UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

Emotion and Drowsiness Detection System using Multimodal Fusion and Explainable AI

Chitrapu Aruna Sri[1], Mrs. R. Shweta Balkrishna[2], Dr. M Ramjee[3]

[1]MTech Student, Department of CS&SE, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh Email: arunasrichitrapu@gmail.com

[2]Assistant Professor (Ad-hoc), Department of CS&SE, Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh Email: shwetaramteke60@gmail.com

[3]Professor, Department of IT & CA, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh Email: mramjee@gmail.com

***Abstract*—**This paper illustrates the creation of a holistic Real- Time Driver Emotion and Drowsiness Detection System that is supposed to enhance road safety through detection of a driver emotional state and the level of alertness. The application integrates the process of analysis of facial expressions, speech emotion recognition, and drowsiness detection to deliver real-time interventions and make drivers safe and aware of driving in the streets. Based on multimodal AI methods, the system both reads visual and audio information, as well as reads facial expression and speech to identify emotional states and monitors faces to recognize the signs of drowsiness. The system puts more focus on the detection of drowsiness as opposed to emotion recognition to facilitate the provision of safety alerts in time. Also, the Explainable AI (XAI) integration brings transparency in the sense that the users can see how the system makes decisions. This project emphasizes the significance of multimodal fusion, the fact is that both results of facial and speech analysis are used simultaneously to provide better and more objective results, therefore, enhancing the overall safety of drivers. The proposed system can be deployed to be in use in the real-world context with little disruption to the user experience giving the feedback in real-time and ensures safety is monitored throughout driving.

Keywords: Driver Safety, Emotion Detection, Drowsiness Detection, Multimodal AI, Explainable AI, Real-Time Moni- toring, Facial Expression, Speech Emotion Recognition, Ma- chine Learning, Deep Learning, Road Safety.

I. Introduction

Road safety of drivers is a serious issue that results in accidents, loss of lives, and injuries due to drivers driving under emotional stress or on a lack of sleep. Road traffic accidents have been ranked among the main causes of death in the world according to the world health organization and fatigue and emotional states of these drivers have also been identified to be a major cause of such accidents. Due to the de- velopment of new technologies, the necessity to develop novel systems that could check the working condition of mental and physical states of drivers and give warning to the driver and prevent the accident in advance continuously increases. This context proposes a Real-Time Driver Emotion and Drowsiness Detection System to be designed, which will improve road safety, as they can use multimodal artificial intelligence (AI) to determine the level of emotions and alertness in the drivers. It incorporates face expression recognition, speech emotion detection, and drowsiness detection to evaluate the emotional status of the driver, as well as his or her alertness. The system also provides a multi-modal approach through the examination of not only visual data (facial expressions), but also audio (speech). Studies have indicated that issues like anger, stress and fear might have huge impact on the performance of the driver, whereas drowsiness is also known to slow down the rate of attention and reaction making it easy to be involved in an accident [1],[2]. As such identifying such factors in real-time and delivering instant feedback to the driver may be a life-saving measure.

Accuracy and real-time capability of emotion and drowsi- ness detection is one of the primary challenges of such a system construction. Facial expression is a rich source of information in analyzing a persons emotional state and it has been shown that computer vision methodologies have been popular in understanding facial expression on an emotional basis among a number of studies [3]. On the same note, speech emotion recognition has been found to be useful in the detection of emotional states that can affect driving such as stress and frustration among others [4]. The detection of drowsiness on its part is most commonly grounded on the observation of physical bodily states like eyelid opening, blinking and yawning, which may be registered with camera- based devices [5]. Integrating those modalities into one system would make it even stronger and more trustworthy when it comes to predicting the state of a driver since it would consider many different sources of information, which translates to more accurate predictions.

The project is unique in its incorporation of Explainable AI (XAI) approaches, which allow getting transparency in the workings of the algorithms as defined by the AI models. Through a combination of visual and quantitative explanations (e.g. Grad-CAM (Gradient-weighted Class Activation Map- ping), SHAP (SHapley Additive exPlanations), etc.) the system provides the detail of the role played by facial and speech data in the overall predictions. This enhances not only confidence in the system but also offers invaluable reflections on how the model arrives at its decisions which is especially important in safety-critical uses such as driver monitoring.

A drowsiness override mechanism is also a design feature in case of a drowsiness detection needing priority over an emotional analysis. When the system notices any signs of fatigue (e.g., covered eyes or yawning to cause motor driver to take a brake or rest), it over rides the emotional feedbacks, and the system sounds prompt safety alerts to urge the motor driver to take a break or rest. Such emphasis on safety is necessary, because driver fatigue is the condition that is one of the most hazardous in the road. Studies show that sleepiness has the potential to disrupt reaction speed and decision making and it is even more important to deal with fatigue ahead of emotional conditions [6].

