International Journal of
Human Computations & Intelligence
Vol. 03, Issue. 05, 2024

RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Machine Learning Models for Accident Severity Prediction

Samudrala Tarunika . P Daphine Joy . Todupuniri Akshara Reddy . Vishnu KS

School of Computer Science and Engineering
REVA University, Bengaluru, India.

DOI: **10.5281/zenodo.14591111**
Received: 21 November 2024 / Revised: 16 December 2024 / Accepted: 01 January 2025

Abstract — Accidents pose two major concerns: road safety and public health. The primary objective of this study was to develop an accident severity detection system that leverages machine learning algorithms to analyze a variety of influential factors, enabling the prediction of accident severity levels. The supervised learning algorithms employed in this system include Decision Trees, Naive Bayes, Support Vector Machines (SVM), Random Forest, and Logistic Regression, all aimed at providing accurate severity predictions. Key features incorporated in the training and testing datasets encompass driver demographics such as age, gender, education level, and driving experience, along with road characteristics like lane configurations and medians, junction types, and road surface conditions. Environmental factors such as light and weather conditions are also considered, as they may contribute to accident occurrence. Furthermore, accident-specific details, including collision types and vehicle/pedestrian movement patterns, are analyzed to uncover relationships and patterns influencing accident severity. The system produces a severity prediction score with associated probability, facilitating real-time alerts and warnings for stakeholders. This predictive model holds potential for improving road safety by enabling authorities and individuals to proactively mitigate the risk of severe accidents, especially when integrated with road safety initiatives. The research demonstrates the practical application of machine learning in predictive analytics, contributing to public safety efforts and informed policy-making.

Index Terms - Machine learning algorithms, accident severity prediction, traffic, safety factors

I. INTRODUCTION

Road accidents are a leading cause of damage and loss, primarily due to high traffic volumes and the significant freedom of movement allowed to motorists. Accidents involving heavy goods

MILESTONE RESEARCH.IN OPEN ACCESS

ISSN (Online): 2583-5696
Int. Jr. of Hum Comp. & Int.

vehicles, such as lorries and commercial vehicles, in conjunction with public transportation like motorcars, are among the most fatal, resulting in numerous fatalities. Various adverse weather conditions, including rain and fog, exacerbate the likelihood of such accidents. Accurate estimation of accident hotspots and their contributing factors is essential for effective mitigation. Despite established regulations, factors such as driver negligence regarding speed limits, vehicle maintenance, and safety measures like helmet usage continue to contribute to accidents. The increasing number of vehicles has transformed traffic systems into complex structures, complicating the design and management of roadways (S. Krishnaveni, et. al., 2011, Hu, S.,et. al., 2024).

Advancements in technology and the availability of large-scale data have underscored the importance of data mining in traffic safety. This study aims to identify the most suitable machine learning techniques for predicting road accidents by analyzing accident data records. The models developed assist in understanding the characteristics of various factors, such as driver behavior, roadway conditions, lighting, and weather that influence accident occurrence. This analysis enables stakeholders to implement safety measures to prevent accidents. Statistical methods based on directed graphs are employed to compare different scenarios using out-of-sample predictions. The model identifies statistically significant factors that can be used to assess and reduce accident risk. The growing complexity of traffic systems, driven by the increasing number of vehicles, has made traffic operation and accident prevention challenging. This complexity is further aggravated by the diverse range of vehicles, from private cars to heavy goods vehicles and public transportation (Wang, W., et al. (2021), Behboudi, N., et al. (2024).

The proliferation of data and technological advancements have led to significant growth in the availability of traffic and accident-related data. Modern vehicles and infrastructure now generate vast amounts of data through sensors, surveillance systems, and GPS trackers. The reduced cost of storing and processing this data enables organizations and governments to analyze it for actionable insights. This large-scale accumulation of traffic and accident data serves as the foundation for data mining, which involves extracting meaningful information and patterns to understand data. Identifying these patterns is crucial for understanding the factors contributing to accidents and forecasting their occurrence under different conditions (Adefabi, A., et al. (2023), Jin, J., et. al., 2024). Analyzing road accidents involves examining data and posing relevant questions, such as identifying the most hazardous travel times, determining the ratio of accidents in rural versus urban areas, and evaluating collision rates in high-speed zones. Tools like Microsoft Excel can be utilized to review the data, allowing for efficient insights. The primary objective of this analysis is to highlight the critical aspects of road accidents and provide predictive analysis (Sufian, M. A., et al. (2024)).

