Severity Classification of Accidents involving Powered Two Wheelers

Divya Gummadi
Department of Computer Science
Kent State University
Kent, OH
dgummadi@kent.edu

Prathyusha Anireddy
Department of Computer Science
Kent State University
Kent, OH
paniredd@kent.edu

Vijay Kumar Reddy Almalachervu
Department of Computer Science
Kent State University
Kent, OH
valmalac@kent.edu

Yugendra Kolluru
Department of Computer Science
Kent State University
Kent, OH
ykolluru@kent.edu

***Abstract*—One of the biggest obstacles to the development of smart cities and transportation systems is road traffic safety. Although there are many ways to minimize the number of fatalities and serious accidents that occur every day on our roads, this reduction is less pronounced than anticipated, and new strategies and intelligent We require systems. The Indian Commission's emergency call program seeks to offer quick assistance to drivers through the use of a specific emergency number. By utilizing machine learning methods based on features that might be fairly obtained at the time of the accident, we investigate the challenge of categorizing the severity of accidents involving powered two-wheelers in this work. The most crucial elements that distinguish accident severity can be found after a thorough analysis of the set of features. The system we create does roughly Using just a handful of eleven features, a huge, publicly accessible corpus may achieve 90**

***Index Terms*—Road accidents, Road safety, PTW techniques, and SVM.**

I. Introduction

More than 1.5 lakh people died on Indian roads in 2016. In 2014, the Indian Union announced its target to halve the number of fatalities by 2020. Despite the fact that the number of fatalities has decreased over the last decade, the EU is likely to fall short of its goal. The development of new safety functionalities and devices to further reduce these numbers, and hence to help the EU in achieving its goal, is thus an urgent need. Since March 2018, all the newly produced passenger cars must be equipped with an emergency Call (eCall) system. The Indian wants to expand the eCall Act to cover additional vehicle types because they are certain that this new method will result in a considerable decrease in traffic fatalities. To prepare the eCall infrastructure, the Indian government supported the I HeERO initiative, which included vehicle manufacturers, suppliers, Public Safety Answering Points (PSAPs), and research organizations. Deriving the unique requirements of an eCall system for buses, coaches, large cargo trucks, including dangerous commodities, was one of the project's main objectives

Thank you for Xiang Lian for giving us this opportunity

II. Related Work

Although largely for incidents involving cars, the subject of determining characteristics that affect the severity of accidents has been extensively studied in the literature. Several studies used logistic regression to assess how each variable affected the seriousness of accidents. They evaluated decision trees, Bayesian networks, and support vector machines for this task and found that their accuracy rates (70–80 percent) were comparable. One approach exploits data from accelerometers, gyroscopes and vehicle speed sensors to classify five main driving behaviors (turn right, turn left, roundabout, straight, and stop). The authors report an accuracy slightly above 60 percent, using real traffic data collected between 2011 and 2016. Results were consistent with previous studies, although they suffered from a high risk of false positives. The method is not suitable for eCall which requires balancing the sensitivity and specificity of the system. An improved nonparametric regression (INPR) algorithm was presented to forecast incidents based on traffic data. The authors found that accuracy in the prediction of accident severity was higher when using neural networks compared to ordered probit models. Neural networks provided the best results, with 87 percent accuracy. A review of the literature shows that the problem of predicting accident severity for PTWs is still largely underexplored, in particular when considering neural networks and other approaches. The authors conclude that social data can be noisy and unreliable so that, this type of analysis, should not replace but rather be seen as complementary.

III. Limitation of Existing System

A major limitation associated with the aforementioned machine learning methods is that they generally work like a black-box, which does not directly report the correlation between crash injury severity and explanatory variables. These

models are not used to predict the type of accident. There is a high probability of incurring in false positives.

IV. Data Collection

It is necessary to gather a set of examples where the accident is described in terms of some features (i.e., characteristics of the accident, vehicles, and drivers), and the severity level is known, in order to build an intelligent system capable of automatically sending a warning message by correctly detecting. We concentrate on information that can be reasonably gathered at the time of an accident using a set of sensors or other easily installed devices on the vehicle

1) Original Data: The Fatality Analysis Reporting System (FARS) is the country's official database for fatal injuries caused by automobile accidents. The Police record all types of crashes, from minor to fatal, to the NASS GES. Since 2009, the National Highway Traffic Safety Administration has combined the two datasets (NHTSA). On the NHTSA website2, the raw data are provided by year and set of variables. The variable body type was used to identify and query motorcycle accidents, whereas the variable person type indicated whether the observation pertained to the driver or the passenger. In this study, we exclusively take into account observations involving motorbike riders. The initial database had 74 characteristics that were either associated with the accident's person (person), accident (accident), or vehicle (vehicle) participants (vehicle). Since the database included all different kinds of vehicle accidents, not all of them were significant for motorcycle accidents.

