Optimized Feature Selection Approaches for Accident Classification to Enhance Road Safety: Case Study Ernakulam

Sobhana Mummaneni[1], Gnana Siva Sai Venkatesh Mendu[1], Nihitha Vemulapalli[1], Kushal Kumar Chintakayala[1]
[1] Department of Computer Science and Engineering, V R Siddhartha Engineering College, Vijayawada, 520007, India

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Coyote Optimization Algorithm
Gated Recurrent Unit
Genetic Algorithm
Long Short-Term Memory
Multilayer Perceptron
Recurrent Neural Network

ABSTRACT

In the modern era, the issue of road accidents has become an increasingly critical global concern, requiring urgent attention and innovative solutions. This investigation has compiled an extensive dataset of 10,356 accident occurrences that occurred between the years 2018 and 2022 in Ernakulam district. By utilizing advanced feature selection methodologies, such as Genetic Algorithm and Coyote Optimization, this research has identified pivotal accident determinants. The study harnesses the potential of deep learning techniques, encompassing Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP) for classifying accidents according to severity (categorized as fatal, grievous, and severe). Eight predictive models are trained using the dataset, and the top two are ensembled. Integrating deep learning and optimization strategies, this research aims to create a robust accident classification system. The system will help in developing proactive policies that can reduce the frequency and severity of accidents in Ernakulam district.

Corresponding Author:

Sobhana Mummaneni
Department of Computer Science and Engineering, V R Siddhartha Engineering College, Vijayawada, 520007, India
Email: sobhana@vrsiddhartha.ac.in

1. INTRODUCTION

India faces a growing road safety crisis, marked by a nine-fold increase in road accident fatalities between 1970 and 2013 [1]. These accidents exhibit age, gender, and time related patterns, with economically active individuals aged 30-59 being the most vulnerable. Males experience higher accident risks. Extreme weather influences accident occurrences during May-June and December-January, while accidents are more frequent between 9 AM and 9 PM. Driver error remains a significant factor in accidents. There's substantial regional variation in accident risks, with certain states facing higher fatality rates. India's fatality rates far exceed those of developed countries, emphasizing the need for comprehensive road safety measures, effective policies, and political commitment.

Road accident classification is a pivotal field of study, focusing on the categorization of accidents according to their severity, root causes, and contributory elements [2]. This analysis holds profound importance for a multitude of purposes, notably in the development of robust road safety strategies, precise insurance assessments, and refined law enforcement approaches. Accidents are stratified into various classes, ranging from minor to fatal, contingent upon the nature of injuries and property damage sustained. The classification process considers factors like weather conditions, driver behavior, road type and vehicle speed, which are vital in determining the severity of road accidents. Statistical models, machine learning methods, and data analysis play a crucial role in revealing patterns and predicting trends in accidents. Precise accident classification empowers authorities to allocate resources efficiently, enact targeted safety measures, evaluate risk, set insurance premiums, and ultimately foster safer road environments, culminating in the preservation of lives.

Choosing the right factors for road accident classification is vital [3]. Picking the most important variables from the many available like driver behavior, road type and weather is crucial for building accurate predictive models. It boosts model performance, making it more efficient and easier to understand. Optimizing feature selection not only enhances classification accuracy but also aids in directing resources and efforts toward improving road safety. It's a significant step in reducing the social and economic impact of accidents.

The Coyote Optimization Algorithm (COA) is a novel metaheuristic approach that draws inspiration from the behavioral patterns exhibited by coyotes (Canis latrans), devised to tackle optimization challenges across scientific, computing, and engineering domains [4]. Employing a population-based approach, COA organizes coyotes into packs, representing potential solutions for optimization problems. By mirroring the objective function's cost in coyotes' social conditions, COA fosters adaptation and interaction among individuals, maintaining diversity and balancing exploration-exploitation trade-offs. Notably, COA introduces mechanisms for birth, death, and cultural exchange led by an alpha coyote within each pack. Numerical evaluations and statistical tests showcase COA's superior performance compared to other nature-inspired metaheuristics, particularly in navigating high-dimensional spaces for efficient feature selection while ensuring adaptability and diversity.

In the domain of feature selection, Genetic Algorithms (GAs) represent a powerful and widely employed optimization technique [5]. These algorithms draw inspiration from the process of natural selection and evolution, mimicking the principles of genetic variation, selection, and recombination to identify the most informative features within a dataset. GAs initiate with a population of potential feature subsets and iteratively evolve them over multiple generations. Through a combination of selection, crossover, and mutation operations, GAs promotes the survival of feature subsets that demonstrate superior performance, ultimately leading to the discovery of an optimal or near-optimal feature combination. Genetic Algorithms offer significant advantages for feature selection, as they are well-suited for high-dimensional datasets and can handle a large search space efficiently. They are versatile and adaptable to various problem domains, making them a valuable tool for optimizing machine learning models, improving data analysis, and enhancing the efficiency of complex problem-solving tasks.

1.1. Literature Review

Mamoudou Sangare et al. [6] utilized a 2017 road traffic accident dataset from data.govt.uk to develop a hybrid forecasting model. It combines Gaussian Mixture Models (GMM) with Support Vector Classifiers (SVC) to predict accident severity. This approach involves data preprocessing, resampling, feature selection, and the radial basis kernel for SVC. The hybrid model significantly improves accuracy compared to the GMM baseline, highlighting its potential for urban traffic accident forecasting.

Peijie Wu et al. [7] employed Crash Prediction Models-Genetic Algorithms (CPM-GAs) to forecast crashes based on road geometry and traffic data, enhancing accuracy by integrating traditional CPM-GAs with machine learning models. The findings highlight the superior accuracy of machine learning models. This study introduces a novel ensemble technique for traffic safety analysis.

D. Devaraj et al. [8] analyzed road accidents in Kerala, aiming to determine accident severity levels based on contributory factors. The analysis encompasses year-wise, day-wise, and district-wise assessments. Various road patterns and accident causes are identified from historical data. A decision tree algorithm is used to classify accidents as severe, medium, or low severity. This research aids in identifying locations in Kerala where the likelihood of serious accidents is elevated.

Md. Farhan Labib et al. [9] addressed the pressing issue of road accidents in Bangladesh, emphasizing the need for precise analysis and accident severity classification. Employing machine learning algorithms like K-Nearest Neighbors, AdaBoost, Naïve Bayes, and Decision Tree, the study classifies accidents into four categories. AdaBoost yields the best performance, with an 80% accuracy. The research also investigates the influence of factors like road class, junction type, surface condition, time, and vehicle type on accident occurrence, offering valuable insights to reduce accidents.

Vahid Najafi Moghaddam Gilani et al. [10] investigated accident severity in Rasht City, utilizing logistic regression and artificial neural networks (ANN). The logistic regression model achieves an 89.17% accuracy rate, identifying variables like time, weather, and vehicle type that impact severity. The ANN model performs even better, with 98.9% prediction accuracy. It highlights the importance of vehicle quality and visibility-enhancing measures to reduce accidents. The study provides valuable insights into improving urban safety and reducing accident rates.

Emmanuel Kofi Adanu et al. [11] conducted an extensive study on severity of in injuries interstate crashes in Alabama, utilizing a random parameters multinomial logit modeling method. The study covers a range of factors, including driver behaviors, roadway features, and crash circumstances. It distinguishes between urban and rural, single-vehicle and multi-vehicle crashes, uncovering significant variations in injury determinants. The findings offer valuable insights for implementing targeted road safety measures and understanding complex relationships influencing crash injury severity.

Md. Kamrul Islam et al. [12] examined the involvement of individuals aged 15 to 44 in serious and severe traffic incidents in Al-Ahsa, Saudi Arabia. To determine the reason behind the particular investigation of this age group, descriptive analyses were performed. Classification and Regression Tree(CART) and Logistic regression models were employed to analyze factors influencing their participation in crashes. Vehicle accidents are the most frequent kind of crashes, and although though they happen less frequently, there is a high severity index. The models confirm that severe crashes with higher injuries and fatalities involve this age group. CART highlights specific scenarios like overturns due to driver distraction and speeding. The CART model, though more accurate, requires longer processing time compared to logistic regression.

Daniel Santos et al. [13] studied the global concern of traffic accidents by analyzing accident data from Setúbal, Portugal (2016-2019) to identify factors influencing accident severity. Utilizing machine learning algorithms such as clustering, logistic regression, decision trees, and random forests, the research successfully identifies key factors for fatal and non-fatal accidents. Predictive models, especially the random forest algorithm, offer valuable insights, although further refinement is needed to enhance accuracy. The findings contribute to the understanding of accident data and have implications for road safety measures.

