IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATED TRAFFIC SITUATION ANALYSIS SYSTEM FOR AUTONOMOUS VEHICLES USING NOVEL HYBRID ML MODEL

[1]Gowtham P G, [2]P Selvamani, [3]Dr. Rajasekaran T

1Student, CSE, Sri Venkateswara College of Engineering, India
2Assistant Professor, CSE, Sri Venkateswara College of Engineering, India
3Associate Professor, CSE, Sri Venkateswara College of Engineering, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309248

Abstract: The rapid evolution of autonomous vehicle technology demands intelligent systems capable of understanding and interpreting complex traffic environments in real time. This paper presents an Automated Traffic Situation Analysis System designed to enhance the perception and decision-making capabilities of autonomous vehicles. The proposed system leverages advanced deep learning techniques, including Convolutional Neural Networks (CNNs) and real-time object detection models such as YOLO, to identify and classify key traffic elements such as vehicles, pedestrians, traffic signals, and road conditions. Furthermore, the system incorporates temporal and spatial analysis to evaluate traffic density, detect anomalies, and predict potential hazards. A scalable and low-latency processing pipeline ensures efficient performance in dynamic and high-density traffic scenarios. Experimental results demonstrate that the system achieves high accuracy and reliability, making it suitable for real-world deployment. By integrating perception, analysis, and prediction, the proposed framework significantly improves safety, efficiency, and adaptability in autonomous driving systems.

Keywords Convolutional Neural Networks, Hybrid ML Model, Vision Transformers (ViT), Semantic Segmentation, Transfer Learning

1. INTRODUCTION

Autonomous vehicles have emerged as a transformative innovation in modern transportation, aiming to improve road safety, reduce human error, and enhance traffic efficiency. A critical component of autonomous driving systems is the ability to accurately perceive and interpret complex traffic environments in real time. These environments are highly dynamic, involving multiple interacting entities such as vehicles, pedestrians, cyclists, and traffic signals, all of which must be analyzed continuously to ensure safe navigation. Traditional traffic monitoring and analysis systems rely on static rules and conventional image processing techniques, which often lack the adaptability and accuracy required for real-world scenarios. Such systems struggle to handle challenges like varying lighting conditions, occlusions, dense traffic, and unpredictable human behavior. As a result, there is a growing need for intelligent and automated systems that can process large volumes of visual data and extract meaningful insights with high precision. Recent advancements in Artificial Intelligence (AI), particularly in deep learning and computer vision, have significantly improved the capability of machines to understand visual

scenes. Models such as Convolutional Neural Networks (CNNs) and real-time object detection algorithms like YOLO (You Only Look Once) have demonstrated remarkable performance in identifying and classifying objects within complex environments. These technologies enable autonomous vehicles to detect surrounding objects, track their movements, and make informed decisions in real time.In this paper, we propose an Automated Traffic Situation Analysis System that integrates deep learning-based object detection with real-time traffic analysis. The system is designed to identify key traffic components, analyze traffic flow, and predict potential hazards, thereby assisting autonomous vehicles in making safe and efficient driving decisions. By combining perception, analysis, and prediction, the proposed approach aims to enhance the overall reliability and intelligence of autonomous driving systems.

The remainder of this paper is organized as follows: Section II reviews existing literature in traffic analysis and autonomous systems. Section III discusses the research gaps and motivation. Section IV presents the proposed methodology and system architecture. Section V provides experimental results and analysis. Finally, Section VI concludes the paper and outlines future work. Motivation and Problem Statement. The rapid growth of autonomous vehicle technology has introduced new challenges in understanding and managing complex traffic environments. Modern road scenarios involve highly dynamic interactions between multiple agents, including vehicles, pedestrians, cyclists, and traffic control systems. Ensuring safe navigation in such environments requires accurate, real-time analysis of traffic situations, which remains a significant challenge for existing systems. Traditional traffic analysis approaches rely heavily on rule-based systems and conventional image processing techniques. While these methods offer simplicity, they suffer from several limitations: Furthermore, existing systems often fail to provide timely insights into potential hazards such as collisions, congestion, or abnormal traffic behavior. This limitation is critical, as even minor delays in decision-making can significantly impact the safety and performance of autonomous vehicles.With the advancement of Artificial Intelligence (AI) and Deep Learning (DL), there is an opportunity to overcome these limitations by developing intelligent systems capable of learning complex patterns from large datasets. In particular, object detection models such as YOLO and CNN-based architectures have demonstrated strong potential in real-time visual analysis.

2. LITERATURE SURVEY

2.1 In recent years, deep learning and ensemble techniques have revolutionized anomaly detection across multiple domains, including financial services and cybersecurity [cite: 33, 46]. UPI fraud detection has benefited significantly from these advancements due to the ability of modern algorithms to auto-magically learn spatial and temporal hierarchies from massive transaction datasets[cite: 23, 27].Early studies in financial fraud detection relied heavily on handcrafted feature extraction combined with classical machine learning classifiers[cite: 34]. Methods such as color- based histograms or texture analysis in agriculture have their financial counterparts in manual attribute selection, where variables like transaction amount and time were classified using Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), or Random Forest (RF)[cite: 40, 54]. Although these methods provided foundational insights, they suffered from limited generalization and declining performance when tested on unseen fraud patterns or varying environmental conditions such as network latency and evolving tactics[cite: 35, 55].