2) Excluded Variables: The following variables were excluded from the dataset because they had over 98 percent of their values missing. The database also contained technical data that the Police had documented, such as whether an alcohol test had been administered. At the time of the incident, it would be quite challenging to know these elements.

3) eCall Relevance: An accident is eCall relevant if the driver involved was taken to the hospital or if the varied severity was assessed as probable, serious, or fatal harm. Only 9 observations had missing hospital and severity fields in the eCall relevance variable, so those observations were eliminated from the dataset.Variables Describing the Moment and the Dynamics of theAccident: We took into account the day of the week, hour, light conditions, and weather conditions while characterizing the variables describing the accident's exact moment. We took into account the driver's age and sex. The motorcycle's brand is indicated by the variable make.We began by taking into consideration the variables harmev (first harmful event) and mharm for the variables characterizing the dynamics (most harmful event). They each include details on the first accident event that caused damage or injuries, as well as the occurrence that caused the most serious injuries. Harmev and mharm are the only two variables that can be encoded as measurable at the time of an accident. Harmev had 52 percent of missing observations, and therefore was included in the final dataset. The original values of this variable have been encoded into seven more categories, so that the information can be reasonably considered as measurable.

4) Variables Related to Location: The location of the car when it crashed (crash) has been encoded, as seen in Table. The proximity of the accident to a junction (reljnc2) and the various types of intersections where the crash happened (typ int) are related by a set of variables.

5) Variables Related to Road Type and Condition: Three additional factors also have anything to do with the state of the roads. Variable vtrafway specifies whether or not the lanes were divided, for example, by a physical barrier, while variable vnum lan provides the number of lanes of the road where the crash has occurred. Once more, GPS technology might be used to obtain this data. Last but not least, the variable vsurcond specifies whether or not the accident was caused by poor road surface conditions (rain, snow, etc.).

6) Speed and Helmet Use: The severity of an accident is often correlated with travel speed. However, variable trav sp, which recorded the real speed of the car prior to the collision, had 59 percent of its values missing. Therefore, we made the decision to create two distinct databases. While a second dataset (Dataset B) comprises fewer samples and includes variable trav sp, the first dataset (Dataset A) does not include variable speed among the features. The database also contains details on the worn helmet (rest use). But the only thing that makes an eCall relevant is the fact that the motorist was taken to the hospital without any details regarding the nature of their injuries or where on their body was struck. This observation led us to ignore this variable in our study, along with the fact that almost 60 percent of the drivers in the database wore either no helmet or one that was non-compliant (or an unidentified type)3.

7) Final Dataset: Two distinct datasets are created, as previously described. Dataset A consists of 16,463 instances that are each defined by 25 variables, with no missing values and no knowledge of the vehicle's speed. 7,736 eCall relevant cases are present in Dataset A overall (positive class), while the other 8,727 occurrences are classified as not relevant (negative class). Instead, the variable trav sp appears in Dataset B, with a total of 7,424 occurrences (3,346 eCall relevant and 4,078 not relevant). The number of observations for each of the two classes and in each dataset under consideration is displayed in Table .Additionally, we will create two new datasets, designated C and D, that only contain a portion of the features discovered using some feature selection techniques.

	Dataset A	Dataset B
eCall Relevant	7736	3346
eCall not Relevant	8727	4078
Total	16463	7424

V. Implementation

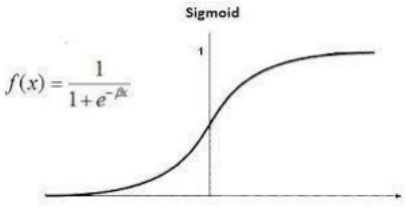
A. Feature Selection

Feature selection is fundamental when the aim is to enhance the interoperability of machine learning approaches. There are several methods that can be exploited to select the variables to be included in a model to predict accident severity. The Pearson Correlation and the Chi-2 evaluate each variable with respect to its capability to classify eCall relevance.

