IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 January 2024, accepted 8 February 2024, date of publication 19 February 2024, date of current version 23 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3366990*

RESEARCH ARTICLE

# Advancing Autonomous Vehicle Safety: Machine Learning to Predict Sensor-Related Accident Severity

**RAHMAN SHAFIQUE**[1], **FURQAN RUSTAM**[2], **SHERIFF MURTALA**[1], **ANCA DELIA JURCUT**[2], **(Member, IEEE), AND GYU SANG CHOI**[1], **(Member, IEEE)**

[1]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38541, South Korea
[2]School of Computer Science, University College Dublin, Dublin 4, D04 V1W8 Ireland

Corresponding author: Gyu Sang Choi (castchoi@ynu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2021R1A6A1A03039493.

**ABSTRACT** Autonomous vehicles (AVs) represent an exciting frontier in transportation, promising increased safety and efficiency on the roads. However, like any technological advancement, they are not immune to accidents. Understanding the severity of accidents involving AVs is crucial for enhancing their reliability and ensuring public trust in this transformative technology. To address this challenge, our study has employed cutting-edge natural language processing techniques combined with machine learning to predict the severity of accidents involving AVs. Our study has contributed significantly by creating a novel dataset derived from post-disengagement accident reports, covering the years 2019-2022. This dataset comprises detailed descriptions of accidents, sensor information, and other critical parameters. Moreover, we have introduced a novel approach called Multi-Distance Synthetic Technique (MDST) to balance the imbalanced nature of our dataset, which included only 334 samples due to the rarity of such accident data. Utilizing MDST for data balancing, we aimed to enhance the robustness of our analysis. Additionally, we employed Recursive Feature Selection (RFS) to extract a valuable feature set that was crucial in predicting accident severity. Leveraging this selected feature set, we trained an ensemble model, which remarkably outperformed expectations, achieving an impressive accuracy score of 0.92.

**INDEX TERMS** Autonomous vehicles, feature selection, machine learning, accident severity prediction, data balancing.

## I. INTRODUCTION

The advent of autonomous vehicles (AVs) has transformed the landscape of transportation, promising increased efficiency and safety on our roads [1]. These vehicles, equipped with state-of-the-art sensors and artificial intelligence, operate with minimal human intervention. However, as the deployment of AVs accelerates, so does the urgency to address safety concerns associated with these innovative technologies. Despite advancements in autonomous driving systems, accidents involving AVs remain a reality [2]. According to the World Health Organization (WHO), an estimated 1.35 million people succumb to road traffic accidents annually [3]. In the United Kingdom, recent statistics indicate that car accidents resulted in 1,460 fatalities, accompanied by a total of 115,584 reported injuries. Among these injuries, 22,069 were classified as severe, while 92,055 were categorized as minor. To ensure the road safety of AVs and reduce the accident rate, it is important to understand the factors contributing to the severity of these accidents. This understanding is crucial for developing effective safety measures and enhancing the overall reliability of autonomous transportation.

AVs have emerged as a transformative solution for establishing a secure and highly efficient urban transportation system. Forecasts by researchers at the US Department

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

of Transportation (DoT) [4] suggest that fully autonomous vehicles, relying on pre-existing street maps and onboard sensors, have the potential to reduce traffic fatalities by as much as 94%, particularly in cases linked to human errors. This notion is undeniably attractive, prompting a critical inquiry: 'What is the current level of safety in AVs?' While extensive research has focused on the safety of automation technology, examining its impact on traffic safety, congestion [5], legal and regulatory aspects of AV deployment in emergency situations [6], and addressing cybersecurity and communication security [7], there remains a significant gap in research addressing specific issues.

Many researchers have been actively engaged in the analysis of AVs road safety, exploring factors contributing to road accidents through various state-of-the-art methods. For instance, the study by [8] focuses on predicting car crash severity. This research draws on data obtained from the California Department of Motor Vehicles (CA-DMV), covering AVs disengagements and crashes spanning the years 2019 to 2022. Similarly, another study, [9], delves into the involvement of vulnerable road users (VRUs) in AV accidents. The primary objective of this research is to gain insights into the factors associated with specific crashes involving AVs and VRUs. Further, the study conducted by [8] contributes significantly to the field through an extensive examination of freeway traffic accident severity prediction. Their approach employs a multi-dimensional and multi-layer Bayesian network to unravel the complexities of accident severity prediction.

Despite these contributions, there is a noticeable gap in research concerning AVs accident severity prediction using machine learning. Addressing this gap could substantially enhance the reliability of AVs systems, contributing to the development of more robust and safer autonomous transportation systems. To shed light on the current safety landscape, this research aims to predict the severity of AV accidents using machine learning techniques. By leveraging data obtained from the CA-DMV and building upon the methodologies explored by previous researchers, our study seeks to fill the existing gap and provide valuable insights into the specific factors influencing the severity of AVs accidents.

By analyzing relevant features such as vehicle speed, weather conditions, road type, and sensor data, our model aims to classify accidents into different severity levels. This classification can provide valuable insights for emergency responders, enabling them to allocate resources more effectively and potentially reduce the consequences of accidents involving autonomous vehicles. Machine learning techniques have the potential to greatly assist in severity classification when it comes to various domains, including road accidents [10]. By leveraging large datasets and advanced algorithms, machine learning can effectively analyze and categorize the severity of incidents. In light of the existing literature review, we summarize our contributions in this paper as follows:

- Developed an innovative machine learning approach for predicting road accident severity, contributing to the advancement of AVs safety.
- Generated a benchmark dataset for road accidents using comprehensive data obtained from the CA DMV database. The data, initially in an unstructured format such as reports, underwent extensive preprocessing to transform it into structured data. This study conducted a thorough analysis and cleaning to prepare the dataset, which now serves as a valuable resource for training and evaluating the proposed machine learning model.
- Proposed a novel oversampling technique, the Multi-Distance-Based Oversampling Technique (MDST), to address potential model overfitting caused by the highlighted imbalanced dataset. Utilized multi-distance-based synthetic techniques to generate diverse and realistic synthetic samples. Implemented MDST to enhance the robustness of the machine learning model and reduce bias in the dataset.

The following sections of the paper are organized as follows: Section II provides a comprehensive literature review, Section III elaborates on the methodology and experiments, Section IV delves into the outcomes and findings, and finally, Section V summarizes the conclusion.

## II. LITERATURE REVIEW

Machine learning has proven to be a valuable tool in predicting the severity of road accidents in recent years. In this section, we review the literature on accident severity prediction using machine learning, deep learning, and statistical approaches. Additionally, we identify gaps in recent literature.

### A. SEVERITY PREDICTION USING MACHINE LEARNING APPROACHES

The study [11] contributes to the existing literature on predicting road accident severity using machine learning algorithms. The data was gathered in New Zealand during the timeframe spanning from 2016 to 2020. Random Forest (RF) exhibited superior performance with an accuracy of 67.67%. To assess the severity of accidents, the research [12] delved into a diverse set of mathematical and statistical tools, complemented by machine learning algorithms. The dataset was sourced from the UK road accident database. Within this study, three feature selection algorithms neighborhood Component Analysis (NCA), Rank Relief F, and the utilization of Partial Dependence Plots along with Individual Conditional expectations were used. For binary classification, the research constructed Support Vector Machine (SVM) models, employing a Gaussian Kernel, also recognized as the Radial Basis Function (RBF) kernel both models demonstrated an accuracy rate of 89.9%.

The study [13] employs descriptive analysis and ML to analyze the determining factors with the potential to cause an effect and predict the severity of road accidents. The study utilized a comprehensive dataset on car accidents, gathered from February 2016 to December 2019, encompassing 49 states of the United States. Among machine learning models, the RF algorithm performs well achieving a high accuracy rate of 97.2%. Similarly, another study [14] contributes to road traffic safety by demonstrating the effectiveness of data mining techniques in analyzing accident data and extracting valuable knowledge. The accident data was collected from the UK road traffic accident repository website which is accessible in the format of an Excel spreadsheet. The model J48 performs well. Another study [15] provides valuable insights into the application of machine learning to predict the severity of road traffic accidents. The authors of the paper utilized an RF model along with various data augmentation and feature selection techniques. No sampling + PCA (80) with 82% shows its exceptional performance as compared with ML algorithms. In the study [16] authors explore the use of machine learning algorithms for severity prediction of traffic accidents. The study utilizes the TRAFFIC ACCIDENTS_2019_LEEDS dataset obtained from the Road Safety Department of Transport. RF achieves an impressive accuracy of 93%. In study [17], authors highlight the limitations of using conventional descriptive statistics in identifying cause-and-effect relationships and developing predictive models for road accidents. The accident data was obtained from the Ministry of Economy and Finance (MEF) through a survey conducted in 2017 by the Financial Services Quality Observatory (OQSF) unit. The study employs supervised learning algorithms out of which (SVM,85.60%) performed exceptionally well.

The study [18] focuses on evaluating ML models to predict road accident severity using the latest road accident dataset from New Zealand. RF emerged as the top performer, exhibiting the highest accuracy(81.45%), precision, recall, and F1 score. The authors [19] highlight the significance of RTAs as a major cause of fatalities, particularly among children and youth, and emphasize the need for effective prediction models to mitigate this problem. The dataset used in this study consists of 11,014 traffic accident records from the Lebanese Internal Security Forces (ISF) for the years 2016-2017. The experimental results presented in the article demonstrate that the SVM model with RBF achieved the highest accuracy of 86%. The core work of the [20] by Miaomiao Yan and Yindong Shen focuses on addressing the challenge of accurately predicting the severity of traffic accidents. Within this study, a hybrid model named Bayesian optimization with random forest (BO-RF) is employed for forecasting traffic accident severity. These findings highlight RF(95.8%) the superior predictive capabilities.

The study [21] effectively addresses the issue of imbalanced crash datasets and provides a practical solution by proposing the use of random under-sampling of the majority class (RUMC). The study utilized several machine learning algorithms to predict crash severity, including random trees, K-nearest neighbor (KNN), logistic regression (LR), and RF. The KNN algorithm demonstrated a true positive rate of 18.3% for predicting fatal crashes and injuries in the imbalanced models. However, the RUMC-based models significantly improved accuracy, with a true positive rate of 57.2% for a KNN. The author in [22] presents a comprehensive study on the application of machine learning algorithms for predicting accident severity in smart transportation systems. The authors address the increasing concern of road accidents and aim to identify specific features that contribute to accident severity. The author utilized various machine learning models such as LR, Artificial Neural Network (ANN), Decision Tree(DT), and KNN to predict accident severity in a smart transportation system. The study found that the DT model achieved the highest mean accuracy of 71.44%, with a standard deviation (SD) of 2.19%. The study [23] provides a comprehensive exploration of a critical research problem and its significance, shedding light on the escalating incidence of road traffic injuries in African countries. By utilizing a motorcycle crash dataset from Ghana, the study employs three machine learning algorithms—J48 DT, RF, and Instance-Based learning with parameter k (IBk). The findings underscore the superiority of machine learning algorithms over the MNLM in terms of accuracy and effectiveness. Notably, the RF-based approach exhibits the highest accuracy of 73.91%.

