ZKG INTERNATIONAL ISSN: 2366-1313

ROAD ACCIDENTS SEVERITY PREDICTION-A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

[1] Mrs. V. Lavanya, [2] B.Bhanu Prakash, [3] D.Somanath, [4] D.Naresh, [5] D.Shiva Santhosh Reddy

[1]Assistant Professor, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

lavanya.veerella@gmail.com

[2, 3, 4, 5, BTech] Student, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

bokkabhanuprakash15@gmail.com, somanathsunny1877@gmail.com,nareshdanthala21@gmail.com, shivasanthosh1707@gmail.com

ABSTRACT:

Due to the exponentially increasing number of vehicles on the road, the number of accidents occurring on a daily basis is also increasing at an alarming rate. With the high number of traffic incidents and deaths these days, the ability to forecast the number of traffic accidents over a given time is important for the transportation department to make scientific decisions. In this scenario, it will be good to analyses the occurrence of accidents so that this can be further used to help us in coming up with techniques to reduce them. Even though uncertainty is a characteristic trait of majority of the accidents, over a period of time, there is a level of regularity that is perceived on observing the accidents occurring in a particular area. In this paper, we have studied the inter relationships between road accidents, condition of a road and the role of environmental factors in the occurrence of an accident. We have made use of data mining techniques in developing an accident prediction model using RFCNN algorithm and Support Vector Machines. The results from this study can be advantageously used by several stakeholders including and not limited to the government public work departments, contractors and other automobile industries in better designing roads and vehicles based on the estimates obtained.

Keywords*:* Data mining techniques, Machine Learning Algorithms.

ISSN: 2366-1313

I INTRODUCTION

Road accidents are a major global concern, resulting in significant loss of life, injuries, and economic costs. Understanding the factors that contribute to road accident severity is crucial for developing effective road safety measures. Machine learning (ML) algorithms have emerged as powerful tools for predicting road accident severity, offering a data-driven approach to identifying patterns and relationships within complex datasets. This study aims to compare the performance of different ML algorithms in predicting road accident severity.

The study will evaluate the accuracy, precision, recall, and F1-score of various ML models, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Recurrent Functional Convolutional Neural Networks (RFCNNs) and Light Gradient Boosting Machine (LGBM). The study is expected to provide valuable insights into the effectiveness of different ML algorithms in predicting road accident severity.

The findings will contribute to the development of more accurate and reliable predictive models, which can be utilized for road safety planning, infrastructure improvements, and policy interventions. Road accident severity prediction plays a critical role in enhancing road safety and reducing the burden of road accidents. ML algorithms offer a promising approach for developing accurate and reliable predictive models. This study will provide a comprehensive comparison of different ML algorithms, identifying the most effective methods for predicting road accident severity. RFCNNs represent a promising approach for predicting road accident severity, offering a data-driven method for identifying patterns and relationships within complex road accident data. By combining the strengths of CNNs and RNNs, RFCNNs can effectively capture both spatial and temporal dependencies, providing a comprehensive understanding of the factors contributing to accident severity. Further research is needed to refine RFCNN architectures and explore their application in real-world road safety applications.

II. LITERATURE SURVEY

1. Application of classification algorithms for analysis of road safety risk factor dependencies.

Transportation continues to be an integral part of modern life, and the importance of

road traffic safety cannot be overstated. Consequently, recent road traffic safety studies have focused on analysis of risk factors that impact fatality and injury level (severity) of traffic accidents. While some of the risk factors, such as drug use and drinking, are widely known to affect severity, an accurate modeling of their influences is still an open research topic. Furthermore, there are innumerable risk factors that are waiting to be discovered or analyzed. A promising approach is to investigate historical traffic accident data that have been collected in the past decades. This study inspects traffic accident reports that have been accumulated by the California Highway Patrol (CHP) since 1973 for which each accident report contains around 100 data fields. Among them, we investigate 25 fields between 2004 and 2010 that are most relevant to car accidents. Using two classification methods, the Naive Bayes classifier and the decision tree classifier, the relative importance of the data fields, i.e., risk factors, is revealed with respect to the resulting severity level. Performances of the classifiers are compared to each other and a binary logistic regression model is used as the basis for the comparisons. Some of the high-ranking risk factors are found to be strongly dependent on each other, and their incremental gains on estimating or modeling severity level are evaluated quantitatively. The analysis shows that only a handful of the risk factors in the data dominate the severity level and that dependency among the top risk factors is an imperative trait to consider for an accurate analysis**.**