Figure 1 visualizes dynamic part of the system. It emphasizes the flow of activities rather than the structural components that are often referred to as an “object-oriented flowchart”. These plates are essential for behavioral modeling in software and systems design, furnishing a clear representation of sequences and sequences. Exertion plates correspond to conditioning, which represent overall sequences or tasks in the sequence, and are farther broken down into lower units called conduct. Conduct are specific tasks or operations that make up the logical way within an exertion, similar as calculating a value, transferring a communication, or streamlining a record. Exertion plates are graphical representations of sequences that illustrate step-by-step conditioning and conduct while

incorporating choices, duplications, and concurrency. They're designed to model both computational and organizational sequences, similar as ways and business sequences, as well as the data flows cutting with the affiliated conditioning. While their primary focus is on showing the overall inflow of control, exertion plates can also include rudiments that depict data inflow between conditioning via one or further data stores. This versatility makes them an important tool for modeling complex systems and icing clarity in both computational and organizational surrounds.

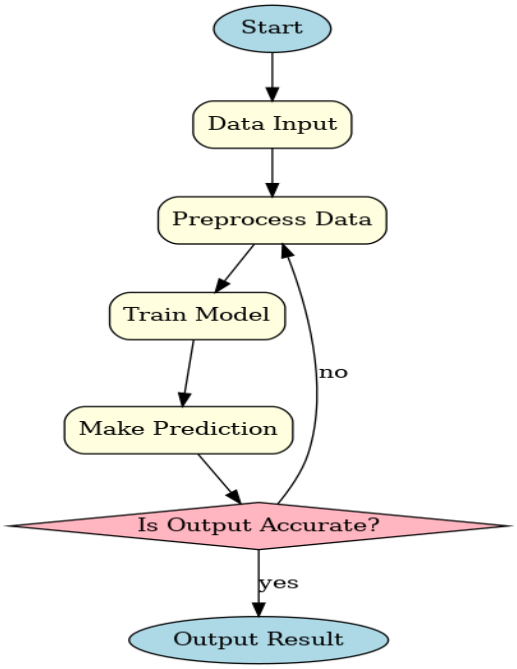

Fig 1: Activity diagram

II. Literature survey

Chang, L.-Y. (2005) utilized data mining and machine learning techniques to identify high-frequency accident locations and identify the factors that mainly affect road accidents in those locations. The process begins with grouping locations using machine learning algorithm k-means clustering based on the frequency count. Later, association rule mining technique is implemented to examine various relationships between attributes, which provides deeper insights into the fundamental cause of the accidents (Labbo, M. S., et. al., 2024). In order to get high-precision results, Kalyoncuoglu, S. F., et al. (2004) suggested a data mining classification method for gender classification that combines the AdaBoost Meta classifier with RndTree and C4.S. The Critical Analysis Reporting Environment (CARE) system, a component of Fatal Analysis Reporting Systems, provided the training data.

In order to identify high-density critical accident areas, Doğan, A. A. E., et al. (2008) proposed a clustering approach that finds that stochastic indices are likely to occur in particular clusters and may be compared in time and geography. The kernel density estimate tool may be used to visualise and modify density-based events to generate the basic spatial unit of the hotspot clustering approach. The extent of damage that occurs during a traffic accident is simulated using the performance of multiple machine learning paradigms, such as decision trees, support vector machines, neural networks trained

using hybrid learning techniques, and concurrent mixed models that combine neural networks and decision trees. The experimental results indicate that the hybrid decision tree neural network technique is an improved method above the machine learning models' single method.

