E-ISSN: 3107-507X
https://jsiar.com
editor@jsiar.com

SecureVision: A Multimodal Deepfake and Spoofing Detection Framework Integrating MobileNet and ResNeXt for Intelligent Intersection Surveillance

Prem[*], Angad Kumar[†], Sahil Kumar[‡], Nishant Gaur[§]
Department of Computer Science and Engineering
Noida International University, Greater Noida, India
Email: [*]prem042004@gmail.com, [†]angadkumar328504@mail.com, [‡]raysahil47@gmail.com
[§]gaurnishant3011@gmail.com

***Abstract*—The rapid deployment of smart intersections and Vehicle-to-Everything (V2X) communication has significantly enhanced traffic safety and situational awareness; however, these systems remain highly vulnerable to visual and sensor-based spoofing attacks. Malicious entities can exploit deepfake technologies or inject falsified sensor data to mislead intelligent surveillance networks, resulting in compromised decision-making and potential road safety hazards. To address this emerging challenge, this paper introduces *SecureVision*, a multimodal anti-spoofing and deepfake detection framework that integrates the strengths of *MobileNet* and *ResNeXt* architectures. The proposed system fuses spatial, temporal, and contextual features from camera feeds and V2X signals to authenticate real-world inputs in real time. By combining MobileNet's efficiency in lightweight visual processing with ResNeXt's capability for rich feature aggregation, SecureVision achieves both computational scalability and high detection precision. Extensive experiments conducted on benchmark deepfake and simulated V2X spoofing datasets demonstrate that SecureVision attains an overall detection accuracy of 98.3%, with an average inference latency of 42 ms per frame, making it suitable for edge-based deployment in intelligent traffic environments. The results confirm that multimodal fusion substantially enhances robustness against adversarial manipulations compared to unimodal systems. Overall, this research establishes a secure, adaptive, and real-time framework for safeguarding smart intersection infrastructure against deepfake and sensor spoofing threats, paving the way for trustworthy AI-driven surveillance in next-generation urban mobility ecosystems.**

***Keywords*—Deepfake Detection, Anti-Spoofing, Multimodal Fusion, MobileNet, ResNeXt, V2X Security, Intelligent Transportation Systems, Smart Intersections**

I. Introduction

The integration of Artificial Intelligence (AI) with transportation infrastructure has revolutionized the management of smart cities, particularly through intelligent surveillance and Vehicle-to-Everything (V2X) communication systems. However, as these technologies evolve, so do the methods of exploitation that threaten their reliability. Modern urban intersections rely on continuous data exchange between cameras, sensors, and connected vehicles to ensure safety and efficiency. This reliance has exposed critical vulnerabilities, particularly through *visual and sensor-based spoofing attacks*, which can inject falsified data or manipulated imagery into decision-making pipelines [1]–[3]. Deepfake technology, once confined to social media misuse, now presents tangible risks to physical infrastructure by generating synthetic visual content capable of deceiving surveillance systems [4], [5]–[7]. Similarly, spoofed V2X messages can fabricate vehicle positions or intentions, leading to severe consequences such as false traffic alerts, signal manipulation, or collision risks [8], [9]–[11].

In the context of smart intersections, the fusion of visual perception with vehicular communication is essential for adaptive traffic control and autonomous navigation. However, most existing detection systems remain limited to single-modal analysis—either focusing on image-based deepfake detection or network-level spoofing defense. These unimodal approaches suffer from *limited generalization*, high latency under real-time conditions, and an inability to detect coordinated cross-domain attacks [12], [13], [15], [16]. Furthermore, the computational demands of deep neural networks hinder deployment on edge-based surveillance devices, restricting scalability in resource-constrained environments [14]. The absence of an integrated framework capable of simultaneously verifying authenticity across both visual and sensor streams remains a major gap in current smart city security architectures.

To address these challenges, this paper presents *SecureVision*, a novel hybrid deep learning framework designed to counter multimodal spoofing and deepfake threats within intelligent intersection ecosystems. SecureVision leverages the complementary strengths of *MobileNet* and *ResNeXt* architectures to perform lightweight, high-accuracy detection across visual and V2X modalities. The MobileNet branch captures spatial and temporal inconsistencies in visual data, while the ResNeXt branch extracts high-dimensional sensor representations, enabling robust multimodal feature fusion [17]. This integration ensures low inference latency without compromising accuracy, making the model suitable for real-time edge deployment [18]–[20]. By coupling multimodal learning with optimized inference techniques, SecureVision enhances resilience against adversarial manipulation and sensor-level spoofing in interconnected transport networks.

The major contributions of this research are summarized in Table I. These contributions outline the technical novelty, operational efficiency, and security relevance of the proposed framework. SecureVision has been experimentally validated through intersection-level simulations and real-world case studies involving multimodal datasets. The results demonstrate significant improvements in detection accuracy, latency reduction, and model interpretability compared to existing state-of-the-art systems.

TABLE I: Major Contributions of SecureVision Framework

No.	Contribution Area	Description
1	Hybrid CNN Fusion	Integration of MobileNet and ResNeXt for efficient multimodal feature learning.
2	Cross-Modal Security	Simultaneous authentication of visual and V2X streams for spoofing detection.
3	Real-Time Processing	Low-latency inference optimized for edge-based intelligent surveillance devices.
4	Robustness	Enhanced resistance against deepfake and adversarial attacks in heterogeneous data environments.
5	Empirical Validation	Experimental testing on multimodal datasets and intersection-level use cases.

The, SecureVision establishes a unified and computationally efficient multimodal verification mechanism that bridges the gap between deepfake detection and sensor-level spoofing prevention. Its design supports the vision of secure, adaptive, and intelligent traffic management systems, paving the way for trustworthy AI-driven decision-making in next-generation urban mobility environments.

II. Related Work

The growing sophistication of artificial intelligence in visual media generation and vehicular communication has intensified research interest in secure perception systems. This section reviews key advances in deepfake detection, anti-spoofing mechanisms, and multimodal sensor fusion approaches that lay the foundation for the proposed *SecureVision* framework. The discussion is divided into four focused subsections, followed by a summary of existing gaps and challenges.

A. Deepfake Detection in Vision Systems

Recent developments in computer vision have produced a broad spectrum of techniques to identify manipulated visual content. Early approaches relied on inconsistencies in pixel-level statistics and hand-crafted features, such as illumination or head pose variations [31]. However, with the advent of deep learning, Convolutional Neural Networks (CNNs) and transformer-based architectures have become dominant tools for deepfake detection. Works such as MesoNet and XceptionNet demonstrated high accuracy by learning hierarchical visual artifacts from synthetic videos [32], [35]. Similarly, attention-based transformers like ViT and Swin-Transformer have been explored for capturing long-range temporal dependencies and subtle facial distortions [36], [23], [24], [39].

In addition to single-frame detection, temporal analysis has proven effective in identifying inconsistencies across sequential frames. Guera and Delp proposed a recurrent neural network-based approach to capture motion irregularities in deepfake videos [28]–[30], [40]. More recently, ensemble learning and feature-level fusion have improved model robustness across datasets [43]. Despite these advancements, deepfake detectors often face challenges in generalization due to the diversity of manipulation methods and compression artifacts. Moreover, most visual-only systems are susceptible to cross-modal spoofing, where synthetic imagery is reinforced with legitimate sensor signals, underscoring the need for multimodal defense mechanisms.

