Comparing Random Forest and Logistic Regression for Accurate Wrong-Lane Accident Detection

Pradyumna B[1], S.Padmakala[2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

[2]*Project Guide, Corresponding Author,Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

ABSTRACT

Aim: The main objective of the work is to perform detection of wrong-lane accidents using Logistic Regression (LR) and compare accuracy with Random Forest (RF) algorithm.

Materials and Methods: A road accident dataset of 1834 records is used for Logistic Regression. For the detection of wrong-lane accidents, a machine learning technique that compares Logistic Regression and Random Forest was proposed and developed. G power was used to determine the sample size, which came out to 21 per group. The sample size was determined using clinical analysis, with alpha and beta values of 0.05 and 0.5, 95% confidence, and pre-test power of 80%. The accuracy of the Wrong-lane accidents were evaluated and recorded. Results: The detection of wrong-lane accidents with Logistic Regression had the highest accuracy (90%) and the minimum mean error when compared to Random Forest (89.77%) (significance value p=0.02; p=0.05). Conclusion: this research shows that when it comes to wrong-lane accident detection, Logistic Regression Algorithm outperforms Random Forest.

Keywords: Logistic Regression, Random Forest, Accident Detection, Wrong-Lane, Data Mining, Classification, Novel Penalty Based Method.

INTRODUCTION

The wrong-lane road accidents are a completely common difficulty in densely populated countries. The capability of the roadway isn't always enough for the developing wide variety of motors and hence imbalance is created (Clarke, Forsyth, and Wright 1998). Many people die and government property is damaged every year due to wrong-lane riding accidents, which costs the countries a lot of money (Zhao et al. 2021). The drivers disregard traffic rules and take advantage of crimson traffic signals to ride on the wrong side of the road (K.m. and Umamaheswari 2020). It will increase visitors on one facet and hamper visitors greatly. It additionally will increase the opportunity of head-on collision in several instances (Alkhorshid et al. 2016). About 355 humans die each 12 months because of the crashes in wrong-manner riding withinside the United States. So, it's very essential to prevent drivers from riding on the incorrect facet. To make certain of it, people who don't comply with visitors' policies want to discover and strict regulation has to be applied (M et al. 2021).

Nearly 10,582 articles in science direct and 6,363 articles in IEEE Xplore on topics similar to this were published in the last five years(Vasavi 2016). Classification algorithms and novel penalty-based methods are widely used to improve accident detection. Based on a proposed Novel Penalty Based Method, the system has great potential for accurately forecasting the detection of accidents, as demonstrated by its classification accuracy. For example, Chen and Chen (2020) proposed a criteria-based spatial grouping of applications as a New Penalty Based Method with Logistic Regression performed by other models that achieved 79.6 percent to predict crashes. This paper proposed a feature selection algorithm with a classifier similar to Novel Penalty Based Method for the design of a high-level intelligent system to predict accidents (Oza and D. Y Patil School of Engineering Lohegaon 2020).

The research gap that is identified from the literature survey is that classification models

adopting Random Forest require lots of training data and don't encode the position and orientation of the object into their predictions. And also, the existing approaches have poor accuracy. The study's goal is to use Logistic Regression to detect vehicles traveling in the wrong lane and improve classification accuracy by comparing its results with those of Random Forest.

MATERIALS AND METHODS

This research work was carried out in the Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS. This study requires an intel i5 processor, a 500 GB HDD, 8 GB RAM, and Jupyter as the hardware and software configurations. An accident severity dataset of 1834 records was used in the study. Two groups are evaluated to see which one is the most accurate at spotting wrong-way drivers. Ten iterations were carried out on each group to ensure the best possible results. The dataset used in this study was obtained from kaggle (Pote 2020) and represents accident severity.

