IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 February 2024, accepted 2 March 2024, date of publication 12 April 2024, date of current version 2 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387972

RESEARCH ARTICLE

Smart City Transportation: Deep Learning Ensemble Approach for Traffic Accident Detection

VICTOR A. ADEWOPO, (Member, IEEE), AND NELLY ELSAYED, (Member, IEEE)
School of Information Technology, University of Cincinnati, Cincinnati, OH 45220, USA
Corresponding author: Victor A. Adewopo (Adewopva@mail.uc.edu)

ABSTRACT The dynamic and unpredictable nature of road traffic necessitates effective accident detection methods for enhancing safety and streamlining traffic management in smart cities. This paper offers a comprehensive exploration study of prevailing accident detection techniques, shedding light on the nuances of other state-of-the-art methodologies while providing a detailed overview of distinct traffic accident types like rear-end collisions, T-bone collisions, and frontal impact accidents. Our novel approach introduces the I3D-CONVLSTM2D model architecture, a lightweight solution tailored explicitly for accident detection in smart city traffic surveillance systems by integrating RGB frames with optical flow information. Empirical analysis of our experimental study underscores the efficacy of our model architecture. The I3D-CONVLSTM2D RGB + Optical-Flow (trainable) model outperformed its counterparts, achieving an impressive 87% Mean Average Precision (MAP). Our findings further elaborate on the challenges posed by data imbalances, particularly when working with a limited number of datasets, road structures, and traffic scenarios. Ultimately, our research illuminates the path towards a sophisticated vision-based accident detection system primed for real-time integration into edge IoT devices within smart urban infrastructures.

INDEX TERMS Traffic surveillance, accident detection, action recognition, smart city, autonomous transportation, deep learning.

I. INTRODUCTION

The interconnection of road networks poses a formidable challenge in detecting and predicting traffic accidents.

The intricate aspect of this challenge lies in the dynamic impact of these accidents, especially at crucial intersections. The evolving field of computer vision, with its focus on analyzing spatial-temporal patterns, plays a critical role in addressing these challenges by enhancing our ability to monitor and respond to accidents in real-time. This technological advancement is especially relevant in the realm of smart city development, where integrating sophisticated accident detection and prediction systems into urban infrastructures can significantly improve safety, reduce traffic congestion, reduce traffic accident frequency, and enhance the overall quality of life for city residents. Road traffic accidents result in 1.35 million fatalities and 50 million non-fatal injuries globally each year [1]. Such alarming statistics underscore the urgent need for advanced traffic management solutions to promote safety and efficiency in urban transport systems.

The advent of intelligent transportation systems has ushered in a demand for intelligent transportation systems capable of identifying and tracking various objects such as vehicles, motorcycles, and buses [2]. Object detection in images has achieved significant performance in detecting and isolating objects in individual frames. However, video-based detection systems increasingly prevalent in diverse applications still face challenges in harnessing spatio-temporal data for enhanced accuracy. Prior research has delved into the use of temporal information for feature extraction in vehicle detection [3], and other techniques have incorporated this information in post-processing stages. Various techniques have been developed to improve traffic safety and reduce accidents including the utilization of sensors for traffic monitoring and accident detection which can provide valuable data for predicting future traffic conditions [2], [4]. Khalil et al. [4] proposed using ultrasonic sensors for automatic road accident detection. The proposed system employs two

The associate editor coordinating the review of this manuscript and approving it for publication was Jason Gu.

ultrasonic sensors to measure distance and sound waves for detecting collisions with obstacles. Despite these advancements, most techniques for prompt road accident monitoring remain expensive and sophisticated [2]. Modern-day technology such as surveillance cameras, GPS, Edge AI, and IoT are increasingly being leveraged to develop and deploy deep learning algorithms for traffic accident detection. However, Recurrent Neural Networks (RNNs), traditionally used in these systems, fall short in extracting spatial information from traffic data due to inherent design limitations in processing sequences from different roads independently. In contrast, Graph Neural Networks offer a promising alternative by integrating sequential and spatial data, enabling a more comprehensive analysis of traffic patterns. Le et al. [5] study highlights the importance of road-level accident prediction, acknowledging that accidents are influenced by a combination of internal factors (environment, road type, structure of the road) and external factors (such as the behavior of drivers, the weather, and the amount of traffic on the road). To bridge these advanced technological approaches with practical aspects of traffic management, it is essential to delineate the distinction between traffic accident detection and traffic anomaly detection. Traffic anomaly encompasses a broader range of irregular traffic movements without collisions, while accident detection is focused on a narrow window of traffic accidents defined by occurrences of vehicle crashes and can be classified as a subset of traffic anomaly [6]. The perspective provided by camera angle plays a crucial role in how traffic accidents are interpreted and analyzed. This research focuses on accidents recorded by traffic surveillance and dash cameras, with a particular emphasis on incidents that involve collisions between different types of vehicles, as well as those where no collision with other vehicles occurs, excluding incidents involving motorcycles. The varied nature of these accident scenes, along with the multifaceted array of viewpoints captured by these cameras, highlights the inherent challenges in accident detection. These challenges are further compounded by external factors such as varying environmental conditions and the dynamic evolution of accident scenes.

A. RESEARCH CONTRIBUTIONS

In smart city transportation systems, where real-time surveillance plays a pivotal role in ensuring safety, the significance of an efficient and accurate accident detection system is paramount. Our study addresses significant gaps in this area and offers the following novel contributions:

- **Advanced Vision-Based Accident Detection System:** We introduce an innovative vision-based accident detection system, specially optimized for real-time implementation on edge IoT devices such as Raspberry Pi. This system is designed to minimize computational overhead, making it highly suitable for smart city infrastructures and traffic surveillance systems. Its lightweight architecture successfully blends computational efficiency with practical applicability, establishing a cost-effective, reliable, and deployable solution for the modern smart city.
- **Novel Model Architecture:** Our research proposes a distinctive model architecture that extracts RGB frames and optical flow information from video sequences. Incorporating transfer learning techniques and the CONVLSTM2D architecture, this model significantly enhances accident detection performance, distinguishing our approach from existing methodologies.
- **Specialized Benchmark Datasets:** Addressing the lack of benchmark datasets in the accident detection domain, we have curated two specialized datasets: the Traffic Camera Dataset and the Dash Camera Dataset publicly available at [7] and [8]. These datasets are specifically de signed for accident detection and offer a diverse range of scenarios and roadway designs, serving as a valuable resource for ongoing research and development in this field.

These contributions aim to fill existing gaps in accident detection and scene recognition within autonomous transportation systems. Despite advances in algorithm development and modeling of spatiotemporal information in road structures [9], [10], [11], [12], the absence of comprehensive benchmark datasets and evaluation metrics has been a significant hurdle. Our contributions bridge this gap and also offer practical solutions for detecting traffic accidents in urban settings.

II. CURRENT ACCIDENT DETECTION TECHNIQUES

The advancement of accident detection systems in autonomous transportation systems is crucial for effective vehicle tracking and identifying anomalies in traffic patterns. The integration of Action Recognition (AR) in auto-navigation systems has brought significant improvements in obstacle detection, accident prevention, and lane departure assistance [13]. This progression highlights the potential of AR in enhancing autopilot technologies. In the domain of traffic flow analysis, Cai et al. [14] explored the detection of abnormal traffic flow using clustering techniques to identify deviations in normal traffic patterns. Earlier studies, like that of Morris and Trivedi [15], applied the Hidden Markov Model for intelligent scene description using spatiotemporal dynamics. More recent research has shifted towards leveraging machine learning and deep learning techniques for capturing spatio-temporal features from video streams [16], [17], [18], with innovations like combining convolution layers with LSTM architectures for improved performance [18], [19], [20]. The exploration of complex networks in accident detection has also been prominent, as seen in Carreira and Zisserman [21] introduction of the two-stream inflated 3D ConvNet (I3D) architectures for enhanced video input classification. This section delves into various methodologies and models that have contributed significantly to detecting and analyzing traffic accidents within smart city frameworks.