Aslam Al-Omari et al. [14] fuzzy logic and employed Geographic Information System (GIS) to predict traffic accident hot spots in Irbid City, Jordan, using accident data from 2013 to 2015. Analyzing occurrence time, accident types, injury, and fatality, the research identifies eight hot spots, including road sections and major intersections. Weighted Overlay and Fuzzy Overlay Methods, aided by the Analytic Hierarchy Process (AHP), successfully reveal these high-risk areas. The findings provide insights into accident-contributing factors and recommendations for safety measures.

Tebogo Bokaba et al. [15] examined various machine learning techniques using real road traffic accident (RTA) data from Gauteng, South Africa. It evaluates various classifiers, including logistic regression, naïve Bayes, k-nearest neighbor, random forest, AdaBoost, and support vector machine, using five different missing data techniques. The study focuses on precision, recall, accuracy, receiver operating characteristic curves, and root-mean-square error. The results show that the random forest classifier works best when used in conjunction with chained equations for multiple imputations. This information can be useful for policymakers and transportation authorities.

Anton Sysoev et al. [16] utilized machine learning techniques to cluster drivers into homogeneous groups based on their violations of road traffic rules. These groupings can inform effective marketing campaigns. The study also explores use of personal features and retrospective statistics for prediction of accident type, likely to involve a driver. Machine learning models including decision trees, random forest, catBoost, neural networks, logistic regression, and ridge regression were applied and discussed for this purpose.

Maria Izabel Santos et al. [17] introduced a novel approach to enhance road safety by utilizing a driver simulator model in a systematic experiment. By analyzing factors such as speed, curve radius, and time of day, the study provides valuable insights into road accidents, offering a potential avenue for improving road safety through informed risk factor assessments.

Imran Ashraf et al. [18] investigated road accidents in South Korea. Analyzing rainfall and accident data, the research identifies several factors contributing to accidents, including traffic volume, limited road expansion, an increasing number of passenger cars, safety violations, and driver characteristics. The results emphasize the need for enhanced traffic regulations and urban road safety measures to address the diverse nature of road accidents.

Joao Mesquitela et al. [19] addressed the issue of urban traffic accidents, utilizing smart city data for in-depth analysis. Multiple information sources are combined through the data fusion method, including accident data, weather conditions, and local authorities' reports, using big data analytics. Geo-referenced accident hotspots are identified through Kernel Density and Hot Spot Analysis (Getis-Ord Gi*) in ArcGIS Pro. This approach helps local municipalities understand the factors influencing accident severity and can be applied in cities with similar data resources.

R. Sathiyaraj et al. [20] addressed the escalating traffic challenges in growing metropolitan areas, emphasizing the need for smart traffic management. The proposed Smart Traffic Prediction and Congestion Avoidance System (s-TPCA) employs Poisson distribution for vehicle arrival prediction. It combines traffic recognition, forecasting, and strategies to prevent congestion, resulting in superior performance, including 20% higher fuel conservation compared to existing systems.

Syed Arshad Raza et al. [21] discussed the rise in global traffic and its impact on accidents, prompting a UN initiative. Saudi Arabia launched a road safety program supported by detailed Eastern Province (EP) data (2010–2020). A meticulous approach is proposed to gather and validate the EP-Traffic-Mortality-and-Policy-Interventions Dataset, involving stakeholder identification, data sources, and cleansing processes. This dataset aids in analyzing accident patterns and policy effectiveness, crucial for researchers and policymakers.

Samuel Olugbade et al. [22] emphasized the persistent issue of road accidents through deploying AI and machine learning for automatic incident detection systems. It reviews the application of these technologies in road transport, emphasizing route optimization, traffic management, and safety measures. The study highlights emerging trends and challenges while serving as a valuable reference for road transport system planning and management.

Romi Satria et al. [23] focused on improving traffic safety through the creation of precise models for crash frequency and the identification of factors that contribute to accidents. It introduces the INLA-CAR model as an alternative to traditional Bayesian models for assessing crash severity on a highway section. The study emphasizes the significance of spatial correlation, with AADT being the most influential factor across all severity types. It also utilizes multiple performance metrics to validate results and provides insights into improving traffic safety management.

2. METHOD

The proposed methodology involves several key modules in the context of Ernakulam. Firstly, the Data Collection module acquires relevant accident data from 2018 to 2022. Next, the Feature Selection module applies advanced techniques like Genetic Algorithm and Coyote Optimization. Then, the Classification module employs deep learning algorithms, including RNN, GRU, LSTM and MLP, to categorize severity of road accidents. Eight predictive models are trained and then best two models are ensembled. The ultimate objective is to develop a robust accident classification system for the implementation of proactive safety policies in Ernakulam district. Figure 1 shows the proposed methodology diagram.

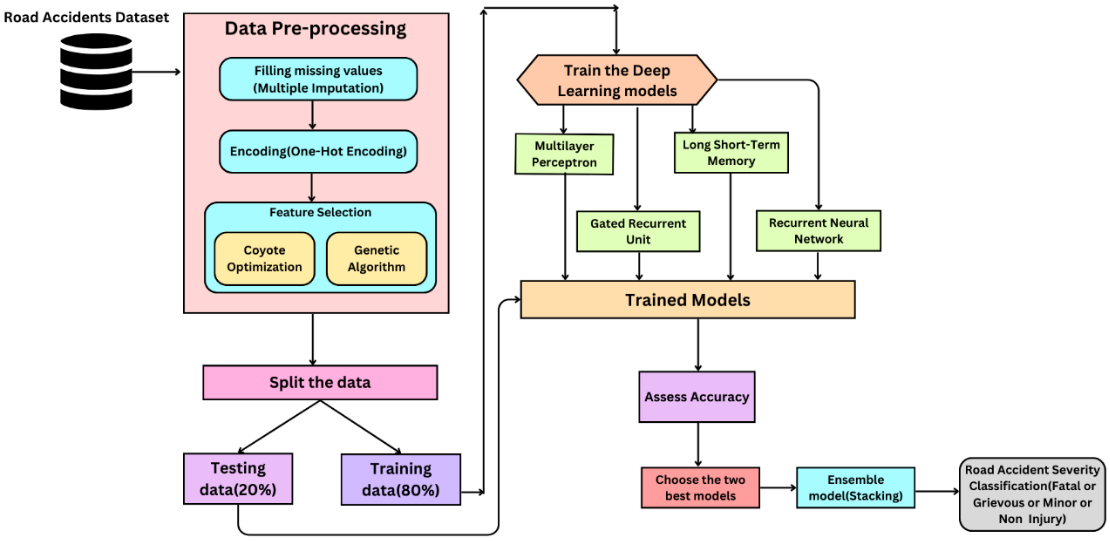


Figure 1. Methodology Diagram

2.1. Handling Missing Values

Multiple Imputation tackles missing data by creating diverse datasets with unique imputations. Each dataset is independently analyzed, and the outcomes are merged, producing a more robust result. This approach considers the variability associated with missing data, contributing to a nuanced understanding. Attention to potential biases is crucial, especially when missingness pertains to categorical values. Statistical procedures are then applied to aggregate results, enhancing reliability in complex datasets for subsequent analyses and modeling tasks.

2.2. Encoding Categorical Data

One-hot encoding transforms categorical data into a numerical form that's compatible with machine learning algorithms. Initially, categorical variables are identified, and unique categories within each variable are determined. For each category, a binary column is created, where '1' signifies the presence of the category, and '0' indicates absence. These binary columns are then appended to the original dataset, effectively expanding the feature space. The original categorical columns are subsequently dropped.

2.3. Genetic Algorithm based Selection of Features

The application of the Genetic Algorithm (GA) for feature selection is a systematic method that aims to enhance model performance by identifying crucial features within a dataset. Beginning with the definition of input parameters, including chromosome length (L), number of features to be selected (N), mutation rate, and maximum generations (Gmax), the algorithm initializes a population of potential feature subsets. Through iterative processes, it evaluates and selects chromosomes based on their positive contributions to model fitness. Crossover operations blend genetic information of selected parents, generating offspring, while mutation introduces small variations. Least fit individuals are replaced with offspring in subsequent generations. The algorithm concludes by extracting the chromosome with the highest fitness, representing an optimal feature subset for enhanced model accuracy and efficiency. Algorithm 1 provides a detailed description of the Genetic Algorithm-driven process for selecting features.

Algorithm 1 Feature Selection using Genetic Algorithm

```
Input: Chromosome length L, Number of features to be selected N, Mutation rate, Maximum generations $G_{max}$
Output: Optimal set of features to be selected
Initialization
while Not converged and $G < G_{max}$ do
    Evaluate fitness of each chromosome
    Select parents based on fitness
    Perform crossover to generate offspring
    Apply mutation to offspring
    Replace least fit individuals with offspring
    $G \leftarrow G + 1$
    end while
    Extract chromosome with the highest fitness
```

2.4. Coyote Optimization based Selection of Features

A metaheuristic technique called Coyote Optimization is used in selecting features to improve the process of finding the best subset of features for classification problems. The algorithm employs a population-based approach, initializing parameters such as population size, feature space dimension, and mutation rates. Through iterative cycles of exploration and exploitation, the algorithm evaluates candidate feature subsets. During the exploration phase, features undergo mutation, while the exploitation phase involves local search operations. Communication among 'coyotes' within the population facilitates information sharing on promising features. The algorithm incorporates termination conditions, such as reaching a specified maximum number of generations. The final result extraction selects the feature subset with the highest fitness, effectively optimizing feature selection for improved classification. Algorithm 2 outlines the process of feature selection using Coyote Optimization, showcasing its power in streamlining data analysis.