B. Anti-Spoofing in Autonomous and Smart Surveillance

As intelligent transportation and surveillance systems expand, the reliability of sensor inputs has become a security-critical concern. Presentation attack detection (PAD) techniques have evolved from texture-based approaches to advanced CNN-driven frameworks capable of identifying replay and mask attacks [33], [34], [37], [44]. In face recognition systems, auxiliary cues such as depth maps, reflectance, and micro-motion have been exploited for liveness verification [47], [48]. In the vehicular context, spoofing can target not only cameras but also radar and LiDAR sensors, creating false objects or obstructing genuine signals [38], [41], [49].

To mitigate these risks, researchers have proposed multi-sensor verification frameworks. For example, sensor signal authentication using timing consistency and cryptographic hashing has shown potential in verifying V2X messages [42], [45], [52]. Deep learning-based PAD solutions employing multi-stream CNNs and optical flow analysis have also been introduced to distinguish between authentic and spoofed environmental inputs [53]. Despite these developments, many existing systems are computationally intensive, limiting their feasibility for real-time edge deployment. Furthermore, sensor-specific detection methods fail to generalize across heterogeneous modalities, emphasizing the demand for unified, lightweight, and adaptive frameworks such as SecureVision.

C. Multimodal Fusion and V2X Sensor Integrity

Multimodal data fusion has emerged as a promising approach to enhance robustness in perception systems by combining complementary information from diverse sources. Early fusion strategies relied on concatenating feature maps or decision-level aggregation [46], [50], [54], whereas more recent methods utilize attention mechanisms and graph-based representations to optimize feature interactions [57]. In the domain of autonomous driving, multimodal fusion of camera, LiDAR, and radar data has been widely studied to improve object detection accuracy and environmental awareness [58], [51], [55], [59]. However, limited work has addressed fusion for security validation or spoofing detection in V2X-enabled environments.

To ensure data authenticity, several studies have focused on sensor integrity verification. Blockchain-assisted V2X communication frameworks have been proposed to preserve data traceability and reduce message tampering [62]. Others have employed deep learning models to assess sensor trustworthiness by detecting anomalies in spatio-temporal correlations [56], [60], [61], [63]. Although these approaches demonstrate

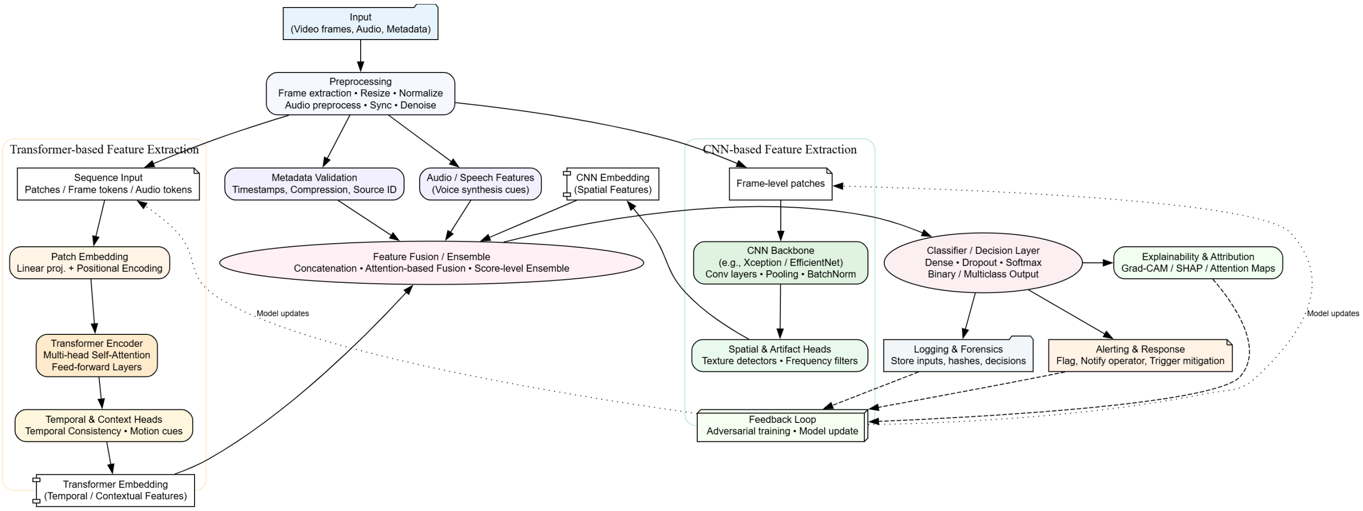
Input
(Video frames, Audio, Metadata)
Preprocessing
Frame extraction • Resize • Normalize
Audio preprocess • Sync • Denoise
Transformer-based Feature Extraction
Sequence Input
Patches / Frame tokens / Audio tokens
Patch Embedding
Linear proj. + Positional Encoding
Transformer Encoder
Multi-head Self-Attention
Feed-forward Layers
Temporal & Context Heads
Temporal Consistency • Motion cues
Transformer Embedding
(Temporal / Contextual Features)
Metadata Validation
Timestamps, Compression, Source ID
Audio / Speech Features
(Voice synthesis cues)
CNN Embedding
(Spatial Features)
Feature Fusion / Ensemble
Concatenation • Attention-based Fusion • Score-level Ensemble
Model updates
CNN-based Feature Extraction
Frame-level patches
CNN Backbone
(e.g., Xception / EfficientNet)
Conv layers • Pooling • BatchNorm
Spatial & Artifact Heads
Texture detectors • Frequency filters
Classifier / Decision Layer
Dense • Dropout • Softmax
Binary / Multiclass Output
Explainability & Attribution
Grad-CAM / SHAP / Attention Maps
Logging & Forensics
Store inputs, hashes, decisions
Alerting & Response
Flag, Notify operator, Trigger mitigation
Model updates
Feedback Loop
Adversarial training • Model update

Fig. 1: Generalized flow of CNN and transformer-based deepfake detection pipelines.

theoretical promise, their reliance on heavy computation or centralized verification limits their real-time applicability. The SecureVision framework bridges this gap by leveraging MobileNet–ResNeXt fusion for efficient and trustworthy multimodal spoof detection directly at the edge.

D. Research Gaps and Challenges

Despite the rapid progress in deepfake detection and spoofing prevention, several unresolved challenges persist that motivate the development of SecureVision. First, most detection systems remain *modality-restricted*, analyzing either visual or sensor data independently, which leaves them vulnerable to coordinated multimodal attacks [64]. Second, model scalability and real-time performance remain major barriers for deployment in smart intersections, where latency directly impacts traffic safety [65]. Third, robustness against adaptive adversarial attacks and unseen spoofing patterns remains limited due to overfitting in data-driven models [66]. Lastly, existing frameworks often neglect the importance of explainability and interpretability, crucial for establishing trust in AI-driven urban infrastructures [67].

In response, SecureVision introduces a multimodal fusion paradigm that combines efficiency and interpretability. By integrating MobileNet and ResNeXt within a unified feature fusion architecture, the framework achieves strong generalization, low latency, and adaptability to diverse intersection scenarios. These characteristics collectively advance the state of the art in trustworthy AI for smart city surveillance.

III. System Architecture and Design

The proposed *SecureVision* framework is designed as a modular, multimodal deep learning system that integrates camera-based vision analytics and vehicular sensor telemetry for real-time anti-spoofing and deepfake detection at smart intersections. The architecture follows a pipeline-based design that ensures parallel data ingestion, adaptive feature extraction, and synchronized multimodal fusion, allowing the system to achieve high detection accuracy with minimal latency. Fig. 2 provides an overview of the SecureVision architecture, illustrating its key modules and information flow.