A. DEEP SPATIO-TEMPORAL CONVOLUTIONAL NETWORK (DSTGCN)

Urban traffic management systems designed for accident detection are essential in mitigating public safety risks and optimizing traffic flow. Le et al. [5] highlighted traffic accidents as significant contributors to fatalities and economic losses in contemporary society. Addressing this challenge, the authors introduce the Deep Spatio-Temporal Convolutional Network (DSTGCN), a model that utilizes deep spatial and temporal data for the prediction of traffic accidents. The field of Graph Neural Networks, a burgeoning area of study, has been effective in mapping the complex, non-Euclidean structures found in graph data. The DSTGCN model, as proposed by Le et al., is structured into three integral layers:

- *Spatial layer learning:* Spatial data was captured from road structures by identifying 20 points of interest (POI) that could capture the characteristics of road structure. Based on the relative proximity of identified point of interest, the probability of accident risk was calculated. Road-related structures are based on a cumulative calculation of points, road segments, and lengths of the road that have been determined by $\chi_{ij}^{P} = \left|\left\{p : p \in \mathcal{D}^{pj} \wedge dist\,(v_i, p) \leqslant d\right\}\right|$, *for* $1 \leqslant j \leqslant 20$; where Xpi is denoted by the distribution of POIs around the road segment, and where $\mathcal{D}^{pj}$ represents the set containing POIs. v_i, $|.|$ represents road segments, $dist\,(v_i, p)$ represents distance between the road segment v_i and POI.
- *Spatial-temporal learning:* Based on the robust data collected, the researcher partitioned road network into several grids and calculated the average speed within each of the grids. Using historical data from each road segment and an average traffic speed in each grid, the authors were able to extract the temporary traffic accident risks, represented as $\chi_i^{temporal} = \left\{\chi_i^{v,t}\right\}_{t=1}^{T}$.
- *Embedding layer for semantics and external representations:* The proposed DSTGCN framework incorporates an embedding layer specifically designed to integrate external factors that potentially influence accident risk. This layer is pivotal in capturing both the semantics and external representations associated with traffic incidents. Recognizing the universality of certain factors, the framework treats meteorological data pertinent to all road structures. External features considered in this model are categorized into two primary groups: calendar features (including day, month, and timestamp) and meteorological features (encompassing weather type, temperature, humidity, and wind speed) represented as $\chi^{external} = \left[\chi^{M,T}; \chi^{C,T+1}\right]$.

The proposed framework combines road network, point of interest (POI), taxi GPS, meteorological data, and traffic accident data for feature extraction and then incorporates them into a spatial convolution layer, spatial-temporal convolution layer, and embedding layer to learn the correlations between different features in accident prediction. Compared to other state-of-the-art methods, such as stack denoising autoencoders (SdAE) and traffic accident prediction methods based on LSTMs (TARPML), the proposed method scaled higher. The DSTGCN proposed model had higher precision, recall, and F1 scores, attributable to the robust information extracted from the three layers in the model architecture.

B. SPATIAL TEMPORAL GRAPH NEURAL NETWORK (STGNN)

The research of Wang et al. [22] investigated the dynamic complexities of road networks, asserting that their influence extends beyond mere proximity analysis. To effectively model long-range global dependencies in traffic flow, they developed the Spatial Temporal Graph Neural Network (STGNN). This innovative network comprises three key components: a positional graph neural network layer designed to capture spatial relationships, a recurrent neural network layer for processing temporal dynamics, and a transformer layer The architecture is based on assumption that traffic $\mathcal{G} = (v, \epsilon)$, *where* $v = \{v_1, \ldots, v_N\}$ is a set of N traffic sensor nodes, where ϵ is a set of edges connecting the nodes. Historically, traffic information is represented as $\mathcal{Y} = (Y_1, \ldots, Y_{\mathcal{T}})$ and $Y_t \in \mathbb{R}^{Nx1}$ is the traffic information of nodes $\mathcal{N}$ at time t. The traffic flow prediction is based on modeling input $\mathcal{X} = (X_1, \ldots, X_{\mathcal{T}})$ of length $\mathcal{T}$ to predict traffic in time $\mathcal{T}'$ [22]. This model leverages a training regime incorporating time $\mathcal{T}'$, where the STGNN is trained using sliding historical data $Y_t + \mathcal{T}'$ [22]. Within its architecture, the Spatial Graph Neural Network (S-GNN) layers are tasked with capturing spatial-temporal information. Concurrently, the GRU layer focuses on sequential temporal relations. The transformer layer plays a crucial role, computing attention across all positions simultaneously based on the queries. The framework proposed has a significant improvement over the baseline method when compared to predicting traffic flow across 15,30, and 60 minute intervals with a MAPE of 15.6%, 14.1%, and 19.3%, respectively underscoring its efficacy in traffic flow prediction.

C. SINGLE SHOT MULTIBOX DETECTOR (SSD)

Taking into consideration the drawbacks of object detector algorithms in single frames that are highlighted in the research of Wang et al. [22], Yang et al. [9] attempts to improve object detection models through the introduction of a novel object detection algorithm developed specifically to handle video data and mitigate the biases of traditional models. This new model can handle object tracking and helps eliminate the limitations of single object detectors. The proposed feature fused Single Shot MultiBox Detector (SSD) can enhance the accuracy of object detection. This detector consists of two stages, namely:

- Feature Fused SSD: Building on the VGGI6 network and additional feature layers, this stage of the SSD addresses its challenges in detecting smaller objects. By fusing the feature maps of FC7 and CONV-2, it captures essential semantic information, thereby improving the detection of smaller vehicles.

The detection is represented as

$$\left\{ \begin{array}{l} B_t^B = \left[b_{t1}^B, b_{t2}^B, \ldots, b_{tm}^B\right] \\ aS_t^B = \left[s_{t1}^B, s_n^B, \ldots, s_{tm}^B\right] \end{array} \right\} \text{[9]}.$$

- Tracking-guided detections optimizing (TDO): The second phase of the algorithm is primarily dedicated to tracking the bounding boxes identified from frame to frame n after the redundant bounding boxes have been removed by using a non-maximum suppression algorithm. A new vehicle is added to the subsequent frames when the distance between the overlapped boxes is greater than the chosen threshold, which is represented as $IoU\left(b_{ti}^B, b_{tj}^T\right) <= threshold$. SSD models have comparative results with the Faster R-CNN. However, in terms of computational complexity and labeled data required, the SSD outperforms Faster R-CNN.

The model was trained on two popular datasets ImageNet and Highway vehicle datasets. A notable observation was the model's predisposition towards detecting cars more frequently than buses, a phenomenon attributed to the higher prevalence of car images in the training dataset, indicative of data imbalance. Comparative performance analysis revealed that the average precision for the feature-fused SSD model stands at 70.5%, surpassing other state-of-the-art models such as Faster R-CNN (63%), the standard SSD (67.5%), and the Tubelet Proposal Network (TPN) at 68.4%. This demonstrates the feature-fused SSD's enhanced capability in object detection accuracy within the context of traffic and vehicle monitoring.

D. SPATIAL-TEMPORAL MIXED ATTENTION GRAPH-BASED CONVOLUTION MODEL (STMAG)

Accidents often stem from hazardous human behaviors, such as illegal parking or driving against traffic. While existing systems like collision alarm systems aid in accident prevention, they typically lack the capability for real-time alerts that could warn of imminent accidents. Addressing this gap, Xiaoyang et al. [12] introduced the Spatial-Temporal Mixed Attention Graph-based Convolution model (STMAG). This model leverages a combination of Convolutional Neural Networks (CNNs) and graph convolution networks, underpinned by a spatial-temporal mixed attention graph, to predict potential accidents. Key to its approach is the extraction of spatial heterogeneity information from traffic data, enabling the model to anticipate and identify possible accident scenarios more accurately. For the purpose of predicting accident at time t, spatiotemporal relationship is modeled by learning a nonlinear relationship between the function $\mathcal{F}$ on the topology of the network G, and the multivariable input sequence $\mathcal{X}$, represented as $\hat{Y}_{T+j} = f\left(G; \left(X_{T-n}, \ldots, X_{T-1}, X_T\right)\right)$. In this case, G is the topological structure of the nodes; $G = (V, E)$ containing a set of nodes, X_T is the concatenation of the traffic feature sets of the N nodes at the time of the target sequence. The main contribution of this research is to establish a mixed attention mechanism that can calculate the spatial and temporal dependencies between each node to the target sequence. The main contribution of this research is to establish a mixed attention mechanism that can calculate the spatial and temporal dependencies between each node to predict incidents when anomalies are detected. Using the attention mechanism, the authors developed a mixed temporal and variable attention model for selecting information that correlates strongly with the target output. The temporal dependencies consist of an encoder, decoder, and intermediate state vector. Historical temporal dependencies were obtained in the encoder using a Gated Recurrent Unit with RNN. The output variable of the mixed attention has three categories of Safety, mildly dangerous, and severely dangerous to trigger early warning for a high probability of traffic accidents.

E. DETECTION TRANSFORMER (DETR)

The critical nature of time in accident scenarios, where mere seconds can be the difference between a minor incident and a catastrophic event, highlights the need for efficient accident detection systems. Srinivasan et al. [23] address this challenge by introducing a novel approach that combines a Detection Transformer (DETR) with a Random Forest classifier, specifically tailored for accident detection using surveillance cameras. The principal aim of their research is to simplify the accident detection process and accelerate the inference time for identifying road accidents. Object detection algorithms like YOLO, Fast R-CNN, and Faster R-CNN have shown success, they often require extensive data annotation and are computationally intensive. In contrast, the DETR algorithm, developed by Facebook, offers an effective means of object detection and tracking. When integrated with a Random Forest Classifier, it enables the detection of accidents postoccurrence. The proposed DETR framework comprises four layers: DETR architecture with a CNN backbone, encoderdecoder blocks, fully connected layers, and a classification algorithm. In the framework, the input image x is processed to learn the residual mapping $F(x) = H(x) - x$. For the classification layer, the researchers selected 500 decision trees with a depth of 40. They found that utilizing entropy, calculated as $Entropy = \Sigma_{i=1}^{N}\left(-p_i \times log\left(p_i\right)\right)$, yields improved results. The framework achieved a detection rate of 78%, although it still trails behind the performance reported in the work of Wang et al. [24].