In study [24], authors provide valuable insights into the analysis of road traffic accidents and the application of classification algorithms in handling accident data. The study aims to develop a decision support system for road traffic accident analysis using traditional machine learning algorithms. The authors used classification algorithms such as SMO, J48, and IBK, implemented in the Weka software, to develop a decision support system. The algorithms are tested on a sample database consisting of over 1,500 accident items, each with 29 attributes. The experimental results indicate that the SMO algorithm provides the most accurate results, achieving 94% accuracy. The study [25] explores the development and application of machine learning models to predict the severity of crash injuries. The study focuses on using 15 crash-related parameters and employs a clustering technique called fuzzy c-means (FCM) to enhance the predictive capability of the ML models. Four ML models are developed: feed-forward neural networks (FNN), SVM, FNN-FCM, and SVM-FCM. The combined use of SVM and FCM resulted in an overall testing accuracy of 74.2%, signifying a higher accuracy using SVM.

### B. SEVERITY PREDICTION USING DEEP LEARNING APPROACHES

The application of deep learning in accident severity prediction has also yielded significant results, as evidenced in [28] where valuable insights into the application of deep learning techniques, specifically ANN, in predicting accident severity were presented. The dataset comprises over 220,000 accident records obtained from the UK's Department for Transport, encompassing the year 2018. The article

**TABLE 1.** Summary of literature review.

| Ref | Year | Classifier | Dataset | Accuracy | Limitation |
|---|---|---|---|---|---|
| [12] | 2021 | LR, KNN, NB, RF, XGB,adaboost | [27] | 67.67% | The study compared feature importance in this study, but utilizing this information for the model training is a potential future step and a study limitation. |
| [13] | 2022 | NCA, KNN, SVM, RBF | [28] | 89.9% | The study does not quantify the factors responsible for the severity of road accidents. While the study identifies the most responsible factors, it does not provide a quantitative analysis or measurement of their impact on accident severity. |
| [29] | 2022 | SVM, GB, LR, RF, NB | [30] | 84.1% | Limited evaluation and comparison of the MLP model's performance against alternative models or techniques. |
| [14] | 2022 | LR, DT, KNN, RF | [31] | 97.2% | Generalizability of the findings to a global context may be limited. |
| [16] | 2022 | RF with (PCA, SMOTE, NearMiss) | [32] | 82.0% | The potential randomness and human influence in predicting accident severity, and the lack of a strong correlation between accident severity and most features used. |
| [17] | 2022 | SVM,RF,ANN | Not open source | 93.0% | Include its narrow focus on severity prediction for traffic accidents |
| [18] | 2022 | RF, KNN, LR, GNB | Not open source | 85% | Need to explore hyperparameter optimization techniques to enhance the efficiency of the RF algorithm for improved estimation of accident severity in Senegal |
| [33] | 2023 | ANN | Not open source | 80% | Include the need to compare the accuracy of different machine learning techniques, validate the model with datasets containing additional variables such as driver velocity or road speed limits, and explore datasets with more than three output classes to assess the impact of increased class diversity on model accuracy |

reports that the multi-layer perceptron (MLP) algorithm outperformed other methods and yields 84.1% accuracy. Similarly, study [32] offers valuable insights into the application of AI-based techniques for predicting and mitigating the accident's severity using ANNs. The data was sourced from the Calderdale government. The authors utilized ANN as the primary framework and achieved an accuracy of approximately 80%. Another study [33] explores the use of advanced ML approaches to predict the severity of traffic crashes. The study incorporated three advanced machine learning algorithms: a standard multi-layer perceptron (MLP) implemented with Keras, an MLP enhanced with embedding layers, and TabNet. In terms of training duration, the Keras MLP model demonstrated superior performance, completing training in 3.45 seconds. This signifies a substantial reduction of 51% and 93% compared to the MLP with embedding layers and TabNet, respectively.

### C. SEVERITY PREDICTION USING STATISTICAL ANALYSIS

The study [34] offers an extensive investigation into forecasting the severity of traffic. The study begins by establishing an XGBoost model and introducing the Shapley Additive explanations (SHAP) values to explain the model's predictions. A Bayesian network-based prediction model (BNA) is developed relative to the selected variables and their values. It implies that BNA is effective in predicting the severity of traffic accidents based on the selected variables and their values. In investigations [35], [36], [37], [38], researchers looked at severity from a medical standpoint. The abbreviated injury scale and injury severity score were used to quantify severity levels. Human behavior, demography, and the utilization of safety facilities were among the exploratory variables. However, certain elements, such as the "driver in the loop" issue in AVs, were neglected in previous medical investigations.

### D. LIMITATIONS AND EXISTING GAPS

Most of the researchers in this field have primarily used statistical modeling and supervised learning techniques in their studies. Their findings and predictions are based on a limited set of variables and data sources. While these studies have identified important factors that affect accident severity, such as weather conditions, lighting conditions, speed limits, and number of lanes, they have not taken into account other potentially influential variables, such as road surface conditions.

Furthermore, the majority of these studies have relied on state-of-the-art datasets extracted from autonomous vehicles at Level 0, where there is no automation. To address real-world crashes across all levels of automation, from Level 0 to Level 6, our research involves collecting information from post-accident crash reports. We then utilize various machine learning and deep learning algorithms, along with a specially designed data balancing method, to predict the severity of future accidents in California. Finally, we evaluate and compare the performance of these models to assess their effectiveness. Some recent studies have investigated accident severity prediction employing machine learning methods. The accuracies and limitations of these recent studies are presented in Table 1.

## III. PROPOSED METHODOLOGY

Figure 1, shows the proposed approach flow for autonomous vehicle accident severity predictions. In the initial step, we gather real-world automobile accident reports. These reports serve as our primary source of data, providing valuable insights into various accident scenarios, and contributing

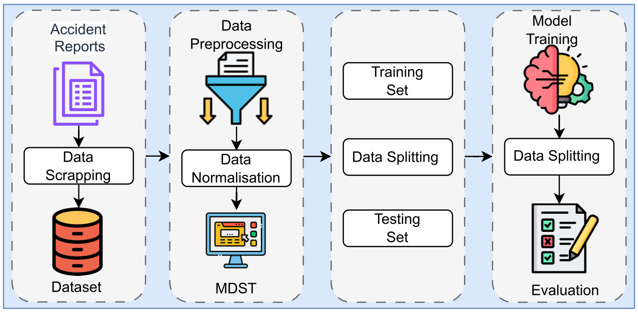


**FIGURE 1.** Proposed methodology diagram.

to the authenticity of our dataset. Utilizing real-world data ensures that our models are trained and tested on realistic accident cases. We extract structured information from the accident reports and organize it into a tabular dataset format. This transformation simplifies data handling, making it suitable for machine learning applications. Converting unstructured data into a structured format facilitates data analysis, model training, and feature engineering. In the preprocessing phase, we address data quality and consistency. We remove any null or missing values, ensuring the dataset's integrity and avoiding errors during model training. Additionally, we employ label encoding to convert categorical data into a numeric form, a necessary step for most machine-learning algorithms. Normalization standardizes the numerical attributes of the dataset, preventing certain features from having undue influence on machine learning models. This process enhances model stability and improves convergence during training.

Addressing class imbalance is vital for developing robust accident prediction models. In this step, we employ a data balancing approach, which could involve either oversampling or generating synthetic data. By balancing the dataset, we ensure that our models do not disproportionately favor the majority class, thus improving model generalization. Data splitting is essential to assess the performance of our machine learning models. We partition the dataset into training and test sets. The training set is used to train models, while the test set is reserved for evaluation. This separation helps in estimating how well our models generalize to unseen data. We train several machine learning models on the training data, leveraging various algorithms and techniques. By doing so, we provide our models with the ability to learn from the patterns and relationships in the dataset. We then evaluate model performance using the test data. The metrics used for evaluation include accuracy, precision, recall, and F1 score, providing a comprehensive assessment of our models' capabilities.

### A. DATASET COLLECTION METHOD

In this study, we curated our dataset related to autonomous vehicle accidents. Initially, we gathered data from an online public repository and subsequently conducted preprocessing to structure the data for model training.

**TABLE 2.** Traing and testing sample values distribution before and after augmentation.

| Class | Before Augmentation | | After Augmentation | |
|---|---|---|---|---|
| | **Training** | **Testing** | **Training** | **Testing** |
| AVdamagelevelMINOR | 209 | 28 | 212 | 25 |
| AVdamagelevelMODERATE | 48 | 4 | 211 | 26 |
| AVdamagelevelNONE | 34 | 1 | 216 | 21 |
| Total | 291 | 33 | 639 | 72 |

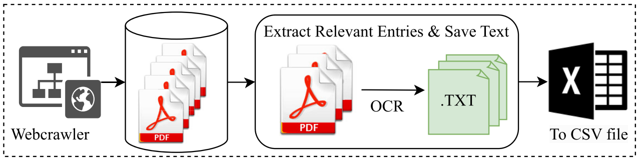


**FIGURE 2.** Our data scraping approach.

### B. DATASOURCE

In a publicly available digital repository, the California Department of Motor Vehicles Database (CDMV) keeps vehicle disengagement records and crash reports [39]. The disengagement reports provide useful information, but they lack unstructured narrative descriptions, rendering them unsuitable for topic modeling studies. As a result, our analyses in this paper focus on crash reports, which provide the following information:

1) Information regarding the automated vehicle's manufacturer, make model, and year, as well as any other cars involved in the incident.
2) Information on the location of the occurrence, such as the city, street, and intersections.
3) Information on the movements of the automated vehicle and other cars involved in the incident.
4) The crash's date and time.
5) Data on the weather, illumination, and road conditions at the time of the collision.
6) Identifying the type of object involved in the incident, such as a car, a bike, or a pedestrian.
7) Documentation of injuries and property damage received.
8) A narrative explaining the events leading up to the occurrence.