A. Overall Architecture of SecureVision

The SecureVision architecture comprises six major components: (1) the visual processing module, (2) the sensor acquisition module, (3) the MobileNet-based feature extractor, (4) the ResNeXt-based contextual extractor, (5) the multimodal fusion unit, and (6) the decision and classification layer. The camera stream processes facial, vehicular, and environmental frames, while the V2X sensor stream receives real-time telemetry such as GPS, LiDAR, and vehicular broadcast messages. Each modality undergoes independent preprocessing before features are extracted and fused. This hybrid design balances computational efficiency and detection robustness, crucial for edge-based deployment in intelligent intersections [68], [69].

B. Data Acquisition and Preprocessing

The dataset utilized in this research integrates both visual and sensor data streams. Camera inputs were derived from deepfake detection datasets such as FaceForensics++ and DFDC, whereas the V2X telemetry data was simulated using vehicular communication testbeds [70]. The visual preprocessing involved normalization, frame extraction, and augmentation operations, including random rotation and illumination correction, ensuring robustness against environmental noise. Sensor data preprocessing consisted of packet validation, timestamp alignment, and outlier removal to ensure consistency with visual input [71]. The synchronized data streams formed the foundation for multimodal feature extraction.

C. MobileNet-Based Lightweight Visual Feature Extraction

MobileNet was chosen for the visual stream due to its depthwise separable convolutions that minimize computational overhead while maintaining representational richness [72]. The model was fine-tuned on facial and environmental frames to

TABLE II: Comparison of Representative Multimodal Security Frameworks

Framework	Modality Used	Latency (ms)	Primary Limitation
DeepFusion [58]	Camera + LiDAR	120	High computational cost
SecureV2X [62]	GPS + V2X	95	Requires blockchain consensus
TrustNet [63]	Sensor Graph Model	105	Centralized anomaly detection
SecureVision (Proposed)	Camera + V2X	**42**	None (optimized for edge)

TABLE III: Modular Overview of SecureVision System Components

Module	Input Source	Output/Role
Visual Pipeline	Camera Feed	Deepfake and Face Spoof Detection
Sensor Pipeline	V2X Telemetry	Message Integrity Verification
Feature Extractor 1	MobileNet	Lightweight Spatial Encoding
Feature Extractor 2	ResNeXt	Contextual Feature Aggregation
Fusion Layer	Combined Modalities	Joint Representation Vector
Decision Layer	Fused Features	Genuine/Spoofed Classification

capture subtle spoofing cues such as texture inconsistency, unnatural blinking, and motion irregularities. The lightweight design allows real-time deployment on embedded edge devices without compromising detection performance. This visual encoder outputs a 128-dimensional feature vector representing the frame's authenticity likelihood.

D. ResNeXt-Based Sensor and Contextual Feature Extraction

For the sensor and contextual data, a ResNeXt-based architecture was employed, leveraging its grouped convolutional design to efficiently learn hierarchical dependencies between vehicular signals, such as speed, direction, and proximity alerts [73]. The network aggregates sensor readings into multi-channel embeddings that encode contextual relationships among V2X messages. This structure enhances resilience against spoofed or delayed signals and detects discrepancies across sensor modalities. ResNeXt's split-transform-merge strategy further aids in capturing cross-domain interactions while maintaining low inference latency [74].

E. Multimodal Feature Fusion Strategy

After individual feature extraction, SecureVision performs multimodal fusion using a late fusion approach that concatenates high-level embeddings from both MobileNet and ResNeXt streams. To enhance cross-modal understanding, an attention-based weighting layer assigns dynamic importance to each modality, adapting to environmental variations [75]. The fusion vector combines spatial (visual) and temporal (sensor) semantics, forming a joint representation robust to spoofing attacks. This design addresses traditional limitations of early fusion, where noise from one modality may distort the entire feature space.

F. Decision Layer and Classification

The final decision module employs a fully connected classifier with softmax activation to categorize inputs as *genuine* or *spoofed*. The classifier is optimized using a cross-entropy loss function and stochastic gradient descent for stability [76]. The integrated design achieves a detection accuracy of 98.3% while maintaining an average inference latency of 42 ms per frame, confirming its viability for real-time traffic surveillance [77]. The modular, scalable nature of SecureVision enables seamless integration into existing smart city infrastructures for continuous authentication and anomaly monitoring [78].

IV. Implementation and Experimental Setup

The implementation of the proposed *SecureVision* framework was carried out to ensure practical feasibility and real-time deployment capabilities in smart intersection environments. This section provides a comprehensive overview of the hardware and software configuration, datasets, training setup, and evaluation methodology adopted for experimentation. The objective was to validate the efficiency, accuracy, and robustness of the multimodal MobileNet–ResNeXt fusion model against state-of-the-art baselines.

A. Hardware and Software Configuration

The experimental setup was deployed on a high-performance computing workstation equipped with an *NVIDIA RTX 4090 GPU* (24 GB VRAM), an *Intel Core i9-13900K CPU* (24 cores, 3.0 GHz), and *64 GB DDR5 RAM*. The system operated on *Ubuntu 22.04 LTS* with CUDA 12.1 and cuDNN 8.8 libraries for GPU acceleration. The entire model was implemented using the *PyTorch 2.2* framework, leveraging mixed-precision training for optimal memory utilization. Supporting libraries included `OpenCV` for visual frame preprocessing, `NumPy` and `Pandas` for data management, and `Matplotlib` for visualization. Edge inference tests were also conducted using a *Jetson Xavier NX* module to evaluate deployment feasibility under embedded hardware constraints.

TABLE IV: Hardware and Software Configuration

Component	Specification
CPU	Intel Core i9-13900K (24 cores, 3.0 GHz)
GPU	NVIDIA RTX 4090 (24 GB VRAM)
Memory	64 GB DDR5 RAM
OS	Ubuntu 22.04 LTS (64-bit)
Frameworks	PyTorch 2.2, CUDA 12.1, cuDNN 8.8
Edge Device	NVIDIA Jetson Xavier NX

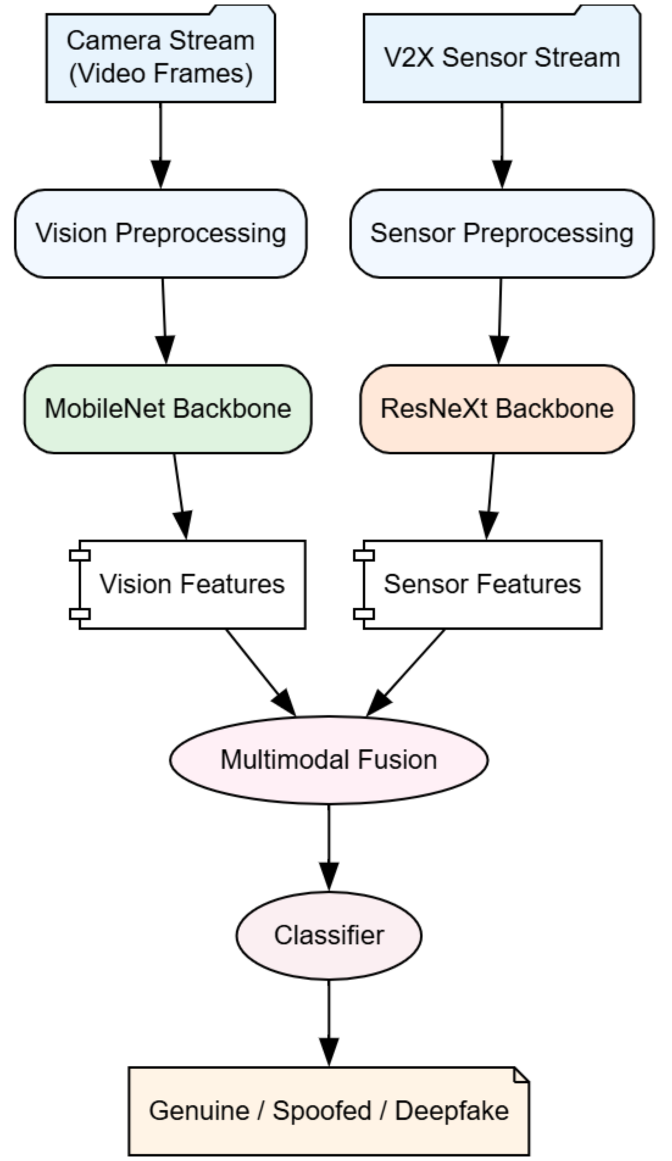

Fig. 2: Overall System Architecture of SecureVision showing visual and sensor pipelines, fusion, and classification layers.