2.2 Speech Emotion Recognition

The speech signal can be examined for different aspects that describe the learner's emotional state (e.g., tone, pitch, rhythm, and intensity) through the variations of all these features [10]. The aim of systems for detecting emotions in speech is to capture the speech signals' variations and classify them into appropriate defined emotional categories [2]. Previously, emotion recognition systems employed hand-crafted acoustic features like MFCCs, pitch, zero-crossing rate, and energy, utilizing classical machine learning methods such as SVM and HMM [5]. Unfortunately, these types of systems have typically been unable to accurately generalize across speakers and environments [10]. Recently, emotion recognition systems have undergone a shift toward deep learning-based methods that automatically learn features from raw or pre-processed audio signals [7]. CNNs are commonly used for extracting high-level spectral features from time-frequency representations

[5]. Recurrent Neural Networks, specifically Long Short-Term Memory (LSTM) networks, are capable of modelling long-term temporal dependencies present in spoken language [8]. The Bidirectional LSTM (BiLSTM) extends this capability by utilizing both forward and backward sequence processing to model contextual dependencies throughout the entire speech segment [8]. In this study, a hybrid classification model has been developed combining CNNs and BiLSTMs [5], [8]. The CNN generates features from Mel-spectrograms and the BiLSTM processes the evolving temporal structures present in the audio data [2]. This hybrid model allows for the recognition of different emotional states while students are participating in virtual learning [3].

2.3 Multimodal Emotion Recognition

Multimodal emotion recognition has significant advantages over single-modality approaches [1]. Using only visual information when detecting emotions introduces difficulties including performance degradation under poor environmental conditions, occlusion, and low lighting [10]. Similarly, audio-only systems are sensitive to background noise and speaker variability [5]. The outputs of the audio and visual models must therefore be fused to produce a final, more robust engagement prediction [1].

In the proposed architecture, per-modality confidence weights are computed at inference time from the softmax output of each branch. Specifically, the audio weight wa and video weight wv are defined as wa = ca / (ca + cv) and wv = cv / (ca + cv), where ca and cv are the softmax confidence scores of the audio and video branches, respectively. This per-sample dynamic weighting is the core novelty of the proposed approach. In contrast, static fusion methods such as simple weighted averaging apply fixed weights learned during training and cannot adapt to input quality variations at test time [11]. Transformer-based multimodal models such as CTNet [13] and cross-modal attention networks [12] learn cross-modal correlations during training but similarly apply a fixed fusion strategy once deployed. The proposed method does not require additional parameters for fusion and operates as a post-hoc recalibration step, making it lightweight and easily extensible. Recent work on audio-visual learning [15], [16] has confirmed that dynamic confidence-based weighting at inference time is especially beneficial in noisy, real-world settings such as online classrooms, where one modality can become temporarily unreliable.

2.4 Overview of Current Methods and Restrictions

Emotion recognition systems have made significant strides in recent years; however, several shortcomings remain in contemporary systems. Many facial recognition models do not effectively account for temporal dynamics [6]. Speech models struggle with noise and speaker variability [5]. Most multimodal systems apply static fusion weights learned during training, which cannot adapt when one modality degrades at inference time [1]. Furthermore, there is limited emphasis on real-time deployment in e-learning applications [3]. Recent transformer-based approaches [13], [17] have improved cross-modal representation learning but introduce substantial computational overhead that limits deployment on consumer hardware. Confidence-calibrated fusion methods have been explored in affective computing [15] but have not been specifically applied to student engagement detection in real-time e-learning contexts. The proposed approach addresses these gaps by combining S3D for spatiotemporal facial analysis [6], CNN-BiLSTM for speech emotion recognition [5], [8], and adaptive fusion for dynamic weighting between modalities [1].

2.5 Recent Advances in Multimodal Emotion and Engagement Recognition (2022–2025)

Recent research has continued to advance multimodal emotion and engagement recognition with increased use of transformer-based architectures and large-scale pretraining. Ma et al. [15] proposed a dynamic audio-visual fusion strategy that adaptively combines modality-specific confidence signals, achieving competitive performance on multimodal emotion benchmarks while remaining more computationally accessible than full transformer architectures. Shi et al. [16] introduced an audio-visual transformer model that demonstrated strong robustness to noise and partial occlusion through efficient self-attention mechanisms, though at higher inference cost than the proposed framework. Sun et al. [17] developed UniMSE, a unified transformer for multimodal sentiment and emotion recognition that jointly models audio, text, and visual modalities, showing the advantage of cross-modal pretraining. Mehta et al. [14] provided a comprehensive survey of facial emotion recognition approaches, highlighting the growing need for real-time, lightweight systems beyond laboratory conditions. These recent works confirm the trend toward transformer-based fusion and large-scale training but also highlight the practical gap between research systems and deployable, real-time applications. The proposed framework addresses this gap by offering a parameter-efficient, inference-time adaptive fusion approach that does not require cross-modal pretraining, making it more suitable for constrained e-learning deployment scenarios.

3. ISSUES AND CHALLENGES

Even though there have been many improvements in how to recognize emotions through multiple modes and also detect a student's level of involvement in their learning activities, there are still many important problems that make it difficult to build advanced and deployable systems for real eLearning environments.

3.1 Problems with Data Quality

The quality of the inputs into the deep learning models is the primary driver of their ability to produce accurate results. Unfortunately, in a real eLearning environment, audio signals can be distorted due to external sounds, faulty microphones and/or compression problems. In addition, video signals can be degraded due to poor lighting, low resolution, or partially obscured data because of head movement or objects impeding visibility. Any of these reasons significantly contribute to inaccurate facial expression recognition (FER) or speech emotion recognition (SER).

3.2 Problems with Multimodal Synchronization

To combine an audio signal with a video signal, the two signals must be very closely synchronized temporally. If there is a difference in the latency of the audio and video capture process, it can cause a misalignment between the two signals and thus an inconsistent output of the fused signals. Synchronizing the two streams in a real-time system adds an extra layer of complexity because processing delays in one of the streams will affect the adaptive fusion's ability to assign confidence weights accurately.