Our dataset comprises three distinct classes, denoted as follows:

- AVdamagelevelMINOR (Class 0)
- AVdamagelevelMODERATE (Class 1)
- AVdamagelevelNONE (Class 2)

The corresponding numerical representations for these classes are 0, 1, and 2, respectively. The distribution of samples for each class is succinctly summarized in Table 3.

### C. DATASET DESCRIPTION

During the study duration, the recorded severity levels of damage sustained by autonomous vehicles were distributed as follows: no damage (7.14%), minor damage (71.43%), severe damage (20.41%), and serious damage (1.02%). The assess-

**TABLE 3. Sample values of dataset.**

| AV Manufacture | AV Make | Accident Country | Accident time | AV damage level MINOR | AV damage level MODEATE | AV damage level MAJOR |
|---|---|---|---|---|---|---|
| Aimotive, Inc | Toyota | Santa Clara | 10:00 | 1 | 0 | 0 |
| Aurora Innovation Inc | Lincoln | San Francisco | 2:52 | 0 | 1 | 0 |
| Waymo LLC | Chrysler | Santa Clara | 6:56 | 0 | 0 | 1 |
| Zoox Inc | Toyota | San Francisco | 11:16 | 1 | 0 | 0 |

ment of crash severity heavily relies on the kinetic energy involved in the collision [40], [41]. Before implementing data augmentation, Table 2 presents the distribution of training and testing samples. Our dataset has been divided into a 90:10 ratio, allocating 90% for training and 10% for testing.

Table 3 showcases a sample of values, providing an initial glimpse into the data's content. This subset offers a starting point for analysis, enabling insights and informed decisions. Figure 2 illustrates the process flow for the extraction of data from digital reports. Using Webcrawler to retrieve the individual report the original character recognition (OCR) is used. to extract relevant information from the report.

### D. DATA ANAYLASIS

Figure 3 focuses on the number of accidents categorized by mode, collision type, injury, and property damage. It reveals that autonomous mode has the highest number of accidents, rear-end collisions are the most prevalent, and most accidents result in no injuries or minor property damage. These findings emphasize the need for preventive measures to address specific collision types and injury prevention efforts. While Figure 4 provides statistics on accidents based on AV manufacturers, AV models, and specific makes. It shows that Waymo has the highest number of accidents among manufacturers, Bolt has the highest number of accidents among AV models, and Chevrolet and Jaguar have the highest number of accidents among specific makes. These statistics highlight the need for further evaluation of safety measures and protocols employed by manufacturers.

Overall, these figures underscore the importance of enhancing safety standards, rigorous testing protocols, and effective risk mitigation strategies in the autonomous vehicle industry. Addressing these concerns is essential for building public trust, ensuring safety, and facilitating the wider adoption of autonomous vehicles in transportation systems.

### E. DATA PREPROCESSING

In this dataset, we also undertake preprocessing steps, including the removal of missing values and the application of encoders for categorical variables. The process of data preprocessing is essential for enhancing the performance and effectiveness of machine learning algorithms. It significantly improves data quality by addressing issues like missing values, data inconsistencies, and outlier detection. In study [42], authors extensively elaborated on how clean and reliable data can minimize the likelihood of noise or bias in the model, ultimately leading to more accurate results.

Following the removal of missing values is the elimination of redundant features, and the dataset's dimensionality is reduced, thereby decreasing the risk of overfitting and enhancing efficiency. This scenario makes feature selection more straightforward, enabling the most impactful features to help the model capture essential patterns and data relationships, as observed in [43].

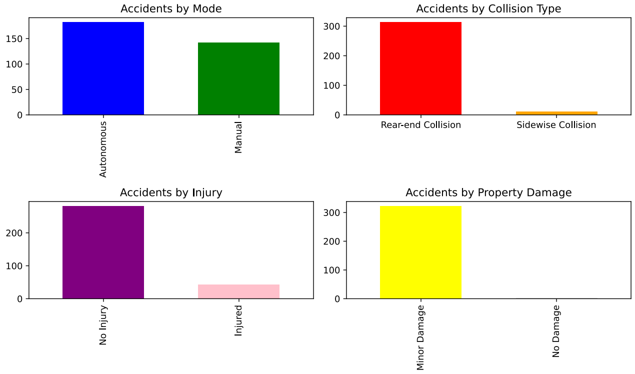


**FIGURE 3. Unveiling accident statistics: Mode, Collision, Property damage, and Injury breakdown.**

As numerous machine learning techniques require numerical inputs, it becomes necessary to represent categorical data effectively. To convert categorical data into numeric formats that can be understood by models, methods like encoding or label encoding are employed. Proper encoding ensures that categorical variables are utilized appropriately by the model, leading to an enhancement in its performance as described in study [44]. The study [45] described the data preparation approaches like oversampling, under-sampling, and the development of synthetic samples that can successfully address class imbalance issues. Equilibrating the dataset improves the model's ability to learn from under-represented classes and forecast accurately across all categories.

### F. MULTI-DISTANCES SYNTHETIC TECHNIQUE (MDST)

MDST, a data augmentation technique, mitigates model overfitting by leveraging Euclidean and Manhattan distances. Euclidean excels in capturing geometric relationships, emphasizing straight-line distances [46], while Manhattan, known for its robustness to outliers, is effective in scenarios with significant axis-aligned movements [47]. This deliberate combination enhances MDST's ability to comprehensively explore datasets, capturing diverse patterns and relationships. Figure 5 illustrates the architecture diagram for the MDST approach.

First, we select one sample from our dataset and denote it as $X$. This sample contains $n$ features. This is the sample for which we want to generate a synthetic counterpart. Next, we calculate the distances between $X$ and all other samples in our dataset. We compute two types of distances: Euclidean distance ($D_E$) and Manhattan distance ($D_M$). The Euclidean distance is computed using the standard Euclidean distance formula for $X$ and each sample in the dataset. The Manhattan distance, on the other hand, uses the Manhattan distance formula [46], [47]. Euclidean Distance ($D_E$) calculation for

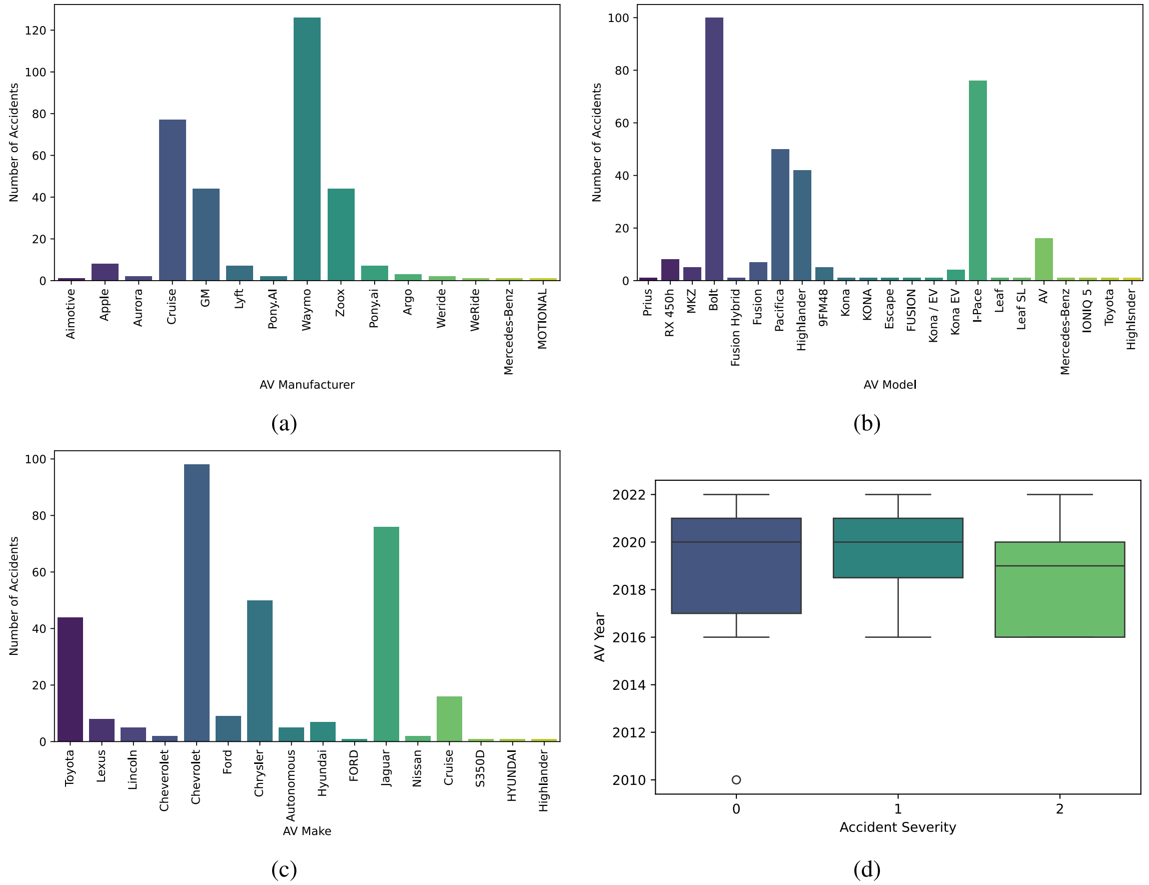


**FIGURE 4.** Accidents statistics by, (a) AV manufacture, (b) AV model, (c) AV make, (d) Severity.

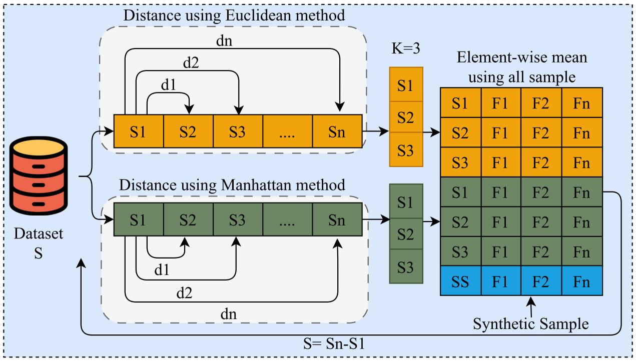


**FIGURE 5.** MDST architecture diagram for augmentation.

$X$ and sample $i$:

$$D_E(X,i) = \sqrt{\sum_{j=1}^{n}(X_j - i_j)^2}$$

where $X_j$ and $i_j$ are the $j$-th components of vectors $X$ and $i$ respectively, and $n$ is the total number of components.

