Journal of Information Systems & Information Technology (JISIT)
ISSN: 2478-0677; Vol. 8 No.1 2023; [Pages 39 - 50]

Road Accident Severity Prediction based on Contributing Factors using Deep Learning Techniques

Thayani Gathirvelou[1], B. Mayurathan[2]
Faculty of Science, University of Jaffna[1],[2]
thayani0521@gmail.com[1], barathym@univ.jfn.ac.lk[2]

Abstract

Road accidents have recently emerged as the ninth most serious threat in terms of fatality causes to both individuals and governments. The cost of traffic related deaths and property damage is disproportionately high in developing countries. Investigating the elements that cause accidents and forecasting the severity of the accident is thus extremely beneficial in reducing future accidents. Existing methods for forecasting the severity of road accidents rely on shallow severity prediction models and statistical models, with only a few researchers using deep learning models. This study proposed an LSTM-based model for predicting road accidents, which forecasts indications of the severity of road accidents by learning data about road accidents. The existing studies have significant shortcomings, including the use of small datasets with limited coverage, reliance on a large set of data, low accuracy, and inability to be used in real-time. To address these challenges, we independently apply the LSTM model, CatBoost-based model and CNN model to the Balanced and Unbalanced US Accident datasets. To enhance the performance of these models, we employ three feature selection algorithms—Random Forest, Decision Tree, and Cat Boost—to extract the most relevant features from the datasets. Subsequently, the effectiveness of these various proposed methods is compared. Cat Boost combined with an LSTM model showcases the synergy between Cat Boost feature selection and LSTM modeling, based on the experimental results, it is clear that this combined approach outperforms the standalone LSTM model and the Cat Boost model, with an accuracy of 98.57.

Keywords: *Deep Learning, Recurrent neural network, Multi-layer perceptron, Cat Boost, Accident Severity level, Road accident severity prediction, CNN, LSTM*

1. Introduction and Motivation

Road accidents can occur unpredictably at any time or any place. Every day, road accidents terribly lead to loss of life, injuries, and substantial property damage. Also, it significantly impacts society and the economy. In the modern world, traffic on the roads increases along with the demand for vehicles, especially during rush hours. Road accidents thus rank among the world's top causes of death and injury. As per the World Health Organization (WHO)[1], every year, traffic accidents cause 50 million injuries and 1.3 million deaths. Every day, some 3,300 people are killed and 137,000 are injured. Traffic accidents not only danger human life and property safety but also give rise to substantial economic losses [2]. In the lack of sustainable transportation, it is also expected that traffic accidents will become the major cause of death by 2030. Hence, Predicting the severity of a road accident is a critical study topic in the field of transportation safety. To prevent and reduce the frequency of

traffic accidents, it is critical to forecast them and identify the elements that contribute to them in various situations. By predicting accidents in advance, traffic safety authorities and organizations can take proactive steps to implement specific actions and preventive strategies. Understanding the factors that contribute to accidents under various conditions such as weather, road conditions, and traffic patterns can provide valuable insights for designing effective safety measures and improving road infrastructure. For these reasons, predictive models for road accidents have been created to identify the important factors that influence road accidents, allowing for the control and/or enhancement of these factors to increase traffic safety.

This paper aims to provide valuable insights and methods for predicting the severity of road accidents. The goal is to help prevent and reduce accidents, ultimately saving lives. Statistical methods [3] and neural networks [2] have typically been the two dominant research methods in studies on predicting the severity of road accidents. Numerous studies have been done using statistical techniques to predict the frequency and severity of road accidents, including K Nearest Neighbors (KNN) [3], [4], Support Vector Machine (SVM) [3]–[5], and Logistic Regression (LR) [5]. In recent years, both researchers and business professionals have shown significant interest in deep learning as a powerful machine learning technique for the application of adaptability, generalization, and powerful forecasting.