F. SCALE-INVARIANT FEATURE TRANSFORM (SIFT) AND SPARSE TOPIC MODEL

Research of Xia et al. [25] present a novel approach to treating videos as documents and trajectories as topics using the scale-invariant feature transform (SIFT) and sparse topic model. The researchers proposed a method for detecting anomalies in traffic surveillance. A deviation from normal traffic flow could indicate an anomaly in the patterns of motion. Several motion pattern analyses have been performed on video streams in order to identify individual anomalies.

The slight difference between normal and abnormal motion patterns makes it difficult to detect abnormal traffic events. The authors presented a sparse topic model based on a probability density function expressed in a non-probabilistic form to describe each trajectory in the video based on the Fisher kernel method. Compared to optical flow methods, SIFT was used to track the trajectory after features were extracted using the dense trajectory method based on its more robust performance. Sparse topic model was created by taking the video as a document and assuming there are K topics present in the video as a topic dictionary represented as matrix $D \in R^{K \times N}$. The whole document is represented as the code set $\alpha_{\lceil} = \left(\alpha_{d,1}, \alpha_{d,2}, \ldots, \alpha_{d,N}\right)^T$ and the dictionary $\mathcal{D}$. Considering that the model was trained on different videos that may differ from the dataset for detecting abnormalities, a constraint term of ℓ_1-norm was applied to the column vector $\mathcal{D}$. Based on the developed model, new evaluation dataset topics are generated by calculating the proportion of each topic to the document d on the k-th topic as: $\Theta_{d,k} = \frac{\sum_{n \in I_d} \alpha_{d,n\ k} D_{kn}}{\sum_{n \in I_d} \sum_{k \in K} \alpha_{d,n\ k} D_{kn}}$. The similarity between the two clips is calculated by taking the distance between clip d_i and d_j using $dis\left(d_i, d_j\right) = -lg\left(\sum_{k=1}^{k} \sqrt{\Theta_{d_i,k} \Theta_{d_j,k}}\right)$. The number of word/clip documents appearing in the test clip must exceed a threshold in order to flag the clip as anomalous.

The framework is designed to scan every topic (trajectories) that appears in the document (video) and classify that video as anomalous if the topic number exceeds a certain threshold. The proposed method was evaluated on the QMUI and AVSS datasets. Compared to other motion analysis methods, the method is effective with an AUC score of 91.2% on the AVSS dataset (JSM is based on motion trajectory and has an AUC score of 80.2%, while STC and GPR are representative video analysis methods for detecting anomalies and their AUC scores are 85 and 84% respectively).

G. MASK R-CNN

Similar to Xia et al. [25] study using a detection algorithm. Ijjina et al. [26] present a framework for detecting car accidents from traffic surveillance by training deep learning models (Mask R-CNN) to detect anomalies in car trajectory or speed after an overlap. The authors aim to identify traffic accidents spontaneously, reducing human bias and the time taken to effectively identify and respond to accidents. Through an automated system for contacting the traffic department and paramedics in cases of accidents, intelligent systems can potentially scale out human performance. Mask R-CNN is an improved state-of-the-art methodology for segmenting images and tracking multiple objects in a single image, which is a difficult task for CNN. Multiple objects can be tracked in an image by using region-based CNN (R-CNN) and fast region-based CNNs. Mask R-CNN offers a number of advantages over Faster R-CNN, including the ability to separate objects from backgrounds in images and identify individual objects by means of semantic and instance segmentation. Additionally, it is a faster network despite being able to localize, segment, and classify objects. Ijjina et al. [26] developed a versatile framework compatible with any CCTV camera system, emphasizing scalability and adaptability. Their methodology encompasses a threephase approach: object (vehicle) detection, vehicle movement tracking, and accident detection based on extracted features. In the initial phase, the Mask R-CNN is employed to delineate bounding boxes, class IDs, and object masks within an image, extending the Faster R-CNN architecture by incorporating masks for detected objects. The second phase focuses on identifying unique vehicles and objects across numerous frames, encompassing object tracking and feature extraction. This is achieved through a centroid tracking algorithm, where vehicles are continuously tracked across frames by noting their coordinates and calculating the Euclidean distances between centroids. Each object is assigned a unique identifier for subsequent frame tracking, with objects being removed from tracking once they exit the frame. The final stage involves making predictions based on various parameters such as Euclidean distance, trajectory changes, intersection angles, and object speed. Anomalies in acceleration, trajectory, and angle are determined against set thresholds to identify potential accidents. The system proposed by Ijjina et al. [26] could significantly reduce the number of false positives, especially in cases where cars travel in opposite directions or different lanes, it faces limitations in detecting accidents occurring in the same direction, as it primarily identifies anomalies post vehicle overlap. This aspect contrasts with Singh et al.'s study [27], which estimates accident probabilities by tracking vehicles moving in opposite directions, utilizing spatial-temporal video volume data. A notable omission in Ijjina et al.'s methodology is the lack of clarity on how thresholds for anomaly detection were established. The model's performance was assessed under various weather conditions, including daylight, low visibility, and adverse weather scenarios. In comparison to other models, their framework achieved a detection rate of 71% and a false alarm rate of 53%, trailing the deep Spatio-temporal model, which reported a detection rate of 77.5%.

III. TYPES OF TRAFFIC ACCIDENTS

Traffic accident is a significant global issue, causing substantial injuries, fatalities, and property damage annually [28]. Understanding the various types of accidents and their causes is essential to formulate effective strategies for reducing their negative impact. This study primarily focuses on the most prevalent and dangerous types of traffic accidents, which pose a considerable threat to road safety: rear-end collisions, T-bone or side impact collisions, and front impact collisions. Our analysis will concentrate on these main categories of accidents. As illustrated in Figure 1, the distribution of collision types indicates that front-to-rear collisions account for 43.9% of traffic accidents, angle collisions for 33.8%, and same-direction side-swipe accidents for 13.6%. This data underscores the importance of understanding the dynamics

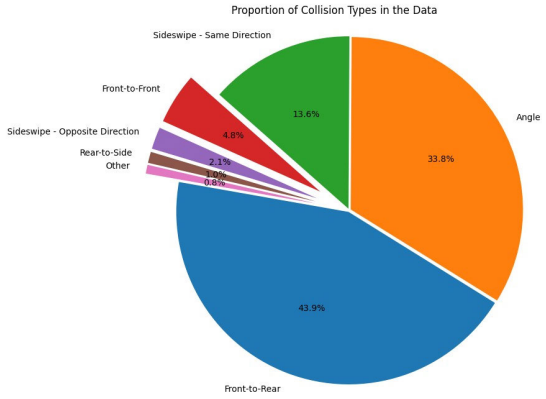


FIGURE 1. Percentage distribution of different types of collision.

and contributing factors of these specific accident types. In the context of our study, the focus on rear-end, T-bone, and front-impact collisions is crucial for developing effective accident detection systems and preventive measures. By comprehensively understanding these common types of accidents, we aim to contribute to the development of strategies and technologies that can effectively reduce their occurrence and improve road safety. The subsequent sections will elaborate on the methodologies and technologies employed in our research, tailored to detect and analyze these predominant types of traffic accidents.

A. REAR-END COLLISION

National Highway Traffic Safety Administrator, highlights Rear-end collisions as the most common type of traffic accident, accounting for 29% of all crashes, causing significant injuries and fatalities each year [29]. This type of accident occurs when one vehicle collides with the back of another vehicle traveling in the same direction. Rear-end collisions can result in injuries, from minor whiplash to severe head/spinal injuries and even fatalities. Several factors contribute to rearend collisions, including distracted driving, following too closely, sudden stops, and poor weather conditions. One of the leading causes of rear-end collisions is driver distraction, such as texting or talking on the phone while driving. Ad ditionally, distracted drivers cannot react quickly enough to avoid a collision in the event of sudden stops or slow down of vehicles ahead [30]. Efforts to address the issue had limited success, with the center high-mounted stop lamp (CHMSL) being one of the most notable initiatives launched in 1986. Although the CHMSL has reduced rear-end collisions by 4%, further improvement is still needed. In 1999, the National Highway Traffic Safety Administration (NHTSA) contracted with the Virginia Tech Transportation Institute (VTTI) to conduct tests and make recommendations for improved rear lighting and signaling systems [29]. To prevent rear-end collisions, drivers can manually employ safety measures such as increasing the distance between vehicles, reducing speed in poor weather conditions, and avoiding distractions while driving. Advanced driver assistance systems (ADAS) can be installed in vehicles to help drivers avoid collisions. These systems use sensors and cameras to detect potential collisions and warn drivers [31]. Developing new intelligent transportation systems (ITS) can help prevent rear-end collisions. These systems leverage connected vehicle technology to enable vehicle-to-everything (V2X) communication and interactions with infrastructure, such as collision avoidance and lane detection systems. They provide drivers with realtime data on road conditions and potential hazards, including indoor monitoring scenarios. For instance, should a vehicle ahead execute an abrupt stop, these systems are capable of alerting surrounding vehicles, thereby prompting drivers to take precautionary measures like decelerating or maneuvering to different lanes [31].