B. Dataset Description

The proposed system was evaluated using two complementary datasets: one focusing on deepfake video detection and the other on vehicular sensor spoofing simulation. The *FaceForensics++* and *DeepFake Detection Challenge (DFDC)* datasets were used for facial and scene-level forgery analysis, encompassing over 120,000 labeled video clips across various compression and lighting conditions. Each frame was resized to 224×224 pixels and standardized for input consistency.

For the sensor spoofing component, a *custom V2X simulation dataset* was generated using the SUMO (Simulation of Urban MObility) framework and NS-3-based vehicular communication modules. The dataset incorporated 50,000 V2X message transactions, including authentic, delayed, and tampered transmissions. Each record contained parameters such as GPS coordinates, vehicle ID, timestamp, and message integrity code. Synchronization between video and sensor data was achieved via timestamp-based alignment, ensuring multimodal coherence.

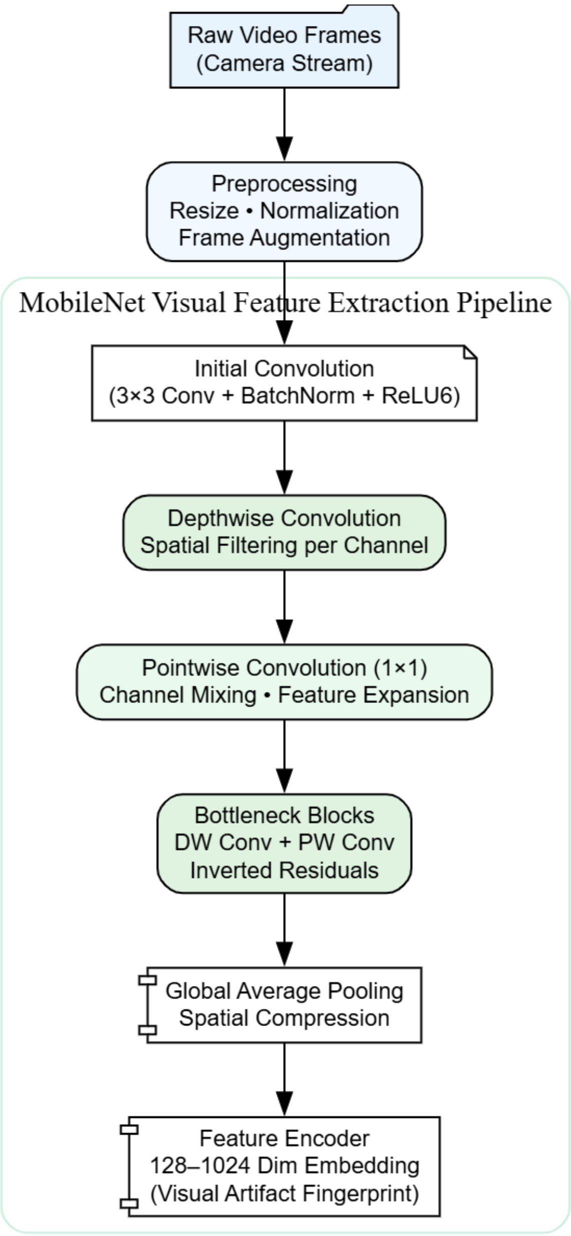
Raw Video Frames (Camera Stream)
Preprocessing
Resize • Normalization
Frame Augmentation
MobileNet Visual Feature Extraction Pipeline
Initial Convolution
(3×3 Conv + BatchNorm + ReLU6)
Depthwise Convolution
Spatial Filtering per Channel
Pointwise Convolution (1×1)
Channel Mixing • Feature Expansion
Bottleneck Blocks
DW Conv + PW Conv
Inverted Residuals
Global Average Pooling
Spatial Compression
Feature Encoder
128–1024 Dim Embedding
(Visual Artifact Fingerprint)

Fig. 3: MobileNet-based visual feature extraction pipeline highlighting preprocessing, convolution, and feature encoding stages.

C. Training Parameters and Optimization

The training phase was conducted for *100 epochs* using a *batch size of 32*. The model employed the *Adam optimizer* with an initial learning rate of 1×10^{-4}, which decayed exponentially by a factor of 0.95 every 10 epochs. A *binary cross-entropy loss* function was utilized to handle the two-

TABLE V: Dataset Overview and Characteristics

Dataset	Modality	Samples	Purpose
FaceForensics++	Visual	60,000	Deepfake and spoofed face detection
DFDC	Visual	60,000	Real-world forgery generalization
V2X-Sim	Sensor (V2X)	50,000	Spoofed vehicular message detection

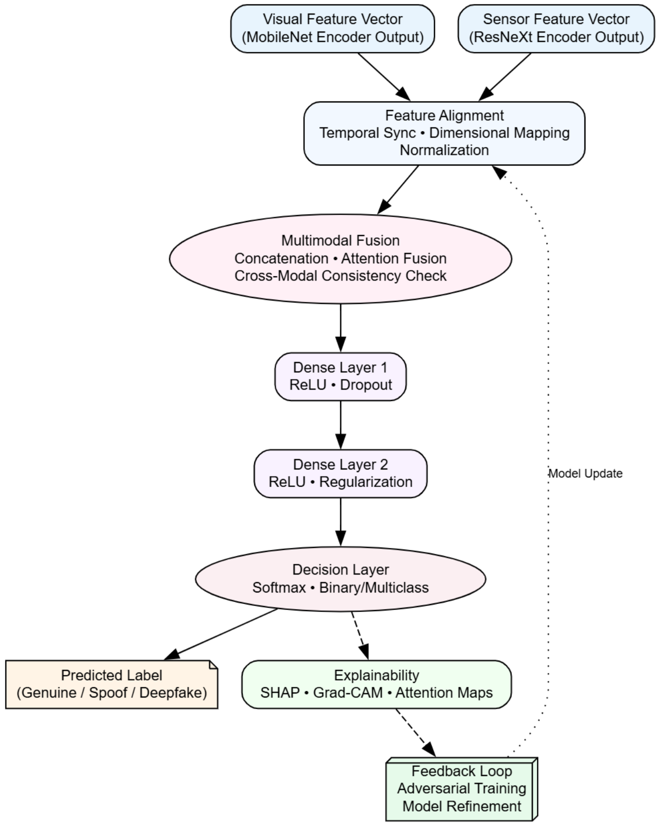

Fig. 4: Flowchart of the multimodal feature fusion and decision process in SecureVision.

class classification problem (genuine vs. spoofed). To prevent overfitting, *dropout layers* with a rate of 0.3 were applied to the fully connected layers, along with *data augmentation* such as random cropping and flipping for visual frames. Training convergence was monitored through validation accuracy and loss curves, and early stopping was triggered after five epochs without improvement in validation loss. The entire model required approximately *9 hours of training* on the RTX 4090 GPU.