Random Forest (RF) - Group 1

Input: Accident dataset
Output: Accuracy
Step 1: The dataset must be loaded and read.
Step 2: Pick features at random from the dataset.
Step 3: The RF classifier criteria can be generated as a parameter.
Step 4: The parameter value "Gini" was employed.
Step 5: Make predictions for each sample by constructing a decision tree using RF classifiers.
Step 6: Voting was conducted for each predicted outcome
Step 7: The predictions with the most votes were selected as the final outcome.
The sklearn ensemble library's Random Classifier class is used in this investigation. Citrian is a required parameter. The parameter value of "Gini" is used. Randomly, the dataset is divided into training (80%) and testing (20%). For each sample, the decision trees were collected and analyzed to predict the outcome. For each predicted outcome, a vote was held and the most popular result was chosen as the final outcome. a novel tree-specific random forest classifier is used in the algorithm (NTSRF).

Logistic Regression (LR)-Group 2

Input: Accident dataset
Output: Accuracy
Step 1: The dataset must be loaded and read.
Step 2: Pick features at random from the dataset.
Step 3: The RF classifier criteria can be generated as a parameter.
Step 4: The parameter value l2 was used.
Step 5: Vary dependent and independent variables to analyze the dataset.
Step 6: LR predicts a categorical outcome.
Step 7: Finally, the log function is used to predict the possibility of an event.
Uses the sklearn.linear model library's logistic regression class. A penalty-based approach is used in this new method. The penalty is used as a metric. The parameter value is "LR." The data set is randomly divided into training dataset (80%) and test dataset(20%). Randomly select samples and run analyses on them, changing the dependent and independent variables as necessary. Finally, the log function can be used to provide the possibility of an event.
A training and a test set were initially separated in the data set. Afterwards, the algorithm is put through its paces on the training and test sets. Depending on the size of the test set, the training and testing sets are changed 10 times. It is shown in Table 1 that the accuracy of the RF and LR is compared for 10 iterations. The following formula can be used to calculate the various analysis parameters:
Accuracy: - It is the number of instances that were correctly classified as shown in following Equation 1.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

Statistical Analysis

The SPSS statistical software was used for statistical analysis in the research. Statistical analyses of the experimental data, including independent sample t-tests and graph construction for two groups and two parameters under investigation were carried out (A. Das, Khan, and Ahmed 2020). Speed Limit and Junction Control are two variables that are independent variables. Accuracy and Prediction are the dependent variables.

RESULTS

The proposed Novel Penalty Based Logistic Regression Method and the standard Random Forest were run at a time for performing detection of wrong-lane accidents. Table 1 displays the classification accuracy achieved by RF algorithm and LR models after various iterations of testing. Two groups are shown in Table 2 based on different parameters. For both RF and LR, accuracy is calculated. Analyses of two groups have shown that the LR is more accurate (90.0%) than the RF. Table 3 shows the statistical analysis of LR and RF with different sets of test data. Fig. 1 shows the weekday of vehicle collisions. Fig. 2 shows the policing area and collision severity. Fig. 3 shows the day and month of collision. Fig. 4 shows the speed limit and junction detail. Fig. 5 shows the collision severity. Fig. 6 shows the mean accuracy of the LR and RF algorithms in comparison to each other. The LR model appears to have a higher overall accuracy than the RF model on average. The LR algorithm outperforms the RF algorithm in terms of performance. The two groups do not differ significantly. Because of this, LR is preferred over RF. An independent study shows that LR's mean accuracy is higher (90%). LR's mean error is a little lower than that of RF's.

DISCUSSION

The work shows that logistic regression is better than RF at detection of wrong-lane accidents in terms of accuracy. From the experimental results performed in Jupyter, the accuracy of LR is 90.0%, while RF provides the accuracy of 89.77%. This shows that LR is better than RF. The different parameters such as the TP and FP rates are also compared. In terms of accuracy (90%), the proposed logistic regression classifier outperforms the RF algorithm, as shown by the SPSS plot.

Accuracy is the most critical factor in detecting wrong-lane accidents. A diagnostic system based on machine learning for the detection of wrong-lane driving was proposed in a study (Zhu, Wang, and Li 2021). Seven classifier performance metrics, including classification accuracy, specificity, sensitivity, Matthews correlation coefficient, and delay execution were used in the study. Popular machine learning algorithms, three feature selection algorithms were also used in this study.

