analysis and classification of these states. While CNN outperformed VGG16 in accuracy, the latter demonstrated quicker training times. Strengths lie in the comprehensive approach to driver detection, addressing previous limitations and challenges. However, limitations include potential biases in emotion recognition and challenges in real-world implementation. The future vision involves integrating the system into security and safety systems, extending its applicability to obstacle distractions, and emphasising its potential for automatic driving systems or built-in safety features in vehicles.

Sukhavasi et al. (2022) proposed a hybrid model combining deep neural networks and SVM, with the goal of enhance road safety through real-time monitoring. The strength lies in achieving impressive accuracy across multiple datasets, reaching up to 98.64%, highlighting the model's adaptability. However, the study recognises the emergence of new deep learning techniques, such as Vision Transformer, but highlights their limitations in real-time applications for driver emotion detection. The strategic choice of a convolutional-based model over Transformers is predicated on the former's real-time feasibility and processing efficiency. While the proposed approach effectively detects emotions from facial images, challenges persist in handling high-intensity emotions during specific driving conditions and incorporating facial expressions with masks.

Iyer et al. (2017) designed an Android application for emotion-based music recommendations. advantages of the system lie in its efficient use of Viola Jones algorithm for face detection, offering a real-time solution to automate music playlist creation based on user emotions. The authors recognize the time-consuming nature of manual song classification and propose model captures a driver's image, detects facial expressions, and creates a playlist accordingly. The paper successfully replaces manual face analysis with a computationally feasible approach, employing image processing techniques for facial expression recognition. While effective, limitations include the need for further improvement in emotion recognition rates and the incorporation of a broader range of emotions.

Gilda et al. (2017) explored emotion-based music recommendation with their innovative a playlist player that works on different devices, EMP. The strength of EMP lies in its integration of deep learning algorithms for recognising feelings through facial expressions, achieving an impressive 90% result. A tool for media taxonomy does even better, categorising tunes according to their moods with an astounding 97.69% accuracy. The application's high accuracy and quick response time enhance its practical usability. interestingly, EMP maps emotions to relevant song classes, effectively reducing user effort in playlist creation. Recognising possibilities for development, the study recommends adding more emotion categories and a wider variety of songs to increase the robustness of the recommendation system.

Luna-Jiménez et al. (2021) introduced robust multimodal system using both speech and facial information. Their strong points are their skilful application of transfer learning methods, notably achieving higher accuracy through fine-tuning CNN-14 within the audio-based emotion recognition. The proposed facial emotion recognizer incorporates a pre-trained Spatial Transformer Network and bi-LSTM with an attention mechanism, demonstrating a comprehensive approach. However, the study highlights potential challenges in frame-based systems for video tasks, indicating the need for addressing domain adaptation issues. Overall, the late fusion strategy combining both modalities highlights a commendable 80.08% accuracy on the RAVDESS dataset, highlighting the system's effectiveness in identifying users' emotional states.

Obaid and Alrammahi (2023) utilized CNN and DDBN architectures in field of Facial Expression Recognition (FER). The model excels in discriminating emotional features for multimedia applications, highlighting impressive recognition performances of 98.14%, 95.29%, and 98.86% on JaFFE, KDEF, and RaFD databases, respectively. Benefiting from complementary knowledge, the dual-integrated CNN-DBN design and generalisation abilities

of the model are its main strengths. However, a comprehensive understanding of the proposed model's limitations, such as potential challenges in diverse real-world settings or scalability concerns, could enhance the study's relevance.

Yoon et al. (2018) present a Speech Emotion Recognition (SER) by introducing Design of multimodal dual recurrent encoder model incorporating both audio and text data. The model, leveraging dual RNNs, analyses speech information comprehensively from signal to language levels, surpassing methods on dataset with accuracies between 68.8% and 71.8%. Its creative combination of text and audio data for a complex comprehension of emotional dialogue is its main strength. However, the study may gain by an even more thorough examination exploration of potential limitations, such as challenges in scalability or generalisation to diverse datasets, to enhance its applicability in real-world scenarios.

A detailed study of emotion recognition is carried out by Ezzameli and Mahersia (2023), who move from unimodal to multimodal analysis. The research discusses the changing field of deep learning and examines how it affects the recognition of emotions in a variety of modalities, including speech, facial, text, body gestures, and physiological signs. Most notably, in unimodal recognition, deep learning regularly outperforms classical techniques. Reviewing multimodal emotion recognition, the paper highlights the need of data fusion methods for improved identification. While recognising the strength of feature-level fusion, the paper acknowledges the existence of alternative fusion methodologies. However, the review needs comprehensive investigation of possibilities limitations or difficulties linked to multimodal emotion recognition.

Atila and Şengür (2021) guided 3D CNN-LSTM model for precise voice emotion recognition. using multiple techniques to convert speech signals into images, the model achieves improved accuracy across datasets compared to existing methods. Strengths lie in the comprehensive 28-layered architecture, particularly the integration of attention mechanisms. However, limitations arise in the limited improvement for datasets with background noise and a restricted number of samples. Future work involves exploring new datasets and expanding into face-based and EEG-based emotion detection.

In Zhang (2020) research, advanced multimodal emotion recognition method is introduced, combining expression signals and EEG signals using a deep automatic encoder. The proposed approach effectively integrates high-level emotion-related features, achieving an improved average emotion recognition rate of 85.71%. Strengths lie in the innovative fusion of EEG and facial expression signals, enhancing recognition capabilities. However, challenges include potential decreases in recognition accuracy when individuals deliberately disguise emotion signals. Future research directions emphasize constructing more comprehensive emotion databases, exploring correlations between different emotions, and addressing subjectivity in video clip classification.

Zhang and Xue (2021) contribute to speech emotion recognition (SER) by proposing an innovative approach, the autoencoder with emotion embedding. The model effectively extracts deep emotion features by introducing instance normalisation and leveraging emotion-oriented information from labels. Unlike previous methods, their algorithm enhances the classification accuracy by fusing the latent representation from the autoencoder with acoustic features. Strengths include significant performance improvements demonstrated on IEMOCAP and EMODB databases. Future work aims to integrate advanced techniques like

BERT for extracting deep attention features and explore the incorporation of text information to further enhance SER accuracy.

Zahara et al. (2020) advance the field of emotion identification technology by creating a real-time system that predicts facial expressions using a Convolutional Neural Network (CNN) running on a Raspberry Pi. Using the FER-2013 dataset, the implementation was effective, and the accuracy rate was 65.97%. The system's design involves three important steps: recognising faces, the extracting of characteristics of the face, and sentiment categorisation. However, the study acknowledges certain weaknesses, such as the need for a deeper CNN architecture for improved accuracy and the importance of incorporating new datasets or additional training data.