Many neural network methodologies have been created with the help of deep learning theory to simplify the process of predicting road accidents. The Feed forward Neural Network (NN) [6], Recurrent Neural Network (RNN) [6], [7], Convolutional Neural Network (CNN) [6], Multi-Layer Perceptron (MLP) [7], [8], and Single Layer Perceptron [7] are a few of these alternatives. However, several researchers also developed novel models for forecasting road accidents, including RFCNN [9], SSAE [10], DAP [11], TAP-CNN [12], TASP-CNN [13], and MVNB [13]. The main goal of deep learning is to create a representation of the actual predictor vector. This transformed data can then be used for tasks like classification or linear regression. The primary aim of this paper is to compare three models: the Cat-boost-based model, the LSTM model and CNN model using accident data to understand the factors influencing accident severity. Our research seeks to uncover accident patterns, identify the most significant factors, and determine the best feature selection method among Random Forest, Decision Tree, and CatBoost. By doing so, we aim to improve road safety through data-driven insights and accident prevention strategies.

After the introduction, the rest of the paper is organized as follows: the next section describes the latest techniques and methods that are relevant to the topic. The

paper describes the proposed methodology for developing a road accident prediction model in Section III. This model forecasts the severity of road accidents by learning from data about previous accidents. Section IV covers the experimental design and the results of the testing phase. Finally, the paper concludes with future work in Section V.

2. Literature Review

In this section, some of the latest research that is relevant to the road accident prediction model are briefly discussed. Many researchers have extensively studied the patterns of road accidents over the years, identifying the factors that contribute to the occurrence of these dangerous events. In [6], the severity of injuries in accidents caused by traffic is examined based on three network topologies: simple feed-forward Neural Networks (NN), Recurrent Neural Networks (LSTM), and Convolutional Neural Networks (CNN). The proposed methodology uses eight variables as input to predict the severity of traffic accidents accurately. It classifies accidents into three categories: property damage, possible or evident injury, and disabling injury or fatality. The proposed network architecture consists of a total of 3225 parameters, which were carefully selected through a combination of grid search and 10-fold cross-validation. Based on their testing results, RNN model outperformed the NN model (68.79%) and the CNN model (70.30%) among the examined algorithms. However, RNN models require complex training algorithms, which can limit their applications, particularly with small datasets or data lacking temporal features (e.g., time of the accident).

The severity of motorcycle accidents is investigated in [7] in terms of prediction accuracy using various deep learning methods: standard recurrent neural networks (LSTM), multi-layer neural networks, and single-layer neural networks. To evaluate the performance of these models, a dataset comprising 2,430 motorcycle accidents that occurred over a ten-year period in mountainous regions of the United States was used. To identify the most significant and relevant features, both the RFE (Recursive Feature Elimination) algorithm and a simple backward selection method were employed in this research. Different metrics, such as the area under the curve and confusion matrix, were employed to compare the performance of the different models. The paper concludes that DNN models are more accurate than single-layer neural networks in predicting the severity of motorcycle crashes. Furthermore, the study discovered that the recurrent neural network (RNN) outperformed the other three neural network models. However, due to the imbalanced nature of the crash datasets utilized in this study, the algorithms could not achieve satisfactory

accuracy for the severe/fatal category.

A comprehensive analytic framework is proposed in [10] that utilizes a deep learning model called the stacked sparse auto encoder (SSAE) to predict the severity of traffic accidents based on contributing factors. The study employs a traffic accident dataset provided by the UK Department. Initially, the paper analyzes the significance and interdependence of the contributing factors to injury severity, eliminating factors with low correlation. Subsequently, the k-means clustering algorithm is utilized to categorize the data into distinct classes based on geographic information. Finally, a deep learning model based on SSAE is trained using highly correlated factors. However, an important limitation of this approach is the potential challenge posed by data integrity, which could impact the implementation of the suggested framework and the attainment of desired outcomes.

According to [12], a unique traffic accident severity prediction-convolutional neural network (TASP-CNN) model was introduced, which considers the correlations between the elements of traffic accidents while predicting their severity. To achieve this, they proposed the feature matrix to gray image (FM2GI) algorithm, which converts individual feature relationships of traffic accident data into gray images containing combination relationships in parallel. These converted gray images serve as the input variables for the model, utilizing the weights of the attributes related to traffic accidents. The study utilized traffic accident data from Leeds City Council in the United Kingdom spanning eight years (2009-2016), resulting in a total of 21,436 accident records. The severity levels were categorized as minor, serious, and fatal.