Algorithm 2 Feature Selection using Coyote Optimization

```
Input: Population size N, Feature space dimension D, Maximum generations $G_{max}$ ,Mutation rate MR, Exploration rate ER, Exploitation rate IR
Output: Optimal feature subset $C_{optimal}$
Initialization
$G \leftarrow 0$
while $G < G_{max}$ do
    Exploration Phase
```

```
$E \leftarrow \{C_i\}$ (size determined by ER)
for each $C_i$ in E do
    $C_i \leftarrow Mutate(C_i, MR)$
    Evaluate the fitness of $C_i$
end for
Exploitation Phase
$X \leftarrow \{C_i\}$ (size determined by IR)
for each $C_i$ in X do
    $C_i \leftarrow LocalSearch(C_i)$
    Evaluate the fitness of $C_i$
end for
Update the Population
for each $C_i$ in E and X do
    $C_i \leftarrow SelectBetter(C_i, C_i')$ based on fitness
end for
Share Information
Enable communication among coyotes to share information on promising feature subsets
Termination Check
Check termination conditions, such as reaching $G_{max}$
$G \leftarrow G + 1$
end while
Result Extraction
Select the feature subset $C_{optimal}$ with the highest fitness as the optimal feature selection
```

2.5. Ensembling

The process begins with the input of a preprocessed road accident dataset, and the ultimate goal is to predict the severity of road accidents. First, the dataset is divided into training data and testing data using an 80:20 ratio. Feature selection is then carried out through a two-step process involving Genetic Algorithm and Coyote Optimization. The relevant features are initially identified using GA, and the selection is further refined using COA to obtain the optimal feature subset. Subsequently, deep learning algorithms like GRU, RNN, LSTM, MLP are trained on the dataset using the identified optimal feature subset. Testing data is used to evaluate the accuracy of every model. To enhance performance, the top two models based on accuracy are selected, and an ensemble model is created by stacking these chosen models. Finally, predictions on accident severity are made using the ensemble model.

2.6. Performance Analysis

The study evaluates the ensemble model's performance by employing classification results like F1-score, precision, accuracy, recall. This analysis offers a comprehensive understanding of the ensemble method's effectiveness in accurately classifying severity of road accidents. Algorithm 3 details the evaluation process, examining the effectiveness of the deep learning ensemble model in classifying severity of road accidents.

$$F1 - score \leftarrow 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

$$Precision \leftarrow \frac{True\ Positives}{Total\ Samples} \quad (2)$$

$$Accuracy \leftarrow \frac{True\ Positives + True\ Negatives}{Total\ Samples} \quad (3)$$

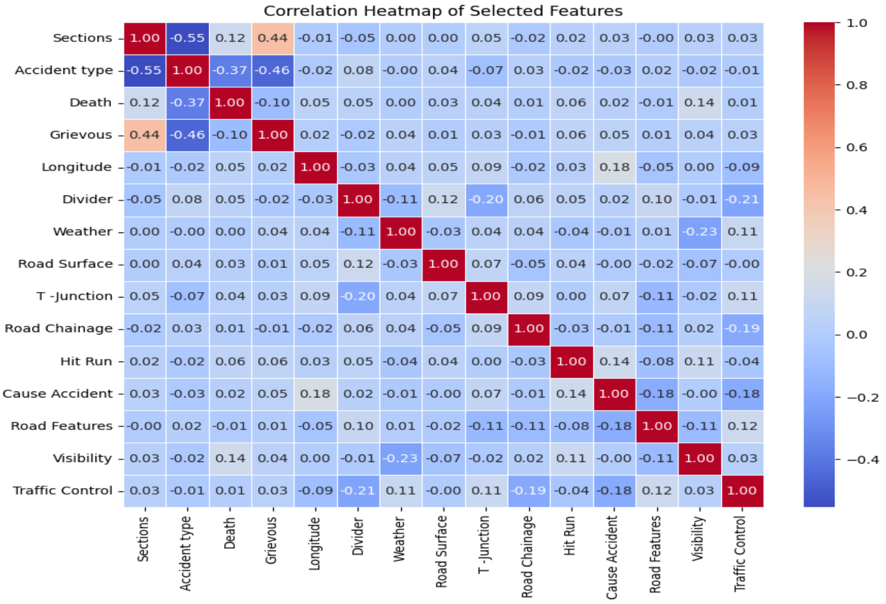
$$Recall \leftarrow \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (4)$$

3. RESULTS AND DISCUSSION

In this section, the feature selection process is conducted by GA and COA on the original 46-feature dataset. The Ensemble model's evaluation includes visualizations such as Model Loss and Model Accuracy Curves, a Confusion Matrix, and a Classification Report. Additionally, the section presents comparisons with other models and highlights a User Interface that showcases predictions and accident hotspots. This interface allows users to input parameters for predicting accident severity, augmenting its real-world applicability.

3.1. Genetic Algorithm Selected Features

The dataset initially comprised 46 features, namely: Zone, Range, District, Subdivision, Circle, PS Name, Firno, Date Report, Date Accident, Time Report, Time Accident, Sections, Accident type, Death, Grievous, Minor, Driver, Passenger, Pedestrian, Cyclist, Other Persons, Motorised, Non Motorised, Latitude, Longitude, Place of Occurrence, Type Area, On going road works, City/Town/ Village, Lanes Road, Divider, Spot Accident, Speed Limit, Weather, Road No, Road Surface, T-Junction, Road Chainage, Hit Run, Collision, Type Road, Cause Accident, Road Features, Visibility, Traffic Control, Vehicle Type. Following a feature selection process utilizing Genetic Algorithm, 15 attributes were identified for further analysis. These selected attributes are: Sections, Accident type, Death, Grievous, Longitude, Divider, Weather, Road Surface, T-Junction, Road Chainage, Hit Run, Cause Accident, Road Features, Visibility, and Traffic Control. Figure 2 shows the Correlation HeatMap of Selected Features using GA.



	Sections	Accident type	Death	Grievous	Longitude	Divider	Weather	Road Surface	T -Junction	Road Chainage	Hit Run	Cause Accident	Road Features	Visibility	Traffic Control
Sections	1.00	-0.55	0.12	0.44	-0.01	-0.05	0.00	0.00	0.05	-0.02	0.02	0.03	-0.00	0.03	0.03
Accident type	-0.55	1.00	-0.37	-0.46	-0.02	0.08	-0.00	0.04	-0.07	0.03	-0.02	-0.03	0.02	-0.02	-0.01
Death	0.12	-0.37	1.00	-0.10	0.05	0.05	0.00	0.03	0.04	0.01	0.06	0.02	-0.01	0.14	0.01
Grievous	0.44	-0.46	-0.10	1.00	0.02	-0.02	0.04	0.01	0.03	-0.01	0.06	0.05	0.01	0.04	0.03
Longitude	-0.01	-0.02	0.05	0.02	1.00	-0.03	0.04	0.05	0.09	-0.02	0.03	0.18	-0.05	0.00	-0.09
Divider	-0.05	0.08	0.05	-0.02	-0.03	1.00	-0.11	0.12	-0.20	0.06	0.05	0.02	0.10	-0.01	-0.21
Weather	0.00	-0.00	0.00	0.04	0.04	-0.11	1.00	-0.03	0.04	0.04	-0.04	-0.01	0.01	-0.23	0.11
Road Surface	0.00	0.04	0.03	0.01	0.05	0.12	-0.03	1.00	0.07	-0.05	0.04	-0.00	-0.02	-0.07	-0.00
T -Junction	0.05	-0.07	0.04	0.03	0.09	-0.20	0.04	0.07	1.00	0.09	0.00	0.07	-0.11	-0.02	0.11
Road Chainage	-0.02	0.03	0.01	-0.01	-0.02	0.06	0.04	-0.05	0.09	1.00	-0.03	-0.01	-0.11	0.02	-0.19
Hit Run	0.02	-0.02	0.06	0.06	0.03	0.05	-0.04	0.04	0.00	-0.03	1.00	0.14	-0.08	0.11	-0.04
Cause Accident	0.03	-0.03	0.02	0.05	0.18	0.02	-0.01	-0.00	0.07	-0.01	0.14	1.00	-0.18	-0.00	-0.18
Road Features	-0.00	0.02	-0.01	0.01	-0.05	0.10	0.01	-0.02	-0.11	-0.11	-0.08	-0.18	1.00	-0.11	0.12
Visibility	0.03	-0.02	0.14	0.04	0.00	-0.01	-0.23	-0.07	-0.02	0.02	0.11	-0.00	-0.11	1.00	0.03
Traffic Control	0.03	-0.01	0.01	0.03	-0.09	-0.21	0.11	-0.00	0.11	-0.19	-0.04	-0.18	0.12	0.03	1.00

Figure 2. Correlation Heat Map of Selected Features using GA

Figures 3, 4, 5, 6 represents the Loss and Accuracy Curves of GA + MLP, GA + RNN, GA + LSTM, GA + GRU Hybrid models respectively.

