Investigating the Casual Effect in Traffic Accident

Badounmara\*, Nandar Win Minb, Ei Ei Moec

*aFaculty of Computer Science, University of Computer Studies (Mandalay), Myanmar*

*b,cDepartment of Computer Science, University of Technology (Yatanarpon Cyber City), Myanmar*

*aEmail: bbadounmar@gmail.com bEmail:nandarwinmin@gmail.com*

*cEmail: eieimoe1989@gmail.com*

# Abstract

An important field of study that attempts to increase road safety and lower the frequency and severity of accidents is the investigation of traffic accidents. For the purpose of creating effective preventative methods and policies, it is imperative to comprehend the underlying causes of traffic accidents. The practice of analyzing the relationship between two or more variables to ascertain whether one has a causal effect on the other is known as causal analysis. Through the integration of mutual information for causality analysis and Support Vector Machine (SVM) for prediction, this system is intended to examine the causative impacts of traffic accidents. The system primarily looks at the reasons behind traffic accidents in Thailand between 2016 and 2019, trying to pinpoint important elements and create practical preventative measures. The system gathers a wealth of information, such as the date, time, and position of each collision as well as information on the type of vehicle, the characteristics of the road, driver demographics, and weather. Mutual information is used to quantify dependencies, highlight important interactions, and study the causal linkages between various components. These analyses show how changes in one variable may have an impact on another. By concentrating on the most important variables, the mutual information and SVM integration improves the system's analytical skills and improves model accuracy and interpretability. As a result, our technology produces thorough reports and visualizations that give stakeholders—such as legislators and traffic safety authorities—actionable insights. These observations aid in the creation of focused initiatives and laws meant to lower the frequency and seriousness of traffic accidents in Thailand.

***Keywords:***  Causal Analysis; Mutual Information; Support Vector Machine; Traffic Accidents.

------------------------------------------------------------------------

*Received: 6/27/2024*

*Accepted: 8/27/2024*

*Published: 9/5/2024*

------------------------------------------------------------------------

\* Corresponding author.

# Introduction

Traffic accidents pose a severe threat to public safety since they have an impact on people's lives and the overall well-being of society. It is crucial to comprehend the underlying factors causing these occurrences in order to enhance road safety and create practical preventative strategies. For the purpose of creating effective preventative methods and policies, it is imperative to comprehend the underlying causes of traffic accidents. The intricate, non-linear interactions between the different elements impacting traffic accidents are sometimes too complex and non-linear for traditional statistical methods to fully capture. The complicated relationships between variables including driver behavior, vehicle attributes, ambient factors, and road infrastructure may be overlooked by these methods, which often rely on linear assumptions. The combination of mutual knowledge and machine learning provides a potent method for identifying and evaluating these complex interactions, which helps to overcome these restrictions.

Conversely, mutual information quantifies the quantity of knowledge gleaned about one random variable via another. It provides information about the potential effects of changing one component on another by quantifying the dependence between variables. Mutual information plays a crucial role in traffic accident investigation by revealing important connections between variables like weather, road parameters, driver behavior, and vehicle kinds. This measure is essential for differentiating between possible causal links and simple correlations. By combining their advantages, machine learning and mutual information integration improve the capacity to analyze data on traffic accidents. By highlighting the most important variables, mutual information helps with feature selection and makes sure that machine learning models concentrate on the most significant elements. Better identification of high-risk scenarios and contributing elements is made possible by this integration, which enhances the prediction models' interpretability and accuracy.

For example, using real-time inputs such as traffic flow and weather, along with historical accident data, machine learning models can identify high-risk zones; mutual information can be used to identify the specific elements that drive these predictions. Additionally, mutual knowledge facilitates the process of drawing conclusions about causality by drawing attention to important relationships that demand greater research. This facilitates the development of models that not only forecast results but also shed light on the causal processes underlying those results. Then, these causal links can be verified using methods like causal graphs and propensity score matching, guaranteeing a thorough grasp of the variables driving road accidents. Policymakers and urban planners who must efficiently allocate resources and prioritize actions will find this greater understanding to be quite helpful.