According to the above, the integrated method should be chosen if overall prediction performance is the primary concern, in which the RF models with only a few significant variables identified by LR or important variables identified by the tree can be entered to achieve greater precision (Gazder, Ahmed, and Shahid 2021). However, if accuracy in large-scale prediction failures is a primary concern, the integrated method may be preferable because it includes RF models with only a small number of statistically significant variables, as well as major random forest variables or limited number of significant LR variables. The size of the training and testing datasets affects the accuracy of the LR classification algorithm. RF appears to have a higher level of accuracy in our study. Our proposed work, on the other hand, appears to have a higher average error, which should be reduced.

Research findings are more robust in terms of experimental and statistical evaluation, although the study has several flaws. It is impossible to get a better result with larger data sets using accuracy evaluation. LR's errors appear to be higher than RF's, at least on average. It would be ideal if the average mistake could be decreased somewhat. As a follow-up project, it is possible to further refine this work by including optimization

algorithm techniques to reduce the mean error. Using feature selection techniques prior to classification can improve the accuracy of a classifier. Due to the use of data mining techniques, computation time is reduced and classification accuracy is improved.

CONCLUSION

The results show that the accuracy is higher in the proposed Logistic Regression than Random Forest. When compared to Random Forest(89.77%), the Proposed Logistic Regression had a better accuracy (90.0%).

DECLARATIONS

Conflicts of interests

No conflicts of interest in this manuscript.

Author Contributions

Images were collected, algorithms were developed and image analysis and manuscript writing were all done by PC.. All aspects of the manuscript's conception, validation, and critical review were carried out by author AG.

Acknowledgements

We thank Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (formerly Saveetha University) for providing facilities and continued assistance to complete this study.

Funding: We thank the following organisations for providing financial support that enabled us to complete the study.

1. Mass Datta Developers , Chennai, India.
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Alkhorshid, Yasamin, Kamelia Aryafar, Sven Bauer, and Gerd Wanielik. 2016. “Road Detection through Supervised Classification.” *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2016.0144.
2. Chen, Mu-Ming, and Mu-Chen Chen. 2020. “Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest.” *Information*. https://doi.org/10.3390/info11050270.
3. Clarke, David D., Richard Forsyth, and Richard Wright. 1998. “Machine Learning in Road Accident Research: Decision Trees Describing Road Accidents during Cross-Flow Turns.” *Ergonomics*. https://doi.org/10.1080/001401398186603.
4. Das, Anik, Md Nasim Khan, and Mohamed M. Ahmed. 2020. “Detecting Lane Change Maneuvers Using SHRP2 Naturalistic Driving Data: A Comparative Study Machine Learning Techniques.” *Accident; Analysis and Prevention* 142 (July): 105578.
5. Das, Subasish, Raul Avelar, Karen Dixon, and Xiaoduan Sun. 2018. “Investigation on the Wrong Way Driving Crash Patterns Using Multiple Correspondence Analysis.” *Accident; Analysis and Prevention* 111 (February): 43–55.
6. Gazder, Uneb, Ashar Ahmed, and Umaira Shahid. 2021. “Predicting Severity of Accidents in Malaysia By Ordinal Logistic Regression Models.” *International Journal of Traffic and Transportation Management*. https://doi.org/10.5383/jttm.03.01.002.
7. K.m., Umamaheswari, and K. M. Umamaheswari. 2020. “Road Accident Perusal Using Machine Learning Algorithms.” *International Journal of Psychosocial Rehabilitation*. https://doi.org/10.37200/ijpr/v24i5/pr201839.
8. Koh, Mirae, Masahito Hitosugi, Eiko Kagesawa, Takahiro Narikawa, and Kohei Takashima. 2021. “Factors Influencing Fatalities or Severe Injuries to Pedestrians Lying on the Road in Japan: Nationwide Police Database Study.” *Healthcare (Basel, Switzerland)* 9 (11). https://doi.org/10.3390/healthcare9111433.
9. Komol, Md Mostafizur Rahman, Md Mahmudul Hasan, Mohammed Elhenawy, Shamsunnahar Yasmin, Mahmoud Masoud, and Andry Rakotonirainy. 2021. “Crash Severity Analysis of Vulnerable Road Users Using Machine Learning.” *PloS One* 16 (8): e0255828.
10. M, Bharath Kumar, Kumar M. Bharath, Abdhul Basit, M. B. Kiruba, R. Giridharan, and S. M. Keerthana. 2021. “Road Accident Detection Using Machine Learning.” *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*. https://doi.org/10.1109/icscan53069.2021.9526546.