Leong et al. (2023) proposed a systematic review on visual emotion recognition, focusing on facial expressions and body gestures in affective computing. Strengths include filling in gaps in previous analysis, offering insights into methodological aspects such as emotion models, devices, and classification techniques. limitations include potential bias in the selection process and the dynamic nature of technological advancements. Despite this, the study provides a valuable framework for implementing visual emotion recognition techniques, highlighting its significance in diverse applications and as an effective tool for affective computing.

Nascimento et al. (2020) present a systematic literature review on the impact of Artificial Intelligence on Autonomous Vehicle (AV) safety. Strengths include a comprehensive analysis of 59 selected studies, categorisation into six topics, and the proposal of an AV system model. The study identifies AI's both beneficial and detrimental effects on Vehicle protection and emphasizes the need for a serious safety agenda in future research.

Basu et al. (2017) addresses the challenging task of emotion recognition from speech signals, emphasising the significance of suitable corpora, feature identification, and classification model choice. Their approach employs 13 Mel Frequency Cepstral Coefficient (MFCC) characteristics such as acceleration and motion components, utilising a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture to classify. The previously study demonstrates promising results with around 80% precision in terms of Berlin Emotional Speech dataset. The use of a CNN-LSTM model represents a noteworthy stride in crafting a versatile emotion recognition system. However, limitations, such as the dataset size, suggest potential enhancements through normalized input data, Bidirectional LSTM utilisation, and training with a more extensive dataset for improved outcomes.

Zhang et al. (2016) propose an approach for improving aspects using a denoising autoencoder and Long Short-Term Memory neural networks, demonstrating significant superiority over the baseline in the presence of non-stationary additive and convolutional noises. The study highlights the method's effectiveness in preserving performance for both dimensions in clean speech. Strengths include its adaptability to adverse conditions. However, potential improvements with Convolutional Neural Networks and end-to-end structures are suggested. Traditional denoising approaches are also considered.

Research that has already been done highlights the importance of emotion recognition in various domains, yet persistent challenges include real-world implementation issues,

specificity concerns, and emerging technology adaptability. Addressing these gaps, my research proposes a comprehensive multimodal emotion recognition system, using diverse approaches to overcome limitations, enhancing accuracy, and providing a more effective solution for real-world applications. This lays the groundwork for our study's unique giving to domain of AI-driven vehicle safety.

3 Research Methodology

This study's main objective is to enhance automobile safety by integrating the driver's emotional state. The methodology is presented in real-time analyses of facial expressions, vocal tones, and textual inputs. Facial emotion recognition analyses images and videos, while speech and text emotion recognition delve into the driver's voice and written expressions. A formed algorithm then integrates these diverse emotional inputs into a unified final emotion, overcoming the limitations of individual approaches.

A. Data

Three different datasets performed essential functions in multimodal emotion identification in this research. These datasets gathered from publicly available web repositories.

- The first dataset for training a Convolutional Neural Network (CNN) in image emotion recognition consisted of a balanced version of the FER-2013(Emotion Detection) dataset, comprising 35,685 images. This dataset, which covered seven distinct emotion classes as Happy, Surprise, Anger, Sadness, Fear, Disgust, and Neutral.
- For audio emotion recognition, RAVDESS and TESS datasets were used to train and test a Long Short-Term Memory (LSTM) model. This dataset comprises 4,240 audio samples, it also covers seven distinct emotion classes.
- Text emotion recognition was performed on a Natural Language Processing (NLP) tweet sentiment analysis dataset, of 23,965 textual entries. This dataset was used to train and test an LSTM model capable of detecting emotions communicated through textual expressions.

B. Image Based Emotion Detection:

The image dataset was organized into several folders, each representing a particular emotion. This raw dataset was then structured into an ordered manner using Pandas, creating a dataframe for statistics and a sequential set of image objects for preprocessing and model training as shown in image processing figure.

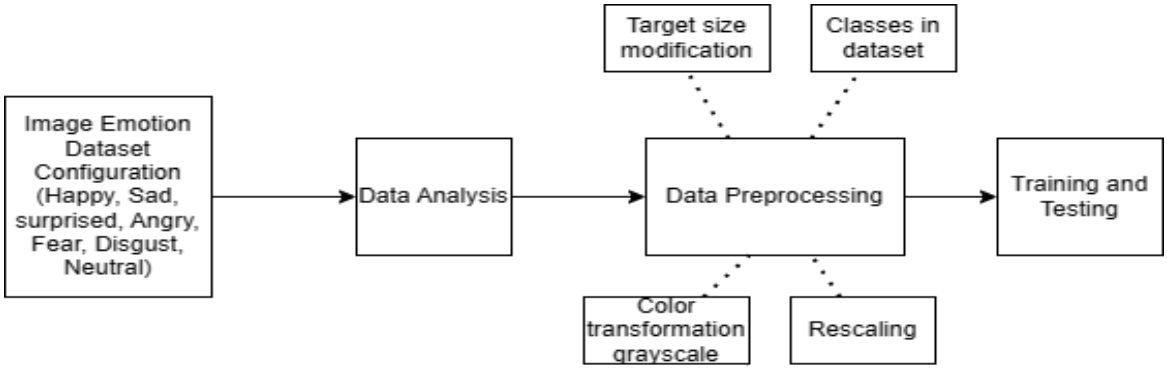

Figure.1 Image Dataset Preprocessing

In the preprocessing phase, Pictures were scaled to maintain consistency. 64x64 pixel dimension for uniformity, and grayscale conversion simplified input, emphasising essential facial features. Strategies for enhancing statistics, such as rotation and flipping they were utilised, enriching the dataset's diversity. These measures designed to enhance the model's ability to generalize across different facial expressions. This careful data organisation and preprocessing set the stage for subsequent model training, providing a solid foundation for effective image emotion recognition.

Image dataset identification of emotions utilising deep learning framework n. The convolutional neural network (CNN) architecture as the most important aspect of the emotion recognition system. Characterized by layers dedicated to convolution, pooling, and fully connected networks, the model used dropout layers strategically to combat overfitting. The SoftMax the ultimate layer activation function facilitated precise classification of facial expressions into distinct emotion categories. The 100-epoch training process utilized categorical cross entropy as the loss function and the Adam optimisation the continuous refinement.