Therefore, the purpose of this system is to identify complicated elements that contribute to accidents and develop practical preventive methods by examining the causes of traffic crashes in Thailand between 2016 and 2019. For a thorough investigation, extensive data on a variety of traffic accident factors, such as time, location, weather, road features, vehicle kinds, and driver demographics, were gathered. This technique uses mutual information to quantify the dependences between variables and find important linkages, exposing the ways in which weather conditions and driving habits affect the incidence and severity of accidents. Next, using patterns that have been learned from past data, support vector machines (SVMs) are used to forecast accident outcomes. SVMs are capable of categorizing, forecasting severity, and estimating likelihood under particular circumstances. By concentrating on the most important variables, the combination of mutual information and SVM improves the analysis and ensures accurate and understandable forecasts. This system is developed as a prediction system of injury severity outcomes based on the causality theory by examining the causal influence in the traffic accident system. Therefore, by offering insights that guide data-driven actions and regulations, this system can help to improve road safety in Thailand by eventually lowering the frequency and severity of traffic accidents.

This system's primary contribution is the development of a mathematical strategy that maximizes causality power with the fewest possible elements, allowing predictions to be made even when distributions are altered and manipulated. The analysis shows that one efficient way to eliminate redundancy and increase causation power is to use approximated multivariate mutual information across explanatory components. The system differentiates between two different processes: first, it uses mutual information to analyze the causal impacts on injury severity; second, it uses Support Vector Machine (SVM) to forecast accident outcomes. This system carries out multiple jobs to produce the multi-level traffic accident outcomes of "Fatal crash," "Severe crash," and "Minor crash" by looking into the casual effect in traffic accident variables.

This paper includes various sections that encompass the introduction, literature review, the proposed system architecture, data collection, data preprocessing, feature selection using mutual information, classification using support vector machine, experimental results and conclusion.

# Literature Review

# This section offers an extensive review of pertinent literature that contributes to our comprehension of the components contributing to traffic accidents, with particular emphasis on research employing Mutual Information or similar quantitative techniques.

Gururat, H. L., Janvavi, V. and Tanuja, U. [1] predicted the traffic accidents and their injury severities using machine learning techniques. The World Health Organization (WHO) says that India saw 5,18,3626 accidents in the year 2019. Inattentive drivers, broken traffic regulations, inadequate road infrastructure, driving in terrible weather conditions, and other factors all contribute to these traffic accidents, road crashes, and injuries. This research project develops models to choose a group of significant characteristics and to construct a model for

categorizing injury severity. The degree of injuries sustained in auto accidents can be predicted using machine learning algorithms.

Ardakani, S. P., Liang, X., and Mengistu, K. T. [2] presented the road car accident prediction system by using a machine learning enabled data analysis. For the sake of saving lives and creating sustainable cities and communities, the number of incidents must be decreased. Techniques for machine learning and data analysis assess the causes of auto accidents and suggest ways to reduce them. However, given that data on traffic accidents is growing larger and moving more quickly, this system must benefit from big data solutions. This system investigated relevant data aspects, such as accident severity, and provides a predictive model by examining road car accident data patterns.According to Baikejuli, M., Shi, J., and Hussain, M. [3], a thorough analysis of the risk factors associated with fatal collisions involving heavy trucks is conducted. The goal of the research is to discover and quantify different risk characteristics in order to evaluate the accident risk associated with such situations. The study analyzes important characteristics that contribute to accident risk and presents additional risk factors based on a nationally representative sample of recent fatal crashes. The study analyzes the interaction among risk variables and reveals the complexity of fatal heavy-truck incidents by using mutual information theory as a primary analytical tool. The bulk of these mishaps are caused by a variety of simultaneous circumstances, according to the study. The study also demonstrates a clear relationship between the quantity of influencing risk variables and the probability of an accident. Moreover, the study pinpoints distinct multi-factor interactions—that is, interactions including environmental and vehicular factors—that gradually increase the risk of an accident.

