Application of machine learning for prediction of road accidents based on indicators: A random forest approach

Miloš Pljakić[1], Osman Lindov[2], Aleksandra Petrović[1], Predrag Stanojević[1], Nebojša Arsić[1]

[1] University of Priština in Kosovska Mitrovica, Faculty of Technical Sciences, Knjaza Miloša 7, 38220 Kosovska Mitrovica

[2]University of Sarajevo - Faculty of Traffic and communications, Zmaja od Bosne 8, 71000 Sarajevo, Bosnia and Herzegvina

Abstract

In this study, efforts were made to examine the extent of the relationship between specific indicators and traffic accidents resulting in fatalities and injuries at the municipal level in the Republic of Serbia through the application of machine learning algorithms. The indicators analyzed relate to the use of seat belts and mobile phones by drivers and passengers in cars, trucks, and buses. Data were observed for urban and rural areas at the municipal level. In this study, the Random Forest model was employed due to its ability to handle high-dimensional data and capture complex relationships between various factors. The results showed that the use of seat belts by drivers on rural roads has significant predictive power for accidents resulting in fatalities, while the use of seat belts by passengers in cars in urban areas has the highest predictive power for accidents resulting in injuries. Based on these findings and those of each individual indicator, efforts can be directed towards the implementation of measures focused on education, technological solutions, and legislative regulations aimed at increasing the use of safety systems in vehicles.

Keywords: *Traffic accidents, Machine learning, Indicators, Random Forest*

1 Introduction

Based on the analysis of traffic accidents from 2010 to 2021, it is possible to conclude that there is a trend of decreasing fatalities in traffic accidents, despite a significant increase in the global motor vehicle fleet and the expansion of traffic networks [1]. These results underscore the importance of efforts aimed at improving traffic safety, but they also highlight the necessity of intensifying these efforts to achieve the United Nations' goal for the Decade of Action for Road Safety 2021-2030, which is to halve the number of deaths by 2030 [2]. This research also indicates significant differences in the success of reducing fatalities between different municipalities in the Republic of Serbia, suggesting the need for further study of factors influencing traffic safety globally to develop effective policies and strategies for reducing accidents in the future.

In order to analyze a traffic safety system in a specific geographical area, it is essential to first assess the existing conditions concerning all elements of the system. System indicators are often classified into direct measures, which constitute traffic accidents, and indirect measures, which consist of various instruments and measures based on which the state of traffic safety in an area can be evaluated [3]. The relationship between traffic accidents and indicators represents a positive correlation depending on the approach to monitoring. In the territory of the Republic of Serbia, a methodology for monitoring the most significant indicators has been developed, which is implemented and measured in municipalities and police administrations. The measured indicators relate to the use of protective systems in vehicles, alcohol, speed, emergency response, etc. Depending on the indicator, many are measured in urban areas, rural areas, and highways. The spatial distribution of traffic

Corresponding author: *Dr. Miloš Pljakić* (milos.pljakic@pr.ac.rs)

accidents and indicators is often uneven when observing spatial units at any spatial level [4]. This unevenness is often highlighted at the level of local communities in the Republic of Serbia due to the diversity in the demographic, sociological, economic, and cultural characteristics of municipalities. Therefore, it is necessary to examine the impact of accidents and individual indicators collected at the municipal level, considering the urban and rural parts. In addition to the impact, all indicators have certain predictive powers in traffic accidents, which need to be identified and explained in order to reduce traffic accidents.

Prediction of accidents is the process of assessing the probability of a specific traffic accident occurring in the future based on the analysis of various factors. This process typically involves the use of statistical models, econometric models, machine learning algorithms, or other data analysis techniques to identify patterns and trends associated with traffic accidents. However, statistical and econometric models can yield unstable results due to difficulties in handling large-dimensional data. A richer set of variables may potentially enhance predictive ability and causal understanding but can introduce additional burden and model complexity [4]. Among various machine learning approaches, artificial neural networks and support vector machines demonstrate high predictive performance. Breiman proposed an ensemble algorithm known as Random Forest (RF), which preserves descriptor importance, maintains tree structure, and generally reduces variance compared to individual decision trees[5]. Random forest has also demonstrated high predictive efficiency [6]. Such characteristics make Random Forest highly efficient for analyzing datasets with numerous features, which is becoming increasingly common. In situations with extensive databases and a large number of variables, traditional classification approaches often face the challenge of too many features and loss of efficiency, while random forest regression algorithms continue to exhibit high performance. As a result, analysts in traffic safety are often forced to carefully balance data volume and the number of factors in traffic accident datasets, as well as clearly define how they will utilize the results of their research.

