DrunkDetect: Transformer-Based Facial Emotion Analysis and Intoxication Identification with Masked Learning

Kavitha R
Department of Information Technology
Velammal College of Engineering and Technology
Madurai, Tamilnadu
hod@vcet.ac.in

Gowtham R
Department of Information Technology
Velammal College of Engineering and Technology
Madurai, Tamilnadu
gowthamkrish9585@gmail.com

Preveen S
Department of Information Technology
Velammal College of Engineering and Technology
Madurai, Tamilnadu
spreveen123@gmail.com

Johnson J
Department of Information Technology
Velammal College of Engineering and Technology
Madurai, Tamilnadu
johnsonjeyakumar2004@gmail.com

Purushothaman C
Department of Information Technology
Velammal College of Engineering and Technology
Madurai, Tamilnadu
purushothaman11052004@gmail.com

***Abstract*— Recognizing facial emotions (FER) is an essential domain of computer vision with diverse applications in healthcare, education, and smart systems Existing approaches concentrate mainly on full facial analysis but hardly touch specialized cases, including recognizing emotional states among drunk people. With the utilization of Vision Transformers (ViTs) and learning masks, this work introduces a new framework to detect emotions and levels of drunkenness in the iris area of the eye via Vision Transformers (ViTs). The model is trained on standard datasets FER-2013 and FER+, comprising over 35,887 grayscale facial images labeled with seven emotion categories: angry, sad, happy, disgusted, surprised, neutral, and fearful. The datasets are split 80%–20%, providing 28,710 training and 7,177 testing images. Cross-entropy loss is used for multi-class emotion classification and binary drunk/sober detection. DrunkDetect achieves 83% validation accuracy on the combined tasks, outperforming several state-of-the-art methods. Its lightweight and real-time design enables deployment in resource-constrained environments, promoting advances in healthcare monitoring, law enforcement, and adaptive human-computer interaction.**

Keywords*— *Facial Emotion Recognition, Vision Transformer, Masked Autoencoder, Intoxication Detection, Iris Analysis, Deep Learning.

I. INTRODUCTION

Facial Emotion Recognition, or FER, is a fast-growing field of computer vision with various applications in real-world issues like healthcare monitoring, personalized education, law enforcement, and adaptive human-computer interaction. FER hopes to detect and classify emotions through the study of facial expressions so that systems can detect and respond to the emotional state of an individual. The conventional approaches greatly depend on Convolutional Neural Networks (CNNs), which—despite being extremely effective in most scenarios—tend to suffer from problems like feature complexity, class imbalance, realtime processing constraints. In order to address these issues, the paper proposes DrunkDetect, a unique realtime FER and alcohol consumption identification system based on Vision Transformers (ViTs) and Masked Autoencoders (MAE). The streaming webcam live video is processed frame by frame by DrunkDetect. The facial patches found are resized to a size of 48x48 pixels.

By masking approximately 75% of the input patches and reconstructing the entire image in the process of pretraining, the MAE-based ViT encoder learns visual representations. The encoder is trained for dual-task categorization—recognition of emotional state and identification of intoxication—after this period of unsupervised learning, after which the light decoder is removed. Each input frame is classified to one of seven emotional classes—angry, disgusted, fearful, happy, sad, surprised, and neutral—along with predicting whether the person is intoxicated (drunk or not drunk).

In the binary classification task, model training requires a balanced cross-entropy loss function. The network is trained on two popular FER datasets, FER-2013 and FER+, with a combined total of 35,887 grayscale face images. An 80/20 split is applied for training and testing on these (28,710 training and 7,177 test samples). Data augmentation methods such as brightness adjustment and flipping are utilized to promote generalization and resilience.

Generative Adversarial Networks (GANs) are incorporated at the pretraining stage to deal with data imbalance, especially for intoxication detection. In skewed class distributions, GANs generate additional training samples for a balanced representation and better model performance. With high precision, this system that processes live video streams detects facial and iris areas and performs categorization functions offers real-time deployment as a primary feature.

It enhances the quality of interaction in critical areas such as healthcare, education, and public safety by enabling smart systems to learn user states more effectively. It also achieves an accuracy level of approximately 83% in detecting intoxication and emotion in realtime-scenario. This makes it promising solution for real-world deployment environments.