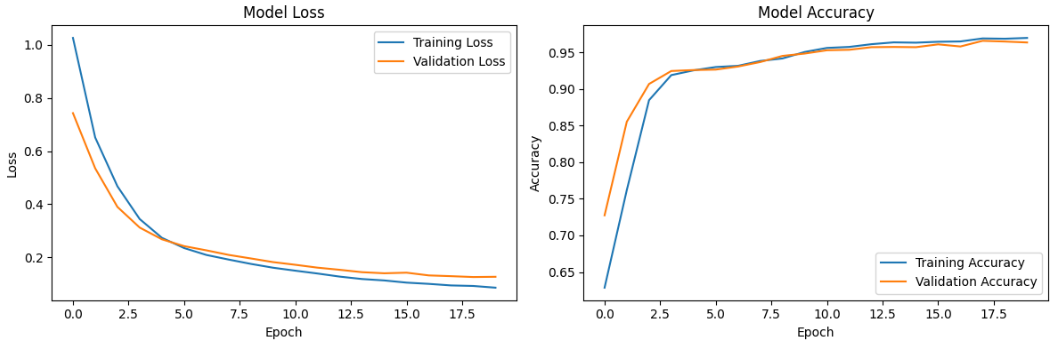


Figure 3. Loss and Accuracy Curves of GA + MLP

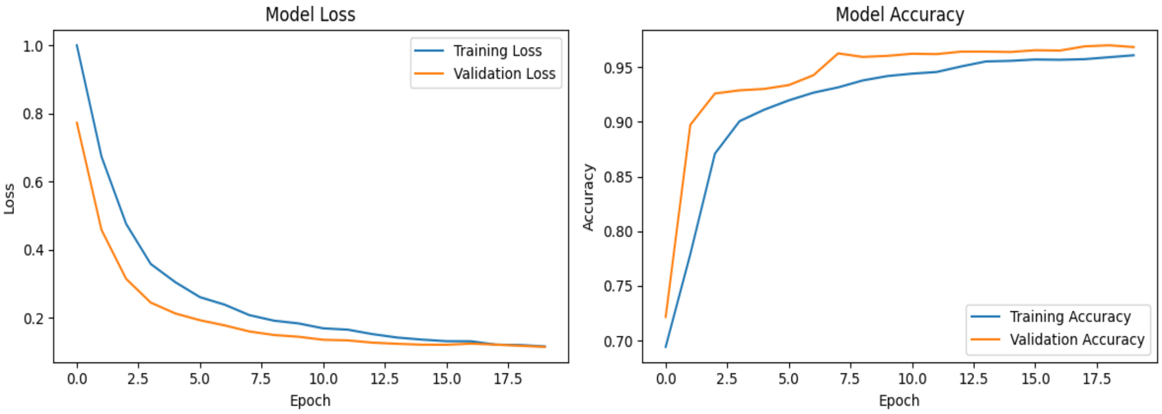


Figure 4. Loss and Accuracy Curves of GA + RNN

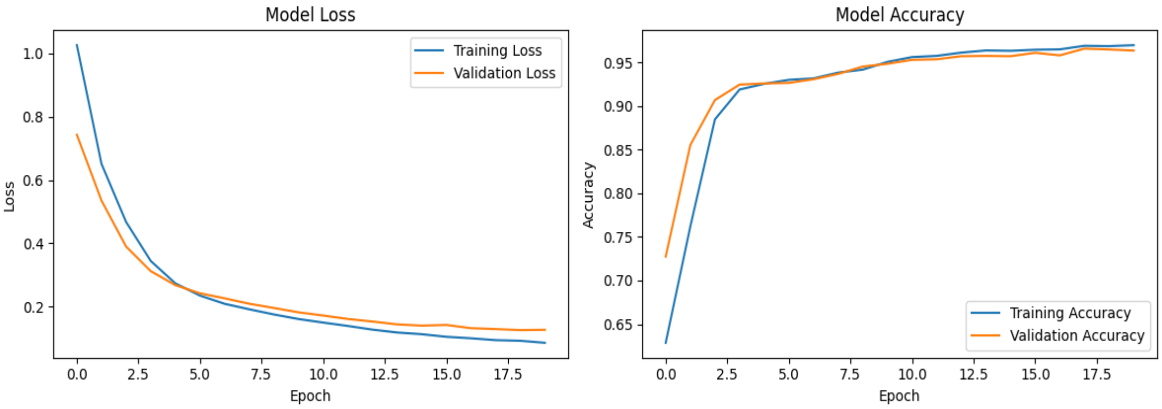


Figure 5. Loss and Accuracy Curves of GA + LSTM

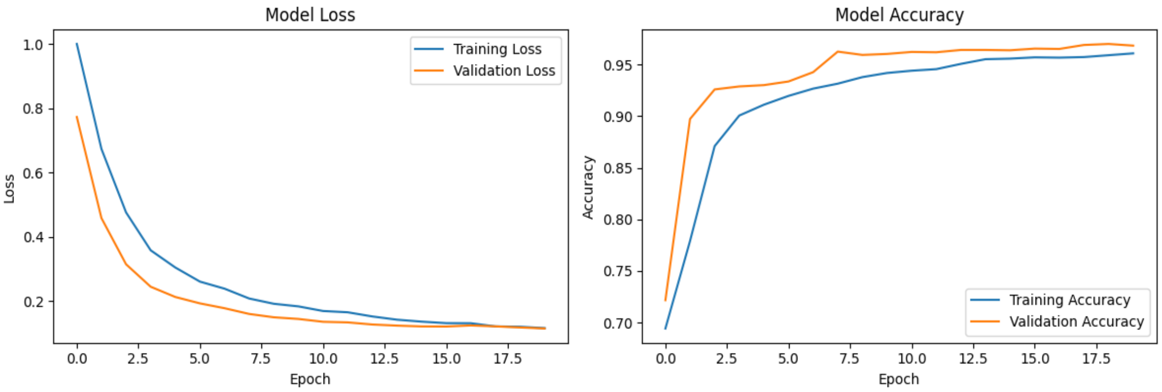
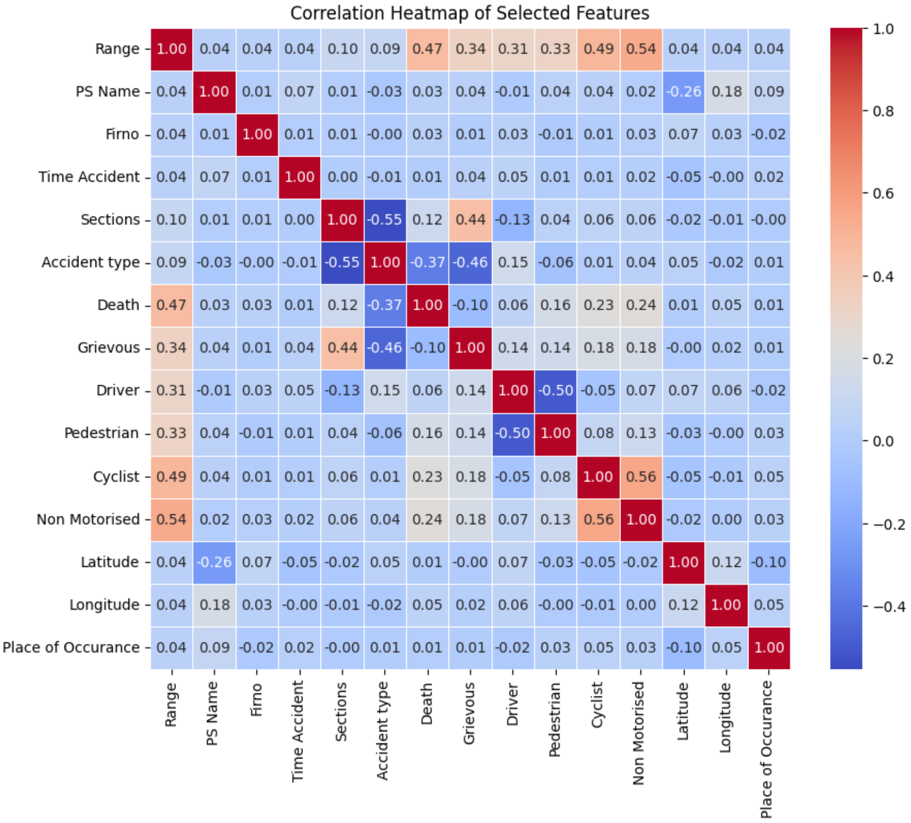


Figure 6. Loss and Accuracy Curves of GA + GRU

3.2. Coyote Optimization Algorithm Selected Features

The original dataset, initially composed of 46 features, underwent a meticulous refinement process guided by the Coyote Optimization Algorithm (COA). This feature selection procedure resulted in a more focused dataset, narrowing down to 14 crucial attributes, including Range, PS Name, Firno, Time Accident, Sections, Death, Grievous, Driver, Pedestrian, Cyclist, Non Motorised, Latitude, Longitude, and Place of Occurrence. The deliberate reduction in features aims to enhance the dataset's efficiency and interpretability for subsequent data analysis and modeling tasks. This strategic curation ensures that the selected attributes align closely with the research objectives. Visual representation of the interrelationships among these chosen features is depicted in Figure 7, showcasing the Correlation HeatMap generated using the COA and providing a comprehensive graphical overview of their associations.