Based on the previous facts, the aim of this study is to identify the fundamental indicators of traffic safety that have predictive power on traffic accidents resulting in fatalities and injuries. The identification is conducted using random forest regression algorithms to systematize all indicators collected in urban and rural areas.

2 Methodology

In this study, data on traffic accidents and indicators in the territory of the Republic of Serbia were observed. The data on traffic accidents include accidents resulting in fatalities and accidents resulting in injuries in the period from 2018 to 2022. Indicator data encompass the use of mobile phones and seat belt usage for different categories of vehicles (cars, freight vehicles, and buses) in urban and rural areas. The data were collected for the same time period and aggregated at the level of local communities to facilitate the examination of the impact of individual indicators on the frequency of traffic accidents. Data on traffic accidents and indicators were obtained from the Traffic Safety Agency of the Republic of Serbia. Table 1 presents a descriptive analysis of the data used in this research.

Table 1. Descriptive Analysis of Variables

Variable	Description	Min	Max	Mean	S.D.
FTA_M	Fatal Traffic Accidents by Municipality	0	97	15,09	14,8
ITA_M	Injury Traffic Accidents by Municipality	8	5018	405,7	612,
MPU_PVU	% Mobile Phone Usage - Vehicle Drivers (Urban)	0,5	9,3	3,59	1,62
SBU_PVU	% Seat Belt Usage - Vehicle Drivers (Urban)	43,5	98,5	83,27	10,5
SBUP_PVU	% Seat Belt Usage - Vehicle Passengers (Urban)	43,8	96,0	78,86	9,33
SBURP_PVU	% Seat Belt Usage - Rear Passenger (Urban)	0	59,4	15,62	11,1
MPU_FVU	% Mobile Phone Usage - Freight Vehicle Drivers (Urban)	2	33,3	9,73	5,08
SBU_FVU	% Seat Belt Usage - Freight Vehicle Drivers (Urban)	2,6	88	50,52	19,4
SBUP_FVU	% Seat Belt Usage - Freight Vehicle Passengers (Urban)	0	100	39,65	27,9
MPU_BU	% Mobile Phone Usage - Bus Drivers (Urban)	2	50	17,38	11,1
SBU_BU	% Seat Belt Usage - Bus Drivers (Urban)	0	83,3	26,88	24,6
SBUP_BU	% Seat Belt Usage - Bus Passengers (Urban)	0	100	20,84	33,6
MPU_PVR	% Mobile Phone Usage - Vehicle Drivers (Rural)	0	11,5	3,72	2,01
SBU_PVR	% Seat Belt Usage - Vehicle Drivers (Rural)	0	97	84,53	15,8
SBUP_PVR	% Seat Belt Usage - Vehicle Passengers (Rural)	0	96,4	82,05	15,0
SBURP_PVR	% Seat Belt Usage - Rear Passenger (Rural)	0	79,8	19,52	13,2
MPU_FVR	% Mobile Phone Usage - Freight Vehicle Drivers (Rural)	0	34,6	10,42	5,75
SBU_FVR	% Seat Belt Usage - Freight Vehicle Drivers (Rural)	0	96,9	57,33	21,1
SBUP_FVR	% Seat Belt Usage - Freight Vehicle Passengers (Rural)	0	88,9	41,95	23,6
MPU_BR	% Mobile Phone Usage - Bus Drivers (Rural)	0	50	16,12	10,2
SBU_BR	% Seat Belt Usage - Bus Drivers (Rural)	0	90	32,35	26,7
SBUP_BR	% Seat Belt Usage - Bus Passengers (Rural)	0	100	27,83	35,1

2.1 Methods

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without being explicitly programmed for them. Machine learning approaches are traditionally divided into three broad categories: Supervised learning, Unsupervised learning, and Reinforcement learning. In this study, the supervised learning technique has been analyzed to investigate the prediction of traffic accidents based on basic indicators of the traffic safety system. Supervised learning represents a category that uses labeled datasets to train algorithms for prediction. One of the most common approaches of supervised learning algorithms is the Random Forest model, which consists of a large number of individual decision trees [5].