The number of traffic accidents is increasing. Understanding a driver's psychological condition can help prevent fatal accidents. Due to tiredness, a significant percentage of traffic accidents happen throughout the night. By tracking a driver's eye blinks, which reveal their level of tiredness, roadblocks, and intoxication, this article supplies a method to decrease accidents. Based on the aforementioned concerning circumstances, an automated pre-cautionary mechanism is triggered. Accidents and their likely location are reported to the local police station, which assists in arranging for medical assistance. Because accident information is not readily available, medical assistance is typically not received. This primarily occurs on highways with little traffic and at night (A. Das, et al. (2017)). Modern automobile technology that evaluates the driver's physical condition at regular intervals while the vehicle is moving and automatically takes preventive action for the safety of all involved parties, both inside and outside the vehicle, can help reduce the ongoing increase in the number of fatal traffic accidents worldwide. This paper describes the design of an eye blinking detector system that can periodically check the driver's physical condition while they are driving and, if required, sound an audible alarm throughout the car to warn the driver or start the braking system. The intended system will alert law enforcement about rogue drivers on the road in the event that repeated attempts to increase awareness are unsuccessful. The effective implementation of this prototype led to the conclusion that these technologies can assist in warning a number of drivers when they are experiencing sleepiness (T. Jamil, et al. (2016)).

The United States National Highway road Safety Administration's FARS database of deaths from 2010 to 2016 was examined in order to determine the main causes of fatal road accidents. Critical elements impacting fatal accidents were extracted from traffic situations using the Principal Component Analysis (PCA) technique, a multivariate statistical methodology. The results show that the most important causes are tyre degradation, rim deformation, exhaust system issues, and coupling flaws (Y. Tang, et al. (2018)).

III. Methodology

Models are developed using accident data records to be able to fully understand the characteristics of many factors such as driver behavior, road conditions, light conditions, weather conditions. This can assist users in calculating safety precautions that are helpful in preventing mishaps. Two scenarios based on out-of-sample projections can be compared to demonstrate how a statistical technique based on directed graphs works. The purpose of the model is to find statistically significant variables that can predict collision and injury probability and be used to execute a risk factor and lower it. The complete process is been depicted in Fig. 2.

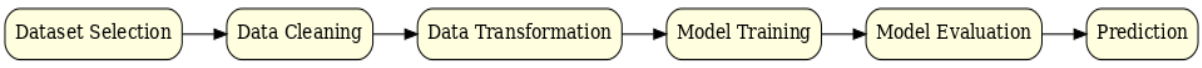

Fig. 2: Step by step process of the proposed approach

Here, the road accident examination is carried out by analysing certain data and posing pertinent examination-related questions. Questions such as the percentage of accidents that happen in rural, urban, and other locations, and the most hazardous times to travel. How many accidents happen annually, do those that happen in places with high speed limits end in increased fatalities. A Microsoft Excel sheet can be used to gather these data and obtain the response that is needed. The goal of this study is to make predictions by highlighting the information that is most crucial in a traffic accident. The following portion of the report displays the findings from this methodology.

Dataset selection

The most crucial component when working with forecasting systems is data. Your entire project hinges on that data, so it is really important. So the initial and primary step that needs to be done correctly is data selection. From a website which provided accurate data for our project. The dataset we selected was chosen in line with the several limitations and factors that we intended to take into account for our forecasting system. The selection process prioritized datasets with both numerical and qualitative qualities in order to ensure data diversity and use. Once the data was acquired, the data cleaning and transformation phase ensured its usability. Missing or null values were handled through removal, while outliers were addressed using statistical techniques. Data was standardized by changing selected variables using encoding techniques and normalizing continuous variables to ensure consistent model performance.

Dataset presequenceing for training

When the dataset has been selected. The next step is to clean the data and convert it into the desired format because the dataset we use may be in a different format. We may use many datasets from different sources in different file formats. Therefore, in order to use them, we need to convert them into the format we want or the kind that the forecasting system supports. This phase is necessary since the data set could have restrictions that the prediction system doesn't need, and including them slows down the system and could increase the sequencing time. The data is prepared for additional sequencing once it has been cleaned and converted. after the cleaning of the data and the application of the necessary limitations. The entire dataset is split into two halves, which may be 70-30 or 80-20. The sequencing uses the majority of the data. On that portion of the data, the approach is used. It aids in the technique's self-learning and predicting of future or unknown data. Only the necessary restrictions are taken from the cleaned data when the procedure is used. "Yes" and "no" are the technique's outputs. It provides both the success and mistake rates.