	Range	PS Name	Firno	Time Accident	Sections	Accident type	Death	Grievous	Driver	Pedestrian	Cyclist	Non Motorised	Latitude	Longitude	Place of Occurance
Range	1.00	0.04	0.04	0.04	0.10	0.09	0.47	0.34	0.31	0.33	0.49	0.54	0.04	0.04	0.04
PS Name	0.04	1.00	0.01	0.07	0.01	-0.03	0.03	0.04	-0.01	0.04	0.04	0.02	-0.26	0.18	0.09
Firno	0.04	0.01	1.00	0.01	0.01	-0.00	0.03	0.01	0.03	-0.01	0.01	0.03	0.07	0.03	-0.02
Time Accident	0.04	0.07	0.01	1.00	0.00	-0.01	0.01	0.04	0.05	0.01	0.01	0.02	-0.05	-0.00	0.02
Sections	0.10	0.01	0.01	0.00	1.00	-0.55	0.12	0.44	-0.13	0.04	0.06	0.06	-0.02	-0.01	-0.00
Accident type	0.09	-0.03	-0.00	-0.01	-0.55	1.00	-0.37	-0.46	0.15	-0.06	0.01	0.04	0.05	-0.02	0.01
Death	0.47	0.03	0.03	0.01	0.12	-0.37	1.00	-0.10	0.06	0.16	0.23	0.24	0.01	0.05	0.01
Grievous	0.34	0.04	0.01	0.04	0.44	-0.46	-0.10	1.00	0.14	0.14	0.18	0.18	-0.00	0.02	0.01
Driver	0.31	-0.01	0.03	0.05	-0.13	0.15	0.06	0.14	1.00	-0.50	-0.05	0.07	0.07	0.06	-0.02
Pedestrian	0.33	0.04	-0.01	0.01	0.04	-0.06	0.16	0.14	-0.50	1.00	0.08	0.13	-0.03	-0.00	0.03
Cyclist	0.49	0.04	0.01	0.01	0.06	0.01	0.23	0.18	-0.05	0.08	1.00	0.56	-0.05	-0.01	0.05
Non Motorised	0.54	0.02	0.03	0.02	0.06	0.04	0.24	0.18	0.07	0.13	0.56	1.00	-0.02	0.00	0.03
Latitude	0.04	-0.26	0.07	-0.05	-0.02	0.05	0.01	-0.00	0.07	-0.03	-0.05	-0.02	1.00	0.12	-0.10
Longitude	0.04	0.18	0.03	-0.00	-0.01	-0.02	0.05	0.02	0.06	-0.00	-0.01	0.00	0.12	1.00	0.05
Place of Occurance	0.04	0.09	-0.02	0.02	-0.00	0.01	0.01	0.01	-0.02	0.03	0.05	0.03	-0.10	0.05	1.00

Figure 7. Correlation Heat Map of Selected Features using COA

Figures 8, 9, 10, 11 represents the Loss and Accuracy Curves of COA + MLP, COA + RNN, COA + LSTM, COA + GRU Hybrid models respectively.