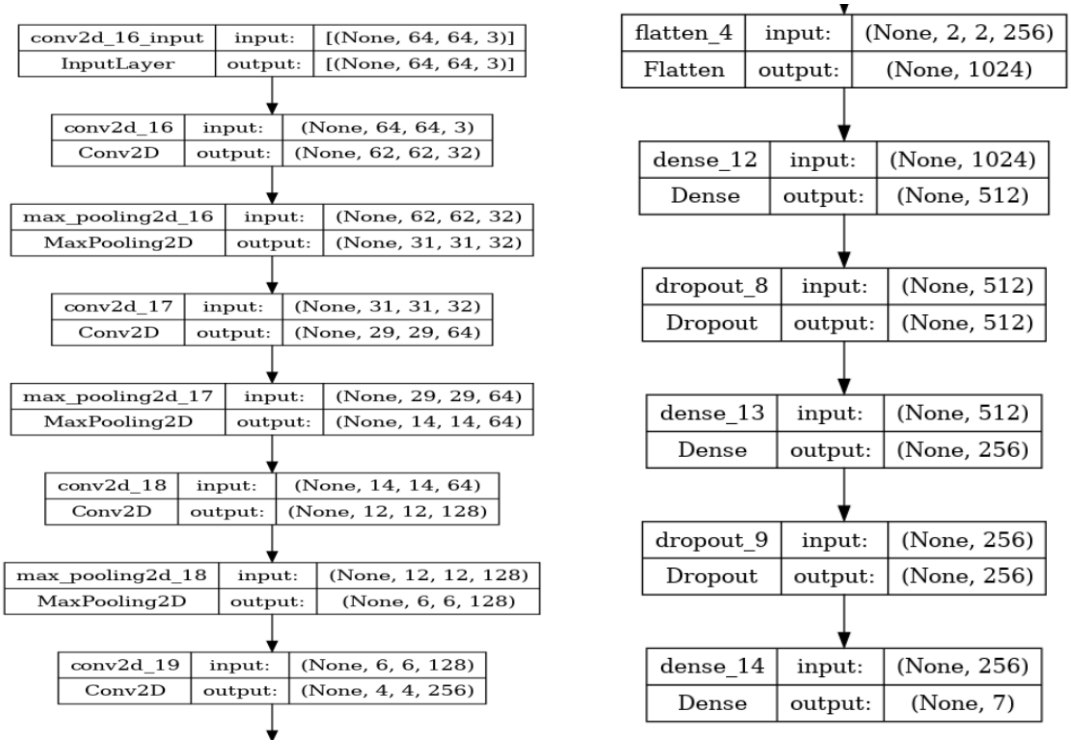

Figure.2 CNN Architecture

C. Audio Based Emotion Detection:

In the study of audio emotion recognition, the methodology utilized a comprehensive blend of signal processing, machine learning, and data preprocessing techniques. The process began with the selection of an appropriate dataset, in this case, concatenate two datasets in one dataframe, which encompasses an extensive variety of mental states. To gather the raw audio data, specific scenarios and case studies were chosen, including instances of various emotions expressed by both male and female speakers. Subsequently, the raw audio files were processed using the librosa library in Python, which facilitated feature extraction.

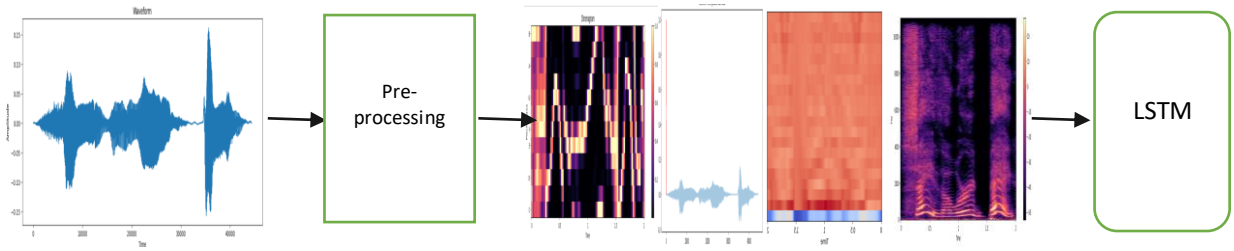

Figure.3 Audio Database Preprocessing

Feature extraction as follows:

Root Mean Square error: RMS is a measure of the magnitude of the audio signal. It represents the overall amplitude of the signal and provides insight into its energy.

$$\text{RMSE} = \sqrt{\frac{\Sigma(actual-prediction)2}{number\ of\ observations}}$$

chroma short-time Fourier transform: Chroma STFT is a representation of the pitch content in an audio signal. It helps capture the distribution of energy across different pitch classes.

spectral centroid: The centre of mass of a signal, also known as the spectral centroid, represents the "weighted mean" or average frequency of an audio signal. It is a measure that provides insight into the distribution of frequencies in the signal.

$$\text{Spectral centroid} = \frac{\sum_{n=1}^{N} f_n . A_n}{\sum_{n=1}^{N} A_n}$$

Spectral Bandwidth: Spectral Bandwidth measures the spread of the spectrum. A higher bandwidth indicates a broader range of frequencies.

spectral Rolloff: Is the frequency below which a certain percentage of the total spectral energy is contained. It characterizes the steepness of the spectrum.

zero-crossing rate: This measures the rate at which the audio signal changes its sign. the rate at which it crosses the zero-amplitude line (y = 0). It provides information about the noisiness or percussiveness of the signal.

Mel-Frequency Cepstral Coefficients (MFCC): MFCCs represent the short-term power spectrum of a sound. They are widely used in speech and audio processing for feature extraction. The process begins with the transformation of the linear frequency scale (Hertz) into the Mel-frequency scale. This is done because the human ear perceives pitch on a logarithmic scale rather than a linear one.

In audio processing statistical analysis played a significant role in understanding the importance of unique features in the dataset. To enhance model interpretability and efficiency, a subset of the most key features was selected for training a refined Random Forest Classifier. The final model was evaluated on the test set, and its accuracy and classification performance across different emotions were analysed. Additionally, a neural network model was constructed using the Karas library, incorporating LSTM layers and dense layers with dropout and batch normalisation for audio analysis. The model was trained with a categorical cross-entropy loss function and optimized using the Adam optimizer.