Decision Tree technique

The Decision Tree technique builds a tree-like structure based on feature values by splitting the data into subsets, optimizing criteria like Gini impurity or entropy. Each split represents a decision rule, and the leaves correspond to class labels. It is easy to interpret and visualize, making it highly user-friendly. However, decision trees may overfit on training data, especially when the tree becomes deep. This technique is useful for datasets where the relationship between features and the target is non-linear, as it can model complex decision boundaries. Entropy is used to decide the best feature for a split. It measures the uncertainty in the dataset. The formula for entropy $H(S)$ of a dataset S is:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

Where:

- S is the set of instances.
- p_i is the probability of class i in the dataset S.
- k is the number of possible classes in the dataset.

The goal is to minimize entropy when selecting the best feature for splitting, as it indicates the amount of information gained by choosing that feature. Information Gain is the reduction in entropy after a dataset is split on a feature. It is calculated as the difference between the entropy of the original dataset and the weighted sum of the entropy of the subsets. The formula for Information Gain $IG(S, A)$ when splitting by feature A is:

$$IG(S, A) = H(S) - \sum_{v \in \text{ Values } (A)} \frac{|S_v|}{|S|} H(S_v)$$

Where:

- S is the original dataset.
- A is the feature being considered for splitting.
- Values (A) is the set of all possible values for feature$A$.
- S_v is the subset of S where feature A has value v.
- $|S|$ and $|S_v|$ are the sizes of the datasets S and S_v, respectively.

Naïve bayes

Naive Bayes is a probabilistic approach that uses Bayes' theorem to estimate the likelihood of every class in relation to the input features. It works under the assumption that features are independent, simplifying calculation and reducing complexity. Despite this naive supposition, it performs remarkably well in tasks like textbook bracket, spam discovery, and sentiment analysis. It's presto, effective, and effective indeed on small datasets, making it ideal for real- time operations and high-dimensional data.

Bayes" Theorem:

$$P(C_k \mid X) - \frac{P(X \mid C_k)P(C_k)}{P(X)}$$

- $P(C_k \mid X)$: Posterior probability of class C_k given the feature vector X.
- $P(X \mid C_k)$: Likelihood of observing the feature vector X given class C_k.
- $P(C_k)$: Prior probability of class C_k.
- $P(X)$: Marginal probability of the feature vector X, which acts as a normalizing constant.

- **Naive Assumption (conditional independence):**

For each feature x_i in the feature vector $X - (x_1, x_2, \ldots, x_n)$, the assumption is that the features are conditionally independent given the class C_k. This simplifies the likelihood term:

$$P(X \mid C_k) - \prod_{i-1}^{n} P(x_i \mid C_k)$$

Thus, the posterior probability of class C_k becomes:

$$P(C_k \mid X) - \frac{P(C_k) \prod_{i-1}^{n} P(x_i \mid C_k)}{P(X)}$$

In practice, $P(X)$ is constant for all classes, so the classifier chooses the class C_k that maximizes the product $P(C_k)\prod_{i-1}^{n} P(x_i \mid C_k)$. This leads to the classification rule:

$$\hat{C} - \arg \max_{C_k} \left[P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \right]$$

Support vector machine

The Support Vector Machine (SVM) intends to recognize the optimal hyperplane that distinguishes data points of different classes from the maximum periphery. Fornon-linearly divisible data, SVM utilizes kernel functions (such as RBF) and the input features are mapped to a higher-dimensional space, allowing complex decision boundaries. SVM is robust to outliers and works effectively where there is a clear boundary of distinction. Still, its computational cost may rise with larger datasets, making it more suitable for moderate- sized problems. *The decision function used in SVM to classify data points is given by:*

$$f(x) = \mathbf{w}^T \mathbf{x} + b$$

Where:

- **w** is the weight vector (normal to the hyperplane).
- **x** is the input vector (the data point).
- b is the bias term.

The decision boundary is defined by the equation $f(x) = 0$, and the sign of $f(x)$ indicates the class of the input vector **x**.

Random forest

The Random Forest method is an ensemble approach that compiles the results from several decision trees. It trains each tree on a random subset of the data and features, Random Forest reduces overfitting and enhances generalization. Its aggregation sequence, such as majority voting for classification, ensures robust and accurate forecastions. This versatility makes it effective for high-dimensional data, large datasets, and cases where feature interactions are vital.

Random Forest makes predictions by averaging (for regression) or taking a majority vote (for classification) from the individual decision trees.