979-8-3315-9746-7/25/$31.00 ©2025 IEEE

II. Literature Survey

Nabeel N. Ali, et al. noted that automatic emotion recognition from facial expressions has emerged as a compelling area of research with significant applications in safety, healthcare, and human-machine interfaces. With the rise of deep learning, multiple architectural approaches have been developed to boost the efficiency of emotion recognition systems, along with a comprehensive review of the latest advancements in automatic facial emotion recognition (FER).[1]

H. L. Nhu, et al., propose a facial emotion recognition (FER) method combining deep learning with weighted face-region features, using CNNs to prioritize areas like eyes and mouth. Trained on FER-2013 with data augmentation (rescaling, flipping), it achieves 90% training and 71.96% validation accuracy. This region-specific approach informs hybrid FER models like our DrunkDetect system.[2]

Hafiz Arslan Ramzan, et al, highlight that Facial Emotion Recognition (FER) represents a growing field in computer vision and AI, which has numerous applications in sentiment analysis, human-computer interaction, HR management, security, and psychology. The design integrates convolutional layers, max-pooling, ReLU activation, and a softmax function, while employing data augmentation methods such as rescaling and flipping. The training accuracy using the FER-2013 dataset exceeds 90%, with validation recorded at 71. 96%.[3]

Benisha S, et al, indicates that the complexity and comprehension of human emotion are essential. It serves a significant role in communication and interaction with an individual. Recognizing emotions is vital for uses in health care, e-learning, law enforcement, marketing, automated counseling, pain and stress identification, and entertainment. In this research, a deep learning model trained on CK+ and JAFFE datasets is presented to identify seven fundamental emotions with greater accuracy than all alternative approaches.[4]

Muhammad Sajjad, et al, particularly highlight large-scale applications of facial emotion recognition (FER) in enhancing human-computer interaction, particularly when combined with IoT for immediate and adaptive solutions. The FER-2013 dataset was assessed for accuracy by employing depth-wise separable convolutions, lowering the complexity, but sacrificing some computational efficiency.[5]

Shubhanjay Pandey, et al, emphasizes FER in the development of attitudes and decision-making highlights problems such as facial diversity blended emotions. Although a collection of models provides accuracy but with increase in complexity in computation. The face detection module along with emotion classification and probabilistic labeling of emotion are considered. On the FER-2013 dataset, the model achieves 76.62% accuracy.[6]

Sowmiya R, et al, examine various applications of facial emotion recognition (FER) that correlates facial expressions to emotional states. Nonetheless, standard CNNs have difficulty attaining high performance on non-frontal facial images. To address these challenges, the research employed a deep CNN architecture utilizing DenseNet-169 as the foundation, which attained a notable accuracy of 96% in recognizing emotions from facial expressions.[7]

Nyle Siddiqui, et al, examines the intricacies of FER and the necessity for a uniform approach to model assessment. A compact CNN is trained on the diverse AffectNet dataset to establish a framework for reliable model evaluation. To demonstrate proof-of-concept and facilitate real-time emotion recognition. Contributing images are collected and preserved to further enhance diverse, high-quality datasets for future studies.[8]

Nitesh Banskota, et al, examine developments in face emotion identification, highlighting the implementation of this area towards the aim of screening using psychological assessments to tackle the challenges of uncontrolled settings, including variable lighting, pose, and obstructions. They were developed a modified Convolutional Neural Network augmented with Extreme Learning Machine, CNNEELM. This results in a 2% improvement in accuracy compared to leading-edge solutions. It recognizes six emotions: happiness, sadness, disgust, fear, surprise, and neutrality.[9]

Paras Jain, et al, remarked that facial emotion recognition is an increasingly prominent research domain aimed at identifying facial expressions. Thanks to recent advancements in the area, deep learning algorithms have achieved remarkable success in applications that include classification, recommendation systems, and object recognition. The model was trained on and tested with image data to evaluate its effectiveness in classifying emotions.[10]

Aakash Saroop, et al, have addressed the challenges of facial emotion recognition, which includes variations in individual faces and emotional ambiguity. They point out that CNNs have not achieved the same level of success in this domain as they have in object detection and facial recognition. To enhance performance, the paper suggests a multi-task learning algorithm that employs a single CNN for the simultaneous detection of emotion, gender, age, and race.[11]