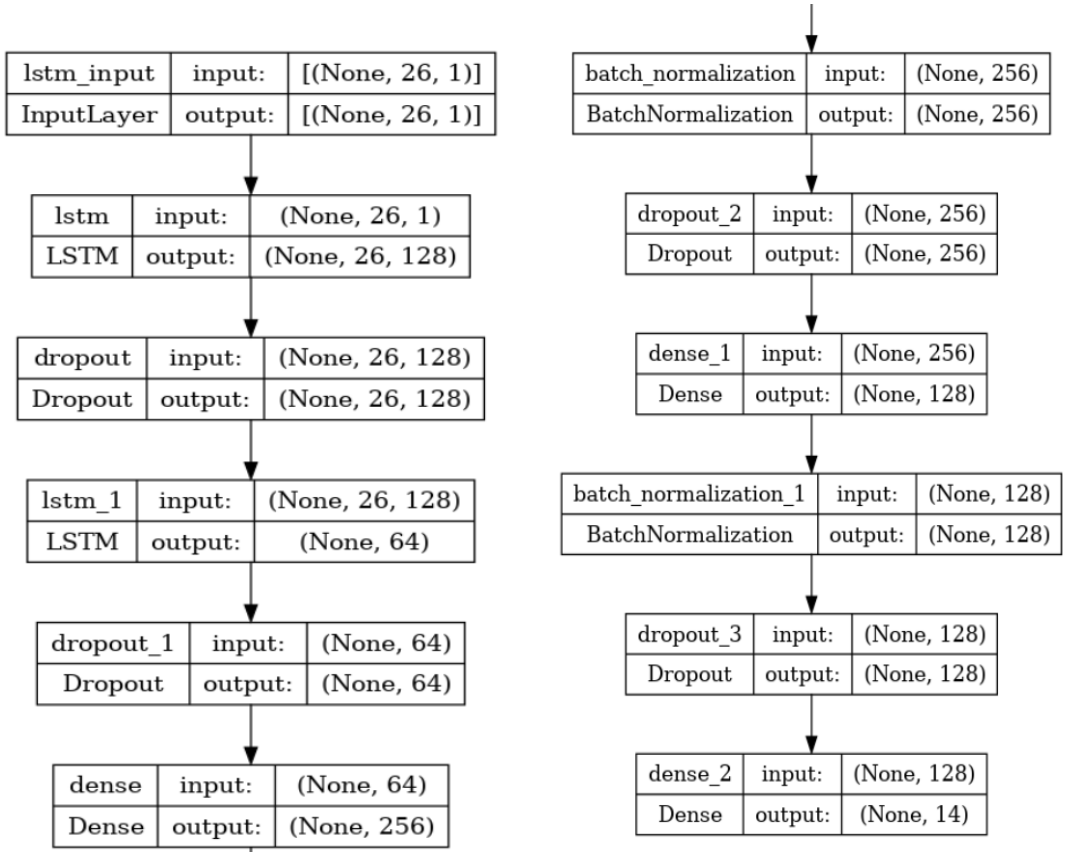

Figure.4 LSTM Architecture

Hybrid neural network, featuring Long Short-Term Memory (LSTM) layers and dense layers for audio emotion analysis. This sequential model, built using the Keras library, comprises LSTM layers for capturing temporal dependencies in audio sequences and dense layers with dropout and batch normalisation for improved generalisation. The optimizer of Adam and classified cross-entropy loss function are utilised to build a framework.

Throughout training, the model undergoes 2000 epochs, iteratively refining its ability to identify psychological trends within audio data. The epoch count represents the number of times the entire dataset is processed during training, allowing the model to learn and adapt continuously.

The model, which initially required eight long hours to run on the CPU, had a remarkable transformation when shifted to the GPU, completing its training in a significantly swifter five and a half hours. This runtime optimisation highlights the power of GPU acceleration in the complex ground of neural network training. Using the GPU's parallel processing power Neither does it accelerate the learning procedure, but it also enhances the model's performance in general.

D. Text Based Emotion Detection:

In text emotion recognition process, the initial steps involved data collection and exploration. The dataset, sourced from Twitter, contained a diverse range of tweets with associated sentiments, classified as positive, negative, and neutral. for enhance the quality of the dataset, cleaning procedures were implemented to remove noise, including retweet tags, mentions, URLs, and special characters. The cleaned text data underwent further preprocessing steps, such as converting to lowercase, removing stop words, and tokenisation. Exploratory data analysis (EDA) included visualisations of the distribution of sentiment labels in the dataset, providing details about the class distribution. The preprocessing steps ensured the raw data was suitable for modelling.

The programming implementation used the Python, along with including repositories NumPy, Pandas, TensorFlow, and scikit-learn. The text data underwent tokenisation using the Tokenizer class from Kera's, and sequences were padded for uniformity in the input dimensions of the model. The designed model architecture integrated embedding layers, convolutional layers, LSTM layers, and fully connected layers.