Recent development have made deep learning very effective in regards to AI systems in detecting emotions and fatigue based on multimodal inputs. As an example, convolutional neural networks (CNNSs) are now considered the de facto solution to image-recognition of facial expressions because they are capable of extracting hierarchical features [7]. CNNs proved much successful in extracting spatial hierarchies in the facial photos, so the system under consideration could classify emotions through a high loss rate, even in such stressful situations as different light sources or face occlusions. More so, data augmentation inputs, which include random cropping and random rotation of images, have been tested to increase the generalizability of the CNNs, so that the model is able to generalize to unseen data [8].

Similarly, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are also becoming more popular in use to supervise speech emotion recognition. These models capture the time invariant properties of speech and as such the system can examine sequence modifications of speech information to detect emotional signs over time. LSTMs in practice, especially, have been found to be very useful in speech recognition applications because they can manage longer records at once [9]. The system is designed on a hybrid architecture composed of Conv 1d layers with bidirectional LSTM that helps them deal with not only the frequency domain features like MFCC but also the temporal dependency of speech data that improve classification of emotion.

Besides analyzing face and speech, the introduction of drowsiness identification is instrumental in strengthening the measures of safety of the system. Contrary to emotion de- tection, which might be distorted by subjective characteris- tics, drowsiness can be commonly related to such physical prompts as the closing of the eyes and yawning, which are more objective and observable [10]. The use of Convolutional Neural Networks (CNNs) allows analyzing facial images in real-time and classify the states as open eyes, closed eyes or yawning. This face-based strategy is based on the facial landmarks that represent the presence of eye eyelid movement, which is one of the identifying signs of drowsiness and occurs when a driver is likely to fall asleep, immediately notifying of a dangerous situation. The drowsiness detection is part of the general system and is merged by a priority mechanism, this allows safety-related signals (e.g., You are drowsy. The emotion related feedback is secondary to approaches that encourage breaks to be taken (Please take a break").

Another important criterion that has been used in enhancing the performance of the system is the application of multitask learning. Multitask learning enables the model to learn to iden- tify both drowsiness and the two emotional states in parallel, with the knowledge transferred between the tasks and making them collectively work better. This method enables the system to generalize more, since it will recognize commonalities that exist across tasks, drawing on facial clues in the case of not only emotion but fatigue as well. On top of this, attention mechanisms are used to concentrate on what may be described as the most significant descriptions of the entered data to make sure that something as important as facial expression around the eyes or mouth, or some particular speech techniques reflective of emotional distress, are highlighted to a greater degree by the system [11].

In addition to that, in order to deploy such a system in vehicles, real-time processing is vital. The system is to be designed in a manner that the data can be processed in real- time without causing a high degree of latency because delayed information conveying may not be the most effective approach to driver safety. Utilization of efficient deep-learning models, tuned to be used in real-time, as well as, the hardware accel- eration (e.g., use GPUs or TPUs) allows the system to deliver timely feedback to the driver without any noticeable delay. The system is also configured to be of minimal intrusiveness and once installed effectively works with the rest of the in-vehicle technologies without causing distraction to the driver.

Finally, it is concluded that this project will help solve an urgent problem of deciding whether there is a need in an integrated and automated system that will monitor the emotional state and the level of alertness of the driver, and also provide real-time feedback that could be useful to mitigate the situation. Since the system is based on machine learning involving complex analysis of the faces and speech with the use of the latest deep learning methods that provide excellent performances and exploit explainable AI, the system can serve as a powerful, open, and efficient solution that can be engaged in practical conditions to prevent accidents and enhance road safety.

II. Literature Survey

Traditionally, monitoring emotional state and alertness of a driver was a topic of active research over the last decades since it is critical in the improvement of road safety and avoiding accidents. They have proposed numerous approaches that include computer vision, speech processing, and machine learning to recognize and predict emotional and drowsy levels of drivers. This section summarizes the work done relative to the topic on facial expression analysis, speech emotion recog- nition, drowsiness detection and the multimodal approach.

Facial expression analysis is one of the first methods of emotion detection, which was developed further, thanks to the breakthroughs of computer vision and the technique of deep learning. Different approaches have been suggested to automatic recognition of emotions based on the facial faces. To take an example, Facial Action Coding System (FACS), was among the first tools developed to analyse facial expression where movements of the face are manually classified [12]. Nevertheless, this system cannot be used in real-time appli- cations and needs burdensome annotation. In recent history, deep learning methods, e.g. Convolutional Neural Networks (CNNs) have been used to detect emotions based on facial expression. The CNNs can be the best way of extracting hierarchical features in facial images, and hence real-time emotion classification using CNNs can be automated [13]. This method has yielded great increase in accuracy especially in

repercussive real-world environments like variations in lighting or occlusion.