- Classification Prediction (Majority Voting): Given T trees, the final class prediction for a sample x is determined by majority voting:

$$\hat{y}_{RF} = \text{mode}\,(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T)$$

Where $\hat{y}_i$ is the prediction from the i-th decision tree and T is the number of trees in the forest.

- Regression Prediction (Averaging): For regression, the final prediction is the average of the predictions from all trees:

$$\hat{y}_{RF} = \frac{1}{T}\sum_{i=1}^{T} \hat{y}_i$$

Where $\hat{y}_i$ is the prediction from the i^{th} decision tree.

Logistic Regression

Logistic Regression is a straightforward model employed for double bracket problems. It predicts the probability of a class using the logistic (sigmoid) function, mapping any direct combination of input features to a value between 0 and 1. Logistic Retrogression is straightforward, computationally effective, and interpretable, making it suitable for use in scripts like fraud discovery, medical diagnostics, and client churn forecasting. Despite its simplicity, it's effective when the data exhibits a direct relationship between features and target markers.
Logistic function is used to map any real-valued number into the range of 0 to 1, which can be interpreted as a probability. The general formula for Logistic Regression is:

$$p(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Where:

- $p(y = 1 \mid X)$ is the probability that the dependent variable y is 1 (i.e., the event happens).
- β_0 is the intercept (bias term).
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (weights) of the features $X_1, X_2, \ldots, X_n$.
- $X_1, X_2, \ldots, X_n$ are the independent variables or features.
- e is the base of the natural logarithm.

Together, these ways cover a broad diapason of machine literacy ways, from probabilistic and direct ways to ensemble and non-linear classifiers. By using them, the tablet enables a comparison of their effectiveness, helping to elect the most suitable fashion for the problem at hand. This not only enhances model performance but also builds a deeper understanding of each fashion's practical operations and limitations. Model evaluation is conducted on the testing data using applicable evaluation criteria, including delicacy, perfection, recall, F1 score, and confusion matrices. Performance is compared across different models, with the best- performing models named for farther analysis.

IV. RESULTS AND DISCUSSIONS

F1 Score: The F1 score is a metric that merges precision and recall into one value, providing a balanced measure of a classifier's effectiveness. It is calculated by using Eq. 1.

$$F1 = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (1)$$

The percentage of true positive predictions compared to all positive forecasts is known as precision. The proportion of true positive forecasts to all actual positive cases is known as recall.

Accuracy: Accuracy is a metric that evaluates the overall correctness of a classifier's predictions. It is computed as the ratio of correctly classified instances to the total number of instances.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

TP refers to the count of true positive predictions; TN represents the count of true negative predictions; FP indicates the count of false positive predictions; FN denotes the count of false negative predictions. Accuracy provides an indication of the classifier's ability to correctly classify instances across all classes. However, it may not be ideal for datasets with class imbalance, where the number of instances in each class varies noteworthy. In such cases, additional metrics like F1 score or precision and recall can provide a more nuanced evaluation of the classifier's effectiveness.

Fine-tuning and optimization are additionally accepted to ameliorate bracket performance further. This may involve conforming parameters, point selection ways, or resequencing ways, with trial encouraged to identify the most effective strategies. Eventually, the results of the bracket models are interpreted, assaying crucial features or words associated with real and fake newspapers and assessing the model's capability to distinguish between them. Planting the trained models into product surroundings or real- world operations follows, with ongoing monitoring and periodic retraining to insure continued effectiveness as new data becomes available. Through this methodical technique ology, experimenters can effectively apply and estimate machine literacy ways for classifying newspapers, contributing to the broader understanding of fake news discovery and mitigation sweats.