# Researchers Rella Riccardi, M., Mauriello, F., and Sarkar, S. [4] investigated the risk variables for serious and deadly car-pedestrian collisions. In order to anticipate crash severity, five non-parametric tools and four parametric models have been developed. Although comparable models have been used in the past to assess the severity of pedestrian injuries, little comparison study has been done with regard to the models' prediction ability. By contrasting these models according to their capacities to pinpoint important explanatory variables and their performance metrics, such as the F-measure, G-mean, and area under the curve, this study thereby closes a gap in the literature on road safety. Data from vehicle-pedestrian collisions that happened between 2016 and 2018 are used in the analysis. It is discovered that parametric models provide results that are simple to understand and exhibit distinct correlations between dependent and independent variables. On the other hand, non-parametric methods show better classification accuracy, find more explanatory variables, and shed light on the relationships between the various elements that affect accident severity. The results of this study imply that the shortcomings of each class of methods can be successfully addressed by combining parametric and non-parametric approaches. This method makes predictions with acceptable accuracy and makes it easier to understand the causes that lead to catastrophic and deadly crashes.

# In order to solve the issue of forecasting traffic accidents, Lee, J., Yoon, T., and Kwon, S. [5] used machine learning techniques, concentrating on Seoul, Korea's Naebu Expressway. The study evaluated the advantages and disadvantages of three distinct machine learning architectures: decision trees, random forests, and artificial neural networks. Three different kinds of data were used: data on road geometry, data on precipitation, and data on past traffic accidents during a nine-year period. Three primary metrics were used to assess the models' performance: the root mean square error (RMSE), mean square error (MSE), and out-of-bag estimate of error rate (OOB). These measurements shed light on each model's accuracy and capacity for prediction. The suggested random forest model demonstrated significantly reduced mean OOB, MSE, and RMSE values, according to the data, demonstrating its better accuracy in forecasting traffic incidents on the Naebu Expressway.

# The application of machine learning-based categorization techniques to model injury severity among vulnerable road users (VRUs), such as walkers, bicyclists, and motorcyclists, was examined by Komol, M. M. R., Hasan, M. M., and Elhenawy, M. [6]. The research makes use of crash data that was obtained from Queensland, Australia, between 2013 and 2019. For the empirical study, Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are used. Thirteen different road crash parameters—which include weather and environment, vehicle and driver circumstances, period, road characteristics and regions, traffic, and speed jurisdiction—are taken into account as input features for training the models. To determine crash severity levels, these classification models are trained and evaluated independently for each of the VRU groups. The findings show that Random Forest classification models perform well in terms of test accuracy for all four modalities of VRU: unified VRU (68.57%), motorcyclists (72.30%), bikers (64.45%), and pedestrians (67.23%).

# In order to identify the critical variables influencing collision severity on Indian state highways, Agarwal, V., Chatterjee, S., and Mitra, S. [7] carried out a study using three non-parametric machine learning approaches: support vector machine (SVM), extreme gradient boosting (XGBoost), and classification and regression tree (CART). Based on their investigation, XGBoost was shown to be the most efficient method, demonstrating remarkable accuracy, quick processing speed, and affordability. The results of the study brought to light the various factors that are important to take into account when assessing how serious accidents are on state roadways. These included the kind of accident, the presence of facilities for pedestrians, the type of weather at the time, the proximity of intersections, the efficiency of traffic management techniques, and the compliance with speed limits.

# Proposed System Architecture

The relationship between two occurrences is governed by causality, also known as causal impact, which establishes how one event can directly affect another. Understanding causality in the context of injury severity is essential for precise assessment and intervention. In this profession, causality analysis has historically been predicated on single-factor analysis, in which each element is looked at separately to ascertain how it affects the outcomes of injuries. Nonetheless, there are a lot of drawbacks to this strategy. Due to the complex nature of traffic accidents, which involve multiple interacting elements, individual factor selection may impair the accuracy of complicated decision problems and cannot improve the assessment of injury severity. As such, applying a multi-factor analysis-based strategy is a more effective approach. We can provide findings that are more thorough and objective by taking into account a wide range of variables and how they interact. The goal of this research is to create a causal analysis method that can effectively predict injury severity outcomes by identifying a subset of minor variables with the highest predictive potential. In traffic accident scenarios, we can make more informed decisions and more successful preventive tactics by utilizing advanced approaches like machine learning and mutual information to discover the most relevant components and their interactions. The block diagram of the system is described at Figure 1.