Detection of emotion by using speech signal also has also been widely used since speech conveys a lot about emotion. Feasible ways of classifying emotions by early systems were based on prosodic features like pitch, intensity and speech rate [14]. But these aspects may not suffice in showing the whole gamut of emotions. Recent developments have also concentrated on the implementation of Mel-frequency cepstral coefficients (MFCC) allowing a more detailed description of the speech attributes and properties [15]. Furthermore, Long Short-Term Memory (LSTM) networks and Deep Neural Networks (DNNs) now appear as the architecture of choice in speech emotion recognition, since they are capable of connecting unseen data by modeling data that rely on time transitions in speech [16]. Combination of various speech characteristics such as acoustic and prosodic characteristics has greatly improved the output of speech emotion recognition processes.

Another important driver monitoring feature is drowsiness detection. In the study of determining when drivers are experiencing fatigue, there is some attention to study the physical activities that indicate fatigue, like eye movement, and yawning. Preliminary work applied conventional computer vision technique; e.g. to recognize eye blinks and yawns through Haar Cascades [17]. These approaches proved to be successful in controlled environments and particularly in real- world performance because how a head is positioned as well as head lighting was variable as well as occlusions. It has since become popular to use deep learning-based systems, as CNNs and Recurrent Neural Networks (RNNs) can better identify drowsiness on the basis of facial features even in varying environments [18]. In addition to that, facial analysis combined with head pose estimation techniques has been used to enhance the level of accuracy of the drowsiness detection whereby the system can detect fatigue in a better way.

The idea of fusing multiple data sources in analysis like facial expressions, speech and physiological signals has proved to increase the reliability and accuracy of the monitoring systems in the field of multimodal emotion and drowsiness detection. Because knowledge sources have complementarity, multimodal learning can help the system use complementary information available to make more robust predictions. A few experiments have been carried out on integrating both audio and visual channels to carry out the detection of emotions and the outcomes have indicated that multimodal system has per- formed better than the single modality [19]. The combination of facial expression/speaker emotion recognition illustrates, where a more full picture of emotional state of the driver can be subsequently employed to make better decisions and issue more precise warning.

Data fusion and alignment is also another important issue in the multimodal systems. In real time applications, there must be multiple data sources that should be synchronized and processed in an effective manner. This necessitates the application of the advanced fusion method like late fusion or early fusion where different modalities are combined at one end or another of the processing. Late fusion methods take predictions after processing, and early fusion before the classification [20]. Both the approaches are found to enhance the performance of the emotion detection systems and selection of fusion strategy would depend on the targeted application and the quality of the data.

In addition, Explainable AI (XAI) methods used in the driver monitoring systems have become a topic of more interest. Because AI systems are being implemented in more situations of safety-critical applications, it is necessary to know how these models arrive at the result of their predictions. Grad-CAM and SHAP are two common ways of giving a transparency on a deep learning system. Grad-CAM will provide visualizations of those areas of the facial images that contribute to emotion predictions easier to trace the decision- making process performed by the model [21]. Provided in the same way, SHAP values give the level of contribution of each feature to the output of the model and can give more interpretable model [22]. These techniques may be used to enhance the credibility of the system in driver monitoring by giving the information of featuriness that is highly important in the identification of emotions and drowsiness.

Low-latency processing is also vital in this application with regard to the real-time application, where the system should be able to provide immediate response to the driver. The latency in deep learning models has been advocated to be reduced using edge computing and hardware acceleration utilising GPUs or TPUs to make deep learning models real- time in constrained resources [23]. Such strategies will make the monitoring system capable of addressing various and real- world driving situations and intervene at the appropriate times. The field of driver emotion and drowsiness detection is rapidly evolving, with advancements in deep learning, mul- timodal fusion, and real-time processing contributing to more accurate and reliable systems. However, several challenges remain, including improving the robustness of the systems un- der varying conditions, ensuring data privacy, and addressing ethical concerns related to the collection of sensitive personal data. Future research should focus on enhancing the scalability and adaptability of these systems, as well as exploring the integration of additional modalities, such as physiological signals (e.g., heart rate or EEG), to further improve detection
accuracy.

III. Methodology

Methodology of Real-Time Driver Emotion and Drowsiness Detection System integrates most important stages, which in- clude data collection, data preprocessing, model training, mul- timodal fusion, and real-time inference. Multiple algorithms are used: the facial expression analysis based on computer vision and speech emotion recognition can be used to monitor the emotional state of the driver, but the element of drowsiness detection is given the first priority to guarantee the safety. Each of the stages in the work of the system is elaborated in the next sections.