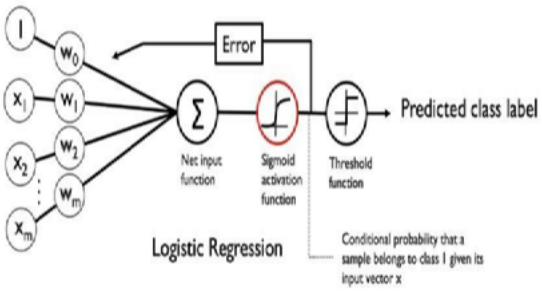
B. Algorithms

- Logistic Regression:

  Logistic regression is a probabilistic approach that provides the conditional probability P(Y—X) for each observation of belonging to one of two classes, given N observed values 37 of the k features. The probability P(Y—X) can be approximated as a sigmoid function applied to a linear combination of input features. This function is called logistic: P(Y—X) = e (0+1X+...+k Xk) / (1+e (0+1X+...+k Xk)). Like linear regression, logistic regression represents data using an equation. To forecast an output value, input values (x) are mixed linearly with weights or coefficient values (referred to as the Greek capital letter Beta) (y). One significant distinction from linear regression is that the result value. Instead of a numeric value, a binary value (0 or 1) is being modelled.

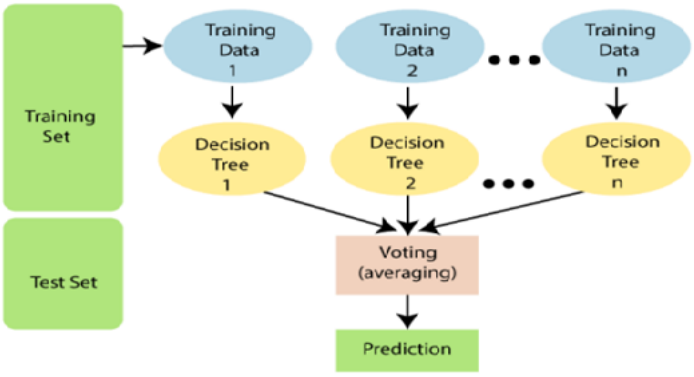


  To assess the significance of attributes in relation to the dependent variable, use logistic regression. The process typically entails evaluating the regression model's estimated coefficients, or k, and then deciding whether to keep or remove the K features from the model.



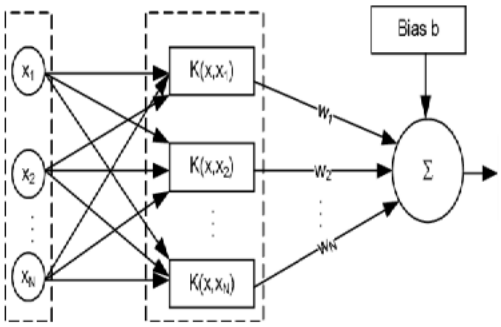
- Random Forest:

  A method for creating a prediction ensemble using a collection of decision trees that grow in randomly chosen subspaces of data was proposed by Leo Breiman in the 2000s. There hasn't been much research on the statistical characteristics of random forests, and little is known about the mathematical principles underlying the method despite growing interest and widespread use. The number of strong features and not the quantity of noisy variables affects the rate of convergence. A group of separate classifiers are built and then combined to produce a final output in an ensemble approach called random forests. More specifically, the Random Forest algorithm creates several Decision Trees and grows them as much as it can. The size of the training set is taken into account when choosing a sample of n samples at random with replacement to generate each decision tree. Only a subset k of all K features are checked 39 for attribute selection at each node in the tree. A ranking of classifiers is generated from each decision tree's classification results based on the number of votes each class received. It then chooses the individual classification with the most votes. As more classifiers are merged, the accuracy of the classification results increases. To increase accuracy, trees must differ from one another in important ways.