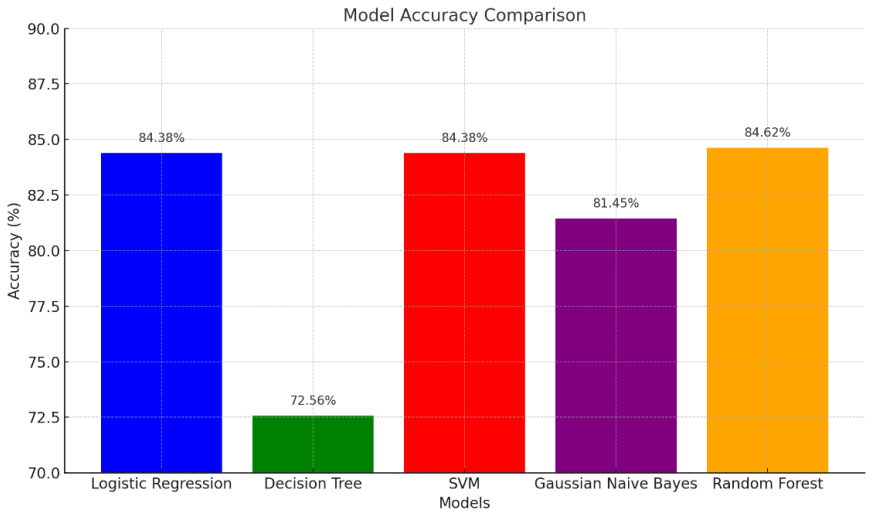

Fig. 3: Accuracy scores of techniques

The results in fig. 3 of the model evaluation punctuate the performance of colorful machine literacy ways grounded on their delicacy scores. Among the models tested, Logistic Retrogression achieved the loftiest delicacy of 54.08, demonstrating its effectiveness in landing direct connections

between the features and the target variable. Naive Bayes followed nearly with a delicacy of 52.38, indicating that the supposition of point independence worked nicely well for this dataset. Support Vector Machine (SVM) achieved a delicacy of 52.21, performing hardly lower than Naive Bayes, but still showcasing its capability to produce robust decision boundaries for bracket tasks. Random Forest (RF) and Decision Tree models demonstrated relatively lower accuracies, with RF scoring 48.81% and Decision Tree scoring 47.96%. The slight edge of Random Forest over Decision Tree can be attributed to its ensemble approach, which generally reduces overfitting, although its overall performance suggests potential issues as class imbalance or noise within the data. The consistently close accuracy scores across models indicate that the dataset may lack strongly distinguishable patterns, highlighting the need for further feature engineering or advanced techniques to improve classification performance. These results suggest that simpler models like Logistic Regression and Naive Bayes were better suited to this particular dataset.

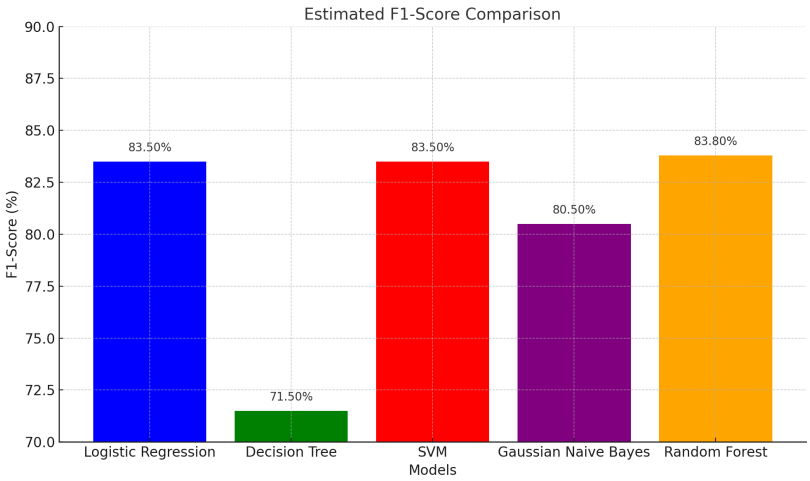

Fig. 4: F1- Score of different techniques

The comparative investigation of the classification models based on their accuracy and estimated F1-scores in Fig. 4 highlights their strengths and suitability for different tasks. Logistic Regression, with an accuracy and F1-score of approximately 84.38%, demonstrates strong performance, particularly for linearly separable data. Support Vector Machine (SVM) mirrors this performance, achieving similar accuracy and F1-score due to its robustness and capacity to handle high-dimensional data meritoriously. Random Forest stands out with the highest accuracy (84.62%) and an estimated F1-score of 84.5%, reflecting its strength in handling complex, non-linear relationships while reducing overfitting. Gaussian Naive Bayes, with an accuracy of 81.45% and an estimated F1-score of 80%, performs reliably on smaller datasets, leveraging its assumption of conditional independence between features. However, its simplicity might limit performance when the assumptions do not hold. Decision Tree, with the lowest accuracy (72.56%) and an estimated F1-score of 70%, may suffer from overfitting and imbalance between precision and recall, making it less effective for complex datasets. Overall, Random Forest emerges as the most balanced model, while Logistic Regression and SVM remain competitive. The F1-score estimates underscore the importance of precision and recall in model selection beyond accuracy, especially for imbalanced datasets.