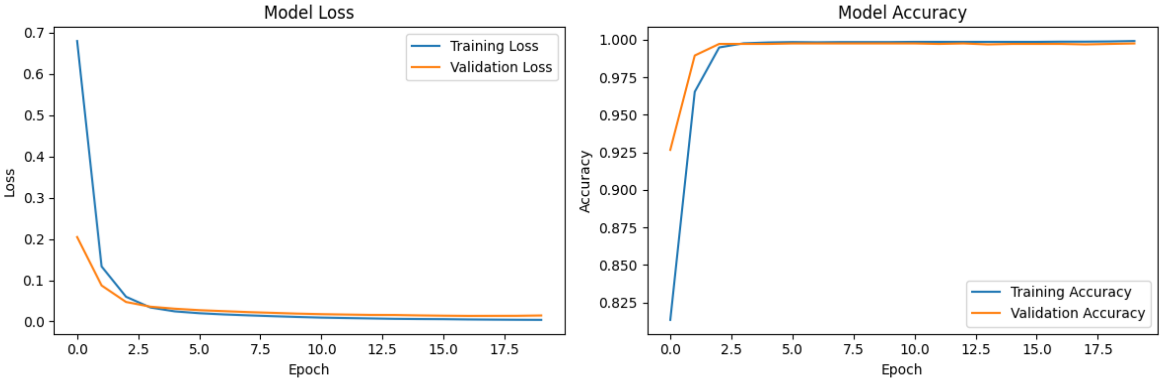


Figure 8. Loss and Accuracy Curves of COA + MLP

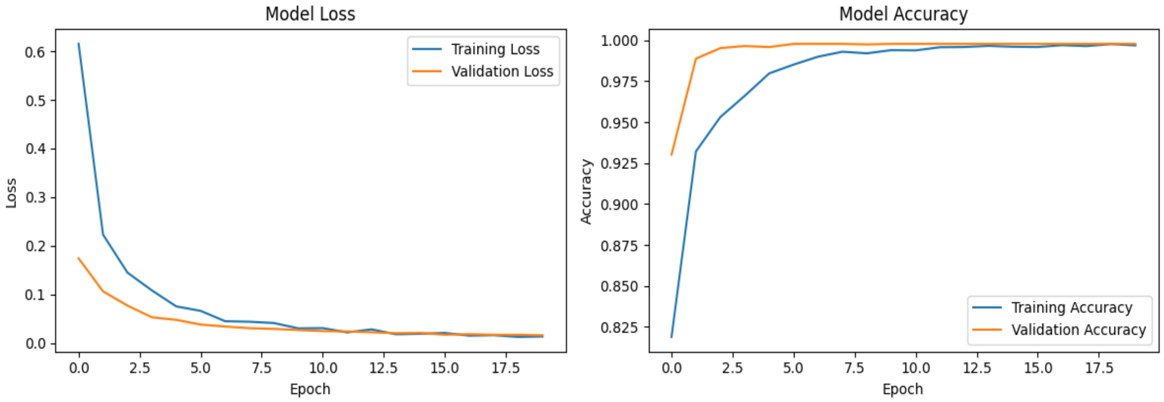


Figure 9. Loss and Accuracy Curves of COA + RNN

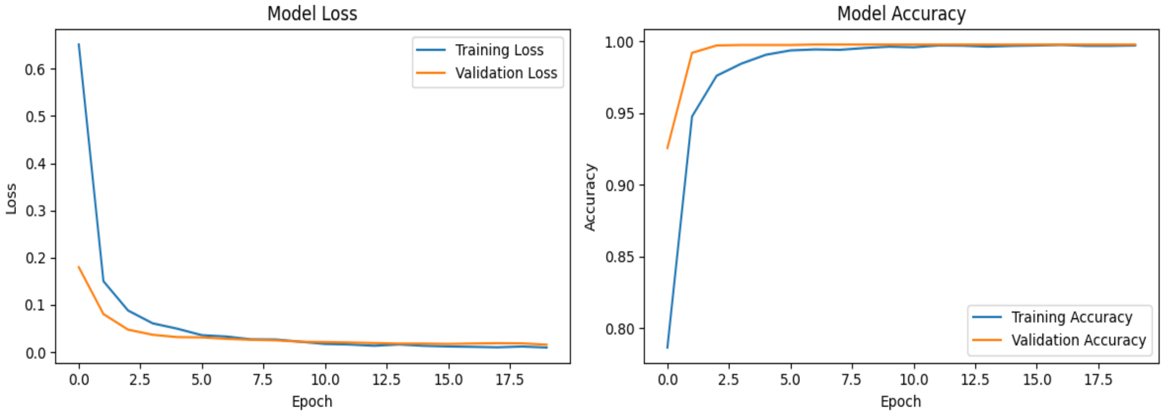


Figure 10. Loss and Accuracy Curves of COA + LSTM

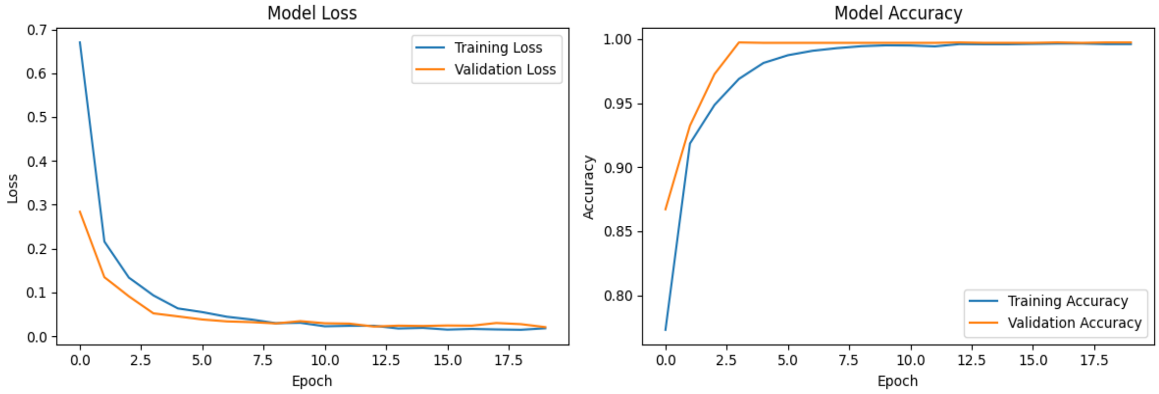


Figure 11. Loss and Accuracy Curves of COA + GRU

In the initial dataset of 46 features, Genetic Algorithm (GA) identified and selected 15 features, while the Coyote Optimization Algorithm (COA) refined the set to 14 features. The outcomes of the selections made by GA and COA are described in Table 1, presenting a detailed overview of the features chosen by each algorithm.

Table 1. Features selected before and after optimization

Initial set of Attributes(46)			Features selected using GA(15)	Features selected using COA(14)
Zone	Minor	Divider	Sections	Range
Range	Driver	Spot Accident	Accident Type	PS Name
District	Passenger	Speed Limit	Death	Firno
Subdivision	Pedestrian	Weather	Grievous	Time Accident
Circle	Cyclist	Road No	Longitude	Sections
PS Name	Other Persons	Road Surface	Divider	Death
Firno	Motorised	T - Junction	Weather	Grievous
Date Report	Non Motorised	Road Chainage	Road Surface	Driver
Date Accident	Latitude	Hit Run	T – Junction	Pedestrian
Time Report	Longitude	Collision	Road Chainage	Cyclist
Time Accident	Place of Occurance	Type Road	Hit Run	Non Motorised
Sections	Type Area	Cause Accident	Cause Accident	Latitude
Accident Type	On going road works	Road Features	Road Features	Longitude
Death	City/Town/Village	Visibility	Visibility	Place of Occurance
Grievous	Lanes Road		Traffic Control	
Traffic Control				
Vehicle Type				

Table 2 displays the accuracies of both deep learning models and hybrid models achieved through feature selection using Genetic Algorithm and Coyote Optimization. Among these models, the two highest-performing ones, COA+RNN and COA+LSTM, with an impressive accuracy of 99.77%, have been chosen for ensemble.

Table 2. Accuracy of the models without optimization and with optimization

Models	Without Optimization	Genetic Algorithm	Coyote Optimization
MLP	83.84	99.77	99.74
RNN	72.15	96.58	99.77
LSTM	72.12	96.84	99.77
GRU	72.15	96.68	99.73

3.3. Ensemble model

The ensemble model results from stacking COA+RNN and COA+LSTM, combining their predictive capabilities. This approach leverages the strengths of each model to enhance overall performance, achieving a robust and highly accurate prediction with improved results. Figure 12 shows the Model Loss and Accuracy Curves of the ensembled model.

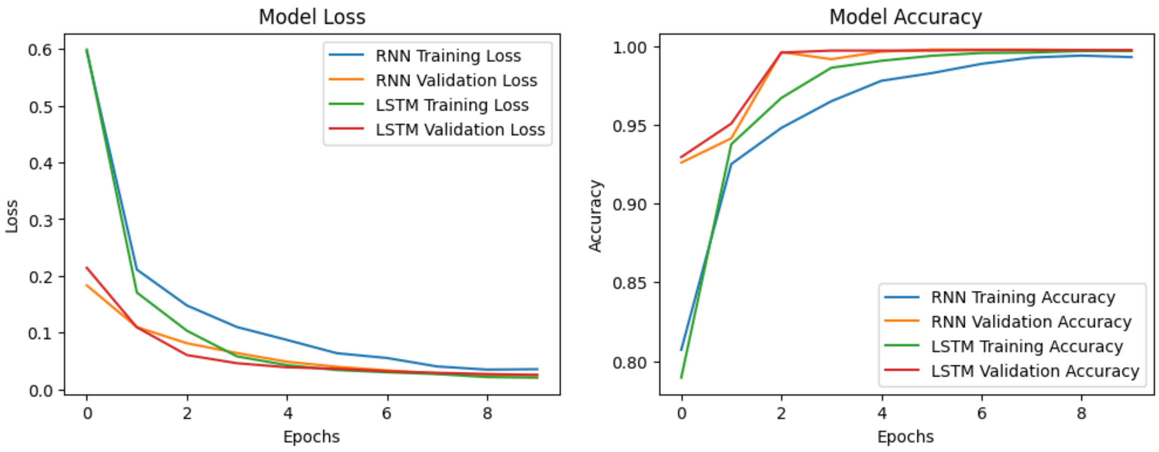


Figure 12. Loss and Accuracy Curves of Ensemble model

In classification tasks, a confusion matrix serves as a valuable instrument, encapsulating the performance of a model. It offers a thorough breakdown of true positives, true negatives, false positives, and false negatives, facilitating the evaluation of F1-score, precision, accuracy, recall. Figure 13 illustrates the Confusion Matrix for the Ensemble model.

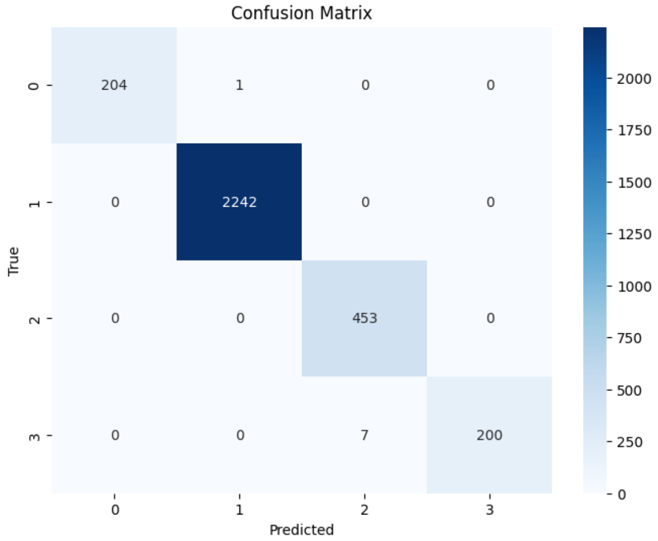


Figure 13. Confusion Matrix of Ensemble model

3.4. Comparision with Other Models

Table 3 displays a comparison with alternative models namely RNN, Random Forest and Convolutional Neural Network (RFCNN), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbours (KNN), MLP, Logistic Regression, Simple CART and PART. It's noteworthy that certain other models utilized with distinct datasets might exhibit higher accuracy than our ensemble model (COA-RNN + COA-LSTM), which achieved 99.77% accuracy.