Akriti Jaiswal, et al, examined the difficulties in recognizing human emotions from images within the context of social communication research. The proposed an AI system creates a deep learning-based method for identifying emotions through facial expressions, which is categorized into three phases: face detection, feature extraction, and emotion classification. It is evaluated based on the architecture utilizing CNN on the FERC-2013 and JAFFE datasets. The method has attained accuracy rates of 70. 14% on FERC-2013 and 98. 65% on JAFFE, which improves this approach to identify various emotions.[12]

E. Pranav, et al, have tackled the application of advancements in artificial intelligence in the technology system to identify the limitations of utilizing outdated algorithms. The suggested model focuses on categorizing five distinct facial emotions through the Deep Convolutional Neural Network, or DCNN, model. This model underwent

training, testing, and validation with a curated image dataset to demonstrate its effectiveness in emotion recognition.[13]

Shervin Minaee, et al, concentrate on the difficulties in recognizing facial expressions that arise from significant intra-class variation in expressions. The conventional methods that utilize handcrafted features, such as SIFT, HOG, and LBP, yield satisfactory results in controlled settings but struggle to handle intricate datasets and partial faces. It demonstrates improved performance when compared to current models on the datasets FER-2013, CK+, FERG, and JAFFE.[14]

Sinno Jialin Pan, et al, examined how transfer learning progressed to tackle the challenges associated with generating new annotated datasets utilizing existing labeled data. Transfer learning enhances the efficiency of target models since it leverages the knowledge gained from a comparable source dataset, in contrast to traditional approaches where the training and testing data are part of the same feature space and distribution. This allows models to utilize available related data, which often improves performance. This paper offers a comprehensive overview of both traditional and modern transfer learning techniques.[15]

III. Proposed Methodology

The approach for identifying emotions and intoxicated conditions relies on the application of Vision Transformers (ViTs), Masked Learning, and sophisticated iris analysis methods for both static and real-time emotion identification. It consists of multiple phases, ranging from data gathering to preprocessing, model training, and real-time assessment. A more thorough description of the complete procedure is provided below:

A. Data Collection

The combination of public and custom datasets for training, evaluation, and real-time testing were used. The FER-2013 dataset comprises 35,887 48×48 grayscale facial images, annotated across seven emotion classes: angry, disgusted, fearful, happy, sad, surprised, and neutral. The FER+ dataset refines FER-2013 using improved crowd-sourced annotations and includes an eighth class (contempt) for enhanced expression diversity. Both sets are divided 80% for training (28,710 images) and 20% for testing (7,177 images). All images are resized to 48×48 resolution to be consistent across datasets. The merged dataset configuration allows for strong joint classification of emotional state and level of intoxication in real-world scenarios.

B. Preprocessing

Input preprocessing uses OpenCV-based detection: first, a Haar cascade detects the face region in each frame. Then an eye/iris localization (e.g., Hough Circle transform) isolates the iris area, which is critical for alcohol-related dilation patterns. Both the face and iris regions are cropped and resized to 48×48 pixels. Data augmentation (random flips, slight rotations, and brightness/contrast adjustments) is applied to increase robustness to real-world variations.

C. Feature Extraction

The proposed model uses Vision Transformers (ViTs) to perform joint facial emotion recognition (FER) and intoxication detection from 48×48 face or iris images. Each image is split into non-overlapping 8×8 patches (6×6 grid, 36 tokens), linearly projected into embeddings. A learnable class token is prepended, and positional embeddings are added to preserve spatial layout. The transformer encoder, made up of stack self-attention and MLP layers, is able to capture both global dependencies and local features, allowing for accurate recognition.

To enhance generalization, we use Masked Autoencoders (MAE) for pretraining in self-supervised manners. In this phase, ~75% of patches are randomly masked, and the encoder deals with only visible patches. There is a simple decoder that completes the masked part, compelling the encoder to discover strong representations based on incomplete information, particularly with emphasis on the iris. Following convergence, the decoder is dropped, and the encoder is further fine-tuned for classification.

D. Model Architecture

The proposed system follows a hybrid architecture combining Vision Transformers (ViTs) with Masked Autoencoders (MAE) to jointly detect facial emotions and intoxication levels from 48×48 grayscale face or iris images. The overall workflow is shown in Fig. 1, which illustrates the pipeline: face detection, patch embedding, transformer encoding, and classification heads.