The application of the Random Forest model involves combining predictions from multiple decision trees to produce a more accurate prediction result. This model does not require specific statistical assumptions about the data distribution, making it suitable for cases of nonlinear relationships among observed factors [7]. The development of the Random Forest model in this research was conducted using the following algorithm:

1. For b = 1 to B:
   a. Create a bootstrap sample N* of size N (75%) from the training data.
   b. Train a random forest tree T_b on bootstrapped data by iteratively executing the following steps for each leaf node of the tree until the minimum node size n_{min} is reached.
      i. Randomly select m variables from the pool of p variables.
      ii. Identify the optimal variable/split-point from the selected m variables.
      iii. Partition the node into two child nodes based on the selected split-point.
2. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$
3. To predict at a new point (Accident) x: $f_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$, where:
   a. $f_{rf}(x)$: This represents the prediction accidents made by the Random Forest model for a new input x. The symbol f typically denotes a predicted value or function.
   b. B: Refers to the number of trees (or base learners) in the Random Forest ensemble. Each tree is denoted by T_b, where b ranges from 1 to B.
   c. $T_b(x)$: Represents the prediction made by the b -th tree in the ensemble for the input x. This is the output of a single decision tree within the Random Forest.
   d. $\sum_{b=1}^{B}$: This summation symbolizes the aggregation of predictions from all the trees in the ensemble.
   e. $\frac{1}{B}$: This term signifies the averaging of predictions across all trees in the ensemble. It's used for regression tasks to compute the mean prediction of all the trees.

Three metrics utilized to assess prediction performance include the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) [8].

$$MAE = \frac{\sum_{i=1}^{N}\left|t_{pi} - t_{oi}\right|}{N}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{t_{pi} - t_{oi}}{t_{oi}}\right| \times 100\%$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left|t_{pi} - t_{oi}\right|^2}{N}}$$

In which, N is the total number of testing data used in prediction, t_{pi} is the i^{th} predicted clearance time, t_{oi} is the i^{th} actual clearance time.

3 Results

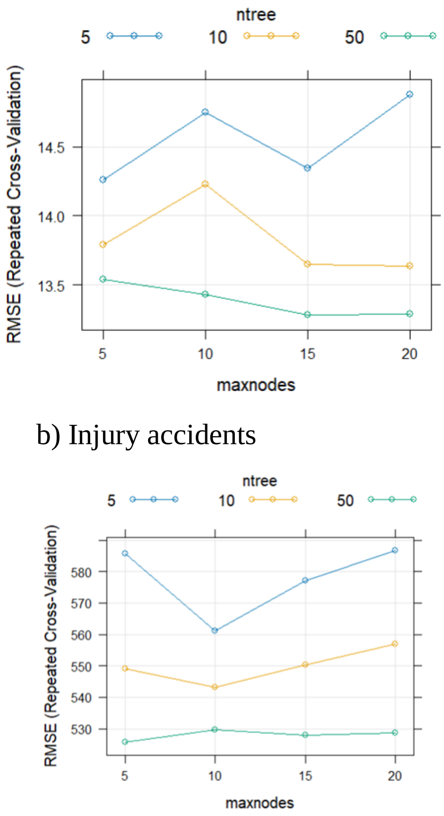
The number of traffic accidents resulting in fatalities and injuries is unevenly spatially distributed across municipalities in the territory of the Republic of Serbia. This spatial unevenness is accompanied by significant differences in demographic, sociological, economic, and cultural characteristics of the observed municipalities. In this research, the predictive power of certain basic indicators on the frequency of traffic accidents resulting in fatalities and injuries was determined.

The analysis process involves generating databases on accidents and indicators, dividing them into training and test datasets, developing predictive functions, and assessing the predictive strength for each individual indicator. The data were divided into 75% training set and 25% test set. Based on the training set, a model was trained to predict values for the test data. The reliability of the model was determined by comparing the predicted data with the values of the test data. The results of the model's reliability are provided in Table 2.