- Manhattan Distance ($D_M$) Calculation for $X$ and sample $i$:

$$D_M(X,i) = \sum_{j=1}^{n} |X_j - i_j|$$

where $X_j$ and $i_j$ are the $j$-th components of vectors $X$ and $i$ respectively, and $n$ is the total number of components.

For both Euclidean and Manhattan distances, we identify the three samples in the dataset that are closest to the selected sample $X$. These nearest neighbors are denoted as $N_{E1}$, $N_{E2}$, $N_{E3}$ for Euclidean distance and $N_{M1}$, $N_{M2}$, $N_{M3}$ for Manhattan distance. After finding the nearest neighbors using both distance metrics, we calculate the element-wise mean for the selected sample $X$ and its nearest neighbors. For the Euclidean neighbors, the element-wise mean is denoted as $\overline{M_E}$, and for the Manhattan neighbors, it is $\overline{M_M}$. The element-wise mean is computed by taking the mean of corresponding feature values for $X$ and its neighbors. Element-wise mean $\overline{M_E}$ is calculated for each feature $j$ as:

$$\overline{M_{Ej}} = \frac{X_j + N_{E1j} + N_{E2j} + N_{E3j}}{4}$$

where $X_j$ is the $j$-th component of vector $X$, and $N_{E1j}$, $N_{E2j}$, $N_{E3j}$ are the $j$-th components of vectors $N_{E1}$, $N_{E2}$, and $N_{E3}$ respectively.

- For Manhattan Neighbors: - Element-wise mean $\overline{M_M}$ is calculated for each feature $j$ as:

$$\overline{M_{Mj}} = \frac{X_j + N_{M1j} + N_{M2j} + N_{M3j}}{4}$$

where $X_j$ is the $j$-th component of vector $X$, and $N_{M1j}$, $N_{M2j}$, $N_{M3j}$ are the $j$-th components of vectors $N_{M1}$, $N_{M2}$, and $N_{M3}$ respectively.

With the element-wise mean calculated, we generate a new synthetic sample. We can use either the element-wise mean $\overline{M_E}$ or $\overline{M_M}$ as a template for the synthetic sample. The synthetic sample is created by replacing the feature values of $X$ with the corresponding values from the element-wise mean.

We repeat the entire process by selecting different samples from our dataset and generating new synthetic samples. This iterative procedure allows us to create synthetic samples based on the characteristics of different data points in our dataset. Throughout this process, we utilize both Euclidean and Manhattan distances to identify nearest neighbors and calculate element-wise means, ultimately generating synthetic samples that capture the characteristics of the original data.

### G. MACHINE LEARNING HYPERPARAMETER SETTINGS

Within this segment, we delve into the region of machine learning and deep learning techniques that were applied to classify the severity of accidents. Our approach involved utilizing a variety of algorithms, including RF, DT, LR, SVM, Naive Bayes (NB), KNN, Recurrent neural network(RNN), convolutional neural network(CNN), and an ensemble of RF and SVM, Each fine-tuned with their respective optimal hyperparameter settings. The fine-tuning process was executed within specified parameter ranges, as outlined in Table 4, which presents the details of parameter configurations and tuning intervals for these machine-learning models. The hyperparameter for RF and DT are n_estimators = 300 signifies that the RF consists of 300 trees. Increasing the number of trees can potentially enhance model performance, but it also leads to higher computational costs. For seeding the random number generator random_state = 52 parameter was used. Fixing this number ensures that the model exhibits deterministic behavior, meaning it will consistently produce the same results across each run max_depth = 50 for RF determines the maximum depth of individual decision trees within the RF. For SVM we've utilized a Linear kernel function, indicating that the decision boundary of the SVM is a hyperplane. In this case, the regularization level is set at C = 1.0, which can be considered moderate. For KNN hyperparameter value n_neighbors = 3 we set the parameter n_neighbors to 3, which determines the utilization of the three closest neighbors when making predictions for a given data point.

Deep learning algorithms, RNN, and CNN are also considered in this study. The CNN model includes 1D convolutional (Conv) layers with a 3 × 3 kernel size, max-pooling layers, a global max-pooling layer, a flattening layer to transform the data into a one-dimensional format, and dense layers. The RNN model includes 3*3 kernel size along with activation type sigmoid and batch size 32 parameter governs the number of training examples used in each forward and backward pass within a single epoch of training.

**TABLE 4. Hyperparameters and their tuned values for experiments.**

| Model | Hyperparamets | Tuning Range |
|---|---|---|
| DT | max_depth =300, random_state=42 | max_depth ={5 to 200} |
| LR | solver='liblinear', multi_class='multinomial', C=3.0 | solver='liblinear', C=[1.0 to 3.0] |
| SVM | kernel=linear', C=1.0 | kernel=[linear, Polynomial, Radial Sigmoid] , C=[1.0 to 5.0] |
| RF | n_estimators = 300, max_depth =50, random_state=52 | n_estimators = {20 to 400}, max_depth ={5 to 200} |
| KNN | n_neighbours=3 | n_neighbours={1 to 10 } |
| RNN | input_shape=(189,1), activaition='relu', activation='sigmoid' ,loss='binary_crossentropy', epochs=10, batch_size=32 | - |
| CNN | input_shape=(189.1), activaition='relu', activation='sigmoid', loss='binary_crossentropy epochs=10, batch_size=32 | - |

## IV. RESULTS & DISCUSSION

The proposed approach was experimentally evaluated on a Windows operating system, specifically on a Core i7 12th generation machine. The machine was equipped with 64 GB of RAM and a 1TB SSD. The implementation of the proposed approach was carried out in Python language using Jupyter Notebook. Various libraries such as TensorFlow, Keras, scikit-learn, and pandas were utilized in the implementation process. To assess the performance of classifier, standard metrics such as accuracy, precision, recall, and F1 score are employed and calculated using the following equations.

$$Accuracy = \frac{T_rP_o + T_rN_e}{T_rP_o + T_rN_e + FP_o + FN_e} \tag{1}$$

$$P = \frac{T_rP_o}{T_rP_o + FP_o} \tag{2}$$

$$R = \frac{T_rP_o}{T_rP_o + FN_e} \tag{3}$$

$$F1 - score = 2 * \frac{P * R}{P + R} \tag{4}$$

### A. RESULTS USING ORIGINAL DATASET

This section presents the outcomes of severity detection using various machine learning and deep learning algorithms on the original dataset. The analysis involved the utilization of a diverse set of machine learning algorithms, including RF, DT, KNN, LR, NB, and SVM. For deep learning, both RNN and CNN were employed, and the results are depicted in Table 5. Notably, among these models, NB and SVM exhibited superior performance with micro-average scores of 0.62 and 0.61, respectively. This can be attributed to the small size of the dataset, which had a suppressive effect on the other

**Algorithm 1** MDST Algorithm

```
Input:
 Dataset with $n$ features and $m$ samples
Output:
 Synthetic samples
while not converged do
    Select a sample $X$ from the dataset                          ▷ Step 1
    Calculate Euclidean Distance Matrix $D_E$                     ▷ Step 2
    Calculate Manhattan Distance Matrix $D_M$                     ▷ Step 2
    Find the 3 nearest neighbors with Euclidean distance:
        $N_{E1}, N_{E2}, N_{E3}$                                  ▷ Step 3
    Find the 3 nearest neighbors with Manhattan distance:
        $N_{M1}, N_{M2}, N_{M3}$                                  ▷ Step 3
    Calculate Element-Wise Mean for Euclidean Neighbors:
        $\overline{M_E}$                                          ▷ Step 4
    Calculate Element-Wise Mean for Manhattan Neighbors:
        $\overline{M_M}$                                          ▷ Step 4
    Generate a new synthetic sample using $\overline{M_E}$ or $\overline{M_M}$   ▷ Step 5
end while
```

models RF, KNN, LR, and SVM, resulting in scores of 0.31, 0.29, 0.31, and 0.31, respectively. As our original dataset is small, it results to challenges such as overfitting, where the models KNN, RF, LR, and SVM may perform well on the training data but poorly on new, unseen data. Overfitting is a concern, especially when the dataset size is small because the model might capture noise or specific patterns that do not generalize well. Additionally, an ensemble model combining SVM and RF achieved an accuracy of 0.31 in both HV and LV scenarios. The deep learning models, RNN and CNN, also yielded a micro-average of 0.31. These findings underscore the inconsistent performance of the models and highlight the imperative need for overall accuracy improvement. To overcome this constraint, we redirect our attention towards feature engineering, aiming to extract optimal features and augment the sample size for improved performance.

### B. RESULTS USING RFE AND CTGAN APPROACH

Feature engineering and augmentation are pivotal strategies in mitigating overfitting and enhancing the performance of machine learning and deep learning models so this section elaborates on the performance of different ML and DL models by extracting the best features using Recursive Feature Elimination (RFE) and then applying augmentation using Conditional Tabular Generative Adversarial Network (CTGAN) on these features to produces more synthetic samples. Table 6 shows results after applying feature engineering to extract 100 best features and the generation of 237 new samples for each class by applying CTGAN. As a result, the performance of machine learning models, RF and SVM significantly improved, achieving accuracy scores of 88% with a micro of 88% and 86% with a micro average of 87%, respectively. The combination of RFE for feature selection and CTGAN for synthetic data generation enhances the robustness and generalization capability of RF and SVM models, contributing to better performance on unseen data and addressing challenges associated with high-dimensional datasets. The performance of RNN and CNN is low and it can be their struggle to effectively capture relevant patterns in the data. RNNs might face challenges in handling long-term dependencies or exploiting sequential information, particularly if the dataset lacks strong sequential patterns. To address this limitation, our focus shifts to data balancing techniques, incorporating a novel MDST technique.

### C. RESULTS USING RFE AND MDST APPROACH

This section details the outcomes achieved by integrating our innovative MDST technique with Recursive RFE. Initially, the RFE technique is applied to select the best 100 features, followed by the application of the MDST technique. After this combined approach, the results from both ML and DL models exhibit a significant improvement. The accuracy scores are presented in 7, where SVM performance stands out, achieving an outstanding micro-average score of 0.92. The exceptional performance of all machine learning models can be attributed to the acquisition of more accurate and synthetic data. CTGAN, by design, does not involve the calculation of a distance formula. Its primary objective is to understand and replicate the underlying distribution of the input data. It generates synthetic samples that reflect the statistical patterns inherent in the real data. In contrast, our proposed technique incorporates the calculation of two types of distance metrics— Euclidean and Manhattan distances. After computing these distances and taking the mean with the original samples, the technique produces new synthetic samples that closely resemble the original ones.