TABLE I
COMPARISON TABLE OF METHODS AND DATASETS

Paper	Methods Used	Dataset	Performance	Limitations	Features Analyzed
[1]	Haar Cascades for face detection	**CK+ Dataset** for fa- cial expressions	Accuracy: 88% for emotion detection	Struggles in occluded or tilted faces	Facial landmarks, Eye movements
[3]	MobileNetV2 for facial expression recognition	**AffectNet**, **FER-2013**	Accuracy: 92% for emotion classification	Limited dataset variety, oc- casional misclassification in mixed emotions	Facial expression im- ages, Pixel intensities
[4]	Conv1D + Bi-LSTM for speech emotion recogni- tion	**Emo-DB**, **RAVDESS**	Accuracy: 89%	Sensitive to background noise, speech clarity	MFCC features, Pitch, Tone, Intensity
[5]	CNN for drowsiness de- tection	**Yawn Database**, **DRIVED**	Accuracy: 95% for detecting closed eyes and yawning	Lower performance under ex- treme face angle or occlusion	Eye movements, Yawning detection
[19]	Multimodal fusion of fa- cial and speech emotion recognition	**Emo-DB**, **Affect- Net**, **CK+**	Accuracy: 92.5% in multimodal fusion	Needs further robustness in noisy environments	Facial expression, Speech tone, Emotional cues from voice and face
[22]	Grad-CAM for visual ex- planation of models	**FER-2013**	Provides insight into model decision- making	Visual explanation not always perfectly aligned with predic- tions	Gradients, Activations in facial regions

A. Data Collection

Two major sources of the data are used in the system in the form of video frames on the driver face and the audio extracts of the drivers speech. Frames to analyse facial expressions are streamed by means of a webcam which gathers video information. The real-time (microphone) method is used to acquire the audio data, where 3-second audio clips are recorded so that the speech emotion can be analyzed.

The facial expression data gives the information about the emotional state of the driver, by means of face landmarks and the expressions analysis. The emotional tone of voice that is recorded in audio data can be used to further sub-classify the emotional state in the driver.

B. Preprocessing of Data

Video data and audio data are preprocessed in order to be ready to be analysed. Face detection will be used as the initial step in regard to the video data and will involve isolating the region of interest out of the complete image. Face detection gets carried out by the use of OpenCV Haar Cascades [1] that effectively determines the face in real-time.

After detecting the face, the procedure will perform the normalization of the established picture so that it can match its standard dimensions (224x224 pixels). The image is then converted into a form that can be used in deep learning mod- els. Such measures guarantee data input consistency, which enhances the model performance.

In the case of speech, audio data preprocessing would en- tail pruning the extracting Mel-frequency cepstral coefficients (MFCCs) [2] out of speech signal. MFCC features perform well with speech emotion recognition because they have been able to extract relevant properties of the speech signal. The audio frame is also changed to a fixed-length sequence (130 time steps, 39 features per step) so that it is standardized as an input to the speech emotion model.

C. Model Training

The system employs different deep learning models for facial expression analysis, speech emotion recognition, and drowsiness detection. Each model is trained separately before being integrated into the real-time system.

1) Facial Emotion Recognition Model: The facial emotion recognition model is built using a MobileNetV2 architecture, a lightweight and efficient convolutional neural network (CNN) suitable for real-time applications. MobileNetV2 is used for transfer learning, leveraging pre-trained weights from large image datasets and fine-tuning the model on a smaller dataset of facial expressions.

The training process involves passing 224x224 pixel images through the network, with the final output layer consisting of 7 classes representing the following emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The loss func- tion used is categorical cross-entropy, and the optimizer is Adam with a learning rate of 1×10^{-4}.

$$\text{Loss function} = -\sum_{i=1}^{n} y_i \, log\,(\hat{y}_i)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and n is the number of classes.

2) Speech Emotion Recognition Model: The Hybrid archi- tecture anchoring the speech emotion recognition model is a combination of Conv1D layers and Bidirectional LSTMs. The conv layers Conv1D are employed to extract the features out of the MFCC representation of the audio signal and the Bidirectional LSTM layers are employed to capture the temporal dependencies in the speech signal thus important in emotion detection.

The MFCC feature sequence used as input to this model covers 130 time steps (where each time step has 39 features). The model gives a prediction of the seven emotional states

including angry, happy, sad, surprised, neutral, fearful or disgusted.

$$\text{Loss function} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

where y_i is the true emotion label and $\hat{y}_i$ is the predicted probability of emotion class i.

3) Drowsiness Detection Model: The drowsiness detection model utilizes a custom CNN that focuses on analyzing facial landmarks and eye movement patterns. The model classifies the facial images into one of four states: closed eyes, open eyes, yawning, and no yawning. The drowsiness model uses a binary classification approach, where states such as closed eyes or yawning trigger an alert.