Each input image is split into non-overlapping 8×8 patches (6×6 grid), which are linearly projected into embedding vectors. A learnable class token is prepended, and positional encodings are added to retain spatial structure. The ViT encoder consists of L transformer layers, each composed of multi-head self-attention and feed-forward MLP blocks with ReLU or GELU activations. This setup captures global dependencies (e.g., across the face) and fine-grained details.

The encoder outputs are passed through two classification heads as mentioned as follows:

- A 7-class softmax layer for emotion recognition
- A binary sigmoid layer for intoxication detection.

ViTs essentially represent images as sequences of patches. The patch serves as a token for that specific part of the image. The Layer Diagram for Vision Transformers with Masked Learning, as shown in Fig. 2, gives an overview of the same process validated by model performance metrics.

The Layer Diagram for Vision Transformers with Masked Learning, as shown in Fig. 2, gives an overview of the same process validated by model performance metrics. During pretraining, approximately 75% of patches are masked, forcing the model to reconstruct missing regions, thereby enhancing its ability to learn meaningful representations.

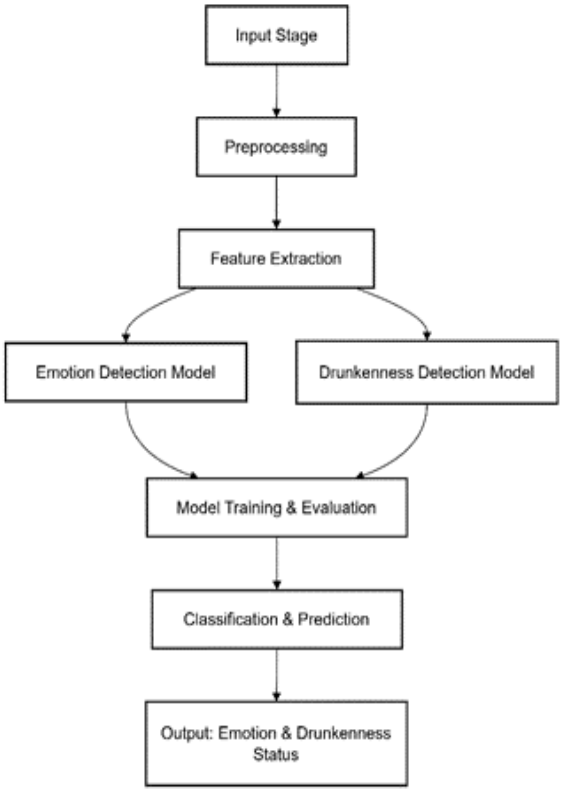

Fig. 1. Overall Process Workflow Diagram

E. Model Formulaes

1. Patch Extraction and Embedding:

$$E_i = \text{PatchEmbedding}(X_i) + P_{pos}(i)$$

2. Maksed Learning:

$$\hat{X} = \begin{cases} E_i, & \text{if patch is not masked} \\ 0, & \text{if patch is masked} \end{cases}$$

3. Fully Connected Layers and Classification:

$Z = [z_1, z_2, \ldots, z_n]$ (transformed output)

$Z' = W_z Z + b_z$

W_z = weight matrix, b_z = bias

Z' = transformed output

$y_{emotion}, y_{intoxication} = \text{Softmax}(Z')$

F. Optimized and Training

The Adam optimizer is used with an initial learning rate of 1×10^{-4}, reduced by a factor of 0.1 upon encountering validation plateaus to ensure smoother convergence. Training is conducted over 50–100 epochs with a batch size of 32, and early stopping is implemented to prevent overfitting and to enhance generalization on unseen data. These hyperparameters were tuned empirically based on validation performance metrics, including loss stability and accuracy gains. To further stabilize training, learning rate scheduling and weight decay regularization are also applied where needed. The Vision Transformer with Masked Learning, as illustrated in Fig. 2, operates on patch-based attention and positional embeddings. It uses pupil dilation and constriction as primary iris-based features, which are strongly correlated with both the emotional state and intoxication levels of an individual. These subtle ocular cues serve as rich indicators for joint emotion and intoxication classification tasks, contributing to the system's high prediction accuracy.