TABLES AND FIGURES

Table 1. Accuracy achieved during the evaluation of RF algorithm and LR models for classification with different iterations.

Iteration No.	ACCURACY	
	RF	LR
1.	89.77	90.00
2.	88.10	89.20
3.	89.50	91.25
4.	90.20	90.30
5.	88.60	89.40
6.	89.40	90.50
7.	88.30	91.08
8.	90.40	89.80
9.	88.32	90.20
10.	89.60	90.15

Table 2. Experimental analysis in Jupyter for Accuracy for RF and LR. LR provides better Accuracy (90.0%) than RF.

MODEL	ACCURACY (%)
LR	90.00
RF	89.77

Table 3. Statistical Analysis of Mean, Standard Deviation and Standard Error Mean and Accuracy of LR and RF algorithms. The standard deviation of RF is (0.830) and for LR is (0.65132). The significance is (0.02). There is a statistical difference in accuracy values between the data mining algorithms. LR had the higher mean accuracy (90.1880%) and RF had mean accuracy of (89.2320%).

ACCURACY (Algorithm)	N	Mean	sig	Std. Deviation	Std.Error Mean	95% Confidence Interval for Mean	
						Lower	Upper
RF	10	89.2320	.02	.83037	.26259	-1.67013	-.26787
LR	10	90.1880		.65132	.20597	-1.67300	-.26500

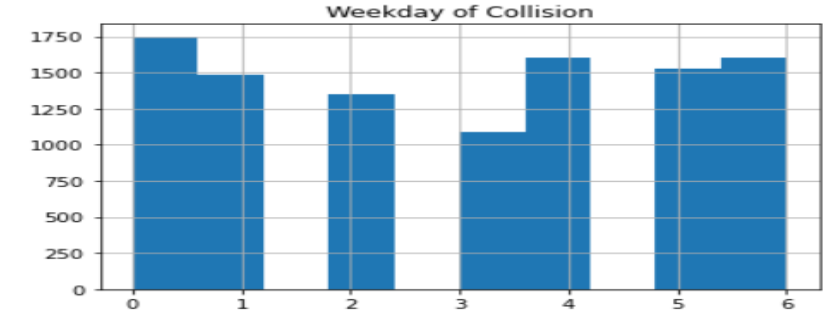
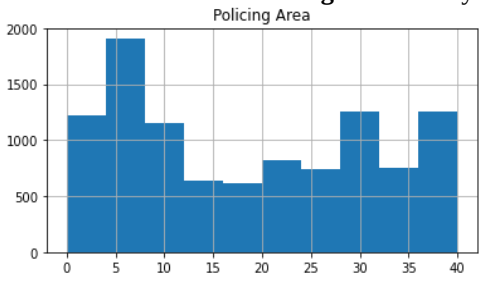


Fig. 1. Weekday of Collision of Vehicles



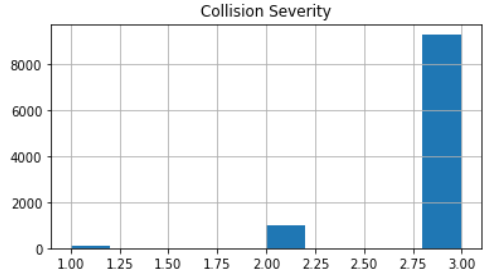
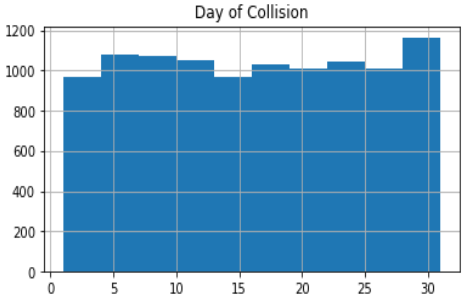