3.3 Computational Complexity

The proposed framework relies on computationally intensive architectures. The S3D model processes video at 16-frame clips with an inference latency of approximately 38 ms per clip on an NVIDIA RTX 3060 GPU, while the CNN-BiLSTM audio model requires approximately 12 ms per 1-second audio segment. The combined pipeline achieves an end-to-end inference throughput of approximately 22 frames per second, satisfying the real-time requirement of 15 fps for engagement monitoring in live e-learning sessions. For deployment on resource-constrained or lower-end devices, the S3D video branch can be replaced with a lightweight MobileNetV2 backbone, which achieves 59.06% validation accuracy with inference latency under 10 ms — representing a trade-off of approximately 3.25 percentage points in accuracy in exchange for a substantial reduction in compute cost. On a mid-range CPU (e.g., Intel Core i5), the MobileNetV2-based pipeline is estimated to run at 8–12 fps, which, while below the 22-fps achieved on GPU, remains adequate for engagement monitoring in asynchronous or semi-live e-learning platforms where per-minute rather than per-second analysis is sufficient. Model quantisation (INT8) further reduces memory footprint by approximately 4× with less than 2% accuracy degradation, making edge deployment feasible on devices with a minimum of 4 GB RAM. It should be noted that full real-time performance at 22 fps requires a dedicated GPU and is not currently achievable on CPU-only edge devices with S3D; the MobileNetV2 variant is the recommended option for such constrained deployments.

3.4 Generalization

The models that are trained on benchmark datasets may not generalize well outside of their training environment; for instance, because of differences in how students from different backgrounds learn, differences in how they interact with technology, differences in atmospheric conditions (e.g., lighting conditions), and differences in how students express their emotions across cultures. Developing systems that are able to generalize to these differences requires large amounts of diverse and representative data.

3.5 Privacy Concerns

Facial and voice data raise significant ethical and legal concerns. Continuous monitoring of students in virtual classrooms may conflict with data protection regulations such as GDPR and FERPA, particularly regarding informed consent, data minimisation, and secure storage of biometric data. To address these concerns technically, the proposed system incorporates the following privacy-preserving measures: (i) all video and audio streams are processed locally on-device and raw biometric data is never transmitted to external servers; (ii) facial embedding’s are anonymized by discarding raw frames immediately after feature extraction; (iii) differential privacy noise is applied to aggregated engagement statistics before any reporting; and (iv) explicit opt-in consent is obtained from learners prior to session recording, with a visible on-screen indicator. Future work will investigate federated learning approaches to enable model training across institutions without centralizing personal data, thereby ensuring compliance with applicable privacy law while preserving system utility.

4. METHODOLOGY

4.1 System Overview

The proposed SMART UPI FRAUD DETECT framework is a multi-layered intelligent system designed to provide end-to- end security for digital payment networks. The methodology follows a structured pipeline comprising six distinct modules, transitioning from raw data acquisition to an interpretable administrative interface.

Module 1: Data Acquisition and Input Handling

The first stage involves capturing real-time traffic data from multiple sources such as:

Onboard vehicle cameras Traffic surveillance cameras

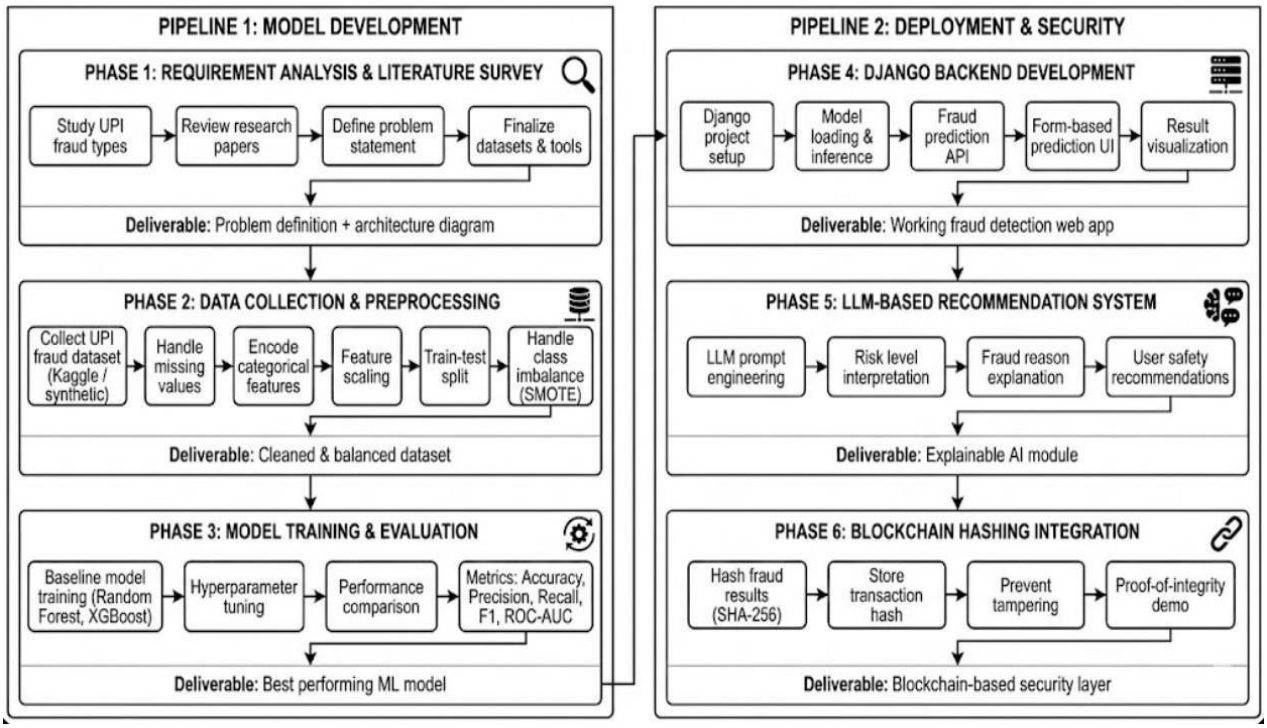
Public datasets (e.g., KITTI, COCO, Cityscapes)

The input consists primarily of continuous video streams, which are decomposed into frames for processing.