Table 2. Results of Model Reliability

	MAE	MSE	RMSE	R^2
Fatal Accidents	10,13	162,11	12,73	0,32
Injury Accidents	374,85	249974,26	499,97	0,396

In order to improve the model, certain parameters have been modified to minimize the predictive error. The parameters that have been modified relate to the number of trees in the forest and the maximum number of terminal nodes trees in the forest can have. Figure 1 depicts the visualization of parameter selection and how they behave under different variations of the model.



a) Fatal accidents

b) Injury accidents

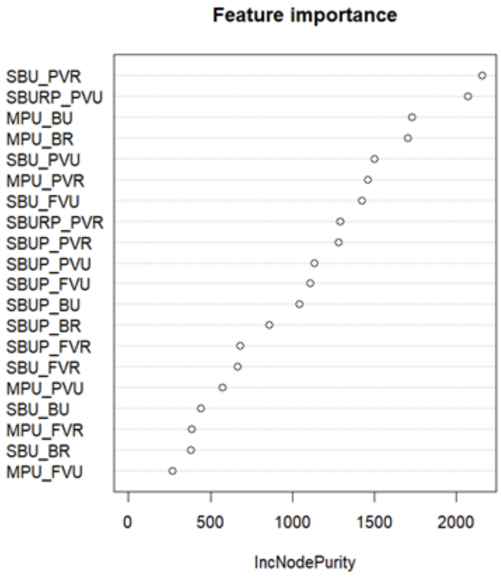
Figure 1. Testing Parameters for the Model

The selection of specific hyperparameters for the Random Forest model was based on preliminary analyses that identified them as the most impactful for our predictive goals. The number of trees in the forest and the maximum number of terminal nodes were chosen because they directly influence the model's complexity and its ability to generalize from the training data. The number of trees affects the variance and bias trade-off, while the maximum number of terminal nodes determines the granularity of the trees, influencing both overfitting and underfitting. Other hyperparameters, such as the minimum number of samples required to split an internal node and the minimum number of samples required to be at a leaf node, were also considered but were found to have less significant impacts on model performance. These settings were optimized through cross-validation to ensure that the model was robust and reliable. Details on the optimization process for these additional hyperparameters can be provided upon request, adding further transparency and reproducibility to our methodology.

Based on the predictive parameters in the model, a list of indicators was created according to their predictive power for accidents resulting in fatalities and accidents resulting in injuries. Figure 2 shows the results. The x-axis displays the list of indicators, which are average values for all trees. This is measured by the residual sum of squares and indicates the predictive power of each individual indicator. In this way, all causal interpretations are excluded.

In Figure 2, various indicators with the highest predicted power for accidents resulting in fatalities and injuries can be identified. In the following section, the top 5 indicators with the highest predictive power are highlighted. For fatalities, the most prominent indicators are: % Seat Belt Usage - Vehicle Drivers (Rural), % Seat Belt Usage - Rear Passenger (Urban), % Mobile Phone Usage - Bus Drivers (Urban), % Mobile Phone Usage - Bus Drivers (Rural), and % Seat Belt Usage - Vehicle Drivers (Urban). For injuries, the indicators with the highest predictive power are: % Seat Belt Usage - Rear Passenger (Urban), % Mobile Phone Usage - Vehicle Drivers (Rural), % Seat Belt Usage - Bus Passengers (Urban), % Mobile Phone Usage - Bus Drivers (Rural), and % Seat Belt Usage - Vehicle Passengers (Urban).

a) Fatal accidents



b) Injury accidents

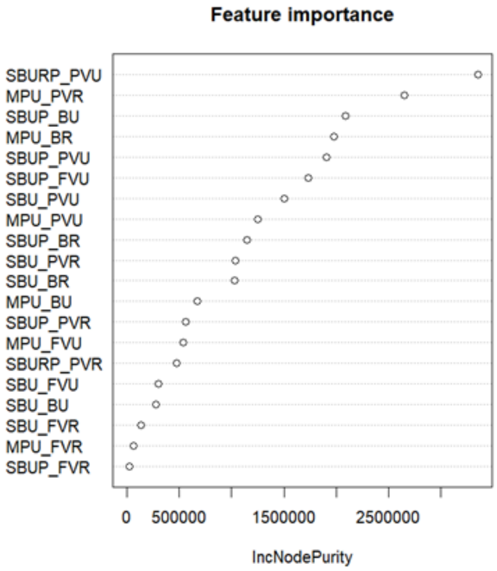


Figure 2. Predictive power of individual indicators on traffic accidents.