Fig. 2. Policing Area and Collision Severity



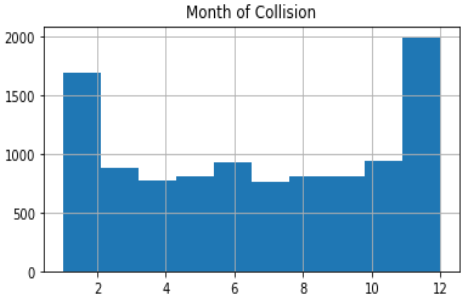
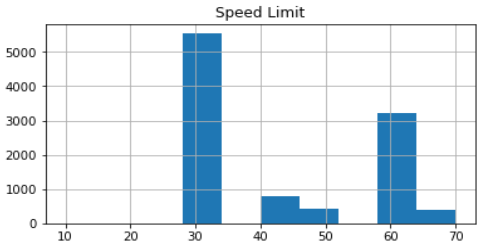


Fig. 3. Day and Month of Collision



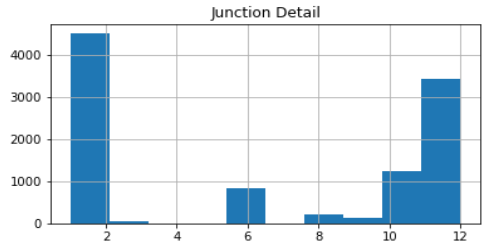


Fig. 4. Speed Limit and Junction Detail

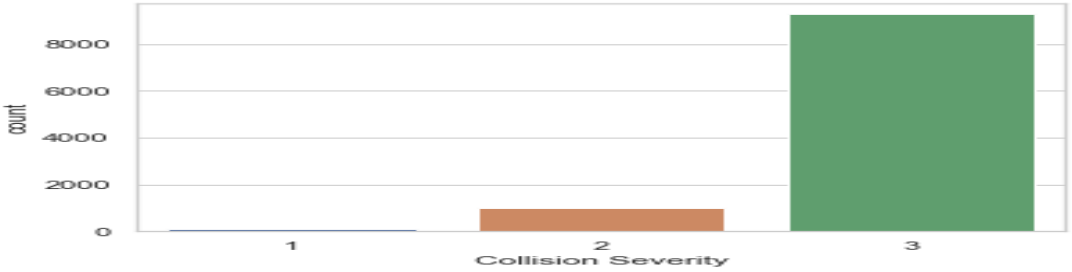


Fig. 5. Collision Severity

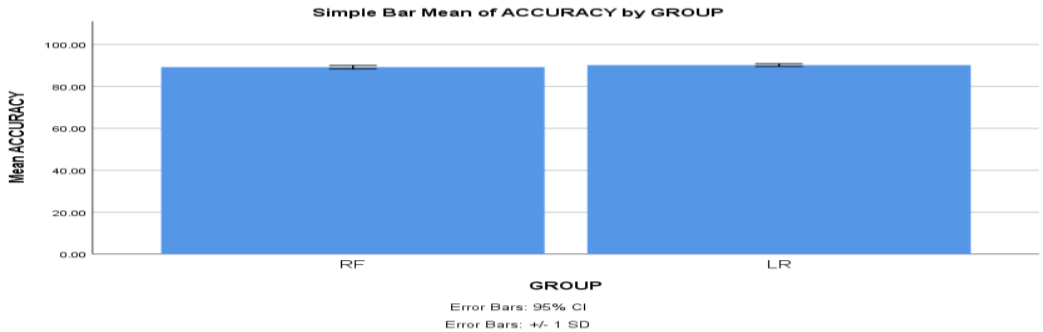


Fig. 6. Comparison of mean accuracy of RF(89.77%) and LR(90.00%) algorithms. LR appears to produce more consistent results with higher accuracy. X-axis: RF vs LR. Y-axis: Mean Accuracy of ± 1 SD.