#

**Figure 1:** Proposed System Design for Investigating the Casual Effect in Traffic Accident

* 1. ***Data Collection***

In Thailand, the Department of Highways (DOH) has meticulously collected and compiled a comprehensive dataset spanning the years 2016 to 2019, encompassing an extensive array of information pertinent to traffic incidents. This dataset is particularly noteworthy due to its scale and depth, featuring 91 distinct attributes and recording a total of 66,113 occurrences of traffic-related events. Each recorded incident is detailed through 90 feature columns, which represent a wide range of variables that might influence the nature and severity of traffic accidents. Data distribution of dataset with detail is shown in Figure 2.



**Figure 2:** Data Distribution of Dataset with Detail

The remaining 90 columns are factors, while the “FATAL\_CRASH”, ‘SEVERE\_CRASH’ and “MINOR\_CRASH” columns are the target variables. Target variables are shown in Table 1.

**Table 1:** Target Variables of Traffic Accident Dataset

|  |  |
| --- | --- |
| Target Variables | Description |
|
| FATAL\_CRASH | At least one person (driver or passenger) killed (within 30 days) by injuries sustained in the crash |
| SEVERE\_CRASH | At least one person injured and admitted to hospital but no fatalities |
| MINOR\_INJ\_CRASH | At least by one person requiring medical care but no fatalities or injuries requiring hospitalization |

There are 26,445 events with a value of 1 and 39,667 events with a value of 0 in the target column. With the exception of the following columns: "Kilometer," "Damage\_pub," and "Damage priv," most feature columns have binary values. Traffic accident factors are shown in Table 2.

**Table 2:** Traffic Accident Factors

|  |  |  |
| --- | --- | --- |
| **Factors** | **Description** | **Value** |
| **1** | **0** |
| kilometer | Kilometer |  |  |
| NO\_MEDI | Road was not divided by median island | Yes | No |
| FLUSH | Road was divided by flush median island | Yes | No |
| RAISED | Road was divided by raised median island | Yes | No |
| DEPRESS | Road was divided by depressed median island | Yes | No |
| BRIDGE | Crash occurred on a bridge | Yes | No |
| TUNNEL | Crash occurred on an underpass road segment | Yes | No |
| … | … | … | … |
| SERIOUS\_M | Number of males with serious injury (>15 year old) | Yes | No |

# *3.2 Data Preprocessing*

As part of the initial stage of data preprocessing, the dataset's "route," "YEAR," "SEVERE\_CRASH," "MINOR\_INJ\_CRASH," and "PDO" columns as well as any others that are superfluous are eliminated. Next, column-by-column categorical variable identification is carried out for standardization. Columns with high values are the feature columns, including "Kilometer," "Damage\_pub," and "Damage priv," while the other columns have binary values (0 or 1). The characteristics in the categorical columns are normalized using the feature scaling method, more precisely standardization. Feature standardization techniques can be used to convert the variables or features in a dataset into a common scale. Ensuring that all features contribute equally to the analysis is crucial to prevent certain variables from dominating due to differences in their original scales. The standardizing formula is as follows:

$X\_{new}=\frac{X\_{i}-X\_{mean}}{σ}$ (1)

where, $X\_{i}$is ith feature value, $X\_{mean}$is the mean of training samples X, and $σ$ is the standard deviation. The values are not limited to a certain range in this case.

# *3.3 Feature Selection using Mutual Information*

Mutual information is a subset of relative entropy, or Kullback-Leibler divergence, which is a more general concept. In addition to measuring how different two probability distributions are from one another, relative entropy also expresses how much information one random variable can tell you about another. In essence, it quantifies the amount of information that is shared between two variables—a critical component in comprehending data interdependence. The degree of uncertainty or unpredictability in a random variable is described by the notion of entropy, which is fundamental to mutual information. The entropy H(X) for a discrete random variable X is formally defined as:

$H\left(X\right)=-\sum\_{x\in X}^{}p\left(x\right)log\_{2}p\left(x\right)$ (2)

where p(x) is the probability mass function of X. This formula sums the products of each probability p(x) and the logarithm of that probability, providing a measure of the average information content or uncertainty associated with the random variable X.