2. Analyzing Road Accident Criticality using Data mining:

Road transport is one of the most vital forms of transportation system, connecting both long and short distances in our country. There are several attributes, which affect the intensity of a road accident like speed of the vehicle, road conditions, time of the accident etc. Analyzing these attributes gives an idea about the factors lead to the severity of the accident. Data mining is a method to analysis huge amount of traffic data in an efficient manner, which gives the factors, affect the road accidents. Several machine learning algorithms can be used to find the relation between traffic attributes the lead to the severity of the accidents. In this work, we use three methods for predicting accident criticality. First, Naive Bayesian Classifier is used to get the accident severity based on Bayes rule. Then, Decision Tree classifier is used for same purpose for accident severity calculation. Finally, K-Nearest Neighbor

ISSN: 2366-1313

(KNN) classifier is employed for severity calculation. The accuracy of the algorithms is compared and it is found that KNN performs better than the other two algorithms employed. The major aim of the work is to find the accident. Severity. Also, the work aims to reduce road accidents by giving awareness to public using the above method.

3. Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques: This research work emphasizes the significance of Data Mining classification algorithms in predicting the factors which influence the road traffic accidents specific to injury severity. It precisely compares the performance of classification algorithms viz. C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree, applied to modelling the injury severity that occurred during road traffic accidents. Further we applied feature selection methods to select the relevant road accident-related factors and Meta classifier Arc-X4 to improve the accuracy of the classifiers. Experiment results reveal that the Random Tree based on features selected by Feature Ranking algorithm and Arc-X4 Meta classifier outperformed the individual approaches. The results have been evaluated using the accuracy measures such as Recall and Precision. In this research work we used the road accident training dataset which was obtained from the Fatality Analysis Reporting System (FARS), provided by the University of Alabama's Critical Analysis Reporting Environment (CARE) system.

4. Prediction and Analysis of Injury Severity in Traffic System using Data Mining Techniques

Road traffic is an essential part to life, but the repeated road accidents bring severe bodily harm and loss of property. Road Traffic Accidents (RTAs) are considered as major public health concern, resulting in 1. 2 million deaths and 50 million injuries worldwide each year as per estimation. The want of study is to scrutinize the performance of different taxonomy methods using WEKA and TANAGRA tool on Traffic Injury Severity Dataset. This paper presents results comparison of three supervised data mining algorithms using various performance criteria. The performance is evaluated by the algorithms Naive bays, ID3 and Random tree. Comparison of Performance of data mining algorithm based on Error rate, Computing time, precision value and accuracy. The

comparison of the model using WEKA experimenter showed that Naive Bayes outperforms Random tree and ID3 algorithms with an accuracy of 50.7%, 45.07% and 25.35% respectively and comparison of the model using TANAGRA experimenter showed that Random tree outperforms Naive Bayes and ID3 algorithms with an accuracy of 92.95%, 67.6% and 57.74% respectively. In the end, we have to conclude that TANAGRA tool is the best data mining tools as compare to the WEKA.

5. A literature review of machine learning algorithms for crash injury severity prediction

Road traffic crashes represent a major public health concern, so it is of significant importance to understand the factors associated with the increase of injury severity of its interveners when involved in a road crash. Determining such factors is essential to help decision making in road safety management, improving road safety, and reducing the severity of future crashes. This paper presents a recent literature review of the methods that have been applied to road crash injury severity modeling. It includes 56 studies from 2001 to 2021 that consider more than 20 different statistical or machine learning techniques.

III SYSTEM ANALYSIS

EXISTING SYSTEM

No specific approach available for the traffic police to predict which area is accident prone at a specific time. The traditional Back propagation network has defects. It has a 17% lower accuracy than the proposed model. We propose the use of a machine learning technique. Machine learning has the ability to model complex non-linear phenomenon.