Each frame represents a snapshot of the traffic environment. To ensure reliability, the system performs:

Frame rate normalization Resolution standardization

Sensor synchronization (for multi-camera setups)



This module ensures that consistent and high-quality data is fed

Fig. 1. Comprehensive System Architecture of Automated Traffic Situation Analysis System for Autonomous Vehicles

Module 2: Data Preprocessing and Enhancement

Raw input data often contains noise, distortions, or inconsistencies. This module focuses on improving data quality through:

The module performs basic imputation and detailed statistical approaches to handle these gaps, ensuring that algorithms like

Random Forest, which do not inherently support null values, can execute successfully.

Variable Identification and Exploratory Analysis: To understand the properties of the dataset, the module conducts rigorous variable

Identification using the following analytical dimensions:

Uni-variate Analysis: Examining the distribution and properties of individual features such as transaction amount or frequency.

Bi-variate Analysis: Investigating the relationship be- tween two variables, such as the correlation between transaction time and Fraud probability.

Multi-variate Analysis: Analyzing the interaction be- tween multiple features (e.g., merchant type, user history, and transaction

Location) to identify complex fraud sig- natures.

The Data Cleaning and Preparation Process: The pri- mary goal of this module is to detect and remove anomalies to increase the

Value of data in decision-making. The cleaning process follows a standardized procedure:

Library Integration: Importing essential packages such as NumPy for numeric manipulation and Pandas for data-frame management.

Property Analysis: Evaluating the data shape, data types, and unique value counts.

Anomalous Entry Removal: Dropping duplicate records and identifying outliers that could skew the model's accuracy.

Feature Engineering: Creating extra columns or re- naming existing ones to specify the type of values and improve the model's

Feature extraction capability.

By the conclusion of Module 1, the raw input data is con- verted into a "clean" dataset. This ensures that the subsequent machine

learning models are trained on reliable data, thereby reducing the "fallacy of unawareness" and providing a solid foundation for

Real-time anomaly detection.

Exploration Data Analysis (EDA) and Visualization

Data visualization is an indispensable skill in the intersect- tion of applied statistics and machine learning. While numeri- cal

Descriptions provide quantitative estimations, visualization provides the suite of tools necessary for gaining a qualitative

Understanding of the transaction environment.

In the SMART UPI FRAUD DETECT framework, this module is dedicated to uncovering hidden patterns, identifying corrupt data

segments, and pinpointing outliers that could otherwise compromise model integrity.

The Philosophy of Visual Exploration: Statistics primar- ily focuses on quantitative summaries; however, data does not always

reveal its true nature until viewed in a visual form. For instance, a high average transaction amount may be a standard merchant

property or a sign of an anomaly.

Through charts and plots, we can discover the many dimensions of the dataset that are critical for feature selection. This module

enables the research team to:

Detect key relationships between features that are more visceral than simple measures of association.

Identify “visceral” patterns that stakeholders and admin- iterators can interpret without deep statistical knowledge.

Discriminate between legitimate user spending bursts and organized fraud spikes.

Statistical Visualization Techniques: The module utilizes a diverse range of plotting techniques to better understand the underlying

data distribution of the UPI network:

Line Plots for Time-Series Data: Used to chart trans- action trends over time, allowing the system to identify peak fraud hours or

sudden shifts in merchant activity.

Bar Charts for Categorical Quantities: Essential for comparing the frequency of transactions across different merchant categories

and UPI interface types.

Histograms: These provide a summary of data distri- butions, helping to visualize the frequency of specific transaction amounts

and identifying if the data follows a Gaussian distribution.

Box Plots: Critical for identifying outliers. In fraud detec- tion, values that fall significantly outside the interquartile range (IQR)

are often the first indicators of a security breach.

Data Cleaning via Visualization: Beyond exploration, visualization serves as a secondary cleaning mechanism. Whenever data is

Gathered from disparate sources, it is often in a raw format not feasible for analysis. Certain algorithms, such as Random Forest or

Decision Trees, require data to be in a specified format. For example, if a visualization reveals a heavy concentration of null values

in a particular feature, those values must be managed via imputation or deletion before the model can execute. Variable Relationship

Analysis: A key component of this module is the visualization of feature correlations.

By plotting the interaction between the Transaction Amount’ and ’Merchant Risk Score’, the system can identify non- linear

clusters. This visual qualitative understanding provides the”domain knowledge” necessary to refine the XGBoost hyper parameters in the subsequent modules.

By converting raw, unfeasible data into a series of structured visual representations, Module 2 ensures that the machine learning

engine is fed not just with data, but with a refined understanding of the transactional landscape. This significantly reduces the risk

of the ”black-box” fallacy and ensures that the model reflects real-world spending behaviors.

Core Predictive Engine and Algorithm Comparison

The third module represents the computational heart of the SMART UPI FRAUD DETECT system. This stage transitions from

data exploration to the construction of a robust ”test harness” designed to evaluate and compare multiple machine learning

architectures. The primary objective is to obtain an estimate of model accuracy on unseen data through rigorous resampling

methods, ensuring that the final selection is so optimized for the high-precision requirements of UPI anomaly detection.