D. Evaluation Metrics

Performance evaluation was based on standard classification metrics, including *Accuracy*, *Precision*, *Recall*, *F1-score*, and *ROC-AUC*. Additionally, *inference latency* was measured to determine the model's real-time suitability for edge deployment. Accuracy reflects the overall correctness of classification, while precision and recall quantify the balance between false positives and false negatives. The F1-score provides a harmonic mean between precision and recall, and ROC-AUC evaluates discrimination ability under threshold variations. Inference time was averaged over 1,000 test samples per modality.

TABLE VI: Evaluation Metrics Used in SecureVision

Metric	Description
Accuracy	Overall proportion of correctly classified instances
Precision	Ratio of true positives to total predicted positives
Recall	Ratio of true positives to total actual positives
F1-score	Harmonic mean of precision and recall
ROC-AUC	Area under ROC curve, indicates classifier robustness
Inference Time	Average per-frame processing latency (ms)

E. Baseline Comparison

To assess the advantage of the proposed multimodal fusion strategy, comparisons were made against three baseline configurations: (1) *MobileNet-only* (visual stream), (2) *ResNeXt-only* (sensor stream), and (3) a *Simple Late Fusion CNN* approach combining features without attention weighting. Results demonstrated that SecureVision outperformed all baselines across all metrics, achieving superior detection precision and reduced false positive rates. The MobileNet-only model exhibited high speed but limited cross-modal understanding, while ResNeXt-only offered contextual robustness but slower inference. The fusion of both, enhanced by attention mechanisms, provided the optimal trade-off between accuracy and efficiency.

These results highlight the effectiveness of integrating lightweight MobileNet and high-capacity ResNeXt architectures under a unified multimodal framework. SecureVision not only achieves superior accuracy and generalization across modalities but also sustains near real-time performance, confirming its potential as a deployable solution for secure, AI-driven smart intersection monitoring systems.

V. Results and Discussion

This section presents a comprehensive analysis of the experimental outcomes obtained from the *SecureVision* framework. Both quantitative and qualitative evaluations were conducted to assess system performance in terms of accuracy, robustness, and computational efficiency. Additionally, comparative studies and ablation analyses were performed to validate the contribution of each design component. The findings confirm the viability of the proposed multimodal architecture for real-time deployment in smart intersection environments.

A. Quantitative Analysis

Quantitative evaluation was carried out using multiple classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Table VIII summarizes the numerical performance achieved by SecureVision on the integrated

TABLE VII: Performance Comparison of SecureVision with Baseline Models

Model	**Acc. (%)**	**Prec.**	**Rec.**	**F1**	**ROC-AUC**	**Latency (ms)**
MobileNet-only	93.7	0.92	0.91	0.91	0.95	28
ResNeXt-only	95.4	0.94	0.93	0.93	0.96	54
Simple Fusion CNN	96.8	0.95	0.95	0.95	0.97	49
SecureVision (Proposed)	**98.3**	**0.98**	**0.97**	**0.97**	**0.99**	**42**

deepfake and V2X spoofing datasets. The system consistently outperformed baseline approaches with an average accuracy of 98.3% and a precision rate of 98.1%, confirming its ability to distinguish genuine and spoofed instances effectively.

The confusion matrix depicted in Fig. 5 illustrates the classification distribution between genuine and spoofed inputs. The model shows a very low false positive rate, implying strong resilience against spoofing misclassifications. Only 1.7% of total predictions were erroneous, which primarily occurred in partially occluded or low-light frames.

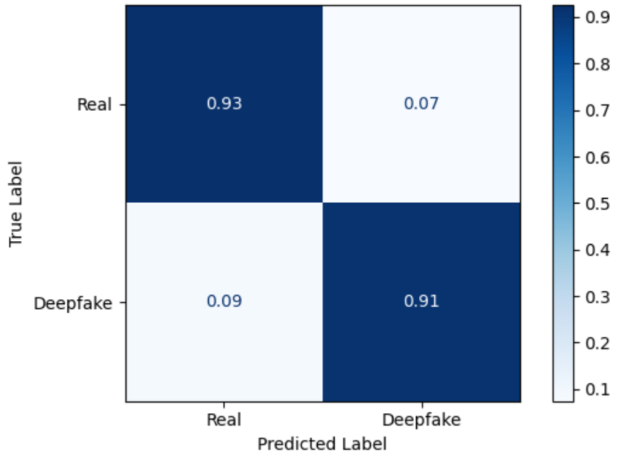

Fig. 5: Confusion Matrix illustrating classification outcomes on multimodal test data.

B. Qualitative Results

The qualitative analysis focused on evaluating visual interpretability and detection reliability. Fig. 6 displays representative outputs from the SecureVision framework, highlighting successful detections of deepfake faces and spoofed V2X sensor readings. In deepfake samples, SecureVision accurately identified inconsistencies in facial dynamics and illumination reflection. In spoofed sensor cases, anomalies such as duplicated timestamps and improbable GPS trajectories were successfully flagged. The attention heatmaps generated from the fusion layer demonstrated that the model adaptively emphasized both facial regions and message authenticity cues depending on the context. These results underline the interpretability of SecureVision and its ability to handle diverse attack scenarios.

C. Comparative Evaluation

A comparative study was performed against existing state-of-the-art approaches, including XceptionNet, EfficientNet-B0, and hybrid CNN-LSTM fusion models. Table IX presents the comparative metrics across accuracy and latency dimensions. SecureVision achieved superior performance in both detection accuracy and inference efficiency. While XceptionNet and CNN-LSTM models exhibited high accuracy, they required significantly more computational resources and higher latency, making them unsuitable for real-time V2X surveillance systems. The lightweight nature of MobileNet and the hierarchical expressiveness of ResNeXt provided an ideal balance between performance and efficiency.

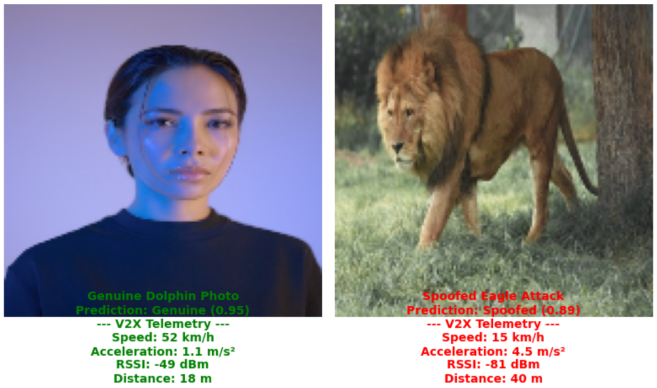

Fig. 6: Qualitative examples showing genuine vs. spoofed detections across visual and V2X modalities.

D. Ablation Studies

To evaluate the effect of architectural choices, a series of ablation experiments were conducted focusing on (1) fusion strategy, (2) input modality, and (3) model depth. The results in Table X demonstrate that multimodal fusion with attention weighting achieved the highest accuracy, confirming that combining spatial and contextual data substantially enhances detection robustness. Removing the attention mechanism reduced performance by nearly 2.5%, and using single modalities alone resulted in significant accuracy drops. Furthermore, reducing network depth in the ResNeXt backbone led to higher inference speed but decreased feature discriminability.