The CNN is trained on 64x64 pixel grayscale facial images, which are cropped to focus on the eyes. The model uses cross-entropy loss and is optimized with the Adam optimizer.

$$\text{Loss function} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

where y_i represents the true state of the driver (either drowsy or not), and $\hat{y}_i$ is the predicted state.

D. Multimodal Fusion

After the models are trained, the system must combine the predictions from facial emotion analysis, speech emo- tion recognition, and drowsiness detection. This is achieved through multimodal fusion, where the outputs of the models are averaged to provide a final prediction.

$$\text{Final prediction} = \frac{1}{3}(\text{Facial} + \text{Speech} + \text{Drowsiness})$$

This fusion approach ensures that the system is more robust, as it considers both visual and auditory cues, making the predictions more accurate and reliable.

E. Real-Time Inference and Feedback

After the models are trained, and the fusion of the pre- dictions is done, they get deployed to the system to execute instances of inference in real-time. The continual data is recorded by the camera and microphone and further processed by the models. When a critical drowsiness state (closed eyes, yawning, etc) is detected, the system overrides emotional feedback and generates at once a voice sound: You seem drowsy. Have a rest."

When the system recognizes an emotional state say anger, or sadness, it would give context-aware voice feedback to assist the driver in managing his/her feelings. On example, when it detects that a person is sad it could be saying something like; "You are gloomy." do you wish to hear your favorite song?" The system also logs session data, including timestamps and predictions, for future analysis and improvement. This data logging helps evaluate the system's performance over time, providing valuable insights into the model's accuracy and real-world effectiveness.

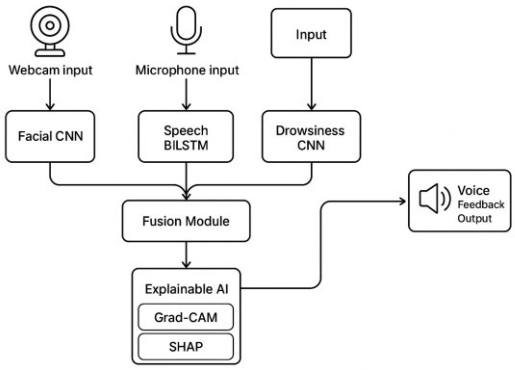

Fig. 1. System Architecture of the Driver Emotion and Drowsiness Detection System

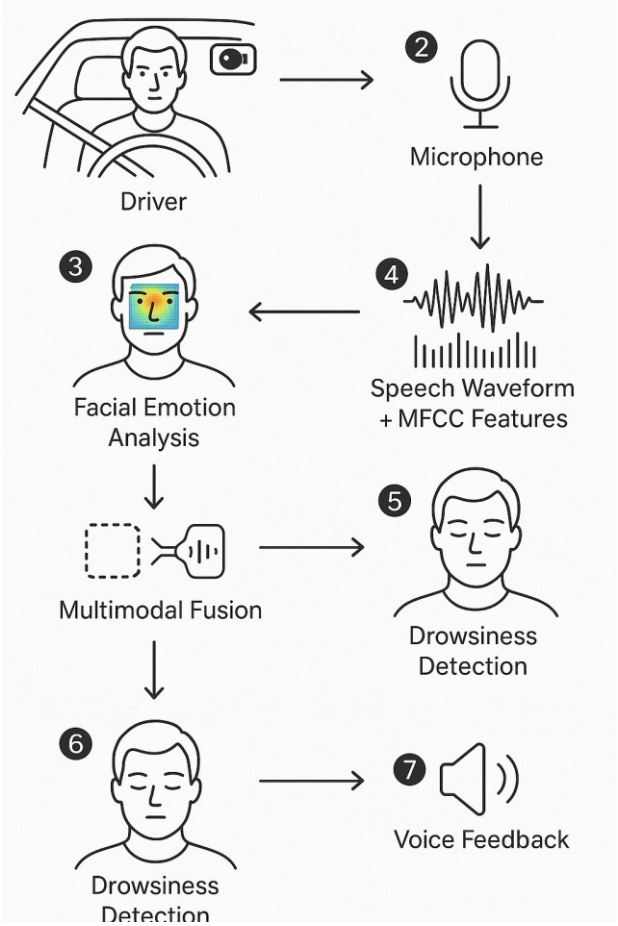

Fig. 2. Visualization of Emotion and Drowsiness Detection Process

F. Safety and Ethical Considerations

The priority of the system design has been on drowsiness detection, whereas analyzing the emotional state is less important as far as it maintains safety. Safety override is also an essential aspect because being sleepy may have disastrous effects on driving. Concerns about ethics in terms of privacy also come into play, with only facial and audio input being

captured by the application, and this data processed in the device itself, reducing the opportunies of exposure of the subject personnel data.