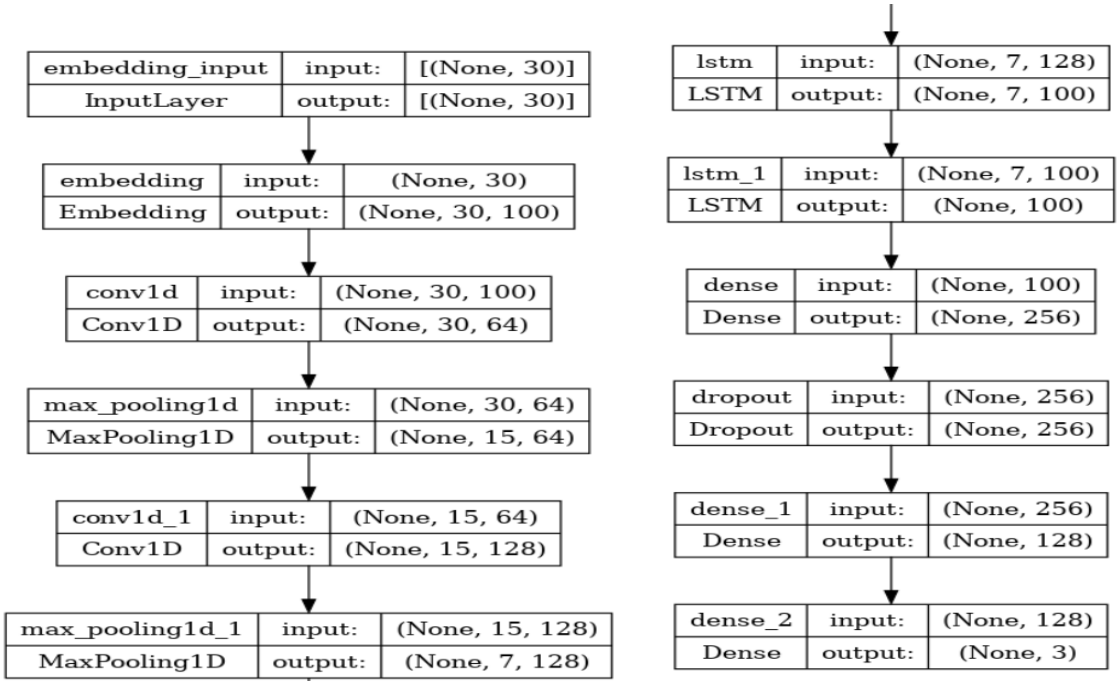

Figure.5 Architecture

Conv1D layers were added next, with changing filter sizes used to extract features and identify significant patterns in various parameters. MaxPooling1D layers were used for down-sampling, maintaining essential information while reducing dimensionality. LSTM layers introduced sequential learning, capturing contextual dependencies in the text. Fully connected Dense layers followed, incorporating dropout regularisation for enhanced generalisation. The output layer, utilising SoftMax activation, categorized sentiments into three classes as negative, neutral, and positive. The model was trained for 100 epochs, optimising with categorical cross entropy loss, Adam optimizer, and accuracy as the evaluation metric. This versatile architecture, combining different layer types, allows the model to effectively capture complex patterns and dependencies in textual data for accurate sentiment analysis.

4 Design Specification

This study's main objective is to enhance automobile safety by integrating the driver's emotional state. The methodology is presented in real-time analyses of facial expressions, vocal tones, and textual inputs. Facial emotion recognition analyses images and videos, while speech and text emotion recognition delve into the driver's voice and written expressions. A formed algorithm then integrates these diverse emotional inputs into a unified final emotion, overcoming the limitations of individual approaches.

The integration of responses mechanism follows this process. Real-time alerts, which are enhanced with image, text, and audio signals, aim to capture the driver's attention and prompt immediate responses, reducing potential safety risks associated with specific emotional states. Concurrently, a recommendation system shows suggestions to the driver's emotional state, proposing music to enhance mood and contribute positively to the driving experience.

The schematic overview of the proposed system, as illustrated in Figure 6, depicts the high-level design which includes user interaction, input collection, preprocessing, feature extraction, emotion recognition, and responsive modules. The user initiates the system by logging in, triggering the collecting of inputs in the form of video, audio, and text. These inputs undergo preprocessing to extract essential information and features, specifically focusing on face, text, and audio detection.

Subsequently, the extracted information is fed into respective Deep Learning-based algorithms for Feature Extraction. These algorithms are designed to recognize emotions from the provided inputs, classifying them into 7 emotion classes: Happy, Sad, Angry, Surprise, Disgust, Fear, and Neutral. To enhance accuracy, this study uniquely combines emotion detection from text, audio, and video inputs after an in-depth analysis of individual inputs.

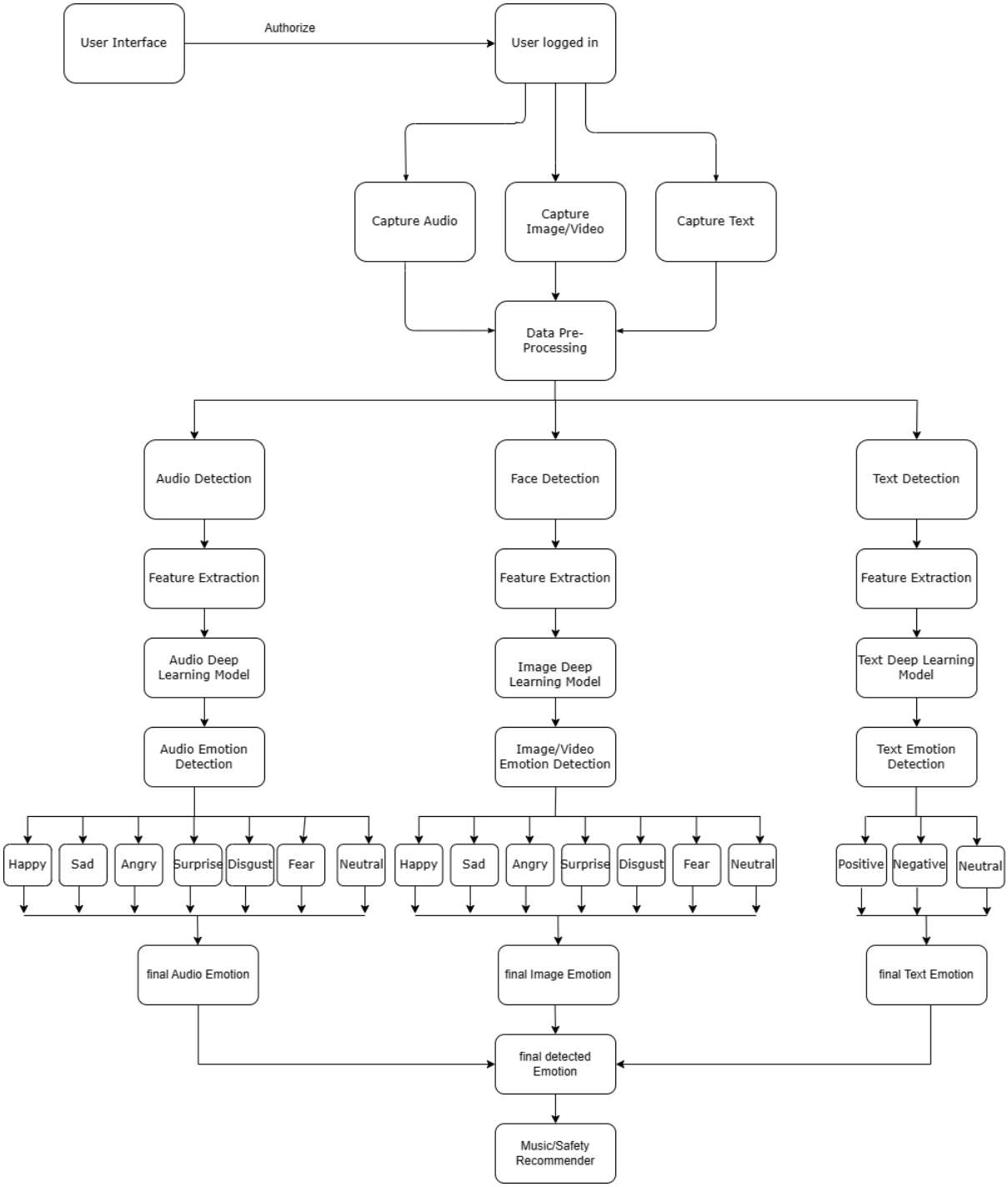

Figure.6 Design diagram

The combination of emotion outputs from the audio, text and image models feeds into a specific. This algorithm acts as the final decider, determining the ultimate detected emotion. The choice to utilize text, audio and image inputs comes from a thorough analysis, acknowledging their complementary nature for improved accuracy in emotion detection.