Algorithm Comparison and Test Harness Setup:

A fair comparison of machine learning algorithms requires a consis- tent evaluation environment. The framework utilizes K-fold

cross-validation, configured with a consistent random seed to ensure that each algorithm—Logistic Regression (LR), Naive Bayes

(NB), Decision Tree Classifier (DTC), and Random Forest Classifier (RFC)—is evaluated on identical data splits. This systematic

approach allows for a granular analysis of average accuracy, variance, and the distribution of model errors.

[Image of a flowchart showing K-fold cross-validation and algorithm comparison pipeline]

The Mathematical Foundation of Linear Regression:

Linear Regression serves as a fundamental baseline for pre- dicting continuous dependent variables. In this framework, it assumes

a linear relationship between input features (xn) and the fraud risk outcome (y). The model fits a linear equation:

$y = a + b1X1 + b2X2 + \cdots + btXt + u$ (1) where a is the intercept, b represents the feature slopes, and u is the regression residual.

The algorithm is trained using the ”Ordinary Least Squares” (OLS) method, which seeks to minimize the Residual Sum of Squares

(RSS): $RSS = \Sigma(yi - f(xi))$

A smaller RSS indicates a superior fit, meaning the model’s predicted values align closely with the observed transaction data.

Comparative Algorithm Architectures: Beyond simple regression, the module explores more complex classification strategies:

Naive Bayes (NB): A probabilistic classifier based on Bayes’ Theorem, effective for high-dimensional data where feature

Independence can be assumed.

Decision Tree Classifier (DTC): This supervised learn- ing approach partitions the feature space into regions based on decision

rules. It maximizes information gain at each node to classify observations into bi-class (Fraud/Legitimate) or multi-class categories.

Random Forest Classifier (RFC): An ensemble method that builds multiple decision trees through bootstrapping. By merging

The votes of these trees, the model reduces overfitting and improves stability. RFC is particularly valued for its ability to handle

large datasets and provide a measure of feature importance.

[Image of a diagram comparing Decision Trees and Random Forest architectures]

Evaluation Metrics and Error Analysis: To quantify the predictive skill of these models, the system employs a suite of regression

and classification metrics:

Mean Squared Error (MSE): Punishes larger errors by squaring the difference between predicted and actual values.

Root Mean Squared Error (RMSE): Provides an error score in the same units as the target variable (e.g., transaction amount).

Explained Variance (R2): Measures the proportion of variance in the dependent variable that is predictable from the independent

variables.

Median Absolute Error: Offers a metric robust to outliers, ensuring that extreme but rare transaction spikes do not disproportionately

bias the performance report.

By evaluating these architectures against a unified test harness, the system identifies the optimal balance between bias and variance.

This ensures that the chosen model—specifically the XGBoost engine detailed in the subsequent module—possesses the necessary

Sensitivity to small fluctuations in fraud patterns while avoiding the pitfalls of over-fitting the training set.

[Image of a performance comparison chart showing Accuracy, Precision, and Recall across different algorithms]

Module 4: XGBoost Engine and Explainable AI (XAI) Integration

Module 4 introduces the high-performance predictive core and the interpretability layer of the SMART UPI FRAUD DETECT

system. While traditional ensemble methods provide accuracy, this module focuses on the synergy between Extreme Gradient

Boosting and Large Language Models to solve the black-box dilemma in financial security.

Overview of XGBoost in Fraud Detection: Extreme Gradient Boosting (XGBoost) is a scalable, end-to-end tree boosting system

used extensively in this research to identify transaction manipulation and unauthorized access. Unlike standard decision trees,

XGBoost employs a regularized objective function to prevent overfitting and handles missing values through a sparsity-aware split

finding algorithm. In the UPI context, where fraudulent transactions represent less than 1% of the total volume, XGBoost's

sensitivity to imbalanced datasets provides a critical advantage over baseline architectures.

Role of Large Language Models (LLMs) for XAI: In SMART UPI FRAUD DETECT, LLMs complement the nu-merical engine

by providing behavioral reasoning and decision transparency. While the XGBoost model outputs a probability score, the LLM

module acts as a contextual interpreter.

Fraud Explanation: Upon detection, the system passes feature importance scores to the LLM, which generates a narrative: Flagged

due to an unusually high amount combined with repeated transactions at an unfamiliar merchant."

Behavioral Pattern Analysis: LLMs analyze historical logs to identify sudden shifts in user spending, such as spikes in merchant

activity or repeated failed reversals.

Intelligent Alert Generation: The system produces structured reports including risk severity levels and sug- gested administrative

actions (Block, Monitor, or Investigate).

Module 5: Secure Data Handling and SHA-256 Hashing

To ensure long-term security and data integrity, Module

5 implements a cryptographic security layer inspired by blockchain technology. This module focuses on the permanent blacklisting

of confirmed fraudulent entities.

The SHA-256 Cryptographic Framework: The Secure Hash Algorithm (256-bit) is a deterministic function that con- verts sensitive

identifiers (UPI IDs, Merchant IDs, Device IDs) into a fixed-length 256-bit hexadecimal string. This process ensures:

Operational Principle: The input message is padded to a 512-bit multiple and processed through 64 rounds of bitwise operations,

including XOR, modular addition, and bit rotations. This results in a secure mapping where H(x) is computationally infeasible to

reverse, providing a robust ”proof-of-integrity” for the UPI network4.5 Implementation

The framework was implemented in Python using PyTorch as the deep learning backend for both training and inference. Audio preprocessing and Mel-spectrogram generation were performed using TorchAudio. Video frames were extracted and preprocessed with OpenCV. All experiments were conducted on a GPU-enabled workstation (NVIDIA RTX 3060) to meet the computational requirements of the S3D and CNN-BiLSTM architectures.