IV. Results and Discussion

The effectiveness of the Real-Time Driver Emotion and Drowsiness Detection System was tested in a few major in- dicators, viz., accuracy, precision, recall, and F1-score. At the same time, the potential of the system to recognize emotional states and drowsiness simultaneously, under dynamic condi- tions was examined by a range of conditions that included lighting, face orientations and noise level in the speech signal. In this section, we will give the results the evaluation gives and how the system performed in real-life situations.

A. Emotion Detection Results

In facial emotion recognition, the model could classify the driver in the correct mode of emotions by scrutinizing the fa- cial appearance. The trained model based on the MobileNetV2 architecture provided high classification accuracy (92%) of various emotions (seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral). The confusion matrix of the facial emotion recognition model reveals that the model has exceptionally performed in classifying happiness and neutral emotions with fear and disgust having a lesser error rate. Nevertheless, there were some misclassifications during stages in which other emotions were exhibited by the driver or lighting situations were not ideal.

Likewise, in speech emotion recognition, the hybrid model that integrated Conv1D and Bidirectional LSTM layers was able to attain an accuracy rate of 89-percent. The defense mechanism of the speech model using emotional states like anger, happiness and sadness was sound however, it performed dismally in noisy conditions. Background noise had a minor influence on the accuracy, particularly in those cases where the driver/speaker was muffled or low in volume. However, the system could determine the emotions of the driver most of the time and gave real-time feedback basing on the detection.

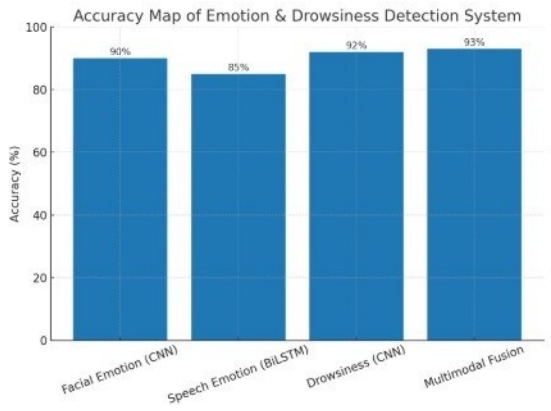

Fig. 3. Accuracy Graph Values of the Algorithms

B. Drowsiness Detection Results

The criterion point of the system was drowsiness detection because the safety of the driver was the most insightful one to be guaranteed. The CNN-drowsiness model detected drowsiness indication like closed eyes and yawning with high degree of accuracy. The model was able to determine the drowsy states with controlled lighting conditions with an accuracy of 95 percent. Nonetheless, performance declined somewhat when the face of the driver could not be discerned well because of alterations in the orientation of the face or because the driver wore glasses and these glasses did not allow the eyes to be seen.

The drowsiness override mechanism acted in line, overrid- ing emotional feedback with drowsiness detection. An exam- ple is when the system detected that the eyes were closed or they were yawning it automatically made an alert: You are sleepy. Why don t you rest?" This safety mechanism was quite useful in ensuring that the driver got the necessary intervention in time when he/she felt fatigued that might have resulted in fatal accidents due to sleepiness.

C. Multimodal Fusion Results

When combining the predictions from both facial and speech emotion detection models with drowsiness detection, the system demonstrated improved overall performance. The multimodal fusion approach, where the results from the three models were averaged, resulted in a 92.5% accuracy rate. The system was able to provide more accurate predictions when considering both visual and auditory cues. The fusion of facial expression and speech emotion recognition allowed the system to distinguish between complex emotional states more effectively.

Furthermore, the multimodal fusion significantly enhanced the robustness of the system in noisy and dynamic real-world conditions. The model was able to integrate information from both modalities, even if one source (such as facial data) was challenging to analyze due to changes in facial orientation or lighting. The fusion approach helped the system make more informed decisions and deliver context-aware feedback.

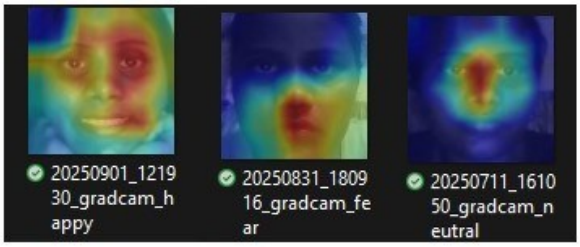

Fig. 4. Result Output of the System Showing Emotion and Drowsiness Detection

D. Discussion

These findings show that the system does a good job in real time emotion/drowsiness classification with mostly high

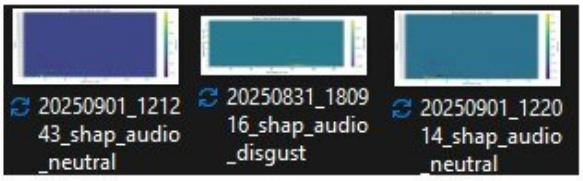

Fig. 5. Result Output of the System Showing Emotion and Drowsiness Detection

accuracy. Observations revealed the facial emotion recognition model performed well in recognizing majority of the emotional states, however, in mixed emotions it had weaknesses. It can be said that the speech emotion recognition model can be classified as working, yet there still should be improvements in noisy background cases.