B. T-BONE COLLISION

T-bone accidents, commonly known as side-impact collisions, represent hazardous class of traffic accidents. These collisions occur when the front end of one vehicle impacts the side of another at a perpendicular angle. The severity of T-bone accidents often stems from the limited structural barrier between the driver and passengers and the colliding vehicle, particularly in cars lacking advanced safety features. Additionally, the sides of most vehicles offer less protection than the front and rear, making occupants more vulnerable in such events. Consequently, T-bone collisions frequently result in severe injuries or fatalities for those in the struck vehicle [28]. According to the National Highway Traffic Safety Administration (NHTSA), T-bone collisions account for approximately 13% of all passenger vehicle occupant fatalities in the United States. In addition, side impact collisions are the second most common type of fatal collision for passenger vehicle occupants, after frontal crashes [32]. Technological advancements have helped in mitigating the severity of side-impact collisions. For example, side airbags are now standard in many new vehicles and can provide additional protection to occupants in a side-impact collision. Moreover, some vehicles are equipped with advanced safety features such as automatic emergency braking and lane departure warning systems, which can help prevent collisions from occurring in the first place. Side impact collisions are a serious safety concern on the roads [31]. Efforts to improve safety through engineering and technology continue to be essential to prevent or mitigate the severity of T-bone collisions. The research of Eboli et al. [30], found some correlation between road structures, driver factors, and types of accident collision through the analysis of road accident type. Surface conditions, such as dry road surfaces, reduce the probability of a front/side collision in serious accidents. Driver-related factors, such as having a car license, increase the probability of a front/side collision in serious accidents. Furthermore, environmental factors, such as sunny weather, increase the probability of a front/side collision. The authors also pointed out that driver-related factors play a more important role in the probability of a front/side collision [30].

C. FRONTAL IMPACT ACCIDENT

Frontal impact accidents constitute a significant portion of inter-vehicular incidents, often resulting in severe injuries and fatalities. Notably, these collisions account for 35% of accidents with high severity levels [33]. A front-end collision typically occurs when the front of a vehicle strikes the front of another, or when a vehicle collides with a stationary object, such as a tree or a wall. Despite their potential for serious consequences, Figure 1 shows that frontal collisions represent only 4.8% of total traffic accidents. These collisions often result in serious injuries or fatalities due to the force of impact. Front-end collisions can occur for a variety of reasons, including driver error, speeding, distracted driving, and poor weather or road conditions. In some cases, faulty vehicle equipment or manufacturing defects may also contribute to front-end collisions. Front-end collisions can be classified into two main categories: offset and full-frontal collisions. In an offset collision, the front of one vehicle collides with the side of another vehicle, resulting in a twisting motion that can cause significant damage to both vehicles. Full-frontal collisions occur when the entire front end of one vehicle collides with the front end of another vehicle, resulting in a direct impact that can cause severe injuries or death [34]. Efforts have been made to reduce the incidence and severity of front-end collisions. This includes using safety features such as airbags, crumple zones, and seat belts, as well as developing advanced driver assistance systems (ADAS) that can alert drivers to potential collisions and even take action to avoid them. The existing forward collision warning (FCW) systems based on kinematic or perceptual parameters have some drawbacks in warning performance due to poor adaptability and ineffectiveness [35]. To address these problems, machine learning and deep learning algorithms have been proposed. However, these models lack consideration for multi-staged warnings, which could distract or startle the driver. A light gradient boosting machine (LGBM) learning algorithm was used to develop a multi-staged FCW model, which was evaluated using a driving simulator by twenty drivers [36]. The study found that the front vehicle acceleration, time-to-collision (TTC), and relative speed strongly affected the warning stages from the proposed model. The authors aim to develop LGBM for developing FCW models that could improve warning performance while considering multi-staged warnings [36]. The study of Jirovsky et al. [33] employed a new approach for determining the probability of collision between two vehicles by defining a 2-D reaction space, which describes all possible positions of the vehicles in the future. This approach enables mitigation of collision by exploring alternative causes of action such as changing direction in addition to braking [33]. Figure 2 shows the analysis of traffic accidents at different intersections and the light conditions. Light conditions play a crucial role in traffic safety, and the data indicates that accidents are more common at intersections, with four-way intersections being hazardous. Whether lighted or not, the number of accidents in dark conditions is high across all intersections. The data

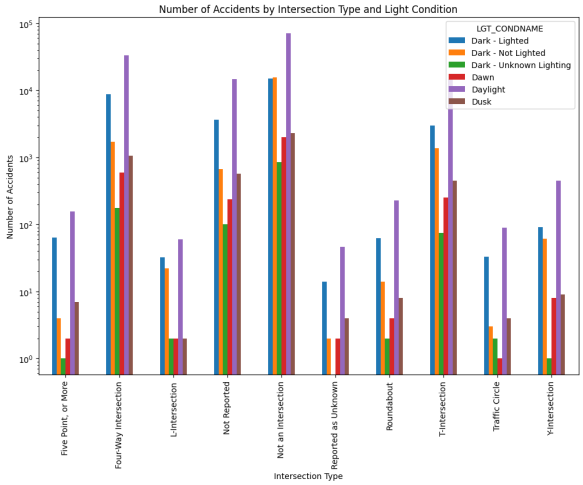


FIGURE 2. Accidents by intersection type and light condition associated.

also shows that accidents in dark-light conditions are more frequent at T-intersections and four-way intersections. Adequate road lighting is crucial for ensuring the safety of drivers during nighttime driving. The study by Alharbi et al. [37] on the performance appraisal of urban lighting systems aligns with the findings in Figure 2. The researchers found that electronic billboard positioning, oncoming vehicle lights, and poor lighting conditions during inclement weather, particularly dust, are significant factors affecting the performance of urban street lighting systems (USLSs) as perceived by road users. The research by Bridger et al. [38] reported that modern LED lighting is a disruptive technology and that decreasing nighttime fatalities and injuries due to modern road lighting has significant cost benefits. On the other hand, the study by Marchant et al. [39] found limited evidence to support road safety improvement through relighting traffic lights with white lamps in the UK. This further highlights the importance of continued research and evaluation of road lighting systems and their impact on road safety.

IV. DATA COLLECTION

A comprehensive data collection process was implemented due to the lack of readily available robust datasets in this domain. Here, we provide a detailed description of our collection methodology and sources.

A. TYPES OF DATA AND SOURCES

Two primary categories of data were collected:

- **Traffic/Surveillance Data (Trafficam):** This data contains videos captured from traffic and surveillance cameras situated at various locations. Trafficam provides a unique vantage point for capturing vehicular movement. These cameras are strategically positioned at various locations to offer an aerial or elevated view of the road, enabling a holistic view of vehicles. This perspective is essential for several reasons, including; Trajectory Angle: Trafficam videos capture the trajectory angle of vehicles. This angle denotes the path and direction

of a vehicle's movement, which is valuable for understanding vehicular dynamics and causal factors leading to an accident. Full Car View: The elevated vantage point of traffic and surveillance cameras provides a full silhouette and profile of vehicles crucial for detecting anomalies or changes in vehicular posture, like tilting during a potential rollover. Datasets from the TrafCam angle are particularly beneficial for accident detection because the overhead view minimizes occlusions, allowing for an unobstructed view of potential accident sites, and monitoring multiple vehicles simultaneously can help detect and analyze multi-vehicle collisions.

- **Dash Camera Video (DashCam):** These videos are typically recorded from cameras installed on vehicle dashboards. DashCam videos offer a ground-level, front-facing perspective from vehicles, capturing the road ahead and occasionally vehicle interior/rear view. While invaluable in many respects, DashCams are beneficial for providing firsthand accounts of incidents and useful for understanding drivers' accident viewpoints, including capturing close-up details of incidents. Some of the challenges of Dashcams are the restricted field of view (focus mainly on the road ahead), Camera view occlusion, and variability in video quality based on different brands and models.

To curate a robust, diverse dataset, we collected videos from a wide array of sources and different geographic locations worldwide. The majority of our video data originates from platforms like YouTube. We employed strategic keyword searches to identify and curate relevant videos. Keywords included terms like ''traffic accident,'' ''traffic camera accidents,'' and ''car accidents.'' To maximize the geographic and linguistic diversity of our dataset, these keyword searches were conducted in multiple languages, including English, French, Spanish, and Russian. Table [1] provides a detailed breakdown of traffic and dashcam videos, categorized by accident occurrences and normal traffic conditions. The data is further divided into training, validation, and testing sets. Table [1] showcases the distribution of these categories across different data types, including Trafficam, Dashcam, and External Data sources. In order to supplement our curated dataset. We utilized an external data source (Car Crash Dataset) publicly available on GitHub [40]. We aggregated approximately 5,700 datasets to train and evaluate our model performance.