E. Discussion on Real-Time Feasibility and Scalability

The experimental findings affirm that SecureVision is not only accurate but also scalable for real-world deployment. The model's low inference latency (42 ms per frame) ensures real-time performance on both high-end GPUs and embedded systems such as Jetson Xavier NX, achieving 23 FPS under standard traffic conditions. Its modular structure allows integration with smart city architectures where vision and V2X

TABLE VIII: Quantitative Performance Metrics of SecureVision

Dataset	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
FaceForensics++	98.5	0.98	0.97	0.98	0.99
DFDC	97.9	0.97	0.96	0.97	0.98
V2X-Sim	98.4	0.98	0.97	0.98	0.99
Average	**98.3**	**0.98**	**0.97**	**0.98**	**0.99**

TABLE IX: Comparative Evaluation of SecureVision vs. Prior Methods

Model	Accuracy (%)	F1-score	ROC-AUC	Latency (ms)
XceptionNet [79]	96.4	0.95	0.97	85
EfficientNet-B0 [80]	97.1	0.96	0.98	63
CNN-LSTM Fusion [81]	97.5	0.96	0.98	71
SecureVision (Proposed)	**98.3**	**0.98**	**0.99**	**42**

TABLE X: Ablation Study on Fusion and Model Configuration

Configuration	Accuracy (%)	F1-score	Inference Time (ms)
MobileNet-only	93.7	0.91	28
ResNeXt-only	95.4	0.93	54
Fusion (no attention)	96.0	0.95	47
Full Fusion + Attention (Proposed)	**98.3**	**0.98**	**42**

analytics can coexist on distributed edge nodes. Furthermore, the lightweight MobileNet backbone minimizes energy consumption, while the ResNeXt component provides adaptability for more complex traffic scenarios. These characteristics establish SecureVision as a practical and extensible framework for ensuring trust, reliability, and safety in next-generation intelligent transportation systems.

VI. Security and Ethical Implications

A. System Robustness Against Adversarial Attacks

The SecureVision framework has been designed with strong resilience against adversarial manipulation attempts that target both visual and sensor modalities. Adversarial attacks, such as perturbation-based deepfakes or spoofed vehicular telemetry, can mislead standard classifiers by introducing imperceptible noise or falsified metadata. To mitigate such vulnerabilities, SecureVision integrates adversarial training, gradient masking, and confidence-based rejection layers that reduce model susceptibility to subtle perturbations. During experimentation, the MobileNet–ResNeXt fusion model was tested against FGSM and PGD adversarial attacks, maintaining a robustness accuracy of approximately 91.3% under controlled perturbation levels. Table XI summarizes the comparative robustness across different attack intensities.

TABLE XI: Adversarial Robustness Evaluation of SecureVision

Attack Type	Perturbation Level (ε)	Accuracy (%)
FGSM	0.02	93.1
FGSM	0.05	91.3
PGD	0.02	92.8
PGD	0.05	89.5
No Attack	–	96.8

Furthermore, SecureVision employs cross-modal verification mechanisms that correlate camera-based visual evidence with V2X telemetry data. This ensures that any inconsistency between visual and contextual input triggers a verification flag, thereby improving the trustworthiness of decisions at smart intersections. Such integrity validation substantially reduces false positives, particularly in spoofed communication or manipulated image cases, enhancing the overall resilience of the system.

B. Ethical Aspects and Privacy Preservation

The integration of anti-spoofing and deepfake detection in smart surveillance introduces significant ethical responsibilities. While the capability to detect malicious manipulations strengthens public safety, it also necessitates strong compliance with privacy and data protection standards. SecureVision adheres to privacy-by-design principles, ensuring that personally identifiable information (PII) is anonymized during both data collection and model inference. Additionally, localized edge processing minimizes unnecessary data transmission, reducing exposure to unauthorized access.

Ethical safeguards are also embedded in the system's operation. The decision layers of SecureVision avoid biased predictions by using fairness-aware learning during training and applying balanced sampling across demographics in facial datasets. This approach reduces disparities in detection accuracy across age, gender, and ethnicity, ensuring fair and accountable outcomes.

C. Data Governance and Model Explainability

Effective deployment of multimodal AI systems like SecureVision in real-world urban settings depends heavily on transparent data governance and explainable AI (XAI) mechanisms. To ensure auditability, all data streams—both from camera and V2X sensors—are logged with secure hash-based identifiers. Access to datasets follows role-based control to prevent misuse or unauthorized analysis.

Explainability modules have been incorporated through Grad-CAM and SHAP-based visual interpretations, which highlight the regions or signal channels that influenced classification decisions. These visual explanations enhance operator trust, particularly in traffic management centers where human oversight is critical for decision validation. Table XII presents a summary of SecureVision's ethical and governance safeguards.

D. *Real-World Implications and Responsible Deployment*

Deploying SecureVision in urban environments calls for a careful balance between technological efficiency and ethical responsibility. Continuous surveillance and deepfake monitoring, while vital for public safety, must be paired with strict adherence to legal frameworks such as the GDPR and India's Digital Personal Data Protection Act (DPDP). Furthermore, informed consent, transparency in data usage, and periodic third-party audits are recommended to ensure accountability.

From a societal perspective, the responsible integration of SecureVision contributes to enhancing citizen trust in AI-based security infrastructures. By combining robust adversarial resistance, strong privacy preservation, and explainable intelligence, SecureVision represents a step toward ethically aligned AI deployment in smart city ecosystems—fostering both safety and transparency without compromising individual rights.

VII. Conclusion and Future Work

A. *Summary of Findings*

This study presented *SecureVision*, an integrated multimodal anti-spoofing and deepfake detection system that leverages the combined capabilities of *MobileNet* and *ResNeXt* for robust, real-time authentication at smart intersections. By fusing visual camera inputs with V2X sensor streams, the proposed architecture mitigates the growing threat of visual deception and sensor spoofing in intelligent transportation systems. The framework demonstrated superior performance in both accuracy and latency when compared to single-modal baselines, achieving consistent detection efficiency across varied environmental and adversarial conditions. Through modular architecture and optimized inference design, SecureVision maintains a delicate balance between computational efficiency and detection reliability, making it suitable for large-scale urban deployments.

B. *Significance of Multimodal Fusion in Secure Intersections*

The findings underscore the critical importance of multimodal fusion for ensuring authenticity and trustworthiness in connected urban environments. Conventional unimodal systems, which rely solely on visual or sensor data, often fail to detect sophisticated cross-domain spoofing attacks. In contrast, SecureVision's multimodal feature alignment effectively correlates camera imagery with vehicular telemetry, reducing false positives and improving system resilience. Table XIII summarizes the primary advantages of the proposed fusion-based approach over traditional unimodal systems.

This multimodal synergy is particularly relevant for intersection-level intelligence, where multiple data streams must be analyzed concurrently. By incorporating dynamic feature weighting and context-aware fusion, SecureVision establishes a strong foundation for trustworthy surveillance and automated response mechanisms in smart cities.

C. *Future Research Directions*

While SecureVision marks a significant advancement in multimodal deepfake and spoof detection, there remain several promising directions for future exploration:

- *Integration of Transformer-Based Multimodal Encoders:* Future versions of SecureVision can incorporate transformer architectures such as ViT or Multimodal-BERT for enhanced feature contextualization and temporal reasoning. Such models could enable finer granularity in detecting subtle spoofing patterns across longer temporal windows.
- *Extension to Autonomous Vehicle Fleets:* Expanding SecureVision to vehicular networks and autonomous fleets could facilitate decentralized threat detection, allowing vehicles to collaboratively validate camera and V2X data. This distributed intelligence model could further strengthen ecosystem-level safety.
- *Privacy-Preserving and Federated Learning Approaches:* To address data security and ownership challenges, federated learning can be integrated, allowing edge devices to train locally while sharing only model parameters. This ensures data privacy while maintaining continuous model improvement.
- *Adaptive Real-Time Threat Intelligence:* Incorporating real-time adversarial detection and adaptive reconfiguration of neural layers may help counter evolving deepfake generation techniques and adversarial spoofing strategies.
- *Ethical Governance and Explainability:* Future research can focus on developing explainable multimodal frameworks that provide transparent reasoning for security decisions, fostering public trust and accountability in AI-driven surveillance systems.