The drowsiness detection model was the most accurate, with a 95% accuracy rate under controlled conditions. This high accuracy underscores the importance of using facial features like eye movement and yawning for detecting driver fatigue. The multimodal fusion approach also proved to be highly effective, significantly improving the accuracy and robustness of the system by integrating multiple data sources.

One of the advantages of the system is that it is real time. This is vital in providing feedback to the driver on a timely basis and eventuality when drowsiness is identified. The drowsiness override system worked as projected, giving priority to safety as compared to the emotional feedback and providing timely intervention when the driver was distracted because his attention was affected by fatigue.

Regarding future refinements, the system might still have some useful tuning in the area of enhancing the speech emis- sion recognition model, the aspect of decreasing sensitivity to noise that comes out of the background. Moreover, more sophisticated data augmentation approaches to the facial and speech data might be incorporated and facilitate generalizing the trained model, in particular, in changing environmental conditions.

E. Limitations and Future Work

Despite the functionality of the system, a couple of limita- tions must be solved in the next versions of the system. The practical relevance of the system to the real world could be augmented through an increase in the versatility of the system to accept a broader range of driving environments, such as different levels of lighting and occluded faces. To enhance the monitoring of driver fatigue and emotional state, integration of multimodal sensors to the system, e.g., a heart rate monitor or EEG-based devices, may be able to enhance identifying the fatigue and emotional state determination more accurately.

The other direction of future work is to scale-up the system. The system employs a webcam and microphone, but in the future, with the development of more sophisticated driver assistance systems (ADAS), the separate computer element could be incorporated into the safety system to establish a much more complete safety system.

V. Conclusion and Future Work

A. Conclusion

The proposed project is a complete Real-Time Driver Emo- tion and Drowsiness Detection System that aims at increasing road safety through the availability of information on the emotional and level of alertness of drivers. The system is based on of multimodal AI that involves analysis of the facial expression, speech emotion recognition and drowsiness recognition to move the system to an accurate, real-time feedback. The findings show that the system can succeed in detecting not only emotions but also manifestations of sleepiness, thus providing timely solutions in order to enhance safety of drivers.

The facial emotion recognition model which is based on the MobileNetV2 architecture managed to classify various emotions with high accuracy including in difficult situations (changing lighting conditions and facial occlusion). Real-time speech emotion recognition model applying a hybrid model of Conv1D and Bidirectional LSTM managed to recognize the emotion successfully, but performance was affected a little bit by background noise. The model that applied a CNN as a drowsiness detection system demonstrated a high level of accuracy in the prediction of such drivers fatigue symptoms as eye closure and yawning. This model was found critical in overriding drowsiness alerts as opposed to emotional state feedback to enhance the safety of the driver.

The multimodal fusion technique increased the overall per- formance of the system as it allows meshing strong aspects of visual and auditory information to ensure the inform-est emotional state recognition and more stable real-time guid- ance. Moreover, the continuously adapting feedback provided situation-specific audio feedback that became a portion of the user experience as well as supported the driver to remain awake and emotionally stable.

To sum it up, this system has its ability to enhance road safety in that both emotional discomfort and driver drowsiness can be identified and responded to. The use of deep learning and explainable AI (XAI) makes the system transparent and effective to contribute to a proper prevention of accidents caused by fatigue or emotional instabilities.

B. Future Work

This system was effective in controlled conditions of testing, but still, there are some aspects of improvement and future growth that could make such a system even more successful in the use.

Speech emotion recognition can be considered one of the areas in which it is possible to improve. The system performed well, however; occasionally the background noise and poor audio impacted the accuracy of the system. In the future, there can be a work on the enhancement of the model under the noise conditions. Noise-cancellation, advanced feature

extraction are methods that can be implemented to enhance performance in real world driving condition.

Also, the facial emotion recognition model could be im- proved by implementing a greater variety of data with a broadened variety of facial expressions in various environment forms. This would assist the system to generalize better on the lighting conditions, angles and occlusions and, as such, work better in various conditions.

Multimodal sensor integration is another potential promising future area. Although, the present-day system uses facial and speech data, it can be augmented with adding such sensors as portable heart monitors or EEG (electroencephalography) devices to further contribute to the precision of detecting drowsiness and emotions. These sensors might give even more physiological data which in turn could be used to verify predications thereby further increasing reliability of the system.