B. VIDEO PROCESSING AND ANNOTATION

To ensure our dataset is concise and relevant, videos were processed based on the rapid dynamics of accidents. We segmented the videos into five-second non-overlapping clips, ensuring that each segment captures a distinct event or scene. This segmentation strategy aids in minimizing redundancy and focusing on the most relevant content for accident recognition. Each 5-second video segment was annotated manually using the Labelbox annotation tool. Annotations provide information about the type of accident depicted in the video. Categories included, but were not limited to, ''all over'', ''front end,'' ''rear end,'' ''side hit,'' and ''normal traffic.'' Due to the limited dataset in each category, our final dataset was grouped into two distinct categories (Traffic Accident and Normal Traffic). The camera resolution impacts the clarity and detail of the captured video, with higher resolution offering finer details. In this study, regardless of the initial resolution, videos are processed and cropped to $224 \times 224 \times 3$ to conform to the I3D model input requirements. This standardized resolution maintains consistency and ensures comparability in results. We also performed video normalization by dividing pixel values by 255 and scaling the data, aiding the model in faster and more stable convergence. To manage computational limitations, we strategically decreased our model's input from 150 to 50 frames, incorporating every third frame and subsequently to 30 frames by utilizing every fifth frame. This consideration was crucial in ensuring the effective detection of swift, brief actions for identifying accidents. The optimized frame selection was calibrated to retain the integrity of rapid actions, balancing the need for computational efficiency and resource conservation.

V. METHODOLOGY

The integration of machine learning (ML) and deep learning (DL) techniques has significantly advanced the field of accident detection. These technologies are particularly adept at processing large accident datasets, enabling the detection and classification of incidents based on crucial parameters such as speed, direction, and vehicle type. Singh et al. [27] proposed a framework that leverages deep representation extraction using autoencoders paired with an unsupervised learning model, like the Support Vector Machine (SVM), to predict the likelihood of accidents. This approach underscores the potential of combining machine learning models with sophisticated feature extraction for enhanced predictive capabilities. Complementing ML, deep learning models offer promising avenues for real-time accident detection. Utilizing camera systems to continuously monitor roadways, these models apply advanced image recognition and video processing techniques to identify potential hazards and accident scenarios. The capability of deep learning models to analyze complex video data in real time is vital for prompt accident detection a key factor in accident prevention and mitigation. Zadobrischi [41] focuses on the integration of traffic monitoring systems into intelligent transport systems (ITS) to properly manage traffic and reduce the negative impact of congestion and road accidents in real-time using video and image data. Chan et al. [42] proposed a Dynamic-Spatial-Attention (DSA) Recurrent Neural Network (RNN) for anticipating accidents in dashcam videos based on the vehicle trajectory and motion. The developed algorithm contains an object detector to dynamically gather subtle cues and the temporal dependencies of all cues to predict accidents two seconds before they occur. Ghahremannezhad et al. [43] introduce a three-step hierarchical framework for detecting

TABLE 1. Distribution of traffic and dashcam footage across different data types.

Data Type	Trafficam Accident	Trafficam Normal-Traffic	Total Trafficam Dataset	Dashcam Accident	Dashcam Normal-Traffic	Dashcam Total Dataset	Ext. Data Accident	Ext. Data Normal-Traffic	Total DataSet
Train	603	511	1114	763	639	1402	1250	146	3912
Val	190	140	330	268	224	492	150	82	1054
Test	134	60	194	196	154	350	100	81	725

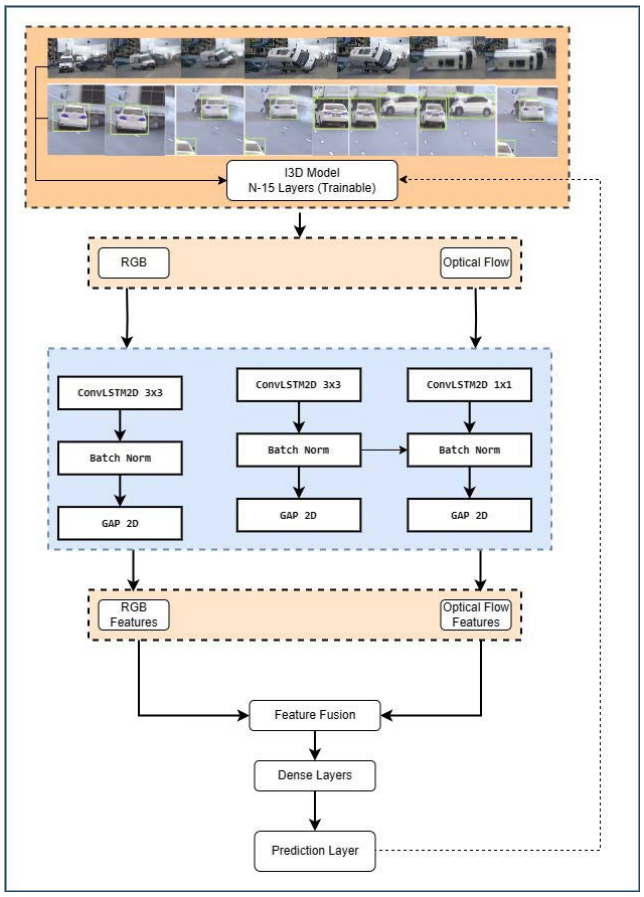


FIGURE 3. Accidents detection framework.

traffic accidents at intersections using surveillance cameras. A unique cost function is utilized during object tracking to handle occlusions, overlapping objects, and object shape changes. Overall, our methodology employs deep learning approach with transfer learning to develop a comprehensive, efficient, and accurate system for traffic accident detection.

A. I3D-CONVLSTM2D MODEL ARCHITECTURE FOR ACCIDENT DETECTION

The architecture of our model, as depicted in Figure [3], is designed to effectively capture the spatiotemporal characteristics of video data for accurate accident detection. The model incorporates a dual-branch approach, processing both RGB and Optical Flow (OF) information through an Inflated 3D ConvNet (I3D) followed by ConvLSTM2D layers.

- **Input Video Frames:** Given a video V with N frames, each frame f_i is of size $224 \times 224 \times 3$ for RGB and $224 \times 224 \times 2$ for Optical Flow.

$$V = \{f_1, f_2, \ldots, f_N\}$$

The frames are segmented into overlapping windows to capture temporal continuity, crucial for detecting motion-related anomalies indicative of accidents. Each window w_j contains a sequence of frames, where T is the temporal depth of the window.

$$w_j = \{f_j, f_{j+1}, \ldots, f_{j+T}\}$$

- **I3D Model:** The I3D model, introduced by Carreira et al. [21], is adept at capturing complex spatiotemporal patterns within video data, making it particularly suitable for accident detection tasks. It consists of multiple 3D convolutional layers, pooling layers, and inception modules that together process both spatial and temporal information. To adapt the pre-trained network to features specific to traffic accidents, the last 15 layers of the I3D model are set to trainable, allowing for fine-tuning on our accident-specific dataset:

$$\text{Trainable layers: } \{l_{total} - 14, l_{total} - 13, \ldots, l_{total}\}$$

where l_{total} is the total number of layers in the I3D model. This fine-tuning step is crucial for the model to learn from the unique temporal dynamics and visual cues present in accident scenarios.

- **ConvLSTM2D Layers:** RGB and Optical Flow branch of the I3D model is passed through three ConvLSTM2D layers. The ConvLSTM2D layer is a combination of convolutional and LSTM layers, designed to capture spatial and temporal dependencies in sequence data [44], [45], [46]. The choice of three layers was empirically determined to offer an optimal trade-off between capturing detailed spatio temporal information and maintaining computational efficiency. The architecture of the ConvLSTM2D layers is carefully configured to process the input tensor X with dimensions $T \times H \times W \times C$, and the operations within these layers are as follows:

$$\begin{aligned} F_t &= \sigma\left(W_f * X_t + U_f * H_{t-1} + b_f\right) \\ I_t &= \sigma\left(W_i * X_t + U_i * H_{t-1} + b_i\right) \\ O_t &= \sigma\left(W_o * X_t + U_o * H_{t-1} + b_o\right) \\ C_t &= F_t \odot C_{t-1} + I_t \\ &\quad \odot \tanh\left(W_c * X_t + U_c * H_{t-1} + b_c\right) \\ H_t &= O_t \odot \tanh\left(C_t\right) \end{aligned}$$

where: σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication and W, U, and b are the

weights and biases of the ConvLSTM2D layer. This configuration enables the model to capture the intricate motion patterns associated with accidents while preserving the critical spatial information of each frame.