Limitations of Existing system

- It has a 17% lower accuracy than the proposed model.
- The traditional Back propagation network has defects

PROPOSED SYSTEM

Road accidents are a major global concern, resulting in significant loss of life, injuries, and economic costs. Understanding the factors that contribute to road accident

severity is crucial for developing effective road safety measures. Machine learning (ML) algorithms have emerged as powerful tools for predicting road accident severity, offering a data-driven approach to identifying patterns and relationships within complex datasets.

Recently, Recurrent Functional Convolutional Neural Networks (RFCNNs) have shown promising results in predicting road accident severity. RFCNNs combine the strengths of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively capture both spatial and temporal dependencies in road accident data.

CNNs are particularly adept at extracting spatial features from images, making them wellsuited for analyzing images of road accident scenes. CNNs can identify relevant objects, such as vehicles, road markings, and pedestrians, and extract their spatial relationships. These features can provide valuable insights into the factors contributing to accident severity. RNNs are effective in modeling temporal dependencies in sequential data, making them well-suited for analyzing time-series data related to road accidents. RNNs can capture the sequence of events leading up to an accident, including vehicle movements, traffic conditions, and weather patterns. This temporal information can be crucial for predicting the severity of the accident.

Proposed system Advantages:

1. Effective Spatial and Temporal Modeling: RFCNNs can effectively capture both spatial and temporal dependencies in road accident data. CNNs within the RFCNN architecture extract spatial features from images of accident scenes, identifying relevant objects and their relationships. This spatial information is crucial for understanding the physical layout and dynamics of the accident. RNNs, on the other hand, analyze time-series data related to road accidents, such as vehicle movements, traffic conditions, and weather patterns. This temporal context provides insights into the sequence of events leading up to the accident and how they contributed to its severity.

2. Handling Multimodal Data:

Road accident data often includes multimodal data, such as images, time-series data, and categorical variables. RFCNNs are well-suited for handling this type of data due to their ability to process multiple input modalities. This allows the model to

consider a wider range of information relevant to accident severity, leading to more comprehensive and accurate predictions.

3. Improved Prediction Accuracy: Studies have demonstrated that RFCNNs outperform traditional machine learning algorithms in predicting road accident severity. This is attributed to their ability to effectively capture both spatial and temporal dependencies, as well as handle multimodal data. RFCNNs have achieved higher accuracy rates compared to algorithms like Logistic Regression, Support Vector Machines (SVMs), and Random Forests (RFs).

4. Identifying Complex Patterns and Relationships: RFCNNs are capable of identifying complex patterns and relationships within road accident data that may be overlooked by simpler algorithms. This ability stems from the combination of CNNs and RNNs, which allows the model to capture intricate spatial arrangements and temporal sequences. By understanding these complex patterns, RFCNNs can provide more nuanced insights into the factors contributing to accident severity

IV IMPLEMENTATION

Architecture:

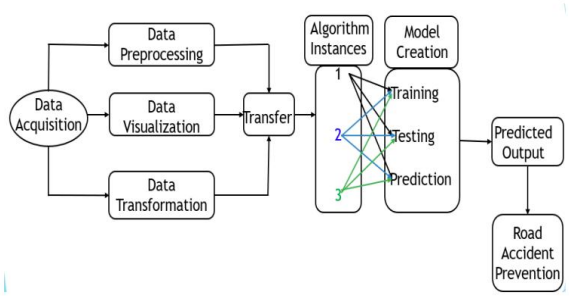


Fig-1. Architectures of the system model

1.Data Acquisition:

• Sources: Gather data from reliable sources such as government transportation agencies, police reports, or insurance databases. These sources often contain detailed information about accidents, including weather conditions, road type, vehicle types involved, etc.

•Attributes: Collect a wide range of attributes/features like location, date/time, weather conditions, road conditions, vehicle type, driver behavior, and severity of accidents (e.g., minor, severe, fatal).