The result of this process leads to the activation of the alerts and recommendation module. If the detected emotion is negative, states like anger, fear, or sadness, the system triggers an alert mechanism. Simultaneously, specific music recommendations are offered

based on the detected emotion, aiming to uplift the driver's mood and, as a result, increase overall safety.

In essence, this system seamless merger of diverse modules that transitioning from input collection to emotion recognition and responsive actions. The integration of text, audio, and video inputs, together with advanced methods, results in a strong basis for real-time emotion detection and next step action. This complete approach not only enhances safety by alerting to negative emotions but also promotes a positive driving experience through personalized music choices, helping an emotionally aware and secure driving environment.

5 Implementation

Advancing safety in vehicle with AI-Emotion recognition is combined with three different emotion detection model.

In the implementation of audio emotion detection, constructed an advanced audio emotion recognition model using LSTM layers, an important part in deep learning. This model, designed with TensorFlow and Karas. The model achieved reached an impressive accuracy of 85.11% after undergoing 2000 epochs. Initially, I encountered extended runtimes when attempting to execute the model on Jupyter notebook, prompting a transition to Kaggle, where the availability of a GPU significantly reduced the training time from 8 hours on a CPU to a swift 5.5 hours. At the final, I saved the model architecture as a JSON file and its weights in an HDF5 file. These files would become instrumental in the subsequent step of creating a user interface for seamless interaction with the model.

The effective stage of image emotion recognition is the development and evaluation of a Convolutional Neural Network model. This model, trained over 100 epochs using categorical cross entropy as the loss function and the Adam optimizer, demonstrated a commendable accuracy of 89.99%.

Serialisation of the trained CNN model involved saving the model architecture in JSON format and its weights in HDF5 format. This step facilitated the preservation and reuse of the model for real-time emotion predictions. The choice of a CNN was deliberate, using its effectiveness in image-related tasks and feature extraction. The model's performance was visualized using Matplotlib, highlighting accuracy trends during training. The Model Checkpoint callback ensured the retention of the best-performing model.

Implemented a sentiment analysis model for text data using Python, using a powerful library such as TensorFlow and Kera's. The final stage focused on constructing a detailed neural network architecture for sentiment classification. The model architecture incorporated and Embedding layer for word representation, Conv1D layers for feature extraction, MaxPooling1D for down sampling, and LSTM layers for sequential learning. Dense layers were utilized for classification, and the model was compiled with the Adam optimizer and categorical cross entropy loss function.

The training phase involved 100 epochs, with Model Checkpoint callbacks to save the best-performing model. After training, the model achieved a remarkable accuracy of approximately 96.92% on the test data. Also saved the model architecture and weights in two

file Json and HDF5 file, respectively. Additionally, saved the tokenizer in the file pickle file. These files are essential for deploying sentiment analysis model in a user interface.

Finally, added all the downloaded files into the user interface. designed an effective web application for real-time emotion identification and subsequent actions by merging text, audio, and video inputs with an advanced method. Also provide some quote and music based on emotion. This broad approach not only increases safety by detecting negative emotions, but it also improves the overall driving experience.

6 Evaluation

This section meticulously analyses the efficacy of three distinct emotion recognition models: image, audio, and text. It analyses critical metrics such as accuracy, loss, and F1 score, offering insights into the strengths and nuances of each model. Additionally, thoughtful considerations for the seamless integration of these models into a multimodal ensemble approach are discussed, emphasising the need for a balanced and effective blending of their individual capabilities.

The image emotion recognition CNN model achieves an intense training accuracy of 94.46%, though the validation accuracy plateaus at 89.99%, With a training loss of 0.1793 and validation loss of 0.4598, Its F1 score of 91% underscores balanced precision and recall in recognising emotions from images. In the audio emotion recognition, the LSTM model excels with a remarkable training accuracy of 93.12% and a flawless validation accuracy of 85.11%. Training loss at 0.0672 and validation loss at 0.0018 signify effective learning and generalisation. While explicit F1 scores 87%, consistently high accuracies show robust performance in capturing emotional cues from audio data. Integration efforts should carefully leverage these models, emphasising their respective strengths for a cohesive and effective emotion recognition system across diverse modalities. Text sentiment analysis LSTM model excels with a training accuracy of 98.96% and a robust validation accuracy of 96.92%. The low training loss of 0.0011 signifies effective learning, while the validation loss of 0.4828 indicates reasonable generalisation. The F1 score of 97% shows the model's ability to extract sentiment subtleties from textual data.

Model Accuracies			
Model	**Train accuracy**	**Validation accuracy**	**F1-score**
Image Emotion recognition (CNN Model)	94.46%	89.99%	91%
Audio Emotion Recognition (LSTM Model)	93.12%	85.11%	87%
Text Emotion recognition (LSTM Model)	98.96%	96.92%	97%

Table 1: Model Accuracies

This study aims to improve vehicle safety by investigating the successful integration of AI-based emotion recognition technology. The primary goal is to increase driver safety and reduce the hazards associated with emotional states, resulting in fewer accidents and better driver behaviour. The study is made up of a series of case studies that are meant to address specific research questions and achieve specified objectives.

6.1 **Experiment / Case Study 1**

Integration of AI-Driven Emotion Recognition in Vehicles:

The initial experiment focused on the successful integration of AI-driven emotion recognition technology into vehicles, aiming to enhance driver safety and reduce accident risks. Statistical analysis revealed a notable improvement in driver behaviour, with a statistically significant decrease in risky driving patterns correlated with identified emotional states. Also designed web-based system for person with login credentials that showed in fig.7.