- Support Vector Machines:
  One of the most popular approaches for supervised classification in machine learning is the use of support vector machines. The main principle of the approach is to select the largest margin hyperplane, which evaluates the similarity between examples, as the decision surface to distinguish between positive and negative examples in binary classification problems. Support vectors are training examples whose corresponding I are different from zero since they are the only ones on which the decision function depends. The resulting decision function is a linear hyperplane if the kernel function is only a dot product (as the sum of the products of the respective vector elements). In contrast, the kernelfunction can also be a non-linear function, enabling non-

linear dependencies between the input variables and the goal to be captured. For some straightforward types of algorithms, statistical learning theory can pinpoint rather exactly the considerations that must be made in order to train well. This is where the SVM algorithm comes into play. However, the employment of more sophisticated models and algorithms, like neural networks, which are much more difficult to theoretically examine, is frequently required for real-world applications. Both are achieved via the SVM method. Radial basis function (RBF) nets, polynomial classifiers, and a sizable class of neural nets are all included in it as special cases because it builds sophisticated enough models to support them. However, it can be demonstrated that it corresponds to a linear approach in a high-dimensional feature space that is nonlinearly connected to the input space, making it simple enough to be mathematically studied. The sets to discriminate are frequently not linearly separable in that space, despite the fact that the original problem may have been expressed in a finite-dimensional space. In order to facilitate the separation in the much higher-dimensional space, it was suggested that the original finite-dimensional space be mapped into. The mappings employed by SVM schemes are created to make sure that dot products of pairs of input data are accurate in order to keep the computational load manageable.



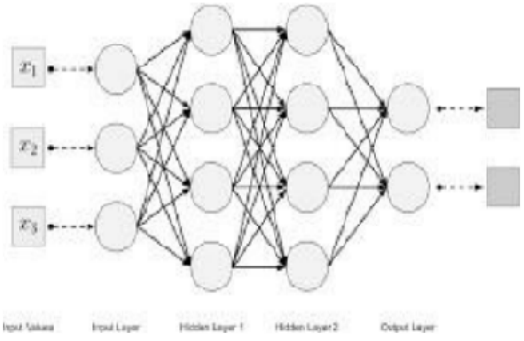
- Deep Neural Network

  Deep learning has transformed machine learning and artificial intelligence over the past ten years, reviving the artificial neural network paradigm and making it the cutting edge in a wide range of applications. In this work, multilayer feed-forward neural networks—the simplest architecture—are taken into consideration. One way to describe such a network is as a stack of L layers, each made up of nl neurons. When a system uses multiple layers of nodes to extract high-level functions from input data, it is using a deep neural network. It entails repurposing the data to become a more imaginative and abstract element. Imagine a picture of a typical man to help you better grasp the outcome of deep learning. Even though you have never seen this image of the guy before, you will always be able to tell that it is a person and set it apart from other animals. This serves as an illustration of how a deep neural network operates. To ensure that the object is appropriately identified, analytical and creative components of the information are examined and organized. The ML system must alter and derive these components because they are not provided to the system directly. The output of neuron j in layer l of a feed-forward neural network is a nonlinear function of the output of every neuron in layer l.

$$o_j^{\ell} = \sigma\left(\sum_{i=1}^{n_{\ell-1}} o_i^{\ell-1}\right)$$

where the neuron's activation function is. The output of the final layer serves as the sole input for the decision function. Due to the issue of vanishing gradients, traditional artificial neural networks that used the sigmoid function as the activation function had inherent difficulties when training networks with several layers . The Rectified Linear Unit (ReLU) has replaced the usual activation function in deep networks because it solves the issue of vanishing gradients and permits networks to be trained using standard backpropagation.

Deep neural network models consume a lot of resources, and this is especially true when numerous model ensembles are used. Even with GPUs available, it may still take a few seconds to examine each retinal image after preprocessing. The total time required soon accumulates to 44 due to the potential requirement to analyze numerous retinal photos for each patient, maybe for various disorders needing different models. When turnaround times are required, this has consequences. Additionally, hosting of commercial servers may not be allowed due to security concerns, which would result in expensive beginning costs. In less developed environments, remote hosting of models can even be completely impossible. Lightweight compressed models might be acceptable in certain situations.



VI. Results

Using the scikit-learn package to create random forests and support vector machines, the keras package with tensorflow back-end for deep neural networks, and the dataset provided in the aforementioned section, we conducted tests on it.

A. Experimental Setup

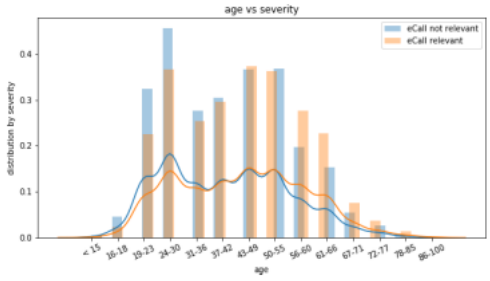
We used the same 20-fold cross-validation split for all of the systems in order to compare their performance, taking advantage of hyper parameter adjustment on the first fold. We utilized 1,000 estimators (i.e., trees) for random forests and the default settings for all other hyper-parameters. With a grid search across the values of C (linear) and (C,) pairs, we tested the linear and radial basis function (rbf) kernel for support vector machines (rbf). We efficiently chose the ideal network design for deep networks by optimizing hyper-parameters using the hyperas package.