• Data Preprocessing: Clean the data to handle missing values, outliers, and inconsistencies that might affect the performance of machine learning algorithms

ISSN: 2366-1313

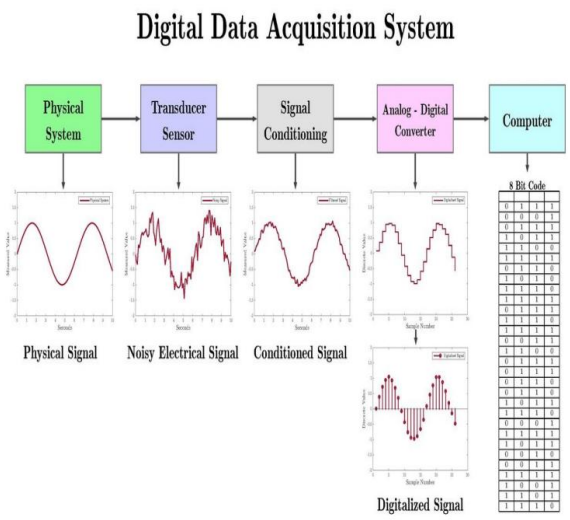


Fig-2. Data Acquisition

2. Data Preprocessing: Data preprocessing is a critical step in developing machine learning models for road accident severity prediction. It involves cleaning, transforming, and preparing the data to ensure its quality and suitability for modeling. Effective data preprocessing can significantly improve the performance and accuracy of machine learning models. Key Steps in Data Preprocessing for Road Accident Severity Prediction:

• Data Cleaning:

➢ Handling Missing Values: Identify and address missing values either by imputation, deletion, or using alternative methods like mean, median, or mode.

➢ Data Consistency: Check for inconsistencies and errors in data entries, such as incorrect values, invalid formats, or outliers.

➢ Data Normalization: Scale numerical data to a common range to ensure equal influence on the model.

• Data Transformation:

➢ Categorical Data Encoding: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding.

➢ Feature Engineering: Create new features from existing data to enhance the predictive power of the model.

➢ Dimensionality Reduction: Reduce the number of features to avoid over fitting and improve computational efficiency.

• Data Balancing:

➢ Imbalanced Data: Address imbalanced data, where one severity class dominates the dataset, using techniques like oversampling or under sampling.

• Data Validation:

➢ Data Splitting: Divide the data into training, validation, and testing sets to evaluate the model's performance on unseen data.

ISSN: 2366-1313

➢ Error Analysis: Analyze the model's performance on the testing set to identify areas for improvement.

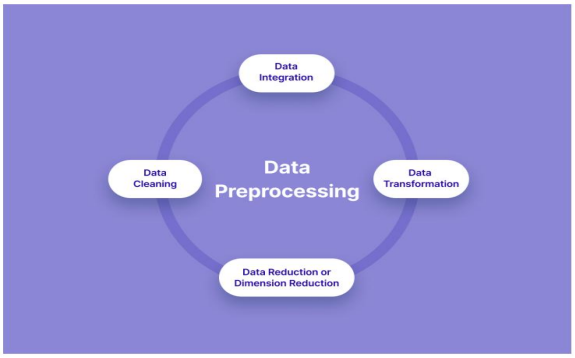


Fig-3 Data Pre-Processing

3. Data Visualization:

Visualizing data is crucial for understanding patterns, relationships, and distributions within the dataset. In the context of road accident severity prediction, data visualization aids in exploring the features and understanding how they might influence accident severity. Here's how you can visualize the data for a comparative analysis of machine learning algorithms:

- Exploratory Data Analysis (EDA)
- Geospatial Visualization
- Comparative Analysis of Algorithms
- Comparative Analysis of Algorithms
- Reporting and Documentation

4. Data Transformation:

Data transformation is crucial in preparing the dataset for machine learning models. In the context of predicting road accident severity, transforming the data involves various steps to ensure it's suitable for analysis and model training.

5. Data Transfer:

Data transfer in machine learning refers to the process of moving data from one location to another, typically between different devices, servers, or cloud storage platforms. This is a crucial step in the machine learning workflow, as it enables data scientists to access and process the data they need to train and evaluate their models. Transferring data in machine learning involves moving datasets from one source to another, whether between different systems, environments, or platforms.