The performance of the proposed TASP-CNN model was compared to several other models, including NBC, KNN, LR, DT, GB, SVC, Conv1D, NN, and LSTM-RNN. The results clearly demonstrate that the TASP-CNN model outperformed the competing models. In this study [8], road accident data was evaluated using the decision tree, ANN, and random forest algorithms to assess the benefits and drawbacks of each approach. In this methodology, 75% of the datasets is used as training data, while the remaining 25% is used as the test data. The classifier was trained using the training samples and its performance was assessed using the testing data. Also, non-critical events (0) and critical incidents (1) are considered as two levels in the accident level (AL) variable. The results revealed that the random forest model produced the most accurate forecasts, whereas the ANN algorithm tended to overestimate the AL compared to the other models. This study provides valuable insights for researchers, enabling them to employ the random forest model for

predicting traffic accidents and achieving more accurate results.

RFCNN, an ensemble model that combines Random Forest and Convolutional Neural Networks (CNN) for accurately predicting the severity of road accidents is introduced in [9]. The effectiveness of the suggested methodology is evaluated using accident data that was gathered from the United States between February 2016 and June 2020. The Random Forest algorithm was employed to identify significant factors strongly associated with the severity of highway accidents. Several key features, such as distance, temperature, windchill, humidity, visibility, and wind direction, were identified as having a significant influence on accident severity.

In order to evaluate the performance of the proposed RFCNN model, several ensemble models such as Random Forest (RF), Ada Boost Classifier (AC), Extra Trees Classifier (ETC), Gradient Boosting Machine (GBM), and a Voting classifier using regression algorithms (LR+SGD) is compared. All these models utilized the 20 significant features identified by the Random Forest algorithm as input. The experimental results presented in the paper show that the RFCNN model performs better than the RF, AC, ETC, GBM, and the voting classifier (LR+SGD) in terms of classification accuracy, precision, recall, and F-score. The RFCNN model achieved an accuracy of 0.991, precision of 0.974, recall of 0.986, and an F-score of 0.980.

3. Methodology

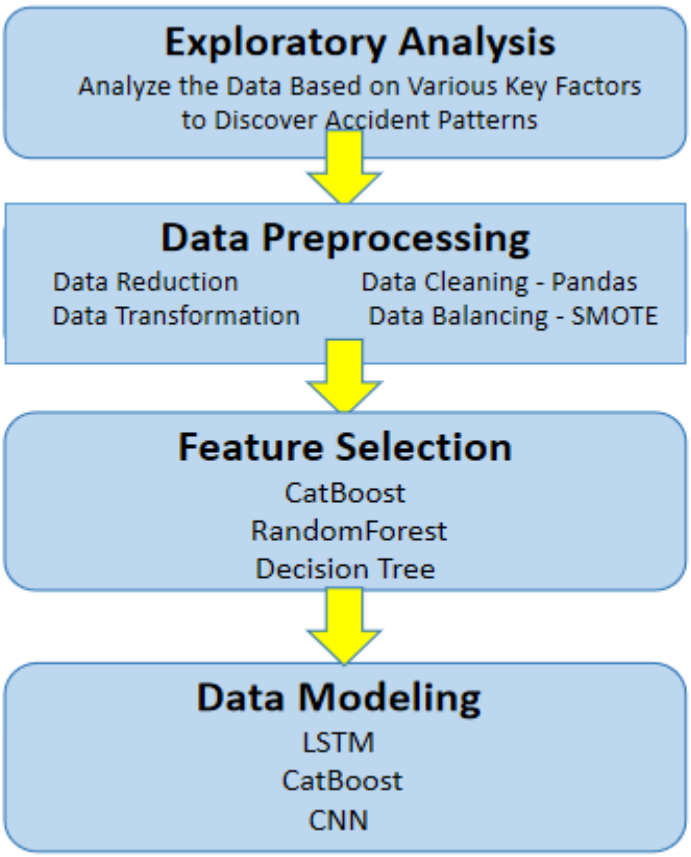

Fig. 1. Proposed Methodology for Road Accident Severity Prediction

The proposed methodology for road accident severity prediction model is visually depicted in Figure 1, which provides a diagrammatic representation. The road accident prediction model consists of the following stages:

A. Data Collection

Data collection is a systematic process of gathering observations or measurements. It is increasingly important in research as it assists in making data-driven decisions, validating theories, and populating databases with relevant datasets. In this study, road accident data was collected from the publicly accessible website Kaggle. A total of over 2.8 million accident reports spanning a six-year period (2016 - 2021)

were retrieved from Kaggle. The collected dataset consists of 48 significant factors, including severity, temperature, wind speed, weather timestamp, distance, wind direction, pressure, traffic signal, sunrise sunset, etc. This dataset categorizes severity into three levels: slight, serious, and fatal.

B. Data Preprocessing

To improve the quality of the data, we must first construct a suitable data structure before we examine the trend of traffic accidents and develop a deep-learning model. As a result, the following step will begin with preprocessing our raw data. The initial data processing involved deleting the incomplete, incorrect, and redundant road accident data, normalizing the road accident datasets, and processing the data in an unbalanced manner. Data pre-processing encompasses several steps, including data cleaning, integration, transformation, and reduction. State-of- the-art methods have employed various methods such as pandas, data profiling techniques [10], and manual approaches [7] to clean the data. The findings indicate that pandas outperform other methods. A popular Python library called Pandas was specifically designed for data pre-processing tasks such as cleaning, manipulation, and analysis. It offers classes for reading, processing, and writing CSV data files. Although there are many data cleaning technologies available, the Pandas library offers a quick and effective approach to handling and studying data. Therefore, Panda's library is utilized in this research for cleaning the road accident data.

C. Feature Selection

After preprocessing the data, it is crucial to analyze and identify the relevant features that influence road accidents. This step serves as a fundamental technique to guide the selection of variables that are most effective and efficient in predicting accidents. Features in the gathered and cleaned road accident data correspond to numerous variables that affect the target variable. The benefits of high-quality features include being useful, relevant, interpretable, and non-redundant. These qualities serve as the foundation for modeling, problem solving, and generating consistent and convincing solutions. In addition, machine learning techniques make it easier to recognize the significance and contribution of factors that influence injury severity in traffic accidents. These approaches explain the nonlinear behavior of specific variables in traffic accidents with varying degrees of injury severity. In previous studies, researchers employed various methods, including cat boost (combined with Sharp Value) [10], Random Forest [8], [9], grid search [6], Decision Tree, and RFE (Recursive Feature Elimination) [7], to analyze contributing factors. The results consistently demonstrated that Cat Boost, Random Forest, and Decision Tree produced nearly identical results. Consequently, we utilized these three

algorithms to compare their performance and determine the most effective one for feature selection. From Cat Boost, we identified seven significant features, while Decision Tree and Random Forest each yielded nine features. These findings further our understanding of discriminant factors in our analysis.

D. Data Modeling

Data modeling involves the representation and communication of data requirements through the task of creating a data model. Road accident prediction relies heavily on modeling, as it is the key to obtaining accurate forecast results. Therefore, after the selection of features in the previous stage, less significant features will be eliminated from the dataset, and the significant features will be utilized as inputs for the deep learning models. In previous research, various deep learning models were commonly used, including RNN [6], [7], CNN [2], [6], MLP [7], and SLP [7]. Additionally, researchers have developed novel models specific to their studies, such as RFCNN [9], SSAE [10], DAP [11], TAP-CNN [2], TASP-CNN [12], and MVNB [13]. Among the pre-existing deep learning models, RNN and MLP demonstrated the highest accuracy levels. As a result, I used three network architectures based on LSTM, CNN and Catboost models to improve the performance of my study. Each model uses the important features chosen in the previous stage as its inputs and predicts the severity of road accidents as fatal, serious and slight. After receiving the output from three models, several techniques were used to compare the performance of these models based on their accuracy. Our model performed better than other approaches for estimating the severity of traffic accidents in the same dataset.