- **Global Average Pooling 2D (GAP2D):** Following the ConvLSTM2D layers, the model incorporates a Global Average Pooling 2D (GAP2D) layer. This layer serves to condense the spatial dimensions of the feature maps by computing the average value of each channel, resulting in a tensor that preserves essential spatial features while significantly reducing the number of parameters. The GAP2D layer helps in mitigating overfitting and simplifies the model, making it computationally efficient for real-time applications:

$$GAP(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j}$$

- **Feature Fusion:** The features extracted from the RGB and Optical Flow branches are combined to form a comprehensive feature vector that encapsulates both appearance and motion information. This fusion is crucial as it leverages the complementary nature of the data streams to improve the model's ability to detect accidents.
- **Dense Layers and Prediction:** The fused feature vector is then fed through multiple dense layers, which introduce non-linearity and further abstraction capabilities to the model with ReLU activation functions. The final prediction layer utilizes a softmax activation function, providing a probabilistic output for the binary classification. Although binary classification tasks typically use a sigmoid function, the softmax function was empirically found to perform better in this specific case, offering a more discriminative probability distribution for the two classes (''Accident'' and ''No Accident''):

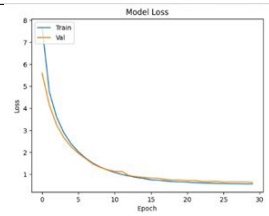
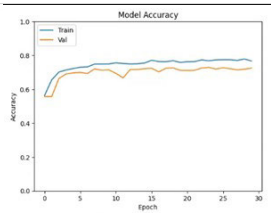
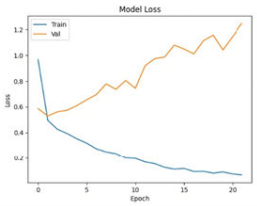
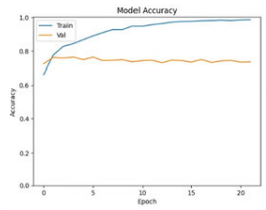
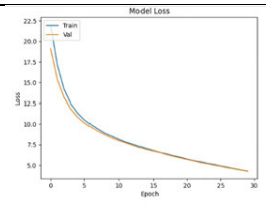
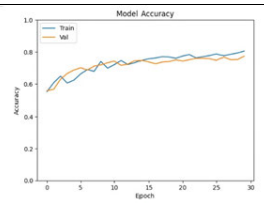
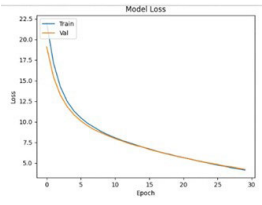
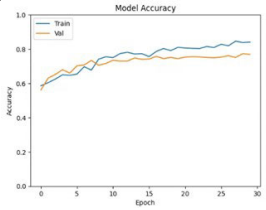
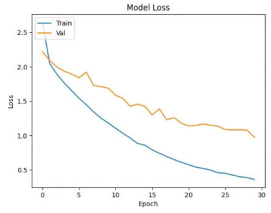
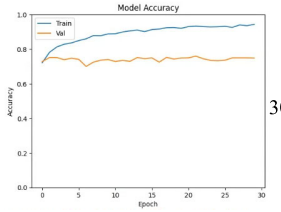
$$P(\text{ Accident }) = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}$$
$$P(\text{ No Accident }) = \frac{e^{z_2}}{e^{z_1} + e^{z_2}}$$

- **Regularization Techniques:** To combat overfitting and enhance the model's generalizability, L1 regularization is employed within each ConvLSTM2D layer. This regularization encourages sparsity in the learned features, which often leads to a model that is both simpler and more robust to the variance in new data. Additionally, batch normalization is applied after each layer to normalize the activations and accelerate the training process.
- **Transfer Learning:** Our model harnesses the power of transfer learning by initializing with weights from a model pre-trained on a large and diverse dataset. This strategy provides a solid foundation of visual features which are then fine-tuned to the specific task of accident detection. Transfer learning not only improves the model's performance but also reduces the training time, which is critical for developing systems that can be deployed in real-world scenarios.
- **Empirical Validation of Model Architecture:** The determination of the optimal number of layers and neurons within the ConvLSTM2D cells was conducted through extensive empirical testing. Various architectures were evaluated for their performance on a validation set, which consisted of diverse traffic scenarios. The chosen architecture represents the best-performing model in terms of accuracy and the ability to generalize beyond the training dataset. These empirical tests ensure that the model is neither underfitting nor overfitting and is capable of capturing the complex spatiotemporal relationships inherent in accident data.
  In our accident detection model, the softmax activation function is employed over sigmoid due to its ability to provide a fine-grained probability distribution across both classes. This choice is pivotal for calibrating the model's predictions, ensuring that output probabilities accurately reflect the confidence levels in real-time scenarios. Empirical testing demonstrated that softmax achieved superior performance in our specific case, offering a more nuanced understanding of the model's probabilistic output between 'accident' and 'no accident' classifications. The proposed architecture effectively captures spatiotemporal information from video frames and uses it to detect accidents. The combination of the I3D model, ConvLSTM2D layers, and feature fusion ensures that both spatial and temporal dependencies are considered, leading to accurate accident detection.

VI. EXPERIMENTAL RESULT

In this section, we present a detailed analysis of the performance of our proposed accident detection models with distinct configurations within the context of action recognition. The comparative analysis, as summarized in Table (2), focuses on the efficacy of these models in accurately detecting traffic accidents. Building upon recent advancements in the field, such as the Hierarchical Accident Recognition Method by Chen et al. [49] and Zhu et al.'s [50] traffic condition-based detection method, our approach introduces a novel perspective. We emphasize the extraction of both RGB frames and optical flow information from video sequences, harnessing the capabilities of the CONVLSTM2D architecture. This model, detailed in Figure [3], is pivotal in our research, as it integrates the strengths of convolutional and LSTM networks. Our architecture effectively captures the dynamic spatiotemporal characteristics of accidents, a critical aspect often overlooked in traditional CNN, RNN, and LSTM approaches. The ConvLSTM2D network is specifically designed to address the challenges associated with motion pattern recognition in accident scenarios. This includes analyzing the intricate interplay of spatial features (as seen in individual frames) and temporal patterns

TABLE 2. Summary of experimental result.

Model Name	Loss	Accuracy	Time	Precision	Recall	F1	Acc.	MAP
I3D-CONVLSTM2D RGB Only			110	0.73	0.72	0.72	0.72	0.78
I3D-CONVLSTM2D Non-Trainable RGB + Optical flow			120	0.75	0.75	0.75	0.75	0.81
I3D-CONVLSTM2D Augmented RGB + Optical flow			150	0.79	0.79	0.79	0.79	0.86
I3D-CONVLSTM2D Trainable RGB + Optical flow			130	0.80	0.80	0.80	0.80	0.87
DenseNet-Transformer RGB Only [47], [48]			300	0.75	0.75	0.75	0.74	0.80

(as observed across sequences of frames). Our experimental setup evaluates various architectural and parameter configurations to determine their effectiveness in isolating features indicative of acci dents. The results demonstrate a marked improvement in accident detection precision, showcasing the efficacy of our model in enhancing autonomous accident detection systems within smart city infrastructures and the ability to provide a more nuanced and accurate detection of traffic accidents.

A. I3D-CONVLSTM2D RGB ONLY

The I3D-CONVLSTM2D RGB Only model aimed to identify accidents by analyzing RGB frames exclusively in the input video. This model utilized a relatively modest configuration of ConvLSTM2D layers with 64, 32, and 32 filters. Additionally, dropouts were applied to mitigate overfitting. The model result showed improved performance in distinguishing accident-related features. It achieved mean average precision (MAP) of 78%, accuracy, precision, recall, and F1 score of 72%, respectively, after training on 30 epochs on extracted features from the I3D Model, signaling its ability to capture some of the critical cues associated with accident recognition.

B. I3D-CONVLSTM2D NON-TRAINABLE RGB + OPTICAL FLOW

The second model iteration adopted a nuanced approach to training by explicitly setting all layers of the I3D component to non-trainable, transforming it into a fixed feature extractor. We removed the classification head of the I3D model, replacing it with multiple layers of CONVLSTM2D designed to process both RGB frames and optical flow. Incorporating a

fixed feature extraction approach through a non-trainable I3D component, the model demonstrated a stable but unimproved performance when evaluated with a modest learning rate of $1\times\ 10^{-4}$. As illustrated in Table (2) row 2, the graphical trends for loss and accuracy provide visual confirmation of the stable performance over training epochs, complementing the numerical constancy of key metrics - accuracy, precision, recall, and F1 score at 75%, with a Mean Average Precision (MAP) of 81%. The steady lines in the loss and accuracy graphs indicate that while the model benefits from the introduction of the non-trainable I3D as a feature extractor, it does not significantly improve beyond a certain point, which could be attributed to the limitations of the non-trainable approach in adapting to the specific nuances of the dataset used for accident detection. This suggests that employing the I3D model in a non-trainable fashion reaches a plateau in effectiveness for new, unseen data. This model performance underscores the assertion that transitioning to a non-trainable I3D configuration, paired with CONVLSTM2D layers, can not succinctly capture relevant spatio-temporal features for accident detection using the I3D transfer-learning explicitly as a feature extractor.