4 Discussion and conclusion

In this study, the correlation and impact of basic indicators of the traffic safety system on traffic accidents in the Republic of Serbia were examined. The analysis was conducted across 161 municipalities during the period of 2018-2022. A database was generated, encompassing accidents resulting in fatalities and injuries as dependent variables and basic indicators of the system as independent variables. Basic indicators were observed in both urban and rural parts of municipalities, which has previously shown that the development of predictive functions varies depending on the observed areas [9]. In order to identify significant indicators with high predictive power, random forest regression models were applied as the most commonly used machine learning model.

When it comes to traffic accidents resulting in fatalities, the highest predictive power was identified in the following 5 indicators: % Seat Belt Usage - Vehicle Drivers (Rural), % Seat Belt Usage - Rear Passenger (Urban), % Mobile Phone Usage - Bus Drivers (Urban), % Mobile Phone Usage - Bus Drivers (Rural), and % Seat Belt Usage - Vehicle Drivers (Urban). The results of this study indicate that the percentage of car drivers not using seat belts in rural areas has the highest predictive power for accidents resulting in fatalities. The lack of seat belt usage can significantly increase the risk of severe injuries or fatalities in the event of a traffic accident [10]. Drivers not wearing seat belts are more prone to serious injuries in collisions, which can lead to a higher number of fatalities. Next in line is the indicator related to the percentage of rear-seat passengers in urban areas using seat belts. Insufficient seat belt usage by rear-seat passengers increases the risk of serious injuries in case of a collision [11]. Such injuries can have a fatal outcome, especially if the rear-seat passenger is unrestrained. The next identified indicator shows how often bus drivers use mobile phones while driving in urban areas. The use of mobile phones while driving can significantly affect the driver's ability to react properly to traffic situations and anticipate hazards on the road [12]. Inattention caused by mobile phone use can lead to fatal accidents. In addition to car drivers, the use of mobile phones by bus drivers was also identified. Similarly to urban areas, using a mobile phone while driving a bus can disrupt the driver's concentration and reactions, increasing the risk of fatal accidents.

In addition to accidents resulting in fatalities, key indicators closely associated with accidents resulting in injuries have been documented. The top 5 indicators with the highest predictive power are: % Seat Belt Usage - Rear Passenger (Urban), % Mobile Phone Usage - Vehicle Drivers (Rural), % Seat Belt Usage - Bus Passengers (Urban), %

Mobile Phone Usage - Bus Drivers (Rural), and % Seat Belt Usage - Vehicle Passengers (Urban). The first indicator highlights the percentage of rear-seat passengers in urban areas using seat belts. Similar to accidents resulting in fatalities, the lack of seat belt usage by rear-seat passengers may result in serious injuries in the event of a traffic accident [13]. In addition to seat belt usage, the predictive power of mobile phone usage also contributes significantly to the occurrence of accidents resulting in injuries in rural areas. Accidents resulting in injuries when drivers use mobile phones on rural roads can occur due to several reasons:

- Reduced attention and concentration of drivers: Using a mobile phone for calls or texting demands the driver's attention and concentration. When the driver focuses on the phone instead of the road and surroundings, their ability to notice potential hazards on the road, such as obstacles, traffic changes, or vulnerable road users, decreases.
- Increased reaction time: When a driver uses a mobile phone, the time required to react to sudden situations on the road may be prolonged. This can be particularly dangerous on rural roads where there may be less traffic but the roads are often longer with fewer warning signs or lighting, meaning the driver must react more quickly to any danger.
- Reduced ability to maintain vehicle control: Using a mobile phone can interfere with the driver's ability to maintain stability and control over the vehicle. This is particularly important on rural roads where there may be sharp curves, uneven surfaces, or other challenging driving conditions.
- Decreased awareness of speed: Drivers using a mobile phone may be less aware of their speed, leading to speeding and increased accident risk. On rural roads where there are often long straight stretches, speeding can be particularly hazardous.