	Dataset A			
Method	A	P	R	F1
Logistic Regression	58.3	54.9	63.1	58.7
Random Forests	90.8	92.5	87.7	90.0
Linear SVM	58.2	54.5	68.2	60.6
RBF SVM	91.4	99.4	82.2	89.9
Deep Networks	83.7	81.1	87.2	83.5

	Dataset B			
Method	A	P	R	F1
Logistic Regression	58.3	54.1	62.7	57.6
Random Forests	91.1	90.5	86.5	87.0
Linear SVM	57.6	53.2	65.1	58.1
RBF SVM	92.1	92.5	82.5	83.4
Deep Networks	62.1	56.2	94.9	70.0

B. Quantitative Evaluation

According to the results, non-linear models (such as random forests, deep networks, and support vector machines with rbf kernel) perform significantly better than linear models. The performance of each r system is displayed in Table II, macro-averaged on the 20-fold cross-validation. Even though the method is straightforward and makes use of ready-made machine learning algorithms, the outcomes are very positive.



C. Feature Evaluation

We outline the six features selection methods used to lower a feature set's cardinality in Section IV-A. We contrasted the outcomes of employing the variables chosen by each of the six methods independently. Table III lists the characteristics that have been chosen by at least five out of the six feature selection techniques.

	Dataset C			
Methods	A	P	R	F1
Logistic Regression	57.5	54.1	63.4	58.4
Random Forests	90.6	91.4	88.4	89.9
Linear SVM	55.7	51.9	80.0	62.9
RBF SVM	92.5	99.3	84.7	91.4
Deep Networks	79.6	73.7	89.8	80.6

	Dataset D			
Method	A	P	R	F1
Logistic Regression	55.1	50.2	58.0	53.8
Random Forests	89.8	87.9	86.7	86.0
Linear SVM	55.0	50.3	58.1	53.8
RBF SVM	91.9	93.7	82.6	83.6
Deep Networks	48.6	46.8	97.7	63.2

We replicated the experiments conducted for Datasets A and B as well as for their condensed variants C and D in order to directly assess the discriminative potential of these reduced sets of features. The results of the classifiers are reported in Table IV, which demonstrates that the selected features do in fact help to reach noteworthy outcomes because

the performance is quite comparable to that obtained with the entire collection of features. By utilizing an interpretation based on two alternative methodologies, we suggest a detailed analysis of varying importance in the remaining portion of this section. Conclusion: An eCall device for a motorcycle might take into account installing sensors to gather pertinent data that enables it to automatically determine the seriousness of an accident and, as a result, to automatically issue an eCall. Future studies could compare other ensemble methods and machine-learning algorithms, as those suggested in this paper.

REFERENCES

[1] N. S. Hadjidimitriou, M. Lippi, M. Dell'Amico and A. Skiera, "Machine Learning for Severity Classification of Accidents Involving Powered Two Wheelers," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 10, pp. 4308-4317, Oct. 2020, doi: 10.1109/TITS.2019.2939624.

[2] Kun Li, Haocheng Xu, Xiao Liu, "Analysis and visualization of accidents severity based on LightGBM-TPE", Chaos, Solitons Fractals, vol.157, pp.111987, 2022.

[3] Ravneet Kaur, Rajendra Kumar Roul, Shalini Batra, "An Efficient Approach for Accident Severity Classification in Smart Transportation System", Arabian Journal for Science and Engineering, 2022.

[4] Xiao Wen, Yuanchang Xie, Liming Jiang, Ziyuan Pu, Tingjian Ge, "Applications of machine learning methods in traffic crash severity modelling: current status and future directions", Transport Reviews, vol.41, no.6, pp.855, 2021.

[5] Laurie Brown, Andrew Morris, Pete Thomas, Karthikeyan Ekambaram, Dimitris Margaritis, Ragnhild Davidse, Davide Shingo Usami, Massimo Robibaro, Luca Persia, Ilona Buttler, Apostolos Ziakopoulos, Athanasios Theofilatos, George Yannis, Alain Martin, Fallou Wadji,