C. I3D-CONVLSTM2D AUGMENTED RGB + OPTICAL FLOW

The third model, I3D-CONVLSTM2D Augmented RGB + Optical Flow, explored the impact of data augmentation techniques on accident detection. Specifically, our augmentation strategy comprised cropping, zooming, and careful rotation. Notably, rotations were cautiously capped to ensure realistic representations and vertical flipping was avoided to prevent generating videos that appeared upside-down. Video augmentation, including flipping and rotation, was applied to the dataset, aiming to augment temporal feature learning and enhance the recognition of accident-related features. Contrary to our initial hypothesis, the integration of these augmentation techniques did not precipitate the anticipated enhancement in accident detection. The model had a MAP of 86%, accuracy, precision, recall, and F1 score of 79%, respectively. This empirical evidence postulates that, within the constraints of our dataset, the data augmentation applied might not be as pivotal in learning accident-associated features.

D. I3D-CONVLSTM2D RGB + OPTICAL-FLOW (TRAINABLE)

Transfer learning is a core machine learning technique that extends knowledge acquired from a source domain to a related target domain, often yielding improved accuracy [51]. In our study, the I3D-CONVLSTM2D RGB + Optical-Flow (Trainable) model leverages this principle by integrating the I3D model, wherein the final N-15 layers are set as trainable [21]. This technique allowed us to utilize the inherent knowledge embedded within the I3D model, substantially improving our system's ability to extract salient features from both RGB and Optical Flow data. Key improvements were incorporated to optimize the model's performance. Batch normalization layers were introduced prior to the Global Average Pooling stage, significantly accelerating convergence during training. To enhance the model's feature extraction and representation capabilities, we expanded the number of neurons in the dense layers to 512, 256, and 256. A crucial modification was the addition of extra filters in the initial ConvLSTM2D layer, a change that markedly boosted the model's efficiency in processing spatiotemporal information. These strategic enhancements culminated in a notable increase in performance metrics. The model demonstrated a mean average precision (MAP) of 87%, with an accuracy, precision, recall, and F1 score of 80% each, as detailed in Table (2). These results underscore the model's enhanced proficiency in accurately detecting accidents from video streams.

E. DENSENET-TRANSFORMER RGB ONLY

In benchmarking our I3D-CONVLSTM2D models, we utilized the DenseNet architecture [47], recognized for its exceptional feature extraction capabilities through densely connected convolutional layers. Our implementation involved resizing videos to 224 $\times$ 224 dimensions, balancing computational efficiency with the extraction of rich features. Subsequently, these features were processed using a Transformer architecture [48], which is recognized for its exceptional ability to process sequential data and understand the context within a sequence. The Transformer's self-attention mechanism is instrumental in comprehending the temporal dynamics and relationships within the video data, a crucial aspect of accurate accident detection In our experimental setup, the DenseNet-Transformer model was trained on our comprehensive dataset, incorporating DenseNet for initial feature extraction, followed by transformer layers for sequence processing. The architecture also included additional dense layers and a fully connected classification layer for final predictions. The model achieved a precision, recall, and F1 score of 75%, as detailed in Table (2). This performance, achieved over a longer training duration, provides a baseline for assessing the effectiveness of our proposed I3D-CONVLSTM2D models. Notably, the DenseNet-Transformer's results offer insights into the computational demands and efficiency of real-time feature processing in accident detection scenarios, thereby establishing a comparative framework for evaluating the state-of-the-art methodologies in this field.

VII. DISCUSSION AND CONCLUSION

Accident detection methods have evolved significantly, transitioning from traditional human-based reporting to modern automated systems. These cutting-edge systems, utilizing sensors, machine learning algorithms, and computer vision, represent a paradigm shift in accident detection. Particularly, computer vision-based systems stand out for their real-time detection capabilities and adaptability to diverse road scenarios. As technology relentlessly advances, these systems are poised to become indispensable tools in traffic safety enhancement. Our research culminated in the development

of the I3D-CONVLSTM2D Trainable RGB + Optical Flow model, which demonstrated outstanding performance, with an accuracy of 0.80 and a mean average precision of 87%. This model's proficiency in isolating traffic accident features amidst complex traffic scenarios marks a significant step forward in automated accident detection Traffic accidents, particularly in densely populated areas, continue to be a major safety concern, accounting for substantial proportion of fatalities. Our study addresses this through the development of a vision-based accident detection system tailored for realtime deployment on edge IoT devices, such as Raspberry Pi. Recognizing the intrinsic challenges of such an approach, notably the massive data requirement, we took the initiative to curate a novel accident dataset. This resource can either complement existing datasets or be employed as a standalone tool, thereby granting researchers the flexibility to extend or modify our foundational framework for accident detection. The I3D two-stream network, trained on the Kinetics dataset with 25 million parameters and its extensive training process across 32 GPU1s for 110k steps, is computationally demanding. In contrast, our model, designed to be efficient and resource-conscious, was trained on an Ubuntu 20.04.2 LTS system leveraging 2 GPUs, each of 11 GB. The model specifications underscore our commitment to efficiency: the RGB model consists of 3 million parameters, and the I3DCONVLSTM2D Trainable RGB + Optical extends to 9 million parameters. In summary, our study bridges the gap between computational efficiency and practical applicability, offering a cost-effective, reliable accident detection system suitable for smart city infrastructures. Our study creates an avenue for more accessible and ubiquitous surveillance solutions. Most notably, our model simplicity, efficiency, and reduced computational demands, especially when juxtaposed against heavyweights like I3D, serve as a testament to our objective of creating lightweight yet effective solutions for critical societal challenges.

VIII. LIMITATIONS AND FUTURE WORK

Our evaluation showcased that stationary vehicles, whether parked or halted along the road, obscure and sometimes interfere with the visibility of an accident scene, potentially causing the detection system to flag such video input as a traffic accident. Differentiating between regular slow-moving traffic and traffic slowed due to an accident presents another challenge, as patterns in the former can often resemble post-accident scenes. Environmental factors, like dust, sand, smoke, or wind, often obscure vital details and make it difficult for the detection system to identify and classify incidents. Our study also acknowledges potential limitations related to the training data used for the accident detection system. The diversity and representativeness of the dataset are crucial for the system's ability to generalize to various real-world scenarios Considering these limitations, our future work would focus on refining the accident detection systems. This includes implementing advanced algorithms that can better discern between stationary obstacles and genuine accidents, as well as traffic patterns. Additionally, building algorithms robust enough to adjust for environmental interference will ensure that the scene remains interpretable, regardless of external conditions. These improvements aim to bolster the reliability and precision of traffic accident detection systems in a myriad of real-world scenarios.

REFERENCES

[1] V. Adewopo, N. Elsayed, Z. Elsayed, M. Ozer, V. Wangia-Anderson, and A. Abdelgawad, "AI on the road: A comprehensive analysis of traffic accidents and accident detection system in smart cities," 2023, *arXiv:2307.12128*.
[2] V. Adewopo, N. Elsayed, Z. ElSayed, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Review on action recognition for accident detection in smart city transportation systems," 2022, *arXiv:2208.09588*.
[3] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7210–7218.
[4] U. Khalil, A. Nasir, S. M. Khan, T. Javid, S. A. Raza, and A. Siddiqui, "Automatic road accident detection using ultrasonic sensor," in *Proc. IEEE 21st Int. Multi-Topic Conf. (INMIC)*, Nov. 2018, pp. 206–212.
[5] L. Yu, B. Du, X. Hu, L. Sun, L. Han, and W. Lv, "Deep spatio-temporal graph convolutional network for traffic accident prediction," *Neurocomputing*, vol. 423, pp. 135–147, Jan. 2021.
[6] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-based traffic accident detection and anticipation: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 1983–1999, Apr. 2024.
[7] V. Adewopo, N. Elsayed, Z. Elsayed, M. Ozer, C. Zekios, A. Abdelgawad, and M. Bayoumi, "Big data and deep learning in smart cities: A comprehensive dataset for AI-driven traffic accident detection and computer vision systems," 2024, *arXiv:2401.03587*.
[8] V. Adewopo, N. Elsayed, Z. ElSayed, M. Ozer, C. Zekios, A. Abdelgawad, and M. Bayoumi, Dec. 28, 2023, "Traffic accident detection video dataset for AI-driven computer vision systems in smart city transportation, *IEEE Dataport*, doi: 10.21227/tjtg-nz28.
[9] Y. Yang, H. Song, S. Sun, W. Zhang, Y. Chen, L. Rakal, and Y. Fang, "A fast and effective video vehicle detection method leveraging feature fusion and proposal temporal link," *J. Real-Time Image Process.*, vol. 18, no. 4, pp. 1261–1274, Aug. 2021.
[10] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1293–1301.
[11] S. Xu, S. Li, R. Wen, and W. Huang, "Traffic event detection using Twitter data based on association rules," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. IV-2, pp. 543–547, May 2019.
[12] J. Wang, Q. Chen, and H. Gong, "STMAG: A spatial–temporal mixed attention graph-based convolution model for multi-data flow safety prediction," *Inf. Sci.*, vol. 525, pp. 16–36, Jul. 2020.
[13] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Comput. Vis. Image Understand.*, vol. 134, pp. 1–21, May 2015.
[14] Y. Cai, H. Wang, X. Chen, and H. Jiang, "Trajectory-based anomalous behaviour detection for intelligent traffic surveillance," *IET Intell. Transp. Syst.*, vol. 9, no. 8, pp. 810–816, Oct. 2015.
[15] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.
[16] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near accident detection in traffic video," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 2, p. 23, Jan. 2019.
[17] N. Saunier and T. Sayed, "Automated analysis of road safety with video data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2019, no. 1, pp. 57–64, Jan. 2007.
[18] S. Robles-Serrano, G. Sanchez-Torres, and J. Branch-Bedoya, "Automatic detection of traffic accidents from video using deep learning techniques," *Computers*, vol. 10, no. 11, p. 148, Nov. 2021.
[19] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.