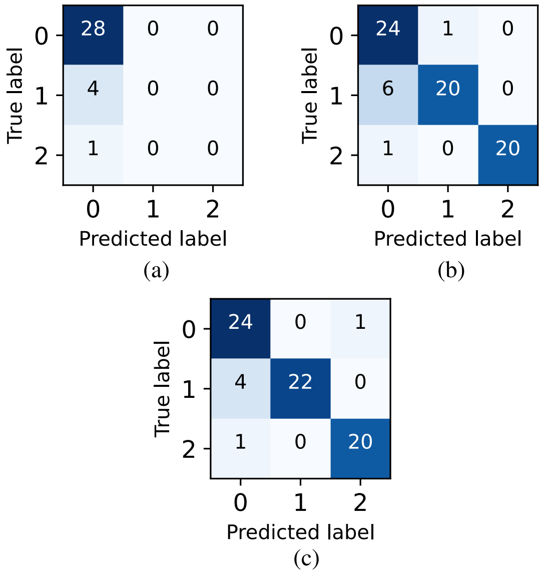
Figure 6 shows the confusion matrices of the best performers using each approach. In the original dataset approach,

**TABLE 5.** Machine learning and deep learning results on the original dataset.

| Model | Class | Precision | Recall | F1-score | Model | Class | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| RF | 0 | 0.85 | 1.00 | 0.92 | NB | 0 | 0.95 | 0.68 | 0.79 |
|  | 1 | 0.00 | 0.00 | 0.00 |  | 1 | 0.27 | 0.75 | 0.40 |
|  | 2 | 0.00 | 0.00 | 0.00 |  | 2 | 0.50 | 1.00 | 0.67 |
|  | Micro. avg | 0.28 | 0.33 | 0.31 |  | Micro. avg | 0.57 | 0.81 | 0.62 |
|  | WAVG. | 0.72 | 0.85 | 0.78 |  | WAVG. | 0.85 | 0.70 | 0.74 |
|  | Accuracy | 0.85 |  |  |  | Accuracy | 0.70 |  |  |
| DT | 0 | 0.93 | 0.89 | 0.91 | RNN | 0 | 0.85 | 1.00 | 0.92 |
|  | 1 | 0.25 | 0.25 | 0.25 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.50 | 1.00 | 0.67 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.56 | 0.71 | 0.61 |  | Micro. avg | 0.28 | 0.33 | 0.31 |
|  | WAVG. | 0.83 | 0.71 | 0.61 |  | WAVG. | 0.72 | 0.85 | 0.78 |
|  | Accuracy | 0.82 |  |  |  | Accuracy | 0.85 |  |  |
| KNN | 0 | 0.84 | 0.93 | 0.88 | CNN | 0 | 0.85 | 1.00 | 0.92 |
|  | 1 | 0.00 | 0.00 | 0.00 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.00 | 0.00 | 0.00 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.28 | 0.31 | 0.29 |  | Micro. avg | 0.28 | 0.33 | 0.31 |
|  | WAVG. | 0.71 | 0.79 | 0.75 |  | WAVG. | 0.72 | 0.85 | 0.78 |
|  | Accuracy | 0.79 |  |  |  | Accuracy | 0.85 |  |  |
| LR | 0 | 0.85 | 1.00 | 0.92 | HV | 0 | 0.85 | 1.00 | 0.92 |
|  | 1 | 0.00 | 0.00 | 0.00 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.00 | 0.00 | 0.00 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.28 | 0.33 | 0.31 |  | Micro. avg | 0.28 | 0.33 | 0.31 |
|  | WAVG. | 0.72 | 0.85 | 0.78 |  | WAVG. | 0.72 | 0.85 | 0.78 |
|  | Accuracy | 0.85 |  |  |  | Accuracy |  | 0.85 |  |
| SVM | 0 | 0.85 | 1.00 | 0.92 | SV | 0 | 0.85 | 1.00 | 0.92 |
|  | 1 | 0.00 | 0.00 | 0.00 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.00 | 0.00 | 0.00 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.28 | 0.33 | 0.31 |  | Micro. avg | 0.28 | 0.33 | 0.31 |
|  | WAVG. | 0.72 | 0.85 | 0.78 |  | WAVG. | 0.72 | 0.85 | 0.78 |
|  | Accuracy | 0.85 |  |  |  | Accuracy |  | 0.85 |  |

SVM performed well in comparison with other models with 28 correct predictions and 5 incorrect predictions. Similarly, when employing the RFE and CTGAN approach, NB emerged as the best performer, yielding 64 correct predictions and 8 incorrect predictions. Regarding the RFE and MDST approach, SVM once again excelled, providing 66 correct predictions and 6 incorrect predictions. These statistics of the confusion matrix show the significance of our MDST approach as compared to others.

Figure 7 illustrates the comparison of model performance concerning various evaluation matrices across different approaches. In Figure 7a, it's evident that while model accuracy appears satisfactory, other evaluation matrices exhibit lower values, indicating a clear sign of model overfitting towards the majority class data due to the imbalanced dataset. However, after employing CTGAN and MDST, models perform equally well across all evaluation matrices. Particularly, models utilizing MDST showcase superior performance across all evaluation matrices compared to other approaches.



(a)

(b)

(c)

**FIGURE 6.** Best performer confusion matrix:(a) Confusion matrix of SVM using original dataset, (b) Confusion matrix of NB using RFE and CTGAN, and (c) Confusion matrix of SVM using RFE and MDST.

## D. EVALUTION IN TERMS OF COMPUTATIONAL COST

In this section, we evaluate model computational costs in terms of training and prediction times, measured in seconds (sec) and milliseconds (ms) respectively. Table 8 summarizes the training and prediction times for models using the original dataset, as well as those incorporating CTGAN and MDST approaches. RF, DT, and LR exhibit shorter training times

**TABLE 6.** Machine learning and deep learning results using RFE and CTGAN methods.

| Model | Class | Precision | Recall | F1-score | Model | Class | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| RF | 0 | 0.75 | 0.96 | 0.84 | NB | 0 | 0.81 | 0.88 | 0.85 |
|  | 1 | 1.00 | 0.77 | 0.87 |  | 1 | 0.95 | 0.81 | 0.88 |
|  | 2 | 0.95 | 0.90 | 0.93 |  | 2 | 0.91 | 1.00 | 0.95 |
|  | Micro. avg | 0.90 | 0.88 | 0.88 |  | Micro. avg | 0.89 | 0.90 | 0.89 |
|  | WAVG. | 0.90 | 0.88 | 0.88 |  | WAVG. | 0.89 | 0.89 | 0.89 |
|  | Accuracy | 0.88 |  |  |  | Accuracy | 0.89 |  |  |
| DT | 0 | 0.77 | 0.80 | 0.78 | RNN | 0 | 0.35 | 1.00 | 0.52 |
|  | 1 | 0.95 | 0.77 | 0.85 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.80 | 0.95 | 0.87 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.84 | 0.84 | 0.83 |  | Micro. avg | 0.12 | 0.33 | 0.17 |
|  | WAVG. | 0.84 | 0.83 | 0.83 |  | WAVG. | 0.12 | 0.35 | 0.18 |
|  | Accuracy | 0.81 |  |  |  | Accuracy | 0.35 |  |  |
| KNN | 0 | 0.74 | 0.92 | 0.82 | CNN | 0 | 0.35 | 1.00 | 0.52 |
|  | 1 | 0.94 | 0.65 | 0.77 |  | 1 | 0.00 | 0.00 | 0.00 |
|  | 2 | 0.83 | 0.90 | 0.86 |  | 2 | 0.00 | 0.00 | 0.00 |
|  | Micro. avg | 0.84 | 0.83 | 0.81 |  | Micro. avg | 0.12 | 0.33 | 0.17 |
|  | WAVG. | 0.84 | 0.82 | 0.81 |  | WAVG. | 0.13 | 0.36 | 0.18 |
|  | Accuracy | 0.82 |  |  |  | Accuracy | 0.35 |  |  |
| LR | 0 | 0.75 | 0.72 | 0.73 | HV | 0 | 0.75 | 0.84 | 0.79 |
|  | 1 | 0.77 | 0.65 | 0.71 |  | 1 | 0.83 | 0.77 | 0.80 |
|  | 2 | 0.81 | 1.00 | 0.89 |  | 2 | 1.00 | 0.95 | 0.98 |
|  | Micro. avg | 0.78 | 0.79 | 0.78 |  | Micro. avg | 0.86 | 0.85 | 0.86 |
|  | WAVG. | 0.78 | 0.78 | 0.77 |  | WAVG. | 0.85 | 0.85 | 0.85 |
|  | Accuracy | 0.78 |  |  |  | Accuracy | 0.85 |  |  |
| SVM | 0 | 0.78 | 0.84 | 0.81 | SV | 0 | 0.78 | 0.84 | 0.81 |
|  | 1 | 0.84 | 0.81 | 0.82 |  | 1 | 0.83 | 0.77 | 0.80 |
|  | 2 | 1.00 | 0.95 | 0.98 |  | 2 | 0.95 | 0.95 | 0.95 |
|  | Micro. avg | 0.87 | 0.87 | 0.87 |  | Micro. avg | 0.85 | 0.85 | 0.85 |
|  | WAVG. | 0.87 | 0.86 | 0.86 |  | WAVG. | 0.85 | 0.85 | 0.85 |
|  | Accuracy | 0.86 |  |  |  | Accuracy | 0.85 |  |  |

(0.009 to 0.44 sec), while RNN and CNN require longer periods (5.64 and 1.22 sec). Prediction times are low for NB, LR, SVM, and DT (0.99 to 2.99 ms), with ensemble methods showing moderate training times (0.087 to 0.376 sec) and prediction times averaging around 0.88 ms. NB, LR, DT, and SVM demonstrate similar computational costs, but the high accuracy of SVM makes it more suitable for our proposed approach with RFE and MDST. The trade-off between accuracy and computational cost is notably higher in deep learning and ensemble models compared to individual machine learning models.

### E. RESULTS USING 10-FOLD CROSS-VALIDATION APPROACH

To validate the proposed approach, we deployed a 10-fold cross-validation. This validation involved dividing the data into 10 folds, using a different fold each time for testing while the remaining nine were utilized for model training. Table 9 presents the mean accuracy and SD resulting from this 10-fold cross-validation experiment.