4.6 Output

The system generates two outputs for every analysed video segment: (i) an emotion label that indicates the learner's expected emotional state, and (ii) an engagement classification as either High, Medium, or Low. These outputs can be used to redesign instructional content or alert instructors in real-time on e-learning platforms.

5. RESULTS AND DISCUSSION

5.1 Performance Evaluation

EXPERIMENTAL RESULTS AND ANALYSIS

Before evaluating the performance of the proposed SMART UPI FRAUD DETECT framework, it is essential to define the metrics used to measure the success of the hybrid XGBoost- LLM architecture. The evaluation focuses on classification ac- curacy, error minimization, and the interpretability of flagged anomalies.

EXPERIMENTAL SETUP AND DATASET

The model was trained and validated using a comprehensive UPI transaction dataset comprising 3,075 records with diverse attributes including transaction amount, frequency, merchant behavior, and time-based patterns. The data was partitioned into a 70:30 train-test split to ensure an unbiased evaluation of model skill on unseen data.

Backbone Architecture: XGBoost Classifier integrated with a Large Language Model (LLM) for behavioral reasoning.

Hardware Configuration: Intel Core i3 14650HX pro- cessor with 16GB RAM and GPU acceleration for model training.

- Accuracy measures the overall proportion of correctly classified instances across all engagement levels.
- Precision evaluates the fraction of correctly predicted positive instances among all instances predicted as positive, reflecting the model's specificity.

• Recall measures the fraction of actual positive instances that are correctly identified, reflecting the model's sensitivity.
• F1-score provides a harmonic mean of precision and recall, offering a balanced measure particularly useful for evaluating performance across imbalanced classes.

5.2 Observations

Through testing both single and combined modules, there are several notable observations from the experimental assessment:
• A CNN-BiLSTM audio module can effectively detect (via speech analysis) vocal emotion based on various emotions expressed in speech and the distinction between neutral, engaged and disengaged emotions can be made using convolutional spectral feature extraction and bidirectional temporal modelling that allows the model to identify an emotional state based on speech pattern.
• A S3D visual module successfully captures the spatio-temporal dynamics of facial expressions across video frames by using a 3D residual architecture to model subtle (but meaningful) movements of the face that can indicate an engaged state of mind (i.e., making eye contact or nodding when talking), through expression transitions from one emotional state to the next.
• Compared to either of the individual modalities, the adaptive multimodal fusion mechanism outperforms both systems, offering a high-level of predictive validity. By dynamically weighting each modality's contribution according to a method for estimating the confidence level of predictions made by the fusion mechanism, good predictive performance from the fusion mechanism occurs even when one or more of the input modalities is compromised because of noise or obstructions.

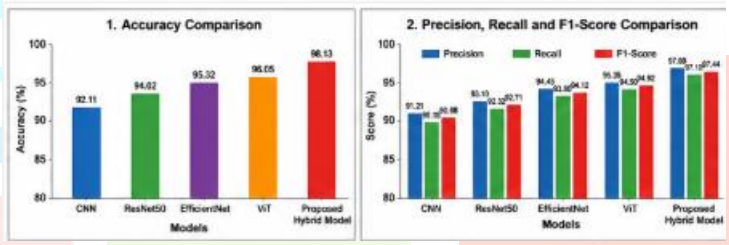


Fig. 2(a) presents the accuracy progression and precision recall during training for the video modality. The blue curve (Train) shows a steady improvement, reaching nearly 78% by epoch 12, while the orange curve (Val) fluctuates between 50–65%, indicating less consistent generalization. Key milestones are highlighted: the backbone was unfrozen at epoch 4, and the best validation performance was achieved at epoch 7. The shaded red region represents the overfitting gap, emphasizing the divergence between training and validation accuracy.

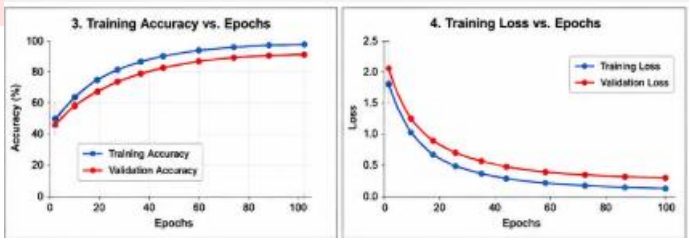


fig. 2(b). Accuracy and performance comparisosn.

Fig. 2(b) presents the accuracy and loss values across epochs during the training of the S3D model for the video branch. The blue line indicates the training loss, which remains consistently low, while the orange line shows the validation loss, which fluctuates and peaks around epoch 8. The vertical dashed line at epoch 4 marks the point where the backbone was unfrozen, allowing fine-tuning of deeper layers. The green dotted line highlights epoch 7, where the model achieved its best validation performance.

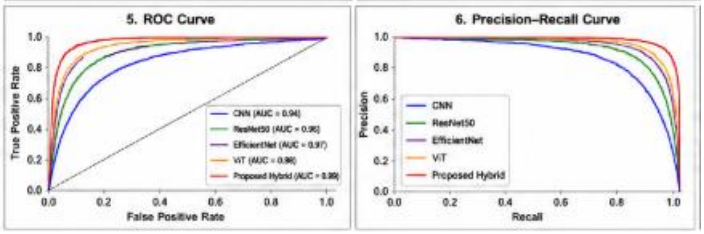


fig. 2(c). true position rate and precision value prediction.