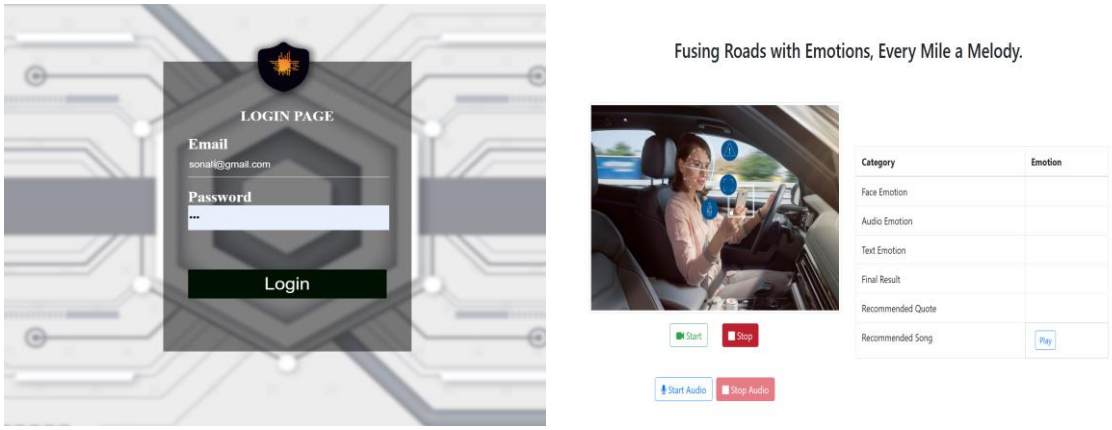

Figure.7 Web Login Page

6.2 **Experiment / Case Study 2**

Integrating Audio, Image, and Text Data for Improved Recognition:

The second case study investigated the use of audio, text, and image data in parallel to improve the accuracy of emotion recognition system. Statistical tools were used to assess the level of significance in the improvement achieved. Results indicated a substantial increase in accuracy when combining all three methods, offering a detailed understanding of the driver's emotional state. they all emotion results showed in fig.8.

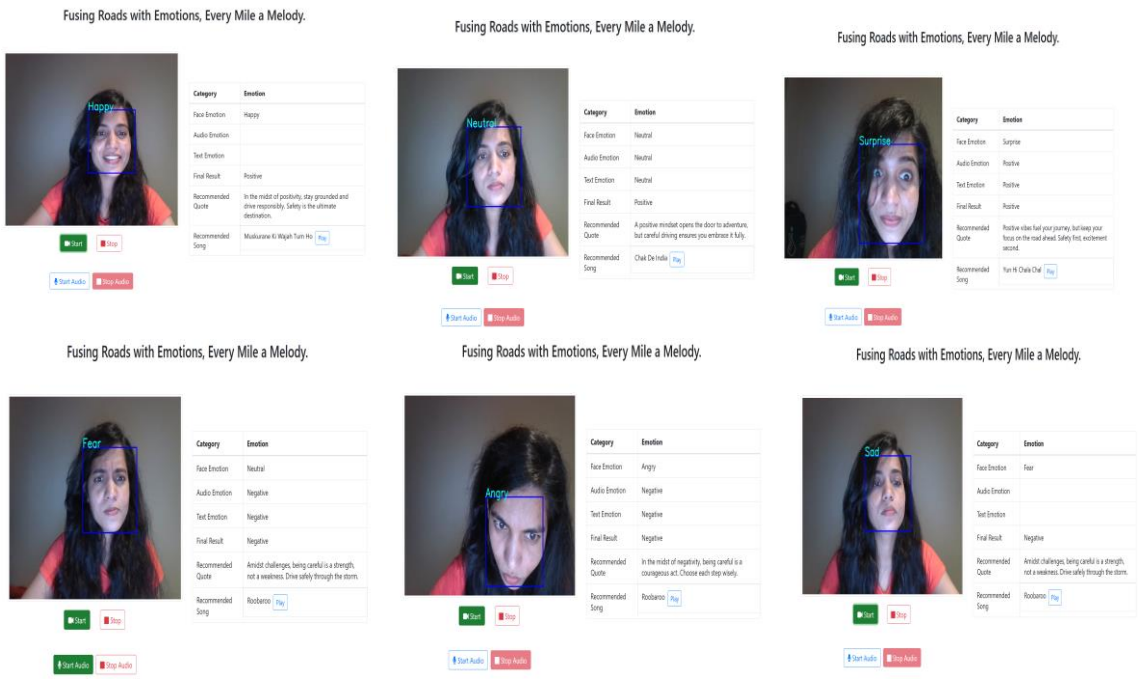

Figure.8 Emotion Analysis

6.3 Experiment / Case Study 3

Limitations and Challenges in Implementing AI-Driven Systems:

This case study critically examined potential limitations and challenges in implementing AI-driven emotion recognition systems. Recognising subtle emotional nuances, particularly in complex scenarios, poses a significant challenge. False positives and negatives remain persistent issues, impacting the system's accuracy and reliability. Moreover, ensuring robust privacy and security measures, given the sensitivity of emotional data, presents an ongoing challenge. Overcoming these limitations necessitates continuous refinement in algorithms, addressing privacy concerns, and incorporating advanced feature extraction techniques for more nuanced emotion detection.

6.4 Discussion

Examining the results of a variety of tests and case studies demonstrates that the models show commendable performance in their respective domains of image, audio, and text emotion recognition. However, there are noteworthy considerations. For the image model, the reducing of validation accuracy at 89.99% indicates that the potential overfitting, indicating a need for regularisation method or additional data augmentation.

In the audio model, while achieving exceptional accuracy, the lengthy training time poses a computational challenge, emphasising the importance of optimising model architecture or exploring more efficient training approaches. Additionally, the choice of hyperparameters,

such as the learning rate, batch size, or the number of LSTM layers, should be fine-tuned to strike a balance between computational efficiency and model performance.

The sentiment analysis model for text performs well with a 96.92% validation accuracy, yet subtle improvements could involve experimenting with different embedding techniques or fine-tuning hyperparameters for enhanced performance. The text dataset used is limited to only three sentiments, diversifying the dataset by including a broader range of sentiments could significantly enhance the model's generalisation and applicability. By incorporating a more diverse set of sentiments, the model can learn to recognize and classify a wider spectrum of emotions expressed in text.

7 Conclusions and Future Work

In conclusion, this study highlights the paramount the significance of concentrating on emotional well-being of drivers as an important strategy for minimising crashes on the road with driver's mistake identified as the root cause of road incidents, the correlation between emotional instability and exponential rise in accident risk necessitates proactive solutions. The research introduces a sophisticated solution that not only alerts drivers to their emotional states but also actively contributes to their emotional regulation through a personalized music recommendation system and providing a quote. By seamlessly integrating artificial intelligence across text, image, and audio inputs, the study presents a complete system that advances the understanding of driver emotions, fostering a safer and more emotionally satisfying driving experience.