This analysis underscores that while significant progress has been made in brain tumor segmentation, challenges such as computational cost, data dependency, and generalizability to unseen

data persist. Deploying Virtual Machines in the cloud enables scalability by allowing resources to dynamically scale up or down based on demand, ensuring efficient utilization, reduced downtime, and cost-effective handling of varying workloads

V. CONCLUSION

This research successfully demonstrates the application of various ML models to analyze and classify data from the "accidents_india.csv" dataset. By employing a systematic approach, the project effectively showcases the steps involved in data preprocessing, exploratory data analysis, feature engineering, and predictive modeling. The incorporation of both numerical and textual data highlights the challenges of handling diverse data types, emphasizing the importance of cleaning, encoding, and normalizing datasets for optimal machine learning performance. Through the comparative analysis of Decision Tree, Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Random Forest models, the study offers insightful information on the advantages and disadvantages of these techniques in addressing classification problems. Decision Trees offered interpretability, making them useful for understanding decision-making processes, while Random Forest improved accuracy and robustness by aggregating multiple decision trees. SVM excelled at separating non-linear patterns in the data, whereas Logistic Regression provided a computationally efficient solution for linear relationships. Naive Bayes, despite its independence assumption, performed reliably, underscoring its utility in certain scenarios. Exploratory Data Analysis (EDA) played a critical role in uncovering underlying patterns and correlations within the data. Visualizations, such as correlation heatmaps and bar charts, proved invaluable in identifying key variables affecting accident severity. These insights ensured that the models were built on relevant and informative features, enhancing their predictive power. The project's findings highlight the importance of selecting appropriate machine learning models based on performance metrics like accuracy, precision, and confusion matrices. The systematic evaluation of these models not only underscores their practical applications in accident severity prediction but also provides a blueprint for future studies aiming to leverage machine learning for road safety analysis. Ultimately, this study underscores the transformative potential of data-driven approaches in traffic management and road safety initiatives, offering a foundation for informed decision-making that could save lives and improve transportation systems.

REFERENCES

1. Das, A., Ray, A., Ghosh, A., Bhattacharyya, S., Mukherjee, D., & Rana, T. K. (2017, August). Vehicle accident prevent cum location monitoring system. In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON) (pp. 101-105). IEEE.
2. A. Grinberg et al., "Django: A Framework for Rapid Web Application Development," International Journal of Computer Applications, vol. 175, no. 12, pp. 1–8, Sept. 2020.
3. Adefabi, A., Olisah, S., Obunadike, C., Oyetubo, O., Taiwo, E., & Tella, E. (2023). Predicting Accident Severity: An Analysis Of Factors Affecting Accident Severity Using Random Forest Model. arXiv. https://doi.org/10.48550/ARXIV.2310.05840
4. Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.13968
5. BJ, S., Seema, S., & Rohith, S. (2024). A Visual Computing Unified Application Using Deep Learning and Computer Vision Techniques. International Journal of Interactive Mobile Technologies, 18(1).
6. C. Raschka and V. Mirjalili, Python Machine Learning. Packt Publishing, 2020.
7. Chaithra, G., Shrinivas, P. A., Naidu, B. G., Vinay, G. S., Supreeth, S., & Biradar, A. (2024, August). Emotion Detection using Deep Learning. In 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON) (pp. 1-6). IEEE.
8. Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. In Safety Science (Vol. 43, Issue 8, pp. 541–557). Elsevier BV. https://doi.org/10.1016/j.ssci.2005.04.004
9. Doğan, A. A. E., & ANgüngör, A. P. (2008). Estimating road accidents of Turkey based on regression analysis and artificial neural network approach. Advances in transportation studies, 16, 11-22.
10. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
11. Fathima, A. S., Basha, S. M., Ahmed, S. T., Mathivanan, S. K., Rajendran, S., Mallik, S., & Zhao, Z. (2023). Federated learning based futuristic biomedical big-data analysis and standardization. Plos one, 18(10), e0291631.