The outcomes from employing both RFE and CTGAN indicate that the models did not perform as well in terms of mean accuracy compared to using RFE and the MDST approach. Particularly, the performance of the RRF model stands out with a mean accuracy score of 0.76 and an SD of 0.06. This could be attributed to RF's ensemble architecture, suitable for both small and large datasets. RF demonstrates robustness against overfitting, especially when contrasted with more complex models like deep neural networks, making it adept at handling smaller datasets by minimizing the risk of memorizing noise.

Conversely, the performance of models is notably more significant with the proposed MDST method as the SVM outperforms significantly with an accuracy score of 0.91 and an SD of 0.09. Our proposed approach generates more linearly separable data through augmentation, aiding the SVM linear kernel in learning patterns more accurately compared to other models. Similarly, the data generated by MDST also improves linear models like LR, elevating its performance from 0.72 to 0.86. These results and statistical analyses underscore the significance of our proposed approach. Overall, the results post-data augmentation exhibit high mean accuracy, indicating minimal chances of overfitting among the models.

### F. COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

In this section, we conducted a comparative analysis between our proposed approach and recent studies in severity prediction to highlight its significance. Given the novelty of our dataset, we aimed for a fair comparison by applying the methodologies of recent studies to our dataset. We repeated

**TABLE 7.** Machine learning and deep learning results using RFE and MDST.

| Model | Class | Precision | Recall | F1-score | Model | Class | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| RF | 0 | 0.79 | 0.92 | 0.85 | NB | 0 | 0.83 | 0.20 | 0.32 |
| | 1 | 1.00 | 0.81 | 0.89 | | 1 | 0.60 | 0.96 | 0.74 |
| | 2 | 0.91 | 0.95 | 0.93 | | 2 | 0.83 | 0.95 | 0.89 |
| | Micro. avg | 0.90 | 0.89 | 0.89 | | Micro. avg | 0.75 | 0.70 | 0.65 |
| | WAVG. | 0.89 | 0.89 | 0.89 | | WAVG. | 0.75 | 0.69 | 0.64 |
| | Accuracy | 0.89 | | | | Accuracy | 0.69 | | |
| DT | 0 | 0.84 | 0.84 | 0.84 | RNN | 0 | 0.35 | 1.00 | 0.52 |
| | 1 | 0.96 | 0.88 | 0.92 | | 1 | 0.00 | 0.00 | 0.00 |
| | 2 | 0.87 | 0.95 | 0.91 | | 2 | 0.00 | 0.00 | 0.00 |
| | Micro. avg | 0.89 | 0.89 | 0.89 | | Micro. avg | 0.12 | 0.33 | 0.17 |
| | WAVG. | 0.89 | 0.89 | 0.89 | | WAVG. | 0.12 | 0.35 | 0.18 |
| | Accuracy | 0.89 | | | | Accuracy | 0.35 | | |
| KNN | 0 | 1.00 | 0.28 | 0.44 | CNN | 0 | 0.35 | 1.00 | 0.52 |
| | 1 | 0.70 | 1.00 | 0.83 | | 1 | 0.00 | 0.00 | 0.00 |
| | 2 | 0.75 | 1.00 | 0.86 | | 2 | 0.00 | 0.00 | 0.00 |
| | Micro. avg | 0.82 | 0.76 | 0.71 | | Micro. avg | 0.12 | 0.33 | 0.17 |
| | WAVG. | 0.82 | 0.75 | 0.70 | | WAVG. | 0.13 | 0.36 | 0.18 |
| | Accuracy | 0.75 | | | | Accuracy | 0.35 | | |
| LR | 0 | 0.86 | 0.72 | 0.78 | HV | 0 | 0.84 | 0.84 | 0.84 |
| | 1 | 0.89 | 0.92 | 0.91 | | 1 | 0.96 | 0.88 | 0.92 |
| | 2 | 0.83 | 0.95 | 0.89 | | 2 | 0.87 | 0.95 | 0.91 |
| | Micro. avg | 0.86 | 0.87 | 0.86 | | Micro. avg | 0.89 | 0.89 | 0.89 |
| | WAVG. | 0.86 | 0.86 | 0.86 | | WAVG. | 0.89 | 0.89 | 0.89 |
| | Accuracy | 0.86 | | | | Accuracy | 0.89 | | |
| SVM | 0 | 0.83 | 0.96 | 0.89 | SV | 0 | 0.84 | 0.84 | 0.84 |
| | 1 | 1.00 | 0.85 | 0.92 | | 1 | 0.96 | 0.88 | 0.92 |
| | 2 | 0.95 | 0.95 | 0.95 | | 2 | 0.87 | 0.95 | 0.91 |
| | Micro. avg | 0.93 | 0.92 | 0.92 | | Micro. avg | 0.89 | 0.89 | 0.89 |
| | WAVG. | 0.93 | 0.92 | 0.92 | | WAVG. | 0.89 | 0.89 | 0.89 |
| | Accuracy | 0.92 | | | | Accuracy | 0.89 | | |

**TABLE 8.** Models Training Time (TT) and Prediction Time (PT).

| Model | Orignal | | RFE&CTGAN | | RFE&MDST | |
|---|---|---|---|---|---|---|
| | PT | TT | PT | TT | PT | TT |
| RF | 0.35 | 20.94 | 0.62 | 21.97 | 0.44 | 21.93 |
| DT | 0.006 | 1.90 | 0.01 | 1.99 | 0.07 | 1.99 |
| KNN | 0.003 | 166.15 | 0.002 | 6.97 | 0.001 | 158.5 |
| LR | 0.01 | 1.99 | 0.014 | 1.99 | 0.009 | 0.99 |
| NB | 0.006 | 2.62 | 0.002 | 0.99 | 0.001 | 0.99 |
| SVM | 0.07 | 1.99 | 0.016 | 4.02 | 0.001 | 2.99 |
| RNN | 5.7 | 182.9 | 10.88 | 187.70 | 5.64 | 0.15 |
| CNN | 1.45 | 180.2 | 1.14 | 71.57 | 1.22 | 0.08 |
| HV | 0.11 | 11.97 | 0.078 | 15.95 | 0.087 | 0.88 |
| SV | 0.38 | 395.3 | 0.36 | 14.95 | 0.376 | 0.88 |

**TABLE 9.** K-fold cross-validation results.

| Method | RF | DT | KNN | LR |
|---|---|---|---|---|
| RFE and CTGAN | 0.74(± 0.06) | 0.70 (± 0.07) | 0.71 (± 0.06) | 0.72 (± 0.07) |
| RFE and MDST | 0.89(± 0.04) | 0.82 (± 0.04) | 0.73 (± 0.04) | 0.86 (± 0.03) |
| **Method** | **SVM** | **NB** | **RNN** | **CNN** |
| RFE and CTGAN | 0.73 (± 0.07) | 0.57 (± 0.06) | 0.35 (± 0.02) | 0.34 (± 0.03) |
| RFE and MDST | 0.91 (± 0.03) | 0.75 (± 0.05) | 0.35(± 0.01) | 0.35 (± 0.04) |

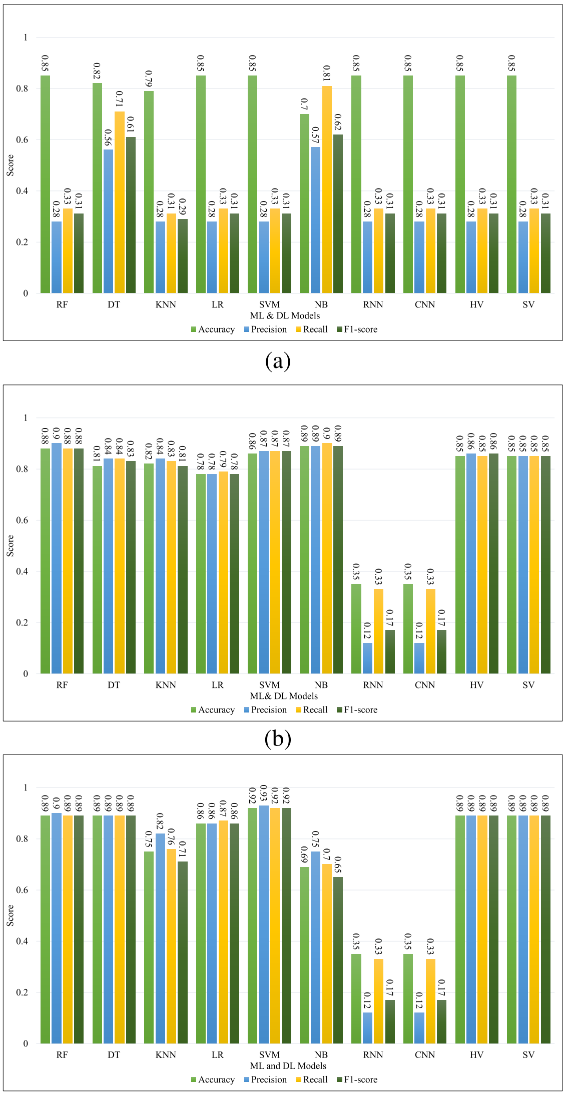
**TABLE 10.** Comparison with recent methods on autonomous vehicle accident severity prediction.

| Re.f | Year | Model | Accuracy | Precesion | Recall | F1 |
|---|---|---|---|---|---|---|
| [12] | 2021 | RF | 0.81 | 0.66 | 0.82 | 0.73 |
| [13] | 2022 | SVM | 0.82 | 0.66 | 0.82 | 0.73 |
| [29] | 2022 | MLP | 0.82 | 0.82 | 0.82 | 0.89 |
| [16] | 2022 | ANN | 0.55 | 0.74 | 0.55 | 0.61 |
| [18] | 2022 | RF,SVM | 0.82 | 0.66 | 0.82 | 0.73 |
| **Our** | 2023 | SVM+MDST | 0.92 | 0.93 | 0.92 | 0.92 |

the approaches of the previous studies in the same experimental environment used for our approach. We selected recent studies for comparison, such as study [11], which utilized RF for severity prediction, achieving significant results in their specific scenario. Similarly, study [12] deployed SVM, while studies [15] and [28] utilized deep learning algorithms like MLP and ANN, respectively. Table 10 presents a comprehensive comparison of these recent studies, showing that our proposed approach outperforms all other models across all evaluation metrics. This superiority stems from the significance of our proposed MDST approach, which facilitates the generation of linearly separable data. This reduction in model overfitting contributes to boosting overall performance across various evaluation metrics.