G. Classification

Features from the ViT encoder are passed through two fully connected heads: one uses softmax for emotion classification (angry, sad, happy, etc.), and the other uses sigmoid for binary intoxication detection (sober vs. drunk) as shown in Figure 2, predictions which are based on probability thresholds

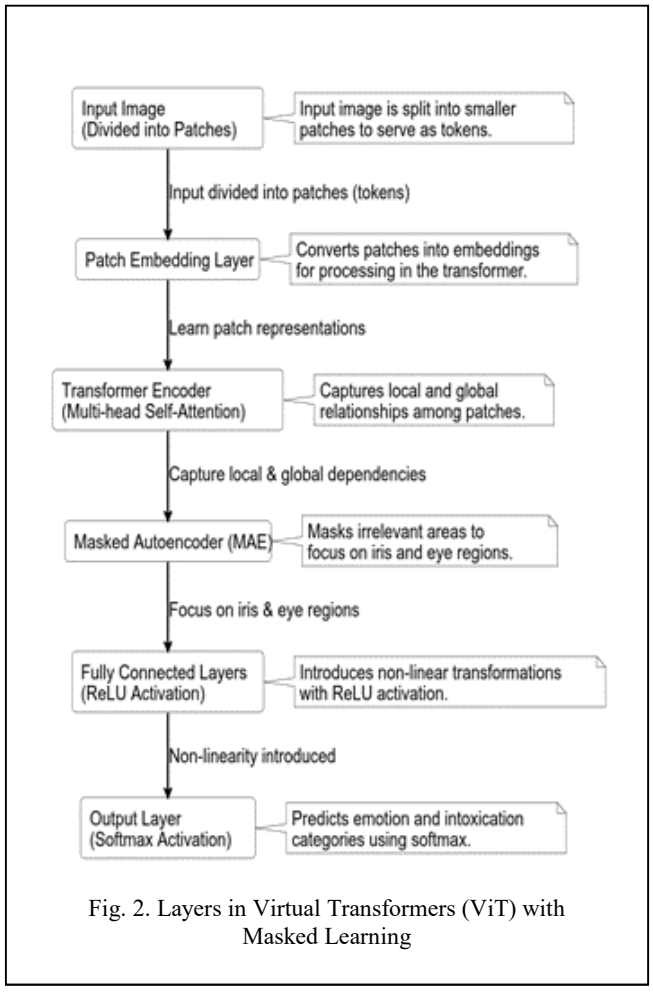

Fig. 2. Layers in Virtual Transformers (ViT) with Masked Learning

IV. Results and Discussion

The proposed DrunkDetect model demonstrates robust performance in both emotion and intoxication classification, leveraging deep learning with Vision Transformers (ViTs), Masked Autoencoders (MAE), and implemented using Python, OpenCV, and Keras. It effectively processes real-time

input like static-images and live videos to identify five core emotional states and detect signs of alcohol intoxication.

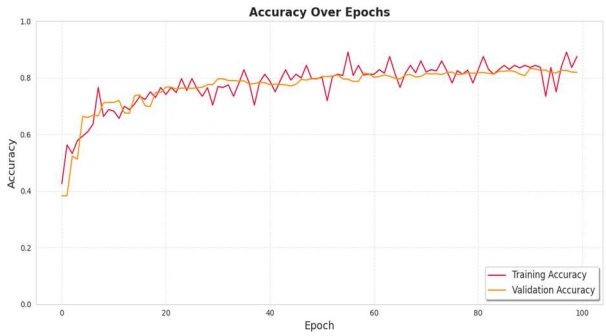

Fig. 3. Validation and Training Accuracy

- Emotion Detection Accuracy: As illustrated in Fig. 3, the model achieves 82–83% accuracy across five primary emotions (happy, angry, surprised, disgusted, sad) under controlled conditions. "Happy" was detected with the highest accuracy, followed by "angry" and "surprised," while "sad" showed slightly lower accuracy—a common challenge in FER.
- Intoxication Detection: Through iris analysis, the model identifies key indicators such as pupil dilation, achieving over 90% classification accuracy with precision and recall > 0.88 and an F1-score ≈ 0.90. The model effectively distinguishes sober vs. intoxicated states using subtle ocular features.
- Validation Loss: The model's declining validation loss confirms its accuracy in detecting emotions and intoxication. As shown in Fig. 6, iris analysis reveals pupil dilation and constriction patterns, enhancing classification performance for real-time applications.