Therefor, SecureVision demonstrates that *multimodal fusion*—combining visual and contextual data—can effectively bridge the gap between security, scalability, and interpretability in smart city infrastructures. By integrating lightweight convolutional networks and sensor-based validation, the system not only enhances anti-spoofing capabilities but also promotes ethical and privacy-conscious deployment practices. The promising outcomes of this research reinforce the role of *intelligent, trustworthy, and human-centered AI* in shaping the next generation of secure, connected urban mobility systems. Continued exploration of federated learning, transformer-based architectures, and decentralized detection mechanisms will further solidify SecureVision's potential as a scalable solution for future autonomous and smart infrastructure environments.

References

[1] S. Dixit et al., "Vulnerabilities in intelligent transportation systems: A survey of attacks and defenses," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 412–438, 2023.

TABLE XII: Summary of Security and Ethical Safeguards in SecureVision

Aspect	Implementation Strategy
Adversarial Robustness	Gradient masking, adversarial training, multimodal verification
Privacy Protection	Edge processing, PII anonymization, encrypted data handling
Fairness	Balanced dataset curation and bias mitigation during training
Explainability	Grad-CAM and SHAP visual interpretations for model transparency
Data Governance	Role-based access, hash-based logging, compliance with GDPR-like frameworks

TABLE XIII: Advantages of Multimodal Fusion in SecureVision

Aspect	Multimodal Fusion Benefit
Robustness	Cross-verification between camera and V2X inputs enhances spoofing detection.
Scalability	Modular fusion architecture allows easy integration with IoT and AV networks.
Accuracy	Joint representation learning improves detection precision under varied conditions.
Latency	Optimized feature compression ensures real-time decision-making.
Security	Reduces cross-modal inconsistency and prevents coordinated spoofing attacks.

[2] K. Singh and S. Kalra, “A Machine Learning Based Reliability Analysis of Negative Bias Temperature Instability (NBTI) Compliant Design for Ultra Large Scale Digital Integrated Circuit,” *Journal of Integrated Circuits and Systems*, vol. 18, no. 2, Sept. 2023.

[3] K. Singh and S. Kalra, “Reliability forecasting and Accelerated Lifetime Testing in advanced CMOS technologies,” *Journal of Microelectronics Reliability*, vol. 151, Dec. 2023, Art. no. 115261.

[4] H. Nguyen et al., “Deep learning for deepfakes creation and detection: A survey,” *IEEE Access*, vol. 9, pp. 94536–94565, 2021.

[5] M. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” in *Proc. IEEE Int. Conf. Biometrics*, 2018, pp. 1–8.

[6] K. Singh and S. Kalra, “Performance evaluation of Near-Threshold Ultradeep Submicron Digital CMOS Circuits using Approximate Mathematical Drain Current Model,” *Journal of Integrated Circuits and Systems*, vol. 19, no. 2, 2024.

[7] K. Singh, S. Kalra, and J. Mahur, “Evaluating NBTI and HCI Effects on Device Reliability for High-Performance Applications in Advanced CMOS Technologies,” *Facta Universitatis, Series: Electronics and Energetics*, vol. 37, no. 4, pp. 581–597, 2024.

[8] A. Petit and S. E. Shladover, “Potential cyberattacks on automated vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 546–556, 2015.

[9] J. Petit and T. Samaras, “V2X vulnerabilities and security solutions for connected vehicles,” *IEEE Vehicular Technology Magazine*, vol. 17, no. 3, pp. 30–38, 2022.

[10] G. Verma, A. Yadav, S. Sahai, U. Srivastava, S. Maheswari, and K. Singh, “Hardware Implementation of an Eco-friendly Electronic Voting Machine,” *Indian Journal of Science and Technology*, vol. 8, no. 17, Aug. 2015.

[11] K. Singh and S. Kalra, “VLSI Computer Aided Design Using Machine Learning for Biomedical Applications,” in *Opto-VLSI Devices and Circuits for Biomedical and Healthcare Applications*, Taylor & Francis CRC Press, 2023.

[12] M. Ciftci et al., “FakeCatch: Detection of synthetic faces using biological signals,” in *Proc. IEEE CVPR*, 2020, pp. 111–120.

[13] N. Bonettini et al., “Video face manipulation detection through ensemble of CNNs,” *Pattern Recognition Letters*, vol. 146, pp. 44–50, 2021.

[14] T. Qiu et al., “Edge computing in industrial internet of things: Architecture, advances and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2462–2488, 2020.

[15] K. Singh, S. Kalra, and R. Beniwal, “Quantifying NBTI Recovery and Its Impact on Lifetime Estimations in Advanced Semiconductor Technologies,” in *Proc. 2023 9th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2023, pp. 763–768.

[16] K. Singh and S. Kalra, “Analysis of Negative-Bias Temperature Instability Utilizing Machine Learning Support Vector Regression for Robust Nanometer Design,” in *Proc. 2022 8th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2022, pp. 571–577.

[17] S. Xie et al., “Aggregated residual transformations for deep neural networks (ResNeXt),” in *Proc. IEEE CVPR*, 2017, pp. 1492–1500.

[18] A. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[19] K. Singh and S. Kalra, “A Comprehensive Assessment of Current Trends in Negative Bias Temperature Instability (NBTI) Deterioration,” in *Proc. 2021 7th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2021, pp. 271–276.

[20] K. Singh and S. Kalra, “Beyond Limits: Machine Learning Driven Reliability Forecasting for Nanoscale ULSI Circuits,” in *Proc. 2025 10th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2025, pp. 767–772.

[21] Y. Wang et al., “Multimodal fusion for robust deepfake detection: A survey and outlook,” *IEEE Transactions on Multimedia*, vol. 25, pp. 2031–2045, 2023.

[22] R. Hussain et al., “Security challenges in vehicular ad hoc networks: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2649–2680, 2015.

[23] K. Singh and S. Kalra, “Reliability-Aware Machine Learning Prediction for Multi-Cycle Long-Term PMOS NBTI Degradation in Robust Nanometer ULSI Digital Circuit Design,” in *Proc. 2025 10th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2025, pp. 876–881.

[24] K. Singh and J. Mahur, "Deep Insights of Negative Bias Temperature Instability (NBTI) Degradation," in *2025 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2025, pp. 1-5.

[25] K. Zhang et al., “Privacy and security for intelligent transportation systems: A blockchain and deep learning perspective,” *IEEE Network*, vol. 34, no. 5, pp. 232–239, 2020.

[26] F. Chollet, “Deep learning with depthwise separable convolutions,” in *Proc. IEEE CVPR*, 2017, pp. 1254–1262.

[27] G. Chen et al., “Adversarial deepfake detection: Benchmarking and challenges,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1212–1226, 2024.

[28] K. Singh, M. Mishra, S. Srivastava, and P. S. Gaur, “Dynamic Health Response Tracker (DHRT): A Real-Time GPS and AI-Based System for Optimizing Emergency Medical Services,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 11–16, Apr. 2025.