### G. DISCUSSION

This study conducts several experiments for predicting accident severity in AVs using machine learning approaches. We deploy data balancing and feature selection methods to achieve significant results. Figure 8 illustrates the impact of the MDST approach for data balancing compared to the original dataset. As shown in Figure 8a, samples for classes 1 and 2 are scattered and limited in number, while samples for



(a)

(b)

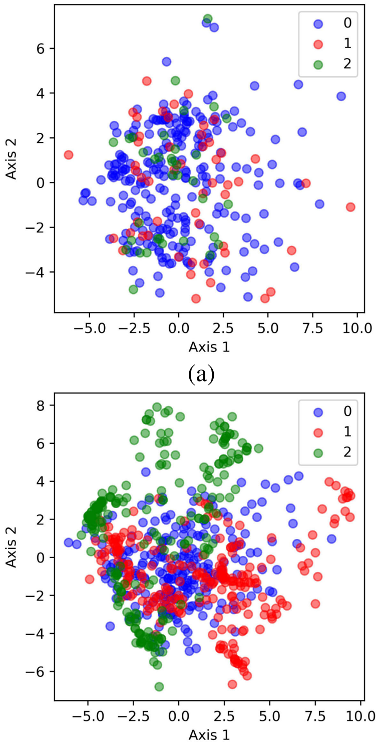
(c)

**FIGURE 7.** Performance comparison using (a) original dataset, (b) RFE and CTGAN, and (c) RFE and MDST.

class 0 overlap with other samples, making it challenging for models to optimize their weights for each class. Conversely, Figure 8b demonstrates a balanced distribution for each class, with linear separability, resulting in an improvement in model performance.

Figure 9 shows the Receiver Operating Characteristic (ROC) curve for the proposed approach, SVM with RFE and MDST. It demonstrates the significance of the proposed approach and the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different threshold values. Class 1 and 2 exhibit high true positive rates at minimal false positive rates, while Class 0 performs slightly behind but remains noteworthy.

Our analysis reveals the importance of numerous features from the dataset in predicting accident severity. We have highlighted the significance of these features in Figure 10. For instance, AVdamagepart in Figure 10a, indicating the damaged part of the autonomous vehicle, proves to be informative in predicting the severity of accidents. Similarly, P1injured in Figure 10c, representing injuries post-accident, emerges as a vital factor, and the lighting condition (Day-Light) in Figure 10d also serves as a feature contributing to severity prediction. In summary, beyond speed and collision types, various other post-accident measured features play a significant role in predicting accident severity.



(a)

(b)

**FIGURE 8.** Class distribution using (a) Orgnail dataset (b) MDST.

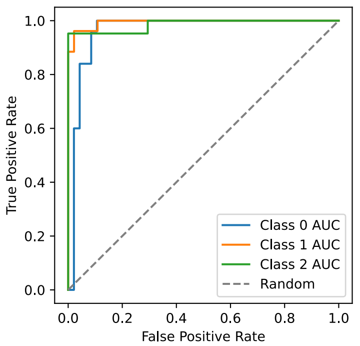
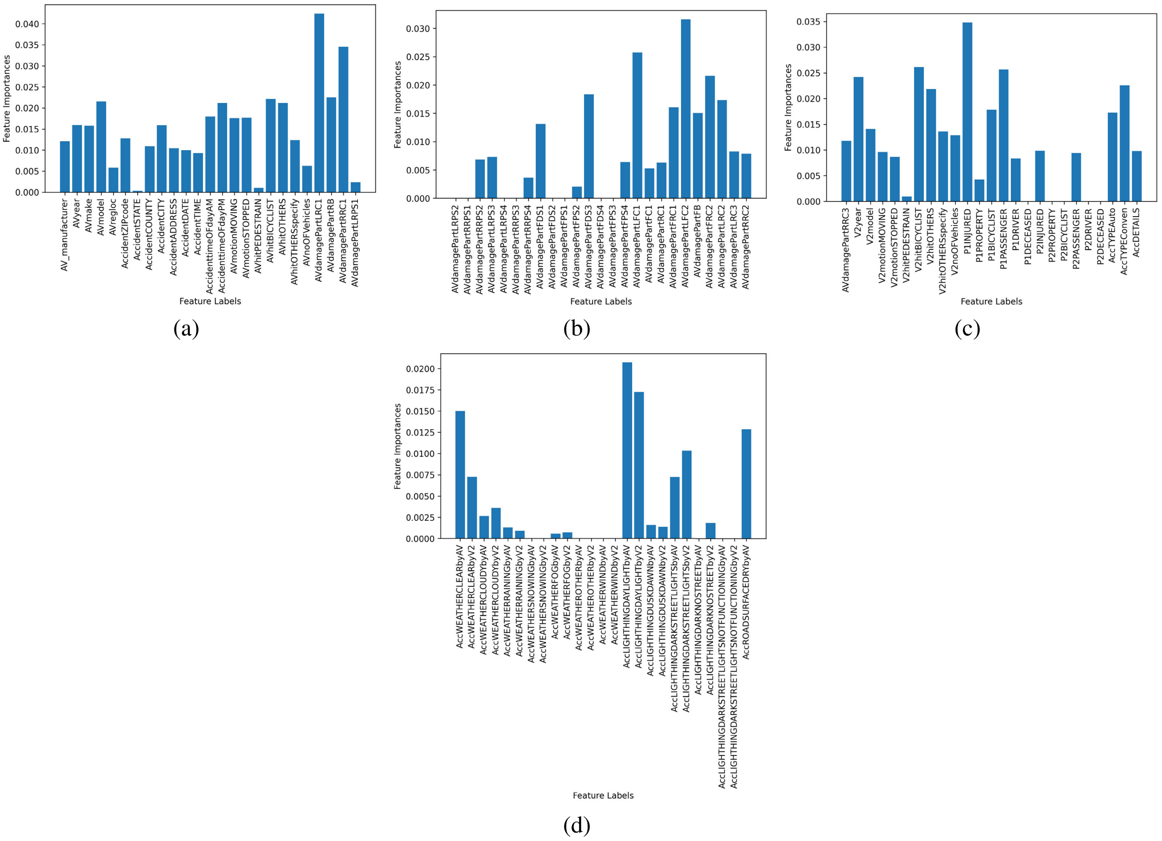


**FIGURE 9.** ROC curve for proposed approach.

Overall, the original dataset was imbalanced and contained numerous meaningless features, leading to a reduction in model performance and overfitting towards the majority classes. In our approach, RFE assists in selecting important features for training the learning models, while CTGAN and MDST contribute to reducing overfitting towards the majority classes. Combining both feature selection and data balancing approaches helps achieve significant results.



(a) (b) (c) (d)

**FIGURE 10.** Feature importance score for features from (a) 0-25(b) 26-50 (c) 51-75 (d) 76-100.

## V. CONCLUSION

In conclusion, our study delves into the critical realm of AVs accidents, recognizing the imperative need to comprehend and address their severity for the advancement and widespread acceptance of this transformative technology. Through the innovative amalgamation of natural language processing techniques and machine learning, we endeavored to predict the severity of AV-related accidents. A pivotal contribution of our study is the development of a novel dataset sourced from post-disengagement accident reports spanning the years 2019-2022. To counter the challenge of limited data, our introduction of the Multi-Distance Synthetic Technique (MDST) aimed to balance the inherent dataset imbalance, a crucial step in fortifying the reliability of our analysis.

Through our analysis, we have determined that various parameters extracted from autonomous vehicle accident reports play a crucial role in predicting accident severity. Factors such as sensor failures or the nature of the vehicle's impact contribute significantly to this prediction. One of the primary conclusions drawn from our study is the inadequacy of the available volume of AV accident data for effective machine learning model training. To overcome this challenge, we introduced the novel MDST. This method generates additional correlated data points to augment the original dataset. The significance of MDST lies in its utilization of diverse distance matrices—employing two instead of relying on a single distance matrix—thereby surpassing the capabilities of conventional approaches.

Additionally, our analysis revealed that many features within the original dataset lacked the predictive power needed for accurate severity estimation. In addressing this issue, Recursive Feature Elimination (RFE) proved invaluable. By selecting only the most meaningful features, RFE significantly improved the system's efficiency. This meticulous feature selection process contributed to achieving an impressive accuracy score of 0.92 in predicting autonomous vehicle accident severity,

**Limitation and Future Work:** This study analyzes collision reports from the California DMV in PDF format, extracting recurring patterns for insights into safety-related decision-making. Leveraging the dataset, future research can explore classification tasks like assessing AV damage, identifying collision types, and evaluating weather, road surface, and motion states. Additionally, future work will explore mapping keywords from crash reports to collection reports for precise sensor-related information. To enhance generalizability, we are actively collecting a new dataset to rigorously test and evaluate the proposed approach in diverse scenarios.

## AUTHOR CONTRIBUTIONS

Rahman Shafique: Conceptualization, methodology, software, investigation, visualization, validation, writing—

original draft, writing—review, and editing. Furqan Rustum: Data curation, methodology, software, investigation, writing—review. Sheriff Murtala: Formal analysis, investigation, validation. Anca Delia Jurcut: validation, investigation, writing—review, and editing. Gyu Sang Choi: Project administration, Software, visualization, validation, investigation.

## DATA AVAILABILITY

The dataset is available on the given link:https://github.com/Arehmans/avaccidents

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

[1] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of autonomous vehicles," *J. Adv. Transp.*, vol. 2020, pp. 1–13, Oct. 2020.

[2] D. Petrović, R. Mijailović, and D. Pešić, "Traffic accidents with autonomous vehicles: Type of collisions, manoeuvres and errors of conventional vehicles' drivers," *Transp. Res. Proc.*, vol. 45, pp. 161–168, Jan. 2020.

[3] *Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World*, World Health Org. (WHO), Geneva, Switzerland, 2018.

[4] R. Philipsen, T. Brell, and M. Ziefle, "Carriage without a driver–user requirements for intelligent autonomous mobility services," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.*, Orlando, FL, USA. Cham, Switzerland: Springer, 2018, pp. 339–350.

[5] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations," *Transp. Res. A, Policy Pract.*, vol. 77, pp. 167–181, Jul. 2015.

[6] S. P. Wood, J. Chang, T. Healy, and J. Wood, "The potential regulatory challenges of increasingly autonomous motor vehicles," *Santa Clara L. Rev.*, vol. 52, p. 1423, Jan. 2012.

[7] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 546–556, Apr. 2015.