The joint entropy H (X, Y) of a pair of discrete random variables (X, Y) with a joint distribution p (x, y) is defined as follow:

$H\left(X,Y\right)=-\sum\_{x\in X}^{}\sum\_{y\in Y}^{}p\left(x,y\right)log\_{2}p(x,y)$ (3)

If (X, Y)∼p(x, y), the conditional entropy H(Y|X) is defined as:

$H\left(Y|X\right) =\sum\_{x\in X}^{}p\left(x\right)H\left(X=x\right)$ (4)

 $=-\sum\_{x\in X}^{}p\left(x\right)\sum\_{y\in Y}^{}p\left(x\right)log\_{2}p\left(x\right) $

 $=-\sum\_{x\in X}^{}\sum\_{y\in Y}^{}p(x,y)log\_{2}p\left(x\right) $

The mutual information I(X; Y) is defined as the relative entropy between the joint distribution and the product distribution p(x)p(y) when two random variables, X and Y, are considered. Mutual information builds on this concept by examining the relationship between two variables, X and Y. It quantifies how much knowing the value of one variable reduces the uncertainty about the other. In mathematical terms, mutual information I(X;Y) between two discrete random variables X and Y is defined as:

$I\left(X;Y\right)=\sum\_{x\in X}^{}\sum\_{y\in Y}^{}p(x,y)log\_{2}\frac{p(x,y)}{p\left(x\right)p(y)}$ (5)

According to these definitions, we have the following theorem:

I(X;Y) =H(X)-H(X│Y) (6) $I\left(X;Y\right)=H\left(Y\right)-H(Y|X)$ (7) $I\left(X;Y\right)=H\left(X\right)+H\left(Y\right)-H(X|Y)$ (8) $I\left(X;Y\right)=I(Y;X)$ (9)where, I(X;Y) is Mutual Information between X and Y, H(X) is Entropy for X, and H(X|Y) is the conditional entropy of X given Y. The outcome is measured in bits on a scale from zero to one. Figure 4.2 shows a Venn diagram that represents the relationship between H(X), H(Y), H(X,Y), H(X|Y), H(Y|X), and I(X;Y). The intersection of the information in X and the information in Y is represented by the mutual information I(X;Y).



**Figure 3:** The relationship between entropy and mutual information

# *3.4 Classification using Support Vector Machine*

Support vector machine (SVM) is a supervised learning model that is employed for binary classification. But, they can be modified to accommodate multi-class problems. Finding the best hyperplane to divide the data points of several classes in a high-dimensional feature space is the fundamental notion behind support vector machines (SVM). Maximizing the margin between classes is the goal of SVM, which improves generalization and robustness to unknown data.

Step-by-step explanation of how SVM works:

* + - Data Representation
		- Feature Space and Hyperplane
		- Margin and Support Vectors
		- Training
		- Soft Margin and C Parameter
		- Kernel Trick
		- Testing and Classification

SVM involves optimizing parameters that are the regularization parameter (C), the width of the Gaussian kernel and kernel-specific parameters. SVM classifier is shown in Figure 4.



**Figure 4:** Support Vector Machine (SVM) Classifier

This system uses SVM to assess and quantify the influence of many factors that lead to traffic accidents. Through the use of a mutual information feature selection method, the most influential features are identified and prioritized, leading to a more effective understanding and prediction of traffic accidents. First, the top attributes with the highest predictive power for traffic accidents are chosen based on mutual information. By ensuring that the most pertinent factors are taken into account, this phase improves the precision and effectiveness of the analysis that follows. These carefully chosen characteristics comprise a range of elements that are essential to comprehending the dynamics of traffic accidents, including weather, driving habits, vehicle kinds, and road conditions. After feature selection, the selected features are subjected to an SVM in order to build a prediction model. Because it has been trained on historical data, this model is able to identify patterns and linkages between different elements and how they relate to traffic incidents. The SVM method looks for the best hyperplane with the largest margin that divides the various classes (such as accident vs. no accident or severe vs. mild accident) during the training phase. The model's capacity for accurate prediction and data generalization is improved by this ideal separation. The resultant SVM model offers insights into the causal links between these elements and accident occurrences in addition to predicting the risk of traffic accidents based on the discovered relevant factors.