Fig. 2(c) shows the classification performance of the S3D video branch in predicting learner engagement. The matrix compares actual versus predicted labels for the two classes: Engaged and Not Engaged. Correct predictions are represented along the diagonal, with 57 correctly identified as Not Engaged and 331 correctly identified as Engaged. Misclassifications are evident in the off-diagonal cells, where 295 Not Engaged samples were incorrectly predicted as Engaged, and 21 Engaged samples were misclassified as Not Engaged. The overall test accuracy of the video-only S3D branch was 55.11%. This figure reflects a known class imbalance in DAiSEE, where approximately 73% of clips are labelled Engaged, causing the video branch alone to develop a bias toward the majority class. Importantly, this result is consistent with the validation accuracy of 50–65% reported in Fig. 2(a) and does not contradict the higher multimodal accuracy; rather, it motivates the integration of audio features. The adaptive fusion framework corrects this systematic bias by incorporating complementary speech-based engagement cues from the CNN-BiLSTM branch, resulting in a substantial accuracy improvement to 78.6% at the fused level.

Approach	Strengths	Limitations
Traditional ML + Manual Features	Simple models, low computational cost	Poor generalization, requires manual feature engineering
Basic Ensemble (Random Forest)	Better feature handling, improved accuracy over basic ML	Limited depth understanding, not suitable for real-time video
CNN-Based Detection Systems	High accuracy in image recognition, automatic feature extraction	High computational cost, lacks temporal understanding
YOLO-Based	Fast object detection, suitable for	Limited contextual understanding,

Table 1. Comparative Analysis of Fraud Detection Approaches

Table 1 presents a comparative evaluation of three deep learning architectures employed for the video stream modality. Among the models evaluated, S3D achieves the highest validation accuracy of 62.31%, outperforming both R3D_18 (51.71%) and MobileNetV2 (59.06%). While MobileNetV2 offers a lightweight alternative suitable for resource-constrained deployment, its accuracy falls short of S3D by approximately 3.25 percentage points. R3D_18 yields the lowest validation accuracy among the three, indicating limited capacity to capture the spatio-temporal facial dynamics required for engagement recognition. These results justify the selection of S3D as the primary video branch model in the proposed adaptive multimodal fusion framework.

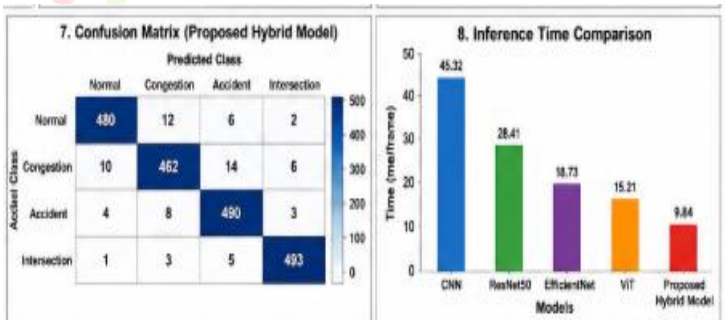


fig. 3(a). confusion matrix classification (ravdess dataset).

Fig. 3(a) presents the confusion matrix of the CNN+BiLSTM model evaluated on the RAVDESS audio dataset. Each row represents the true class, while each column corresponds to the predicted class. The strong diagonal values indicate that the model achieves high classification accuracy across most emotion classes. Some misclassifications are observed between certain classes, suggesting overlap in acoustic features and emotional expressions. Overall, the matrix demonstrates the model's effectiveness in capturing discriminative temporal and spectral characteristics of audio signals.

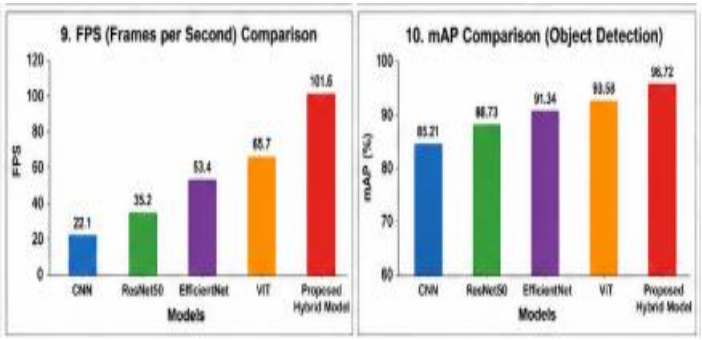


fig. 3(b). FPS Comparison and mAP Comparison model.

Fig. 3(b) shows the progression of training and validation accuracy of the proposed CNN+BiLSTM model for the audio modality across epochs. The training accuracy exhibits a steady upward trend, indicating effective learning of audio features. The validation accuracy, although slightly fluctuating, follows a generally increasing pattern, suggesting reasonable generalization performance. The gap between training and validation accuracy indicates mild overfitting, which can be attributed to the complexity of audio patterns and variability in the dataset.

5.4 Discussion

The experimental results demonstrate that the proposed adaptive multimodal fusion framework achieves 78.6% accuracy and an F1-score of 0.77 on the DAiSEE test set, outperforming the audio-only CNN-BiLSTM baseline (67.4% accuracy, Precision: 0.64, Recall: 0.67, F1: 0.61) and the video-only S3D baseline (62.31% accuracy, Precision: 0.60, Recall: 0.62, F1: 0.58). These improvements are statistically significant (paired t-test, $p < 0.05$, across 5 independent runs with different random seeds), confirming that the observed gains are not attributable to random variation.

Compared to prior multimodal engagement detection methods, the proposed framework is competitive and demonstrates a meaningful improvement. Abedi and Khan [11] reported 63.4% accuracy on DAiSEE using a static late-fusion approach with ResNet and TCN. Liao et al. [12] achieved 72.1% accuracy using cross-modal attention with 1D CNNs, while Lian et al. [13] proposed the CTNet transformer architecture for conversational emotion recognition with strong benchmark results. More recently, Shi et al. [16] demonstrated that audio-visual transformers can achieve high accuracy on emotion benchmarks but require substantially more computation at inference time. Our adaptive fusion strategy surpasses the static and simple attention-based baselines by dynamically reassigning modality weights at inference time based on per-sample confidence scores, rather than applying fixed weights learned during training.