4. Experimental Results

The performance of the proposed methodology including three models such as LSTM, CatBoost, and LSTM paired with CatBoost is evaluated using Balanced and Unbalanced US Accident dataset. The LSTM model was designed to capture temporal dependencies in the data, while CatBoost is a gradient boosting algorithm that handles categorical features effectively. The dataset was divided into training and testing sets at a ratio of 70:30. To tackle the issue of imbalanced data in a road accident dataset, SMOTE (Synthetic Minority Over-sampling Technique) algorithm [19] is used in this research. SMOTE algorithm balances the severity levels of accidents within the dataset. The severity levels were categorized into four classes, namely 1, 2, 3, and 4. Class 1 denoted the least severe accidents, while class 4 represented the most severe ones. By generating synthetic instances of the minority classes (i.e., severity levels with fewer occurrences) SMOTE keeps the dataset as balance. Furthermore, for identifying the most influential features for predicting

accidents, CatBoost, is employed in this research. CatBoost is equipped with the capability to automatically assess the importance of each feature in relation to its impact on accident severity. Leveraging this feature ranking, seven most significant attributes are selected from a list of 47 features. These attributes encompassed Airport Code, County, City, and State, as well as Weather Timestamp, Distance (mi), Wind Speed (mph), and Timezone. These variables were identified as crucial factors in the analysis.

The following table I details the performance of the proposed methodology

TABLE I
PERFORMANCE OF THE PROPOSED METHODOLOGY.

Proposed models	Performance (in %)
Cat boost model with imbalanced dataset	98.14
Cat boost model with balanced dataset	95.86
LSTM Model	98.16
CNN Model	88.53
Cat boost + LSTM (Catboost for feature selection)	98.57
RandomForest + LSTM (RandomForest for feature selection)	98.48
DecisionTree + LSTM (DecisionTree for feature selection)	98.52
Catboost feature selection a nd model with imbalanced dataset	93.39
Catboost feature selection and model with Balanced dataset	90.43

Based on the testing results mentioned in Table I, the combination of catboost and LSTM model achieves the highest accuracy of 98.17 %. Based on the experimental results, it can be confidently concluded that the combination of CatBoost and LSTM models delivers the best performance, exhibiting remarkable accuracy.

Furthermore, the performance of our proposed methodology is compared against state-of-the-art approaches to assess its efficiency and effectiveness. Table II displays the classification performances of various state-of-the-art methods, as well as our own proposed method.

TABLE II
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART WORK.

	Methods	Performances (in %)
[9]	RFCNN	99.10
[18]	LR	95.30
	Decision tree	95.03
	RF	96.50
	XG-Boost	96.38
Ours	catboost + LSTM model	98.17

It is also noted that, both studies [9] and [18] utilized the same dataset as the one used in this paper. Based on the data presented in Table II, it is evident that my model demonstrated superior accuracy compared to [18], outperforming its performance. However, it is worth noting that the research conducted by [9] slightly outperformed our results. This margin can be attributed to the fact that, the authors in [9] considered a dataset spanning 5 years, incorporating 20 significant features, whereas our research utilizes a 6-year dataset encompassing 22 significant features.

5. Conclusion and Future work

In conclusion, this research emphasizes the significance of addressing road accidents as a critical threat to individuals and governments. Developing countries suffer from disproportionately high costs associated with traffic-related deaths and property damage. Therefore, investigating the factors contributing to accidents and accurately forecasting their severity becomes crucial in mitigating future incidents. While existing approaches predominantly rely on shallow severity prediction models and statistical methods, this paper highlights the need to explore deep learning models, such as LSTM, CNN and CatBoost, for more accurate predictions.

Also, this paper experiments the effectiveness of the LSTM, CNN and CatBoost models in predicting the severity of road accidents. Based on the comprehensive experiments using large dataset reveals that both models exhibited promising outcomes in accurately assessing accident severity.

Moreover, when the LSTM model was enriched with CatBoost-selected features, the results improved further. This combination leveraged the strengths of both models and generated more reliable predictions. The LSTM model with CatBoost feature selection surpassed the individual models in terms of accuracy, highlighting the

advantages of integrating different techniques within a unified framework. The synergy between CatBoost feature selection and LSTM modeling contributed to the model's remarkable success, showcasing the power of this approach.
The findings of this paper have important implications for road safety initiatives and accident prevention strategies. The proposed models can play a crucial role in reducing the severity of accidents, saving lives, and enhancing overall road safety.