# Experimental Results of the System

In this system, we optimize the performance of the SVM model by focusing on tuning particular hyperparameters. The hyperparameters that we used in our investigation are the kind of kernel, C, and gamma. A single training example's influence is defined by its gamma value, where larger values indicate that closer points have a greater influence on the decision border. In order to avoid overfitting, the parameter C manages the trade-off between decreasing the weights' norm and obtaining a low error on the training set. In order to keep things simple and efficient, we set gamma to 1, C to 1, and select a linear kernel for our analysis.

The Department of Highways (DOH) created the Highway Accident Information Management System (HAIMS) dataset, which is split into two sections: 80% (59,501 samples) for training and 20% (6,612 samples) for testing. This division offers a distinct dataset to objectively assess the model's performance in addition to ensuring that it has enough data to learn from. "FATAL\_CRASH" is the goal variable and all 91 columns are used as features in the SVM classifier's training process. With the given hyperparameters, the linear kernel makes it easier for the model to identify the ideal hyperplane for successfully dividing the classes (such as deadly crashes from non-fatal crashes). Based on the provided features, the trained model predicts traffic accidents with an accuracy of 88% on the test dataset, indicating its robustness and dependability. Figure 5 presents the model's performance metrics together with the results, supporting the linear kernel's efficacy in this particular context and the selection of hyperparameters. This high accuracy highlights how crucial it is to adjust hyperparameters and choose the right kernels in order to improve the predictive ability of SVM models for the investigation of traffic accidents.



**Figure 5:** Assessment Report on The Test Dataset using SVM with All Features

While the model's accuracy is good, we think our suggested approach can be made even better by using the mutual information algorithm to remove superfluous characteristics. Even if the accuracy of our existing model is 88%, we believe that adding mutual information to feature selection can improve and streamline our process even more. This strategy should increase the model's accuracy, lower its processing requirements, and make it more useful for predicting traffic accidents overall. Our goal is to create a more robust and dependable traffic safety analysis tool through constant improvement of our approach, which will ultimately lead to more informed decision-making and successful accident prevention tactics. We are able to prioritize and choose features according to their informative value by figuring out the MI score for every feature connected to the target variable.



**Figure 6:** Mutual Information (MI) Scores of Each Feature

The mutual information scores for all features are graphically displayed in Figure 6, providing a clear visualization of the relative importance of each feature. This figure helps to identify which features should be prioritized in the modeling process and which can be safely omitted. Overall, the use of mutual information for feature selection is a strategic method that enhances the effectiveness of our model. By focusing on the most significant features, we can develop a more precise and optimized predictive tool for traffic accident analysis, ultimately contributing to more accurate and reliable predictions and better-informed traffic safety interventions.

The results indicate that the mutual information (MI) scores for the features range from 0 to 0.0130, highlighting varying degrees of relevance among the features in relation to the target variable. Specifically, certain features such as "MAIN," "CONSTRUC," "road\_lane," "road\_isle," "DEPRESS," "road\_surface," "CON\_PAVE," and "STRAIGHT" (corresponding to features 2, 5, 6, 8, 12, 15, 16, and 18) display MI scores of zero. This suggests that these features do not provide any significant information about the target variable and, therefore, can be considered irrelevant for the prediction model. On the other hand, features such as "REAR\_END," "NIGHT\_LIGH," "TREE\_WE," "damage\_pub," and "BIKE" (matching features 67, 56, 70, 46, and 60) exhibit the highest MI scores, with values of 0.00787142, 0.008462429, 0.009157378, 0.012548651, and 0.013015467, respectively.

These high MI scores indicate a strong relationship between these features and the target variable, implying that they are highly informative for predicting traffic accidents. The feature "BIKE," with the highest MI score of 0.013015467, emerges as the most significant predictor among the analyzed features. This underscores its critical role in the model and highlights the necessity of including it in the feature set to improve the predictive accuracy of the model. Similarly, other features with high MI scores also contribute valuable information that enhances the model's ability to make accurate predictions.