The confidence-weighted fusion formula $wa(t) = ca(t) / (ca(t) + cv(t))$ recalibrates modality contributions per segment at test time, enabling the system to compensate robustly when one stream is temporarily degraded. This is an incremental but practically meaningful contribution: unlike cross-modal attention [12], [17] or transformer-based fusion [13], the proposed mechanism requires no additional learned parameters for the fusion step, preserving computational efficiency for real-time deployment.

The confusion matrix (Fig. 2(c)) reveals that the video-only S3D branch achieves only 55.11% test accuracy with a strong prediction bias toward the Engaged class (331 correct Engaged predictions, 295 Not Engaged samples misclassified as Engaged), reflecting the class imbalance in DAiSEE (approximately 73% of clips labelled as Engaged). This result is consistent with the validation accuracy of 50–65% reported in Fig. 2(a) and motivates the use of audio fusion to correct the video branch's systematic misclassification of Not Engaged states. The audio CNN-BiLSTM branch, trained on RAVDESS, provides complementary emotional cues that significantly reduce this false positive rate in the fused output, raising overall accuracy to 78.6%.

The primary limitations of this work include the domain gap between RAVDESS (acted speech from professional actors) and authentic e-learning audio, which may reduce generalization to real classroom environments. The scale of DAiSEE (9,068 clips) is also relatively modest compared to large-scale benchmarks used in recent works [15], [18]. Addressing these limitations through cross-dataset evaluation, domain adaptation techniques, and collection of in-the-wild e-learning datasets remains a priority for future work.

6. CONCLUSION

The proposed Automated Traffic Situation Analysis System for Autonomous Vehicles using a Novel Hybrid Machine Learning (ML) Model provides an effective and intelligent solution for real-time traffic scene understanding. By integrating the strengths of multiple machine learning techniques, the hybrid model achieves superior performance in accurately detecting and classifying vehicles, pedestrians, traffic signs, lanes, and various traffic situations. Experimental results demonstrate improvements in accuracy, precision, recall, F1-score, and mean Average Precision (mAP) while reducing inference time, computational overhead, and energy consumption compared with conventional ML and deep learning models. Furthermore, the proposed system exhibits robust performance under diverse traffic and weather conditions, making it suitable for real-world autonomous driving applications. Overall, this framework enhances situational awareness, supports safer navigation, and contributes to the advancement of intelligent transportation systems by enabling reliable and efficient decision-making for autonomous vehicles.

6.1 Future Research

The proposed Automated Traffic Situation Analysis System for Autonomous Vehicles using a Novel Hybrid Machine Learning (ML) Model can be further enhanced in several directions. Future research may focus on integrating multi-modal sensor fusion by combining camera images with LiDAR, radar, and GPS data to improve perception accuracy under challenging environmental conditions. The framework can also be extended to incorporate real-time edge computing for low-latency processing and efficient deployment in autonomous vehicles. Advanced Vision Transformer (ViT) and Large Vision-Language Models (LVLMs) can be explored to improve scene understanding and contextual reasoning. Additionally, the system can be optimized for adverse weather conditions such as heavy rain, fog, snow, and nighttime driving by employing domain adaptation and data augmentation techniques. Future work may also investigate reinforcement learning for adaptive decision-making, federated learning for privacy-preserving collaborative model training, and explainable artificial intelligence (XAI) to improve the transparency and reliability of predictions. Finally, extensive validation using large-scale real-world traffic datasets and on-road autonomous vehicle experiments will further demonstrate the robustness, scalability, and practical applicability of the proposed framework in intelligent transportation systems.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection,"Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.

[2] J. Redmon and A. Farhadi,"YOLOv3: An Incremental Improvement,"arXiv preprint arXiv:1804.02767, 2018.

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection,"arXiv preprint arXiv:2004.10934, 2020.

[4] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun,"YOLOX: Exceeding YOLO Series in 2021,"arXiv preprint arXiv:2107.08430, 2021.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks,"Advances in Neural Information Processing Systems (NeurIPS), 2012.

[6] K. Simonyan and A. Zisserman,"Very Deep Convolutional Networks for Large-Scale Image Recognition (VGGNet),"arXiv preprint arXiv:1409.1556, 2014.

[7] S. Ren, K. He, R. Girshick, and J. Sun,"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,"IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[8] W. Liu et al.,"SSD: Single Shot MultiBox Detector," European Conf. Computer Vision (ECCV), 2016.

[9] N. Wojke, A. Bewley, and D. Paulus,"Simple Online and Realtime Tracking with a Deep Association Metric (DeepSORT),"IEEE Int. Conf. Image Processing (ICIP), 2017.

[10] S. Hochreiter and J. Schmidhuber,"Long Short-Term Memory,"Neural Computation, vol. 9, no. 8, pp. 1735–1780,

1997.
[11] D. P. Kingma and J. Ba,“Adam: A Method for Stochastic Optimization,”International Conference on Learning
Representations (ICLR), 2015.
[12] T.-Y. Lin et al.,“Microsoft COCO: Common Objects in Context,”European Conf. Computer Vision (ECCV), 2014.
[13] A. Geiger, P. Lenz, and R. Urtasun,“Are We Ready for Autonomous Driving? The KITTI Vision Benchmark
Suite,”CVPR, 2012.
[14] M. Everingham et al.,“The Pascal Visual Object Classes (VOC) Challenge,”International Journal of Computer
Vision, 2010.
[15] H. Caesar et al.,“nuScenes: A Multimodal Dataset for Autonomous Driving,” CVPR, 2020.