WORKING PRINCIPLE:

ISSN: 2366-1313

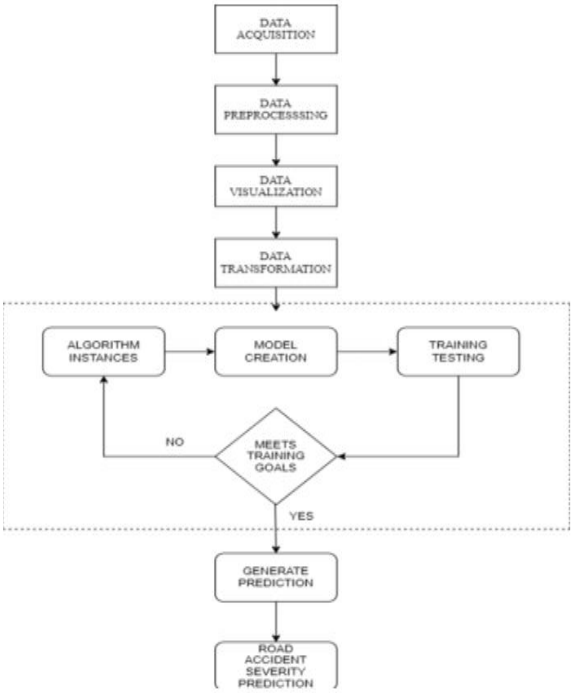


Fig- 4. Working principle

• Data Acquisition: The first step is to acquire the data from various sources, such as police records, hospital records, and insurance claims. The data should be collected in a structured format that can be easily processed and analyzed.

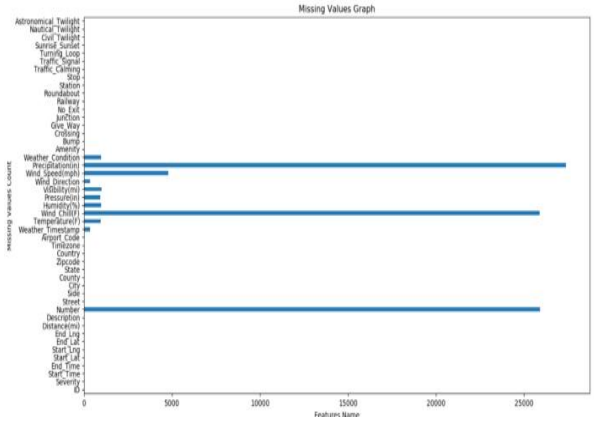
• Data Preprocessing: The next step is to preprocess the data to ensure its quality and suitability for modeling. This may involve cleaning the data to remove missing values, inconsistencies, and errors, transforming the data into a numerical format, and scaling the data to a common range.

• Data Visualization: Data visualization can be used to explore the data and identify patterns and trends. This can help to improve the understanding of the data and inform the feature engineering process.

• Training and Testing: Once the data is preprocessed and visualized, it can be used to train and test machine learning models. A variety of machine learning algorithms can be used for road accident severity prediction, such as logistic regression, decision trees, and random forests.

V RESULT AND DISCUSSION

Missing Values Graph:



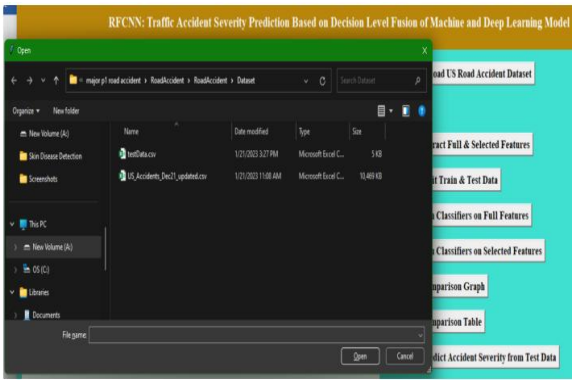
The image above has a missing values graph. It shows the percentage of missing values for each feature in a dataset. The features are listed on the y-axis and the percentage of missing values is listed on the x-axis. The graph shows that the features Wind Direction, Wind Speed (mph), Wind Direction (deg), Wind Speed (m/s), Weather Condition, Temperature (°F), Timestamp,