Among the features, only 37 surpass the average mutual information (MI) score of 0.001684824, representing the mean MI score across all features. This indicates that these 37 features have higher than average informational value relative to the target variable and are thus deemed most relevant for inclusion in the predictive model. To assess the impact of these selected features on the model's performance, we employ a Support Vector Machine (SVM) classifier, evaluating the accuracy across different sets of features with k values ranging from 30 to 37. The results show that using these top features, the SVM classifier achieves an accuracy of approximately 90%, demonstrating the effectiveness of the mutual information-based feature selection in enhancing the model's predictive capability. This high accuracy indicates that the selected features are highly informative and contribute significantly to the model's ability to accurately classify fatal traffic incidents. Figure 7 illustrates the relationship between the number of features (k values) and the model's accuracy, showing a consistent performance as the number of features increases up to 37. This figure provides visual confirmation that incorporating more than 30 features and up to 37 does not significantly degrade the model's performance, suggesting a robust selection of features that optimally balance complexity and predictive power.



**Figure 7:** The Model Accuracy with The Number of Features (k) Ranges from 30 to 37

The 37 selected features for analysis are as follows: ‘kilometer’, ‘road\_spare’, ‘MAIN’, ‘PARAL’, ‘road\_cond’, ‘road\_lane’, ‘road\_direction’, ‘road\_isle’, ‘NO\_MEDI’, ‘RAISED’, ‘DEPRESS’, ‘BARRIER’, ‘road\_traffic’, ‘road\_surface’, ‘char\_horizontal’, ‘STRAIGHT’, ‘char\_vertical’, ‘char\_intersection’, ‘char\_openaccess’, ‘char\_connect’, ‘char\_other’, ‘env\_surface’, ‘WET\_SURF’, ‘env\_status’, ‘env\_weather’, ‘env\_light’, ‘NIGHT\_LIGH’, ‘DA\_RIKNESS’, ‘cause’, ‘EXCESS\_SPEED’, ‘SUDDEN\_CROS’, ‘damage\_private’, ‘VLOOKUP’, ‘PEDESTRIAN’, ‘REAR\_END’, ‘SIN’, ‘province’, ‘holiday’, and ‘BIKE’. These features cover a comprehensive range of variables that influence traffic conditions and accident scenarios.



**Figure 8:** Assessment Report on The Test Dataset using SVM with Selected Features

Furthermore, the detailed analysis of the experiment using 37 features to classify fatal traffic incidents with SVM is depicted in Figure 8. This figure presents a comprehensive evaluation of the model's performance metrics, including precision, recall, F1-score, and the overall accuracy. The high accuracy rate achieved with these 37 features validates the effectiveness of mutual information as a feature selection method and underscores the reliability of SVM as a classification tool in this context.

**Table 3:** Accuracy Results of the System

|  |  |  |
| --- | --- | --- |
| Crash Type | Correct Rate  | Error Rate |
|
| FATAL\_CRASH | 96.1% | 3.9% |
| SEVERE\_CRASH | 92.5% | 7.5% |
| MINOR\_INJ\_CRASH | 96.4% | 3.6% |

The accuracy result of the system is shown in Table 3. The table lists three types of crashes: FATAL\_CRASH, SEVERE\_CRASH, and MINOR\_INJ\_CRASH, along with their respective correct and error rates. The correct rate indicates the percentage of times the crash type was correctly identified, while the error rate shows the percentage of misidentification. FATAL\_CRASH has a high correct rate of 96.1%, indicating a high level of accuracy in identifying fatal crashes. The error rate is relatively low at 3.9%, suggesting few instances of misidentification. SEVERE\_CRASH has the lowest correct rate among the three types, at 92.5%. This lower correct rate indicates that severe crashes are more challenging to identify accurately compared to the other types. The error rate is the highest at 7.5%, suggesting more frequent misidentifications. MINOR\_INJ\_CRASH has the highest correct rate at 96.4%, making it the most accurately identified crash type. The error rate is the lowest at 3.6%, indicating the least number of misidentifications among the three types. Experimental results of the system are shown in Figure 9.