Loss Functions:

1. Emotion Loss:

$$L_{emotion} = -\sum_{i} y_i \log(\hat{y}_i)$$

Where y_i is the ground truth emotion label, and $\hat{y}_i$ is predicted probability for emotion i.

2. Intoxication Loss:

$$L_{intoxication} = -[\, y_{drunk} \log(\hat{y}_{drunk}) + y_{sober} \log(\hat{y}_{sober}) \,]$$

Where y_{drunk} and y_{sober} are the ground truth labels for intoxication, and $\hat{y}_{drunk}$ and $\hat{y}_{sober}$ are the predicted probabilities for drunk, sober states.

3. Total Loss:

$$L_{total} = L_{emotion} + L_{intoxication}$$

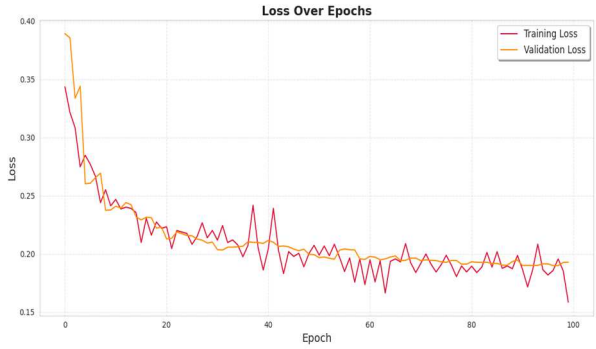

Fig. 4. Validation and Training Loss

The Confusion Matrix in Fig. 5, demonstrates the model's strong classification performance, strong diagonal dominance, validating consistent and reliable classification performance across emotional categories. The output layer, supported by activation functions like ReLU and softmax, classifies emotions such as angry, happy, sad, etc., and intoxication states as drunk or sober.

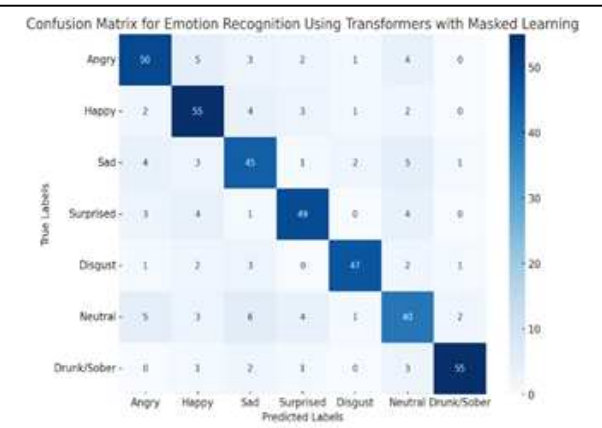

Fig. 5. Confusion Matrix

This capability is reflected in Fig. 6, which presents data from the Emotion Recognition and Detection input and output processes, demonstrating seamless real-time integration and effective classification of both emotional states and intoxication levels.

Fig. 6. Emotion Recognition and Detection

TABLE I. COMPARISON WITH OTHER ALGORITHMS

Algorithm used	Datasets used	Accuracy (%)
Proposed: DrunkDetect (facial emotion, masked learning)	FER-2013, FER+	83.0
CNN with weighted average fusion	FER-2013	74.14
Emotion classification + probabilistic labelling	FER-2013, VGG16, VGG19	76.62
Attentional Convolutional Network	FER-2013, CK+, FER	70.02
Deep Convolutional Neural Network (DCNN)	FER-2013	70.14

Metrics Calculations:

For example, to calculate precision and recall for "Drunk":

- Precision for Drunk: Precision is the ratio of correctly predicted drunk/sober instances out of all instances predicted as drunk/sober.

$$\text{Precision (Drunk)} = \frac{55}{55+7} = \frac{55}{62} \approx 0.887$$

- Recall for Drunk: Recall is the ratio of correctly predicted drunk/sober instances out of all actual drunk/sober instances.

  True Positives (TP) = 55
  False Negatives (FN) = 0 + 1 + 2 + 1 = 4 (Actual drunk/sober instances that were incorrectly classified as other emotions)

$$\text{Recall (Drunk)} = \frac{55}{55+4} = \frac{55}{59} \approx 0.931$$

- F1 Score for Drunk: The F1 score represents the harmonic mean of precision and recall, balancing both indicators.