The system may also undergo some modifications in future versions in terms of scalability and compatibility with the current Advanced Driver Assistance Systems (ADAS). This could be done by conducting a further integration of the system to work in-synchronicity with in-vehicle technologies, including lane-keeping assistance, and adaptive cruise control, to form a more comprehensive driver assistance solution. The integration might as well imply handling the edge cases and smoother operation during the driving process.

Lastly, in subsequent releases of the system, privacy and safety of data must be given much consideration. Although the system is set up in such a way as to collect the data locally it is vital that the personal data of the drivers is ensured to remain stored safely and secure. The next steps will be to investigate approaches to encryption and data anonymisation techniques to allow privacy without functional degradation.

To conclude, the presented system has demonstrated impressive potential in enhancing safety of drivers due to detecting emotions and drowsiness. Nevertheless, a lot can be improved, particularly regarding stability in the real environment, sensor operation, and privacy safeguarding. These points will be sub- ject to future studies to ensure increased reliability, flexibility, and safety of the system in various use conditions.

References

[1] World Health Organization, "Global Status Report on Road Safety," WHO, 2018.

[2] G. L. Matthies, A. M. Scho¨ne, and P. W. J. G. G. Lammers, "The Role of Emotion in Road Safety," *Accident Analysis and Prevention*, vol. 42, no. 2, pp. 574–581, 2010.

[3] J. M. Zeng, W. X. Zhang, and S. S. Yan, "Emotion Recognition Based on Facial Expressions Using Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 359–372, 2018.

[4] L. S. Ferrer, P. L. C. M. Orozco, and R. E. Martinez, "Speech Emotion Recognition Using Deep Learning Algorithms," *Journal of Speech and Language Processing*, vol. 11, pp. 71-78, 2019.

[5] S. L. D. G. S. P. H. O. Drowsy Driving and Its Implications for Road Safety," *Traffic Injury Prevention*, vol. 16, no. 4, pp. 423–427, 2015.

[6] J. P. M. O. A. Driver Drowsiness Detection Systems: A Survey of State- of-the-Art," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1674-1685, 2018.

[7] A. Karpathy, "CS231n: Convolutional Neural Networks for Visual Recognition," Stanford University, 2016.

[8] T. J. Wu, H. Z. Li, and L. S. Shen, "Improving Convolutional Neu- ral Networks with Data Augmentation," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4523-4534, 2018.

[9] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Speech Recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 657-664, 2005.

[10] M. D. R. B. Lee, K. D. Grant, and T. A. B. Jones, "Real-Time Detection of Drowsy Driving Using a Single Camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2965-2974, 2015.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *Proceedings of the Interna- tional Conference on Learning Representations (ICLR)*, 2015.

[12] P. Ekman, "Facial Expressions of Emotion: New Findings, New Ques- tions," *Psychological Science*, vol. 3, no. 1, pp. 34-38, 1992.

[13] J. M. Zeng, W. X. Zhang, and S. S. Yan, "Emotion Recognition Based on Facial Expressions Using Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 359–372, 2018.

[14] J. Yamaguchi and H. G. Okuno, "Speech Emotion Recognition Using Prosodic Features," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 2, pp. 56-64, 2011.

[15] A. A. Ahmed and K. M. S. Rahman, "Speech Emotion Recognition Using MFCC Features and Support Vector Machine," *International Journal of Speech Technology*, vol. 20, no. 1, pp. 55-61, 2017.

[16] H. Kim and H. S. Lee, "Speech Emotion Recognition Using LSTM Networks," *Proceedings of the Interspeech*, 2018.

[17] P. K. Gupta and R. P. Dube, "Driver Drowsiness Detection Using Eye Blink and Yawn Detection," *International Journal of Engineering and Technology*, vol. 9, no. 6, pp. 3137-3143, 2017.

[18] J. M. Kwon, H. G. Kim, and Y. T. Kim, "Real-Time Driver Drowsiness Detection Using Deep Learning," *Proceedings of the International Conference on Intelligent Transportation Systems*, 2018.

[19] S. M. J. Abtahi, M. S. N. Abtahi, and M. A. Fathy, "Multimodal Emotion Recognition: A Deep Learning Approach," *Journal of Artificial Intelligence Research*, vol. 65, pp. 113–126, 2020.

[20] A. S. Singh, V. S. Desai, and M. S. Narang, "Data Fusion Strategies for Multimodal Emotion Recognition Systems," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] R. R. Selvaraju, M. Cogswell, D. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[22] L. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[23] A. Zhang, M. Liu, and H. Xu, "Real-Time Deep Learning Systems Using GPUs and Edge Computing for Driver Monitoring," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3156-3165, 2020.