**Figure 9:** Experimental Results of the System

MINOR\_INJ\_CRASH has the highest correct rate (96.4%), followed by FATAL\_CRASH (96.1%), and SEVERE\_CRASH has the lowest correct rate (92.5%). The correct rate difference between MINOR\_INJ\_CRASH and FATAL\_CRASH is minimal (0.3%), while the difference between MINOR\_INJ\_CRASH and SEVERE\_CRASH is more significant (3.9%). SEVERE\_CRASH has the highest error rate (7.5%), indicating it is the most challenging to identify accurately. FATAL\_CRASH and MINOR\_INJ\_CRASH have similar error rates, with MINOR\_INJ\_CRASH slightly better at 3.6% compared to FATAL\_CRASH's 3.9%.

# Conclusion

This system demonstrates how well mutual information and SVM work when examining the causative relationship in traffic accidents. This method not only improves predictive model accuracy, but it also provides a data-driven foundation for creating all-encompassing traffic safety plans. By offering a methodological framework that is flexible and adaptable to different situations, this approach advances the field of traffic safety by eventually attempting to lower the frequency and severity of traffic accidents. Through the reduction of complexity and enhancement of the classification model's performance, this method allows for a more effective use of computational resources by concentrating on the most informative features. By addressing the underlying causes of traffic accidents, targeted interventions and policy actions can be informed by the insights obtained from this methodology. This system is a flexible tool for road safety research since it can be tailored to different datasets and circumstances. This approach aids in the creation of thorough and successful road safety policies, which eventually seek to lower accident rates and improve overall traffic safety, by offering a data-driven basis for comprehending and preventing traffic incidents.

# References

[1] Gururat, H. L., Janvavi, V. and Tanuja, U., "Predicting Traffic Accidents and Their Injury Severities using Machine Learning Techniques", WIT Press, vol. 6, no. 4, pp. 363-377, 2022.

 [2] Ardakani, S. P., Liang, X., and Mengistu, K. T., "Road Car Accident Prediction using a Machine-Learning-Enabled Data Analysis", Sustainability, March, 2023.

 [3] Baikejuli, M., Shi, J., and Hussain, M., “A study on the probabilistic quantification of heavy-truck crash risk under the influence of multi- factors”, Accid Anal Prev. 2022 Sep; 174:106771. doi: 10.1016/j.aap.2022.106771. Epub 2022 Jul 13. PMID: 35841687.

 [4] Rella Riccardi, M., Mauriello, F., and Sarkar, S., “Parametric and Non-Parametric Analyses for Pedestrian Crash Severity Prediction in Great Britain”, Sustainability, 2022.

 [5] Lee, J., Yoon, T., and Kwon, S., “Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms”, Seoul city study. Appl. Sci. 2020, 10, 129.

 [6] Komol, M. M. R., Hasan, M. M., and Elhenawy, M., “A. Crash severity analysis of vulnerable road users using machine learning”, PLoS ONE, 2021.

 [7] Agarwal, V., Chatterjee, S., and Mitra, S., “Crash severity analysis through nonparametric machine learning methods”, Journal of Eastern Asia Society for Transportation Studies, 2019.

 [8] Mohire, S., "A Machine Learning Based Comparative Analysis of Accident Severity Prediction Mechanism in USA", National College of Ireland Project Submission Sheet School of Computing, December, 2022.

 [9] Gururat, H. L., Janvavi, V. and Tanuja, U., "Predicting Traffic Accidents and Their Injury Severities using Machine Learning Techniques", WIT Press, vol. 6, no. 4, pp. 363-377, 2022.

 [10] Ardakani, S. P., Liang, X., Mengistu, K. T., "Road Car Accident Prediction using a Machine-Learning-Enabled Data Analysis", Sustainability, March, 2023.

 [11] Bhuiyan, H., Ara, J. and Hasib, K. M., "Crash Severity Analysis and Risk Factors Identification based on an Alternate Data Source: A Case Study of Developing Country", Scientific Reports, 2022.

 [12] Mao, Y., Zhang, J., and Robbins, G., "Factos Affecting the Severity of Motor Vehicle Traffic Crashes involving Young Drivers in Ontario", Research Gate, pp. 183-189, 2011.

 [13] Rahman Komol, M. M, Hasan, M. M, and Elhenawy, M., "Crash Severity Analysis of Vulnerable Road Users using Machine Learning", PLoS ONE, vol. 16, no. 8, August, 2021.