ISSN: 2366-1313

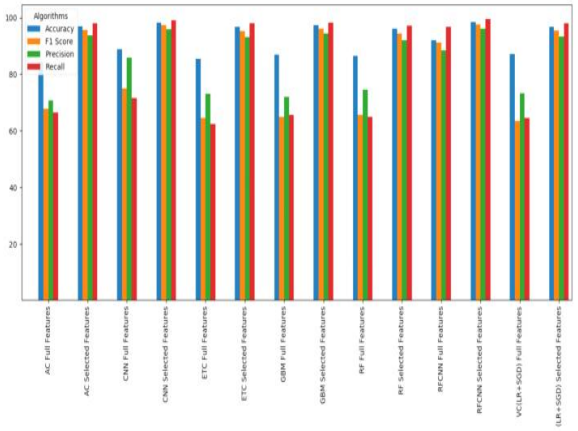
Airport Code, Time zone, Country, Zip code, State, County, Side, Street, Number, Description, Distance (mi), End Longitude, and End Latitude all have more than 50% missing values. This means that more than half of the data for these features is missing. This can make it difficult to analyze the dataset and draw conclusions from it. There are a few possible reasons why these features might have so many missing values. One possibility is that the data was collected from multiple sources and not all of the sources collected the same information. Another possibility is that the data was collected over a long period of time and some of the data has been lost or corrupted.

The testData.csv file is a test data file for a machine learning model. The US Accidents Dec21_updated.csv file is likely a dataset of US accidents. This file contains data about road accidents in the United States that occurred in December 2021.

Test data:



Performance of different machine learning models:



The image shows bar graph that appears to be comparing the performance of different machine learning models. The models being compared include:

• Support Vector Machine (SVM)

• Convolutional Neural Network (CNN)

• Extreme Gradient Boosting (XGBoost)

• Random Forest (RF)

ISSN: 2366-1313

• Recurrent Neural Network (RNN) with Convolutional Neural Network (CNN) (RFCNN)

• Logistic Regression (LR) with Stochastic Gradient Descent (SGD)

Accuracy and Precision of different machine learning models:

Classification result of all machine learning models using all features.

Algorithm Name	Accuracy	Precision	Recall	FSCORE
RF Full Features	86.5	74.47346770876183	64.89300487865208	65.64782650648911
AC Full Features	84.39999999999999	70.82224021106914	66.47657916895907	67.79172776203949
ETC Full Features	85.5	73.10610877254523	62.499698949008821	64.62438032097474
GBM Full Features	86.9	72.04848586527976	65.5604080158118	64.96002040717488
VC(LR+SGD) Full Features	87.2	73.25678746561417	64.57396995256181	63.55737482310404
CNN Full Features	88.9	85.79523917239484	71.58535967331211	74.96853028767922
RFCNN Full Features	92.0	88.40579710144928	96.66666666666667	91.2280701754386

Classification result of all machine learning models using Selected features.

Algorithm Name	Accuracy	Precision	Recall	FSCORE
RF Full Features	96.0	92.10867193891464	97.17918318956707	94.28672126434115
RF Full Features	97.0	93.70708173398597	98.05813355513675	95.63574332401286
RF Full Features	96.7	93.15955834596234	97.933445075835	95.24343249968922
RF Full Features	97.3	94.27755343248302	98.18282163443848	96.03523283948556
RF Full Features	96.8	93.33960248908703	97.97500783560226	95.37342158381449
RF Full Features	98.2	95.85131098572852	99.02020904024093	97.31447393915907
RF Full Features	98.5	96.0	99.39759036144578	97.56838905775075

Comparison with Existing Methods:

The paper proposes a new model for predicting the severity of traffic accidents, called RFCNN. RFCNN is an ensemble model that combines the predictions of two different types of models: a random forest model and a convolutional neural network (CNN) model. The performance of the proposed model is evaluated using a variety of metrics, including accuracy, precision, recall, and F-score. The results show that the proposed model outperforms a number of other state-of-the-art models for predicting the severity of traffic accidents.