Possibility of signal coverage interference: On rural roads, where mobile network coverage is often poorer than in urban areas, drivers may be more prone to distraction trying to establish or maintain a connection. This further increases the risk of accidents.

Overall, using a mobile phone while driving on rural roads can be extremely dangerous and can lead to accidents resulting in injuries due to reduced attention, increased reaction time, and decreased ability to maintain vehicle control. Therefore, it is crucial for drivers to completely avoid using mobile phones while driving to reduce the risk of traffic accidents and injuries.

Overall, all of these indicators have a direct impact on reducing the severity of injuries and the number of casualties in traffic accidents. Increasing seat belt usage rates among passengers and drivers, as well as reducing mobile phone usage while driving, is crucial for improving traffic safety and reducing the number of injuries in traffic accidents. It is important to emphasize that these indicators are essential for the implementation of effective policies and measures to enhance traffic safety. Educating drivers and passengers about the importance of seat belt usage and restricting the use of mobile phones while driving should be priorities in reducing the number of traffic accidents in municipalities throughout the Republic of Serbia.

References

[1] World Health Organization "Global status report on road safety 2023."

[2] "European Parliament P9_TA(2021)0407 EU Road Safety Policy Framework 2021-2030-Recommendations on next steps towards 'Vision Zero,'" 2019.

[3] М. Пљакић, *Анализе у безбедности саобраћаја : методе, модели и алати*, 1st ed. Факултет техничких наука, 2023. Accessed: Mart 20, 2024. [Online]. Available: https://plus.cobiss.net/cobiss/sr/sr/bib/nbs/123361033

[4] M. Pljakić, D. Jovanović, B. Matović, and S. Mićić, "Macro-level accident modeling in Novi Sad: A spatial regression approach," *Accid Anal Prev*, vol. 132, Nov. 2019, doi: 10.1016/j.aap.2019.105259.

[5] L. Breiman, "Random Forests," 2001.

[6] M. M. R. Komol, M. M. Hasan, M. Elhenawy, S. Yasmin, M. Masoud, and A. Rakotonirainy, "Crash severity analysis of vulnerable road users using machine

learning," *PLoS One*, vol. 16, no. 8 August, Aug. 2021, doi: 10.1371/journal.pone.0255828.

[7] P. Wu, X. Meng, and L. Song, "A novel ensemble learning method for crash prediction using road geometric alignments and traffic data," *Journal of Transportation Safety and Security*, vol. 12, no. 9, pp. 1128–1146, Oct. 2020, doi: 10.1080/19439962.2019.1579288.

[8] J. Tang *et al.*, "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review," *Anal Methods Accid Res*, vol. 27, Sep. 2020, doi: 10.1016/j.amar.2020.100123.

[9] M. Пљакић, "УНИВЕРЗИТЕТ У НОВОМ САДУ ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У НОВОМ САДУ."

[10] J. D. Febres, S. García-Herrero, S. Herrera, J. M. Gutiérrez, J. R. López-García, and M. A. Mariscal, "Influence of seat-belt use on the severity of injury in traffic accidents," *European Transport Research Review*, vol. 12, no. 1, Dec. 2020, doi: 10.1186/s12544-020-0401-5.

[11] M. Shimamura, M. Yamazaki, and G. Fujita, "Method to evaluate the effect of safety belt use by rear seat passengers on the injury severity of front seat occupants," *Accid Anal Prev*, vol. 37, no. 1, pp. 5–17, Jan. 2005, doi: 10.1016/j.aap.2004.05.003.

[12] R. Elvik, "Effects of mobile phone use on accident risk: Problems of meta-analysis when studies are few and bad," *Transp Res Rec*, no. 2236, pp. 20–26, Dec. 2011, doi: 10.3141/2236-03.

[13] K. Mizuno and Y. Matsui, "Mizuno 1 EFFECTIVENESS OF SEATBELT FOR REAR SEAT OCCUPANTS IN FRONTAL CRASHES."