In conclusion, both the LSTM and CatBoost models, whether used individually or in combination, show significant potential in predicting the severity of road accidents. This research has some limitations in certain areas due to the complexity of traffic accidents. Firstly, traffic accidents data are only utilized for prediction, but other factors like traffic flow, human mobility, road conditions, and special occasions could also play a crucial role in predicting traffic accidents. Secondly, in this paper, we utilized road accident data from the United States due to limitations in accessing specific road accident data from Sri Lanka.

In the future, we aim to apply our model to Sri Lanka data and draw conclusions relevant to our country. Therefore, future work that combines the layout of the urban road network with comprehensive parameters related to traffic accidents holds promise for generating improved prediction results.

References

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
[2] Wenqi, L., Dongyu, L. and Menghua, Y., "A model of traffic accident prediction based on convolutional neural network," 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE), pp. 198-202, 2017.
[3] Zhang, J., Li, Z., Pu, Z. and Xu, C., "Comparing prediction performance for crash injury severity among various machine learning and statistical methods", IEEE Access, vol. 6, pp.60079-60087, 2018.
[4] Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M. and Rakotonirainy, A., "Crash severity analysis of vulnerable road users using machine learning", PLoS one, 16(8), pp. 1-22, 2021.
[5] Sowdagur, J.A., Rozbully-Sowdagur, B.T.B. and Suddul, G., "Road Accident Severity Prediction in Mauritius using Supervised Machine Learning Algorithms".
[6] Sameen, M.I., Pradhan, B., Shafri, H.Z.M. and Hamid, H.B., "Applications of deep learning in severity prediction of traffic accidents", In Global Civil Engineering Conference, pp. 793-808, Springer,2019
[7] Rezapour, M., Nazneen, S. and Ksaibati, K., "Application of deep learning techniques in predicting motorcycle crash severity", Engineering Reports, 2(7), p.e12175
[8] Lee, J., Yoon, T., Kwon, S. and Lee, J., "Model evaluation for forecasting traffic accident

severity in rainy seasons using machine learning algorithms", Seoul city study. Applied Sciences, 10(1), p.129, 2019.

[9] Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H.A. and Bisogni, C., "RFCNN: traffic accident severity prediction based on decision level fusion of machine and deep learning model", IEEE Access, 9, pp.128359-128371, 2021.

[10] Ma, Z., Mei, G. and Cuomo, S., "An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors", Accident Analysis & Prevention, 160, p.106322, 2021.

[11] Alkheder, S., Taamneh, M. and Taamneh, S., "Severity prediction of traffic accident using an artificial neural network", Journal of Forecasting, 36(1), pp.100-108, 2017.

[12] Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z. and Wang, Z., "Traffic accident's severity prediction: A deep-learning approachbased CNN network", IEEE Access, 7, pp.39897-39910, 2019.

[13] Dong, C., Shao, C., Li, J. and Xiong, Z., "An improved deep learning model for traffic crash prediction", Journal of Advanced Transportation, 2018.

[14] Ibrahim, A.A., Ridwan, R.L., Muhammed, M.M., Abdulaziz, R.O. and Saheed, G.A., "Comparison of the CatBoost classifier with other machine learning methods", International Journal of Advanced Computer Science and Applications, 11(11), 2020.

[15] Shaik, M.E., Islam, M.M. and Hossain, Q.S., "A review on neural network techniques for the prediction of road traffic accident severity", Asian Transport Studies, 7, p.100040, 2021.

[16] Ren, H., Song, Y., Wang, J., Hu, Y. and Lei, J., "A deep learning approach to the citywide traffic accident risk prediction", In 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3346-3351, 2018.

[17] Puneethraj N,Dr. Girish, "Analysis and Prediction of Road accident using machine learning and deep learning approach", International Research Journal of Modernization in Engineering Technology and Science(IRJMETS), Vol 04(02), 2022.

[18] Liu, H. and Shetty, R.R., "Analytical Models for Traffic Congestion and Accident Analysis", 2021.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique.", Journal of Artificial Intelligence Research, vol. 16, June 2002, pp. 321–7. Crossref, https://doi.org/10.1613/jair.953.