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

$$\text{F1 Score (Drunk)} = 2 \times \frac{0.887 \times 0.931}{0.887+0.931} \approx 0.909$$

V. CONCLUSION

The DrunkDetect is a Vision Transformer-based model with masked learning for facial emotion and intoxication redness, outperforming CNN-based approaches. The architecture incorporates face and iris detection, MAE pretraining, and dual-task optimization. Designed for real-time performance, the system holds strong potential for healthcare applications. detection. Leveraging iris-region features such as dilation & the system holds strong potential for public safety and healthcare applications. Future enhancements include edge deployment and temporal sequence analysis.

REFERENCES

[1] Nabeel N Ali, Adnan M Abdulazeez: "Facial Emotion Recognition Based on Deep Learning: A Review", in Journal of Soft Computing and Data Mining conference Vol. 2, No. 1, 2024, doi: 10.30880/jscdm.2021.02.01.006

[2] Hien Le Nhu; Huy Vu Dang; Huan Hoang Xuan: "Facial Emotion Recognition by Combining Deep Learning and Averaged Weight of Face-Regions", in International Conference on Knowledge Science and Engineering (KSE), 2023, doi: 10.1109/KSE59128.2023.10299482

[3] afiz Arslan Ramzan, Ahmed Sohaib, Sadia Ramzan: "Facial Emotion Recognition using Deep Learning (FERDL)" in 25th International Multitopic Conference (INMIC), 2023, doi: 10.1109/INMIC56179.2023.10466203

[4] Benisha S, Mirnalinee TT: "Human Facial Emotion Recognition using Deep Neural Networks", The International Arab Journal of Information Technology, Volume 20, Issue 3, (pages 303-309), 2023, doi: 10.34028/iajit/20/3/2

[5] Muhammad Sajjad, Mohib Ullah: "A comprehensive survey on deep facial expression recognition; challenges, applications and future guidelines", Alexandria Engineering Journal 68(6):817-840, April 2023, doi: 10.1016/j.aej.2023.01.017

[6] Shubanjay Pandey, Sonakashi Handoo, Yogesh: "Facial Emotion Recognition using Deep Learning", 2022 International Mobile and Embedded Technology Conference (MECON), doi: 10.1109/mecon53876.2022.9752189

[7] Sowmiya R, Sivakamasundari G, Archana V: "Facial Emotion Recognition using Deep Learning Approach", 2022 International Conference on Advanced Computing and Robotics Systems (ICACRS), doi: 10.1109/ICACRS55517.2022.10029092

[8] Nyle Siddiqui, Thomas Reither, Rushit Dave, Dylan Black: "A Robust Framework for Deep Learning Approaches to Facial Emotion Recognition and Evaluation", 2022 Asia Conference on Algorithms, Computing, and Machine Learning (CACML), doi : 10.1109/CACML55656.2022.00019

[9] Nitesh Banskota, Abeer Alsadoon, P.W.C. Prasad: "A Novel Enhanced Convolution Neural Network with Extreme Learning Machine: Facial Emotional Recognition in Psychology Practices", Multimedia Tools and Applications in 2022, doi: 10.1007/s11042-022-13567-8

[10] Paras Jain, M. Murali, Amaan Ali: "Face Emotion Detection Using Deep Learning", 2021 International Conference on Computer Communication and Informatics (ICCCI), doi: 10.1109/ICCCI50826.2021.9402581.

[11] Aakash Saroop, Pathik Ghugare, Sashank Mathamsetty: "Facial Emotion Recognition: A multi-task approach using deep learning", arXiv on October 28, 2021, doi: 10.48550/arXiv.2110.15028

[12] Akriti Jaiswal, A. Krishnama Raju, Suman Deb: "Facial emotion Detection Using Deep Learning", in 2020 International Conference for Emerging Technology (INCET), doi: 10.1109/INCET49848.2020.9154121

[13] E. Pranav, Suraj Kamal, M.H. Supriya: "Facial Emotion Recognition Using Deep Convolutional Neural Network", 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, doi: 10.1109/ICACCS48705.2020.9074302

[14] Shervin Minaee, Amirali Abdolrashidi: "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network", in Sensors, 2019, doi: 10.3390/s21093046

[15] Sinno Jialin Pan; Qiang Yang: "A Survey on Transfer Learning", in IEEE Transactions on Knowledge and Data Engineering, 2010, doi: 10.1109/TKDE.2009.191