[8] A. Theofilatos, D. Graham, and G. Yannis, "Factors affecting accident severity inside and outside urban areas in Greece," *Traffic Injury Prevention*, vol. 13, no. 5, pp. 458–467, Sep. 2012.

[9] B. Kutela, S. Das, and B. Dadashova, "Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors," *Accident Anal. Prevention*, vol. 165, Feb. 2022, Art. no. 106473.

[10] I. C. Obasi and C. Benson, "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents," *Heliyon*, vol. 9, no. 8, Aug. 2023, Art. no. e18812.

[11] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A comparative study of machine learning algorithms to predict road accident severity," in *Proc. 20th Int. Conf. Ubiquitous Comput. Commun. (IUCC/CIT/DSCI/SmartCNS)*, Dec. 2021, pp. 390–397.

[12] A. K. Paul, P. K. Boni, and Md. Z. Islam, "A data-driven study to investigate the causes of severity of road accidents," in *Proc. 13th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Oct. 2022, pp. 1–7.

[13] R. Vijithasena and W. Herath, "Data visualization and machine learning approach for analyzing severity of road accidents," in *Proc. Int. Conf. Advancement Technol. (ICONAT)*, Jan. 2022, pp. 1–6.

[14] T. K. Bahiru, V. S. Manjula, T. B. Akele, E. A. Tesfaw, and T. D. Belay, "Mining road traffic accident data for prediction of accident severity," in *Proc. Int. Conf. Intell. Data Commun. Technol. Internet Things (IDCIoT)*, Jan. 2023, pp. 606–612.

[15] M. Iveta, A. Radovan, and B. Mihaljević, "Prediction of traffic accidents severity based on machine learning and multiclass classification model," in *Proc. 44th Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, 2021, pp. 1701–1705.

[16] I. E. Mallahi, A. Dlia, J. Riffi, M. A. Mahraz, and H. Tairi, "Prediction of traffic accidents using random forest model," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, May 2022, pp. 1–7.

[17] Y. Dia, L. Faty, M. D. Sarr, O. Sall, M. Bousso, and T. T. Landu, "Study of supervised learning algorithms for the prediction of road accident severity in Senegal," in *Proc. 7th Int. Conf. Comput. Intell. Appl. (ICCIA)*, Jun. 2022, pp. 123–127.

[18] S. Ahmed, M. A. Hossain, S. K. Ray, M. M. I. Bhuiyan, and S. R. Sabuj, "A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance," *Transp. Res. Interdiscipl. Perspect.*, vol. 19, May 2023, Art. no. 100814.

[19] Z. Farhart, A. Karouni, B. Daya, P. Chauvet, and N. Hmadeh, "Traffic accidents severity prediction using support vector machine models," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 7, pp. 1345–1350, May 2020.

[20] K. Sattar, F. C. Oughali, K. Assi, N. Ratrout, A. Jamal, and S. M. Rahman, "Transparent deep machine learning framework for predicting traffic crash severity," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1535–1547, Jan. 2023.

[21] N. Fiorentini and M. Losa, "Handling imbalanced data in road crash severity prediction by machine learning algorithms," *Infrastructures*, vol. 5, no. 7, p. 61, Jul. 2020.

[22] B. K. Mohanta, D. Jena, N. Mohapatra, S. Ramasubbareddy, and B. S. Rawal, "Machine learning based accident prediction in secure IoT enable transportation system," *J. Intell. Fuzzy Syst.*, vol. 42, no. 2, pp. 713–725, Jan. 2022.

[23] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," *PLoS ONE*, vol. 14, no. 4, Apr. 2019, Art. no. e0214966.

[24] A. Priyanka and K. Sathiyakumari, "A comparative study of classification algorithm using accident data," *Int. J. Comput. Sci. Eng. Technol.*, vol. 5, no. 10, pp. 1018–1023, 2014.

[25] K. Assi, S. M. Rahman, U. Mansoor, and N. Ratrout, "Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol," *Int. J. Environ. Res. Public Health*, vol. 17, no. 15, p. 5497, Jul. 2020.

[26] *Sonnguyen129*. [Online]. Available: https://github.com/sonnguyen129/Accident-Severity-Prediction

[27] *Road Safety Transport Data*. [Online]. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[28] J. A. Sowdagur, B. Tawheeda. B. Rozbully-Sowdagur, and G. Suddul, "An artificial neural network approach for road accident severity prediction," in *Proc. IEEE Zooming Innov. Consum. Technol. Conf. (ZINC)*, May 2022, pp. 267–270.

[29] *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[30] *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[31] *Road Safety Data*. [Online]. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[32] S. M. S. Hamdan, S. Barakat, K. H. Mahfouz, and K. A. Ghuzlan, "Traffic accident severity prediction model using AI," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, Feb. 2023, pp. 1–5.

[33] M. Yan and Y. Shen, "Traffic accident severity prediction based on random forest," *Sustainability*, vol. 14, no. 3, p. 1729, Feb. 2022.

[34] C. Li, X. Wu, Z. Zhang, Z. Ma, Y. Zhu, and Y. Chen, "Freeway traffic accident severity prediction based on multi-dimensional and multi-layer Bayesian network," in *Proc. IEEE 2nd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2022, pp. 1032–1035.

[35] M. Bédard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accident Anal. Prevention*, vol. 34, no. 6, pp. 717–727, Nov. 2002.

[36] J. Langley and S. W. Marshall, "The severity of road traffic crashes resulting in hospitalisation in new Zealand," *Accident Anal. Prevention*, vol. 26, no. 4, pp. 549–554, Aug. 1994.

[37] C. Conroy, G. T. Tominaga, S. Erwin, S. Pacyna, T. Velky, F. Kennedy, M. Sise, and R. Coimbra, "The influence of vehicle damage on injury severity of drivers in head-on motor vehicle crashes," *Accident Anal. Prevention*, vol. 40, no. 4, pp. 1589–1594, Jul. 2008.

[38] J. Lee, C. Conroy, R. Coimbra, G. T. Tominaga, and D. B. Hoyt, "Injury patterns in frontal crashes: The association between knee–thigh–hip (KTH) and serious intra-abdominal injury," *Accident Anal. Prevention*, vol. 42, no. 1, pp. 50–55, Jan. 2010.

[39] C. DMV. (2018). *Report(Traffic Collision) Involving an AV*. [Online]. Available: https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/autonomousveh_ol316

[40] B. Corben, T. Senserrick, M. Cameron, and G. Rechnitzer, "Development of the visionary research model: Application to the car/pedestrian conflict," Tech. Rep., 2004.

[41] R. Elvik, "To what extent can theory account for the findings of road safety evaluation studies?" *Accident Anal. Prevention*, vol. 36, no. 5, pp. 841–849, Sep. 2004.

[42] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets classification on the base of sentiments for U.S. airline companies," *Entropy*, vol. 21, no. 11, p. 1078, Nov. 2019.

[43] R. Khan, F. Rustam, K. Kanwal, A. Mehmood, and G. S. Choi, "U.S. based COVID-19 tweets sentiment analysis using TextBlob and supervised machine learning algorithms," in *Proc. Int. Conf. Artif. Intell. (ICAI)*, Apr. 2021, pp. 1–8.

[44] R. Shafique, A. Mehmood, and G. S. Choi, "Cardiovascular disease prediction system using extra trees classifier," Tech. Rep., 2019.

[45] R. Shafique, F. Rustam, G. S. Choi, I. D. L. T. Díez, A. Mahmood, V. Lipari, C. L. R. Velasco, and I. Ashraf, "Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning," *Cancers*, vol. 15, no. 3, p. 681, Jan. 2023.

[46] P.-E. Danielsson, "Euclidean distance mapping," *Comput. Graph. Image Process.*, vol. 14, no. 3, pp. 227–248, Nov. 1980.

[47] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean distance and Manhattan distance in the K-means algorithm for variations number of centroid K," *J. Phys., Conf. Ser.*, vol. 1566, no. 1, Jun. 2020, Art. no. 012058.

**RAHMAN SHAFIQUE** received the M.S.C.S. degree from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. He is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si, South Korea. He was a Research Assistant with the Fareed Computing and Research Center, KFUEIT. His research interests include data mining, machine learning, and artificial intelligence.

**FURQAN RUSTAM** received the M.C.S. degree from the Department of Computer Science, The Islamia University of Bahawalpur, Pakistan, in October 2017, and the Master of Computer Science degree from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. He is currently pursuing the Ph.D. degree in computer science with University College Dublin, Ireland. He was a Research Assistant with the Fareed Computing and Research Center, KFUEIT. His research interests include data mining, machine learning, and artificial intelligence, mainly involved on creative computing and supervised machine learning.

**SHERIFF MURTALA** received the bachelor's degree in electrical engineering from the University of Ilorin, Ilorin, Nigeria, in 2010, the master's degree in communication engineering from the Federal University of Technology, Minna, Nigeria, in 2017, and the Ph.D. degree in information and communication engineering from the Korea University of Technology and Education (KOREATECH), Republic of Korea, in 2021. Since 2021, he has been a Postdoctoral Researcher with the Department of Information and Communication Engineering, Yeungnam University, Republic of Korea. His current research interests include prognostics and health management for automotive sensors, wireless communications, and artificial intelligence and machine learning methods for autonomous vehicles.

**ANCA DELIA JURCUT** (Member, IEEE) received the B.Sc. degree in computer science and mathematics from the West University of Timişoara, Romania, in 2007, and the Ph.D. degree in security engineering from the University of Limerick (UL), Ireland, in 2013, funded by the Irish Research Council for Science Engineering and Technology. She was a Postdoctoral Researcher with UL, a member with the Data Communication Security Laboratory, and a Software Engineer with IBM, Dublin, Ireland, in the area of data security and formal verification. She has been an Assistant Professor with the School of Computer Science, University College Dublin (UCD), Ireland, since 2015. Her research interests include security protocols design and analysis, automated techniques for formal verification, network security, attack detection and prevention techniques, security for the Internet of Things, and the applications of blockchain for security and privacy. She has several key contributions to research focusing on the detection and prevention techniques of attacks over networks, the design and analysis of security protocols, automated techniques for formal verification, and security for mobile edge computing (MEC). For more information, visit the link https://people.ucd.ie/anca.jurcut.

**GYU SANG CHOI** (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA, in 2005. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been a Faculty Member with the School of Software Convergence, Yeungnam University, South Korea. His research interests include data mining, natural language processing, and reinforcement learning.