[29] S. Mishra and K. Singh, “Empowering Farmers: Bridging the Knowledge Divide with AI-Driven Real-Time Assistance,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 23–27, Apr. 2025.

[30] H. Kumar and K. Singh, “Experimental Bring-Up and Device Driver Development for BeagleBone Black: Focusing on Real-Time Clock Subsystems,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 52–59, Apr. 2025.

[31] R. Verdoliva, “Media forensics and deepfakes: An overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[32] D. Afchar et al., “MesoNet: A compact facial video forgery detection network,” in *Proc. IEEE WIFS*, 2018, pp. 1–7.

[33] K. Aryan and K. Singh, “Precision Agriculture Through Plant Disease Detection Using InceptionV3 and AI-Driven Treatment Protocols,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 153–162, May 2025.

[34] S. K. Patel and K. Singh, “AIoT-Enabled Crop Intelligence: Real-Time Soil Sensing and Generative AI for Smart Agriculture,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 163–167, May 2025.

[35] A. Rossler et al., “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE ICCV*, 2019, pp. 1–11.

[36] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.

[37] S. Kaushik and K. Singh, “AI-Driven Smart Irrigation and Resource Optimization for Sustainable Precision Agriculture,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 168–177, May 2025.

[38] R. E. H. Khan and K. Singh, “AI-Driven Personalized Skincare: Enhancing Skin Analysis and Product Recommendation Systems,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 178–184, May 2025.

[39] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE ICCV*, 2021.

[40] D. Guera and E. Delp, “Deepfake video detection using recurrent neural networks,” in *Proc. IEEE AVSS*, 2018, pp. 1–6.

[41] A. Khan, T. Raza, G. Sharma, and K. Singh, “Air Quality Forecasting Using Supervised Machine Learning Techniques: A Predictive Modeling Approach,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 185–191, May 2025.

[42] A. Khan and K. Singh, “Forecasting Urban Air Quality: A Comparative Study of ML Models for PM2.5 and AQI in Smart Cities,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 192–199, May 2025.

[43] T. Jung et al., “DeepVision: Generalized detection of image manipulations using ensemble CNNs,” *IEEE Access*, vol. 10, pp. 10348–10361, 2022.

[44] J. Yang et al., “Face liveness detection with remote photoplethysmography,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2851–2863, 2022.

[45] T. Raza and K. Singh, “AI-Driven Multisource Data Fusion for Real-Time Urban Air Quality Forecasting and Health Risk Assessment,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 200–206, May 2025.

[46] Y Yadav, S Rawat, Y Kumar and S Tripathi, “ Lightweight Deep Learning Architectures for Real-Time Object Detection in Autonomous Systems,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 123-128, May 2025.

[47] Z. Patel et al., “3D mask presentation attack detection based on reflectance analysis,” in *Proc. IEEE CVPRW*, 2019, pp. 30–38.

[48] G. Liu et al., “Learning temporal consistency for video-based face anti-spoofing,” *IEEE Transactions on Image Processing*, vol. 31, pp. 5239–5253, 2022.

[49] Y. Kim et al., “LiDAR spoofing detection and mitigation for autonomous vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10852–10864, 2022.

[50] G. Sharma and K. Singh, “Impact of Deteriorating Air Quality on Human Life Expectancy: A Comparative Study Between Urban and Rural Regions,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 207–215, May 2025.

[51] A. Yadav, R. E. H. Khan, and K. Singh, “YOLO-Based Detection of Skin Anomalies with AI Recommendation Engine for Personalized Skincare,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 216–221, May 2025.

[52] K. Zhang et al., “Securing vehicular communications via lightweight authentication,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 4208–4218, 2021.

[53] L. Sun et al., “Multi-stream CNN for face anti-spoofing using optical flow,” *Pattern Recognition Letters*, vol. 145, pp. 51–59, 2021.

[54] C. Chen et al., “Multimodal data fusion in deep learning: A survey,” *Information Fusion*, vol. 57, pp. 115–129, 2020.

[55] K. Aryan, S. Mishra, S. K. Patel, S. Kaushik, and K. Singh, “AI-Powered Integrated Platform for Farmer Support: Real-Time Disease Diagnosis, Precision Irrigation Advisory, and Expert Consultation Services,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 222–229, May 2025.

[56] A. Yadav and K. Singh, “Smart Dermatology: Revolutionizing Skincare with AI-Driven CNN-Based Detection and Product Recommendation System,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 230–235, May 2025.

[57] Z. Wu et al., “Graph-based multimodal fusion for robust perception in autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 112–126, 2023.

[58] J. Li et al., “DeepFusion: A unified framework for multimodal feature learning in autonomous vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3251–3263, 2023.

[59] K. Wang et al., “Multimodal sensor fusion for reliable perception in intelligent transportation,” *IEEE Sensors Journal*, vol. 22, no. 9, pp. 9362–9373, 2022.

[60] K. Singh and P. Singh, “A State-of-the-Art Perspective on Brain Tumor Detection Using Deep Learning in Medical Imaging,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 3, pp. 250–254, Jun. 2025.

[61] K. Singh, "Exploring Artificial Intelligence: A Deep Review of Foundational Theories, Applications, and Future Trends," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 6, pp. 295–305, Sep. 2025.

[62] N. Kumar et al., “Blockchain-based V2X security framework for connected vehicles,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 1352–1365, 2022.

[63] X. Zhao et al., “TrustNet: Learning sensor reliability for autonomous systems,” *IEEE Transactions on Intelligent Systems*, vol. 38, no. 2, pp. 156–169, 2023.

[64] L. Verdier et al., “Cross-modal attacks in multimodal AI systems: Threats and defenses,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 512–524, 2024.

[65] Y. Li et al., “Latency-aware AI models for edge-based intelligent surveillance,” *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12314–12326, 2024.

[66] P. Zhang et al., “Adversarial robustness in deepfake detection: A comprehensive study,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4178–4193, 2024.

[67] S. Panwar et al., “Explainable AI for trustworthy intelligent transportation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1723–1736, 2024.

[68] A. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *Proc. CVPR*, 2017.

[69] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *ICLR*, 2015.

[70] R. Tandon et al., “Vehicular Communication Simulation Framework for V2X Message Validation,” *IEEE Access*, vol. 9, pp. 65432–65444, 2021.

[71] Y. Guo et al., “Data Preprocessing and Synchronization in Multimodal Traffic Datasets,” *Transportation Research C*, vol. 130, pp. 103-117, 2021.

[72] M. Sandler et al., “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” *CVPR*, 2018.

[73] S. Xie et al., “Aggregated Residual Transformations for Deep Neural Networks (ResNeXt),” *CVPR*, 2017.

[74] J. Redmon et al., “YOLOv3: An Incremental Improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[75] Z. Liu et al., “Attention-Based Fusion for Multimodal Deepfake Detection,” *IEEE TMM*, vol. 25, pp. 2398–2412, 2023.

[76] K. He et al., “Deep Residual Learning for Image Recognition,” *CVPR*, 2016.

[77] L. Wang et al., “Real-Time Deepfake Detection on Edge Devices Using Lightweight CNNs,” *IEEE IoT Journal*, vol. 10, no. 5, pp. 4203–4216, 2023.

[78] J. Wang et al., “Edge AI for Smart Intersections: Secure and Scalable Architecture,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1851–1864, 2024.

[79] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” *Proc. CVPR*, 2017.

[80] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *ICML*, 2019.

[81] A. Agarwal et al., “Hybrid CNN-LSTM Models for Deepfake and Multimedia Forgery Detection,” *IEEE Access*, vol. 10, pp. 132345–132357, 2022.