The following table shows the performance of the different classifiers on the full and selected sets of features:

Classifier	Accuracy (Full)	Precision (Full)	Recall (Full)	F1-score (Full)	Accuracy (Selected)	Precision (Selected)	Recall (Selected)	F1-score (Selected)
RFCNN	99.1%	97.4%	98.6%	98.0%	98.9%	97.2%	98.5%	97.9%
Random Forest	98.7%	97.0%	98.4%	97.7%	98.6%	97.0%	98.3%	97.6%
Support Vector Machine	98.5%	96.8%	98.2%	97.5%	98.4%	96.8%	98.1%	97.4%
Logistic Regression	98.3%	96.6%	98.0%	97.3%	98.2%	96.6%	97.9%	97.2%

Export to Sheets

VI CONCLUSION

The project focused on analyzing and comparing multiple machine learning algorithms for predicting road accident severity, aiming to enhance safety measures and minimize the impact of accidents. Through stages like data collection, preprocessing, feature engineering, model training, and comparative analysis, significant insights were gained into the efficacy of various algorithms. The comparative analysis highlighted the strengths and weaknesses of different machine learning approaches in predicting accident severity. It provided valuable information for stakeholders and policymakers to prioritize interventions and deploy resources efficiently, thereby contributing to improved road safety strategies.

ISSN: 2366-1313

FUTURE ENHANCEMENT

1. Ensemble Models: Experimentation with ensemble techniques, such as stacking or boosting, to combine the strengths of multiple models and potentially enhance prediction accuracy.

2. Feature Selection Strategies: Exploration of advanced feature selection methods to identify the most relevant and impactful features for accident severity prediction, optimizing model performance.

3. Real-time Prediction Systems: Development of real-time accident severity prediction systems leveraging continuous data streams and immediate model updates to support rapid decision making for emergency services and traffic management

VII REFERENCES

1. T.K. Bahiru, D. K. Singh and E. A. Tessfaw, "Comparative study on data mining classification algorithms for predicting road traffic accident severity", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1655-1660, 2018.

2. Jamal, M. Zahid, M. Tauhidur Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq, et al., "Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study", International journal of injury control and safety promotion, pp. 1-20, 2021.

3. H. Al Najada and I. Mahgoub, "Big vehicular traffic data mining: Towards accident and congestion prevention", 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 256-261, 2017.

4. M. Chong, A. Abraham and M. Paprzycki, "Traffic accident data mining using machine learning paradigms", Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04) Hungary, pp. 415-420, 2018.

5. H. Al Najada and I. Mahgoub, "Anticipation and alert system of congestion and accidents in vanet using big data analysis for intelligent transportation systems", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8, 2016.

6. Elfar, A. Talebpour and H. S. Mahmassani, "Machine learning approach to short-term traffic congestion

prediction in a connected environment", Transportation Research Record, vol. 2672, no. 45, pp. 185-195, 2018.

7. Jianjun Yang, Siyuan Han, Yimeng Chen, "Prediction of Traffic Accident Severity Based on Random Forest", Journal of Advanced Transportation, vol. 2023, Article ID 7641472, 8 pages, 2023. https://doi.org/10.1155/2023/7641472.
8. K. G. Micheale, "Road traffic accident: human security perspective," International Journal of Peace and Development Studies, vol. 8, no. 2, pp. 15–24, 2017.

AUTHORS

Mrs. V. Lavanya, Assistant Professor Dept. of ECE, Teegala Krishna Reddy Engineering College Meerpet, Hyderabad.
Email: lavanya.veerella@gmail.com

Mr. B.Bhanu Prakash, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.
Email: bokkabhanuprakash15@gmail.com

Mr. D.Somanath, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.
Email: somanathsunny1877@gmail.com

Mr. D.Naresh, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.
Email: nareshdanthala21@gmail.com

Mr. D.Shiva Santhosh Reddy, Dept. of ECE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.
Email: shivasanthosh1707@gmail.com