[20] N. Elsayed, A. S. Maida, and M. Bayoumi, "Reduced-gate convolutional LSTM architecture for next-frame video prediction using predictive coding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–9.

[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[22] X. Wang, Y. Ma, Y. Wang, W. Jin, X. Wang, J. Tang, C. Jia, and J. Yu, "Traffic flow prediction via spatial temporal graph neural network," in *Proc. Web Conf.*, vol. 11. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1082–1092.

[23] A. Srinivasan, A. Srikanth, H. Indrajit, and V. Narasimhan, "A novel approach for road accident detection using DETR algorithm," in *Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Oct. 2020, pp. 75–80.

[24] C. Wang, Y. Dai, W. Zhou, and Y. Geng, "A vision-based video crash detection framework for mixed traffic flow environment considering low-visibility condition," *J. Adv. Transp.*, vol. 2020, pp. 1–11, Jan. 2020.

[25] L.-M. Xia, X.-J. Hu, and J. Wang, "Anomaly detection in traffic surveillance with sparse topic model," *J. Central South Univ.*, vol. 25, no. 9, pp. 2245–2257, Oct. 2018.

[26] E. P. Ijjina, D. Chand, S. Gupta, and K. Goutham, "Computer vision-based accident detection in traffic surveillance," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–6.

[27] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 879–887, Mar. 2019.

[28] A. A. Mohammed, K. Ambak, A. M. Mosa, and D. Syamsunur, "A review of the traffic accidents and related practices worldwide," *Open Transp. J.*, vol. 13, no. 1, pp. 65–83, Jun. 2019.

[29] *Analyses of Rear-End Crashes and Near-Crashes in the 100-Car Naturalistic Driving Study to Support Rear-Signaling Countermeasure Development*, document DOT HS 810 846, 2007.

[30] L. Eboli, C. Forciniti, and G. Mazzulla, "Factors influencing accident severity: An analysis by road accident type," *Transp. Res. Proc.*, vol. 47, pp. 449–456, Jan. 2020.

[31] V. K. Kukkala, J. Tunnell, S. Pasricha, and T. Bradley, "Advanced driver-assistance systems: A path toward autonomous vehicles," *IEEE Consum. Electron. Mag.*, vol. 7, no. 5, pp. 18–25, Sep. 2018.

[32] *Newly Released Estimates Show Traffic Fatalities Reached a 16-Year High in 2021*, NHTSA Media, Washington, DC, USA, May 2022.

[33] V. Jirovský, "Classification of road accidents from the perspective of vehicle safety systems," *J. Middle Eur. Construction Design Cars*, vol. 13, no. 2, pp. 1–9, Nov. 2015.

[34] J. M. Nolan and A. K. Lund, "Frontal offset deformable barrier crash testing and its effect on vehicle stiffness," SAE Tech. Paper 2001-06-0109, 2001.

[35] T. Sheorey and N. V. Shrivas, "Development of sensor based front end collision avoidance system for highways," in *Proc. IEEE Int. Conf. Inf. Autom.*, Aug. 2015, pp. 594–598.

[36] J. Ma, J. Li, Z. Gong, and H. Huang, "An adaptive multi-staged forward collision warning system using a light gradient boosting machine," *Information*, vol. 13, no. 10, p. 483, Oct. 2022.

[37] F. Alharbi, M. I. Almoshaogeh, A. H. Ibrahim, H. Haider, A. E. M. Elmadina, and I. Alfallaj, "Performance appraisal of urban street-lighting system: Drivers' opinion-based fuzzy synthetic evaluation," *Appl. Sci.*, vol. 13, no. 5, p. 3333, Mar. 2023.

[38] G. Bridger and B. King, "Lighting the way to road safety: A policy blindspot?" in *Proc. Australas. Road Saf. Res. Policing Educ. Conf.*, Wellington, New Zealand, 2012, p. 22.

[39] P. R. Marchant and P. D. Norman, "To determine if changing to white light street lamps improves road safety: A multilevel longitudinal analysis of road traffic collisions during the relighting of leeds, a U.K. city," *Appl. Spatial Anal. Policy*, vol. 15, no. 4, pp. 1583–1608, Dec. 2022.

[40] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2682–2690.

[41] E. Zadobrischi, "Intelligent traffic monitoring through heterogeneous and autonomous networks dedicated to traffic automation," *Sensors*, vol. 22, no. 20, p. 7861, Oct. 2022.

[42] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 136–153.

[43] H. Ghahremannezhad, H. Shi, and C. Liu, "Real-time accident detection in traffic surveillance using deep learning," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Jun. 2022, pp. 1–6.

[44] R. K. Yadav, S. G. Neogi, and V. B. Semwal, "Human activity identification system for video database using deep learning technique," *Social Netw. Comput. Sci.*, vol. 4, no. 5, p. 600, Aug. 2023.

[45] V. Adewopo, N. Elsayed, and K. Anderson, "Baby physical safety monitoring in smart home using action recognition system," in *Proc. SoutheastCon*, Apr. 2023, pp. 142–149.

[46] N. Elsayed, A. S. Maida, and M. Bayoumi, "Empirical activation function effects on unsupervised convolutional LSTM learning," in *Proc. IEEE 30th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2018, pp. 336–343.

[47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2023, *arXiv:1706.03762*.

[49] M. Chen, J. Liang, J. Shu, and R. Zhu, "A hierarchical accident recognition method for highway traffic systems," *J. Phys., Conf. Ser.*, vol. 2456, no. 1, Mar. 2023, Art. no. 012034.

[50] L. Zhu, B. Wang, Y. Yan, S. Guo, and G. Tian, "A novel traffic accident detection method with comprehensive traffic flow features extraction," *Signal, Image Video Process.*, vol. 17, no. 2, pp. 305–313, Mar. 2023.

[51] V. S. Saravanarajan, R.-C. Chen, C. Dewi, L.-S. Chen, and L. Ganesan, "Car crash detection using ensemble deep learning," *Multimedia Tools Appl.*, vol. 83, no. 12, pp. 36719–36737, Jun. 2023.

VICTOR A. ADEWOPO (Member, IEEE) was born in Ibadan, Nigeria. He received the B.Sc. degree in computer science from Lead City University Ibadan, Nigeria, in 2019, and the M.S. and Ph.D. degrees in information technology from the University of Cincinnati, OH, USA, in 2021. From 2019 to 2022, he was a Research Assistant with the Applied Machine learning Laboratory, School of Information Technology, University of Cincinnati. His research interests include action recognition (AR), mainly on activity detection and incident prediction based on video streams. His awards and honors include the Data Science Fellowship (Lawrence Berkeley National Laboratory), the Google Generation Scholarship, and Research Fellowship Award (Graduate Student Government—University of Cincinnati).

NELLY ELSAYED (Member, IEEE) received the B.S. and M.S. degrees in computer science from Alexandria University and the M.S. (Eng.) and Ph.D. degrees from the University of Louisiana at Lafayette. She is currently an Assistant Professor with the University of Cincinnati, where she is also the Director and the Founder of the Applied Machine Learning and Intelligence Laboratory. She received the Love of Learning Award from the Honor Society Phi Kappa Phi, in 2018 and 2021. She has served as a principal investigator and a co-principal investigator in different federal, educational, and industrial level research projects. She received the Faculty Incentive Award for Research and Scholarship from the University of Cincinnati, recognizing her research contributions, journals and conference peer-reviewed publications, and professional presentations, in 2020 and 2021. She received the UCAADA Sarah Grant Barber Outstanding Advising Faculty Award for the academic year (2021–2022) University of Cincinnati.