Table 3. Comparision with other models

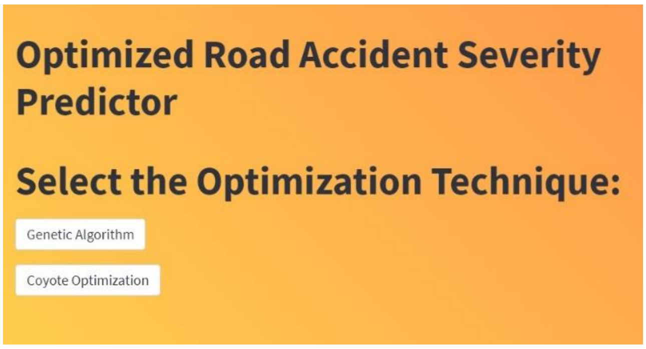
S. No	Author	Dataset	Methodology	Metric Score (Accuracy)
1	Maher Ibrahim Sameen et al. [24]	North-South Expressway (NSE)	RNN	71.77%
2	M. Manzoor et al. [25]	Car Accident dataset(USA)	RFCNN	RFCNN = 99.1%

3	T. Vaiyapuri et al. [26]	Web data from data.gov.in	SVM, RF, KNN, MLP, Logistic Regression	SVM = 60%
				RF = 88%
				KNN = 85%
				MLP = 90%
				Logistic Regression = 54%

4	Lukuman Wahab et al. [27]	Road traffic crash files	Simple CART, PART and MLP	Simple CART = 73.81%
				PART = 73.45%
				MLP = 72.16%

3.5. User Interface

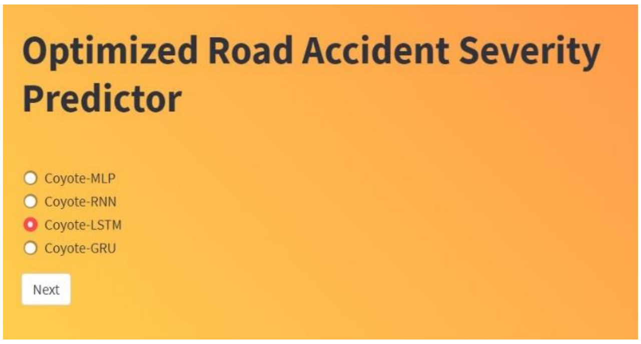
The user interface guides users to choose between two optimization techniques, Genetic Algorithm and Coyote Optimization, offering flexibility. In the subsequent screen, users can select from four models (MLP, RNN, LSTM, and GRU). The chosen model collaborates with the previously selected optimization technique, forming a hybrid model. Based on the optimization technique chosen, the interface displays selected features for user input. For instance, if a user selects 'Coyote Optimization' with the 'LSTM' model, Figures 14 and 15 showcase the chosen optimization technique and the corresponding algorithm.



Optimized Road Accident Severity Predictor

Select the Optimization Technique:

Figure 14. Selection of optimization technique

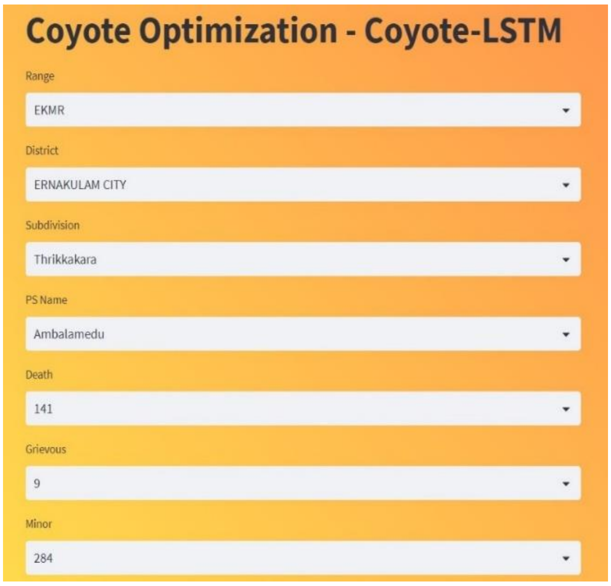


Optimized Road Accident Severity Predictor

○ Coyote-MLP
○ Coyote-RNN
● Coyote-LSTM
○ Coyote-GRU

Figure 15. Selection of hybrid model

The interface collects a range of inputs, including Range, District, Subdivision, PS Name, Death, Grievous, Minor, Passenger, Motorized, Non-Motorized, Longitude, and Place of Occurrence. Using these inputs, the trained model exhibits remarkable accuracy in predicting the corresponding output. In a specific case, input values include EKMR for Range, ERNAKULAM CITY for District, Thrikkakara for Subdivision, Ambalamedu for PS Name, 141 for Death, 9 for Grievous, 284 for Minor, 338 for Passenger, 7 for Motorized, 37 for Non-Motorized, 75.125897 for Longitude, and Bypass Junction Palarivation for Place of Occurrence. The model accurately predicts the output as 'Fatal.' Hence, if a given accident record involves these specified feature values, the accident severity is unequivocally labeled as 'Fatal.' Figures 16 and 17 visually depict the interface, capturing the values for selected features by COA. The COA-LSTM model classifies the accident record as 'Fatal.'



Coyote Optimization - Coyote-LSTM

Range: EKMR

District: ERNAKULAM CITY

Subdivision: Thrikkakara

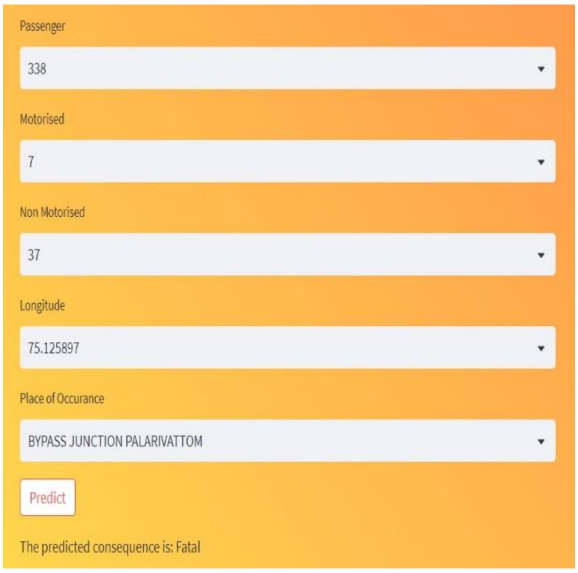
PS Name: Ambalamedu

Death: 141

Grievous: 9

Minor: 284

Figure 16. Inputting the feature values



Passenger

338

Motorised

7

Non Motorised

37

Longitude

75.125897

Place of Occurance

BYPASS JUNCTION PALARIVATTOM

The predicted consequence is: Fatal

Figure 17. Classification of accident record

3.6. Accident Hotspots

Ernakulam district is under consideration, encompassing over 10,000 accident records spanning from 2018 to 2022. All these accident locations are mapped on the Ernakulam district map using ArcGIS. Areas with a higher number of accidents are highlighted in intense red, while those with fewer accidents are marked with a lighter red color. Figure 18 illustrates the Ernakulam district map, showcasing the representation of accident hotspots.

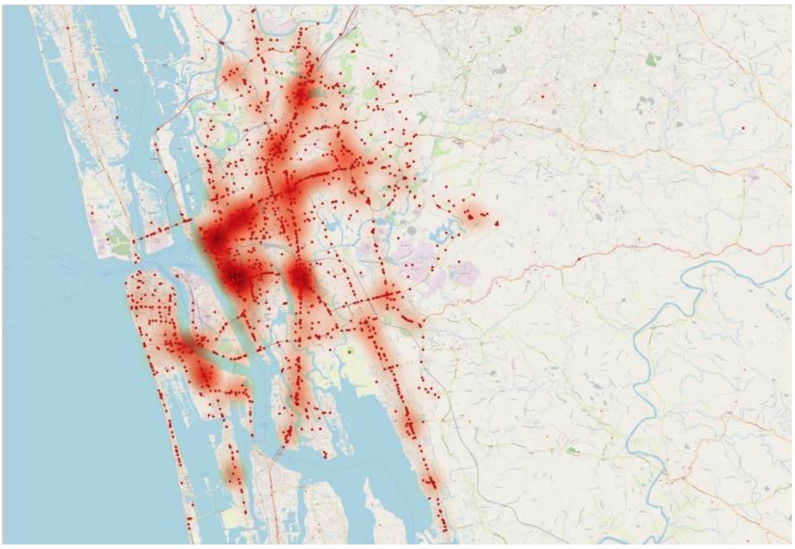


Figure 18. Accident Hotspots in Ernakulam District

4. CONCLUSION

The research is confined to Ernakulam district, focusing on its distinctive road accident challenges. The ensemble model categorizes accident severity into Grievous Injury, Minor Injury, Non-Injury, and Fatal classes. The ensemble model (COA-RNN + COA-LSTM) achieved an accuracy of 99.77%. Further efforts center on refining classification models, involving algorithm fine-tuning, performance optimization, and dataset expansion. The research plans to expand the system beyond Ernakulam, aiming to make a bigger impact on road safety efforts. Through the expanded deployment of the system, the objective is to contribute to a broader reduction in accidents, thereby alleviating the financial burdens experienced by victims and their families.

REFERENCES

[1] S. K. Singh, "Road Traffic Accidents in India: Issues and Challenges," Transportation Research Procedia, vol. 25, pp. 4708-4719, 2017, ISSN 2352-1465, doi: 10.1016/j.trpro.2017.05.484.

[2] B. Kumeda, F. Zhang, F. Zhou, S. Hussain, A. Almasri and M. Assefa, "Classification of Road Traffic Accident Data Using Machine Learning Algorithms," 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN), Chongqing, China, 2019, pp. 682-687, doi: 10.1109/ICCSN.2019.8905362.

[3] Z. O. Hamad, "Review of Feature Selection Methods Using Optimization Algorithm (Review Paper for Optimization Algorithm)," Polytechnic Journal, vol. 12, no. 2, Article 24, pp. 203-214, 2023, doi: 10.25156/ptj.v12n2y2022.pp203-214.