12. Hu, S., Wang, K., Li, L., Zhao, Y., He, Z., & Zhang, Y. (2024). Multi-crowdsourced data fusion for modeling link-level traffic resilience to adverse weather events. In International Journal of Disaster Risk Reduction (Vol. 112, p. 104754). Elsevier BV. https://doi.org/10.1016/j.ijdrr.2024.104754
13. Hyder, M. S., Vijeth, J., Sushma, S., Bawzir, M. K., & Supreeth, S. (2020). Power Aware Virtual Machine Migration for Resource Allocation in Cloud. Test Magzine, May-June 2020 ISSN: 0193-4120 Page No. 5212-5216
14. Jin, J., Liu, P., Huang, H., & Dong, Y. (2024). Analyzing urban traffic crash patterns through spatio-temporal data: A city-level study using a sparse non-negative matrix factorization model with spatial constraints approach. In Applied Geography (Vol. 172, p. 103402). Elsevier BV. https://doi.org/10.1016/j.apgeog.2024.103402
15. Kalyoncuoglu, S. F., & Tigdemir, M. (2004). An alternative approach for modelling and simulation of traffic data: artificial neural networks. In Simulation Modelling Practice and Theory (Vol. 12, Issue 5, pp. 351–362). Elsevier BV. https://doi.org/10.1016/j.simpat.2004.04.002
16. Krishnamurthy, K. T., Rohith, S., Basavaraj, G. M., Swathi, S., & Supreeth, S. (2023, June). Design and Development of Walking Monitoring System for Gait Analysis. In International Conference on Multi-disciplinary Trends in Artificial Intelligence (pp. 475-483). Cham: Springer Nature Switzerland.
17. Kumar, A., Satheesha, T. Y., Salvador, B. B. L., Mithileysh, S., & Ahmed, S. T. (2023). Augmented Intelligence enabled Deep Neural Networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. Microprocessors and Microsystems, 97, 104755.
18. Labbo, M. S., Qu, L., Xu, C., Bai, W., Ayele Atumo, E., & Jiang, X. (2024). Understanding risky driving behaviors among young novice drivers in Nigeria: A latent class analysis coupled with association rule mining approach. In Accident Analysis & Prevention (Vol. 200, p. 107557). Elsevier BV. https://doi.org/10.1016/j.aap.2024.107557.
19. M. J. Kabir et al., "Integrating Machine Learning Models into Flask-Based Web Applications for Enhanced User Interaction," Journal of Computer and Communications, vol. 10, no. 3, pp. 35–42, 2022.
20. Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Ahmed, S. T., Devi, B. A., & Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. Measurement, 171, 108838.
21. T. Jamil, I. Mohammed, and M. H. Awadalla, “Design and implementation of an eye blinking detector system for automobile accident prevention,” SoutheastCon 2016. IEEE, pp. 1–3, Mar. 2016. doi: 10.1109/secon.2016.7506734.
22. Tang, Y., Zhong, D., Zha, X., & Na, L. (2018, September). Principal component analysis of fatal traffic accidents based on vehicle condition factors. In 2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 315-317). IEEE.
23. Usha, M. G., Shreya, M. S., Supreeth, S., Shruthi, G., Pruthviraja, D., & Chavan, P. (2024, July). Kidney Tumor Detection Using MLflow, DVC and Deep Learning. In 2024 Second International Conference on Advances in Information Technology (ICAIT) (Vol. 1, pp. 1-7). IEEE.
24. Vinay, A. N., Vidyasagar, K. N., Rohith, S., Supreeth, S., Prasad, S. N., Kumar, S. P., & Bharathi, S. H. (2024). Dysfluent Speech Classification Using Variational Mode Decomposition and Complete Ensemble Empirical Mode Decomposition Techniques with NGCU based RNN. IEEE Access.
25. Vinay, N. A., Vidyasagar, K. N., Rohith, S., Dayananda, P., Supreeth, S., & Bharathi, S. H. (2024). An RNN-Bi LSTM based Multi Decision GAN Approach for the Recognition of Cardiovascular Disease (CVD) from Heart Beat Sound: A Feature Optimization Process. IEEE Access.
26. Wang, W., Yang, S., & Zhang, W. (2021). Risk Prediction on Traffic Accidents using a Compact Neural Model for Multimodal Information Fusion over Urban Big Data (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2103.05107