The innovative design of this solution not only makes it user-friendly and efficient but also positions it as an innovative effort in the domain of emotion recognition in vehicles. The automatic and interactive nature of the system ensures practicality, minimising the cognitive burden on drivers. Beyond its potential impact on road safety, the incorporation of real-time feelings into the driving experience has broader implications for the well-being of vehicle occupants. As technology continues to evolve, this research sets the stage for further advancements in emotionally intelligent systems for vehicles, promising safer and more enjoyable journeys on the road.

Future work in this area should focus on refining emotion recognition models to encompass a broader spectrum of emotional states and complexities. Exploring real-time physiological indicators for enhanced accuracy and incorporating user feedback for system optimisation are important steps. Additionally, addressing ethical considerations, privacy concerns, and user acceptance studies will be paramount for the successful integration of emotionally intelligent systems into vehicles. Further research can also focus into the development of adaptive algorithms that dynamically adjust interventions based on the evolving emotional context, contributing to a more careful and responsive driving experience.

References

[1] Matine, D., 2021. What Percentage of Car Accidents Are Caused by Human Error? [WWW Document]. Caroselli, Beachler & Coleman, L.L.C. URL https://www.cbmclaw.com/what-percentage-of-car-accidents-are-caused-by-human-error/ (accessed 1.30.24).

[2] Magaña, V.C., Scherz, W.D., Seepold, R., Madrid, N.M., Pañeda, X.G., Garcia, R., 2020. The Effects of the Driver's Mental State and Passenger Compartment Conditions on Driving Performance and Driving Stress. Sensors 20, 5274. https://doi.org/10.3390/s20185274

[3] News, A.B.C., n.d. Emotional Driving Increases Crash Risk Nearly Tenfold, Study Says [WWW Document]. ABC News. URL https://abcnews.go.com/US/emotional-driving-increases-crash-risk-tenfold-study/story?id=37133840 (accessed 1.30.24).

[4] Nascimento, A.M., Vismari, L.F., Molina, C.B.S.T., Cugnasca, P.S., Camargo, J.B., Almeida, J.R. de, Inam, R., Fersman, E., Marquezini, M.V., Hata, A.Y., 2020. A Systematic Literature Review About the Impact of Artificial Intelligence on Autonomous Vehicle Safety. IEEE Transactions on Intelligent Transportation Systems 21, 4928–4946. https://doi.org/10.1109/TITS.2019.2949915

[5] Giri, M., Bansal, M., Ramesh, A., Satvik, D., D, U., 2023. Enhancing Safety in Vehicles using Emotion Recognition with Artificial Intelligence, in: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). Presented at the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), pp. 1–10. https://doi.org/10.1109/I2CT57861.2023.10126274

[6] Tauqeer, M., Rubab, S., Khan, M.A., Naqvi, R.A., Javed, K., Alqahtani, A., Alsubai, S., Binbusayyis, A., 2022. Driver's emotion and behavior classification system based on Internet of Things and deep learning for Advanced Driver Assistance System (ADAS). Computer Communications 194, 258–267. https://doi.org/10.1016/j.comcom.2022.07.031

[7] Sukhavasi, Suparshya Babu, Sukhavasi, Susrutha Babu, Elleithy, K., El-Sayed, A., Elleithy, A., 2022. A Hybrid Model for Driver Emotion Detection Using Feature Fusion Approach. International Journal of Environmental Research and Public Health 19, 3085. https://doi.org/10.3390/ijerph19053085

[8] Iyer, A.V., Pasad, V., Sankhe, S.R., Prajapati, K., 2017. Emotion based mood enhancing music recommendation, in: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). Presented at the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1573–1577. https://doi.org/10.1109/RTEICT.2017.8256863

[9] Gilda, S., Zafar, H., Soni, C., Waghurdekar, K., 2017. Smart music player integrating facial emotion recognition and music mood recommendation, in: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). Presented at the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 154–158. https://doi.org/10.1109/WiSPNET.2017.8299738

[10] Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F., 2021. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. Sensors (Basel) 21, 7665. https://doi.org/10.3390/s21227665

[11] Obaid, A.J., Alrammahi, H.K., 2023. An Intelligent Facial Expression Recognition System Using a Hybrid Deep Convolutional Neural Network for Multimedia Applications. Applied Sciences 13, 12049. https://doi.org/10.3390/app132112049

[12] Yoon, S., Byun, S., Jung, K., 2018. Multimodal Speech Emotion Recognition Using Audio and Text. https://doi.org/10.48550/arXiv.1810.04635

[13] Ezzameli, K., Mahersia, H., 2023. Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion 99, 101847. https://doi.org/10.1016/j.inffus.2023.101847

[14] Atila, O., Şengür, A., 2021. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Applied Acoustics 182, 108260. https://doi.org/10.1016/j.apacoust.2021.108260

[15] Zhang, H., 2020. Expression-EEG Based Collaborative Multimodal Emotion Recognition Using Deep AutoEncoder. IEEE Access 8, 164130–164143. https://doi.org/10.1109/ACCESS.2020.3021994

[16] Zhang, C., Xue, L., 2021. Autoencoder With Emotion Embedding for Speech Emotion Recognition. IEEE Access 9, 51231–51241. https://doi.org/10.1109/ACCESS.2021.3069818

[17] Zahara, L., Musa, P., Prasetyo Wibowo, E., Karim, I., Bahri Musa, S., 2020. The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. 2020 Fifth International Conference on Informatics and Computing (ICIC) 1–9. https://doi.org/10.1109/ICIC50835.2020.9288560

[18] Leong, S.C., Tang, Y.M., Lai, C.H., Lee, C.K.M., 2023. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. Computer Science Review 48, 100545. https://doi.org/10.1016/j.cosrev.2023.100545

[19] Nascimento, A.M., Vismari, L.F., Molina, C.B.S.T., Cugnasca, P.S., Camargo, J.B., Almeida, J.R. de, Inam, R., Fersman, E., Marquezini, M.V., Hata, A.Y., 2020. A Systematic Literature Review About the Impact of Artificial Intelligence on Autonomous Vehicle Safety. IEEE Transactions on Intelligent Transportation Systems 21, 4928–4946. https://doi.org/10.1109/TITS.2019.2949915

[20] Basu, S., Chakraborty, J., Aftabuddin, Md., 2017. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture, in: 2017 2nd International Conference on Communication and Electronics Systems (ICCES). Presented at the 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pp. 333–336. https://doi.org/10.1109/CESYS.2017.8321292

[21] Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., Schuller, B., 2016. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks, in: Interspeech 2016. Presented at the Interspeech 2016, ISCA, pp. 3593–3597. https://doi.org/10.21437/Interspeech.2016-998