[4] S. S. S. Reddy, A. Kumar, K. Z. Ghafoor, V. P. Bhardwaj, and S. Manoharan, "CoySvM-(GeD): Coyote Optimization-Based Support Vector Machine Classifier for Cancer Classification Using Gene Expression Data," Journal of Sensors, vol. 2022, Article ID 6716937, pp. 1-9, 2022. doi: 10.1155/2022/6716937.

[5] J. Zhou and Z. Hua, "A Correlation-Guided Genetic Algorithm and Its Application to Feature Selection," Applied Soft Computing, vol. 123, p. 108964, 2022, ISSN 1568-4946, doi: 10.1016/j.asoc.2022.108964.

[6] M. Sangare, S. Gupta, S. Bouzefrane, S. Banerjee, and P. Muhlethaler, "Exploring the Forecasting Approach for Road Accidents: Analytical Measures with Hybrid Machine Learning," Expert Systems with Applications, vol. 167, p. 113855, 2021, ISSN 0957-4174, doi: 10.1016/j.eswa.2020.113855.

[7] P. Wu, X. Meng, and L. Song, "A Novel Ensemble Learning Method for Crash Prediction Using Road Geometric Alignments and Traffic Data," Journal of Transportation Safety & Security, vol. 12, no. 9, pp. 1128-1146, 2020, doi: 10.1080/19439962.2019.1579288.

[8] D. Devaraj, D. Chandrasekaran, B. Pandian, B. Binitha, M. Dipikhasre, and R. Anil, "Road Accident Analysis in Kerala and Location-Based Severity Level Classification Using Decision Tree Algorithm," Paid. J., vol. 14, no. XIV, pp. 41-50, 2021.

[9] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843640.

[10] V. N. M. Gilani, S. M. Hosseinian, M. Ghasedi, and M. Nikookar, "Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models," Mathematical Problems in Engineering, vol. 2021, Article ID 9974219, pp. 1-11, 2021, doi: 10.1155/2021/9974219.

[11] E. K. Adanu, W. Agyemang, R. Islam, and S. Jones, "A Comprehensive Analysis of Factors That Influence Interstate Highway Crash Severity in Alabama," Journal of Transportation Safety & Security, vol. 14, no. 9, pp. 1552-1576, 2022, doi: 10.1080/19439962.2021.1949414.

[12] Md. K. Islam, U. Gazder, R. Akter, and Md. Arifuzzaman, “Involvement of Road Users from the Productive Age Group in Traffic Crashes in Saudi Arabia: An Investigative Study Using Statistical and Machine Learning Techniques,” Applied Sciences, vol. 12, no. 13, p. 6368, Jun. 2022, doi: 10.3390/app12136368.

[13] D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, “Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction,” Computers, vol. 10, no. 12, p. 157, Nov. 2021, doi: 10.3390/computers10120157.

[14] A. Al-Omari, N. Shatnawi, T. Khedaywi, et al., "Prediction of Traffic Accidents Hot Spots Using Fuzzy Logic and GIS," Appl Geomat, vol. 12, pp. 149-161, 2020, doi: 10.1007/s12518-019-00290-7.

[15] T. Bokaba, W. Doorsamy, and B. S. Paul, “Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents,” Applied Sciences, vol. 12, no. 2, p. 828, Jan. 2022, doi: 10.3390/app12020828.

[16] A. Sysoev, V. Klyavin, A. Dvurechenskaya, A. Mamedov, and V. Shushunov, “Applying Machine Learning Methods and Models to Explore the Structure of Traffic Accident Data,” Computation, vol. 10, no. 4, p. 57, Mar. 2022, doi: 10.3390/computation10040057.

[17] M. I. Santos, P. T. M. S. Oliveira, and A. P. C. Larocca, "Investigation the Influence of Risk Factors on the Occurrence of Road Accidents Using the Driver Performance Model," Transp. in Dev. Econ., vol. 8, p. 10, 2022, doi: 10.1007/s40890-021-00148-x.

[18] I. Ashraf, S. Hur, M. Shafiq, and Y. Park, "Catastrophic Factors Involved in Road Accidents: Underlying Causes and Descriptive Analysis," PLoS ONE, vol. 14, no. 10, p. e0223473, 2019, doi: 10.1371/journal.pone.0223473.

[19] J. Mesquitela, L. B. Elvas, J. C. Ferreira, and L. Nunes, “Data Analytics Process over Road Accidents Data—A Case Study of Lisbon City,” ISPRS International Journal of Geo-Information, vol. 11, no. 2, p. 143, Feb. 2022, doi: 10.3390/ijgi11020143.

[20] R. Sathiyaraj, A. Bharathi, Sikandar Khan, Tayybah Kiren, Inam Ullah Khan, and Muhammad Fayaz, "A Genetic Predictive Model Approach for Smart Traffic Prediction and Congestion Avoidance for Urban Transportation," Wireless Communications and Mobile Computing, vol. 2022, Article ID 5938411, pp. 1-12, 2022, doi: 10.1155/2022/5938411.

[21] S. A. Raza, A. W. Siddiqui, F. M. Butt, M. A. Elahi, and K. S. Minhas, "Saudi Arabian Road Accident Mortality and Traffic Safety Interventions Dataset (2010–2020)," Data in Brief, vol. 44, p. 108502, 2022, ISSN 2352-3409, doi: 10.1016/j.dib.2022.108502.

[22] S. Olugbade, S. Ojo, A. L. Imoize, J. Isabona, and M. O. Alaba, “A Review of Artificial Intelligence and Machine Learning for Incident Detectors in Road Transport Systems,” Mathematical and Computational Applications, vol. 27, no. 5, p. 77, Sep. 2022, doi: 10.3390/mca27050077.

[23] R. Satria, J. Aguero-Valverde, and M. Castro, "Spatial Analysis of Road Crash Frequency Using Bayesian Models with Integrated Nested Laplace Approximation (INLA)," Journal of Transportation Safety & Security, vol. 13, no. 11, pp. 1240-1262, 2021, doi:

10.1080/19439962.2020.1726542.

[24] M. Sameen and B. Pradhan, “Severity Prediction of Traffic Accidents with Recurrent Neural Networks,” Applied Sciences, vol. 7, no. 6, p. 476, Jun. 2017, doi: 10.3390/app7060476.

[25] M. Manzoor et al., "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model," in IEEE Access, vol. 9, pp. 128359-128371, 2021, doi: 10.1109/ACCESS.2021.3112546.

[26] T. Vaiyapuri and M. Gupta, "Traffic accident severity prediction and cognitive analysis using deep learning," Soft Computing, 2021. [Online]. Available: https://doi.org/10.1007/s00500-021-06515-5.

[27] L. Wahab and H. Jiang, "Severity prediction of motorcycle crashes with machine learning methods," International Journal of Crashworthiness, vol. 25, no. 5, pp. 485-492, 2020. DOI: 10.1080/13588265.2019.1616885.

BIOGRAPHIES OF AUTHORS

	Dr. M. Sobhana is currently working as an associate professor in the Department of Computer Science and Engineering, V. R. Siddhartha Engineering College, Vijayawada, India. She received Ph.D. degree in Computer Science and Engineering in 2018 from Krishna University. She has 16 years of teaching experience. Her research interests lie in areas such as Artificial Intelligence, Machine Learning, Data Analytics, Cyber Security, and Software Engineering. She published 35 papers in National and International journals and published 7 patents. She can be contacted at email: sobhana@vrsiddhartha.ac.in.
	Gnana Siva Sai Venkatesh Mendu is a final-year B. Tech. student, specializing in Computer Science and Engineering at V. R. Siddhartha Engineering College, Vijayawada, India. He is passionate about Artificial Intelligence and Machine Learning. He achieved Bronze in the Cisco NetAcad Riders 2023 competition. He holds the position of Institute of Electrical and Electronics Engineers (IEEE) Chair of the Geoscience and Remote Sensing Society (GRSS) Student Chapter. He can be contacted at email: sivasaivenkatesh.m@gmail.com.
	Nihitha Vemulapalli is a final-year B. Tech student specializing in Computer Science and Engineering at V. R. Siddhartha Engineering College, Vijayawada, India. She is passionate about Deep Learning and has been recognized as a Google Women Engineers Scholar. Additionally, she holds the position of IEEE Chair at the WIE (Women in Engineering) student chapter. She can be contacted at email: nihithavemulapalli@gmail.com.
	Kushal Kumar Chintakayala is a final-year B. Tech. student, specializing in Computer Science and Engineering at V. R. Siddhartha Engineering College, Vijayawada, India. He is passionate about Artificial Intelligence and Machine Learning. He is IEEE core member of Computer Society of India (CSI) student chapter. He can be contacted at email: kushalkumarchintakayala12@gmail.com.