Micro-Level Driving Behavior Analysis for Accident Prediction Using Machine Learning

Gautam Karma [1*], Himanshu Bagwaiya [2]
[1, 2] Dept. of Computer Science & Engineering Prestige Institute of Engineering Management & Research Indore Indore, India

* Corresponding Author: **Gautam Karma**

Article Info

ISSN (Online): 2582-7138
Impact Factor (RSIF): 8.04
Volume: 07
Issue: 03
Received: 27-04-2026
Accepted: 25-05-2026
Published: 23-06-2026
Page No: 1145-1154

Abstract
Road accidents do not generally take place because of a single isolated parameter; in contrast, they occur as a result of the combination of various micro-level driving parameters. The comprehension of the collective influence of such minute parameters assumes paramount importance for the development of a credible predictive model. This paper introduces a machine learning model for the prediction of road accidents employing micro-level driving parameters. The performance of the model is tested on a publicly available database of 2,000 samples of a synthesized driving pattern featuring five driving variables and a target variable defining the accident label [21].
As the dataset has serious class imbalance problems, the Synthetic Minority Over-sampling Technique (SMOTE) is only employed on the training dataset to alleviate bias towards the dominant class [10, 11]. A series of traditional machine learning algorithms, such as Logistic Regression, K-Nearest Neighbor Classifier, Support Vector Machines, and Random Forest Classifier, are trained and compared under the same experimental conditions. The performance of various classifiers is evaluated in terms of accuracy, minority class F1-score, confusion matrix evaluation, and five-fold stratified cross-validation to obtain consistent results [12, 13].
The results of experiments prove that ensemble-based learning with better performance by Random Forest algorithm outperforms baseline models with a remarkable ability to model interaction effects of micro-level driving features [4, 5]. Nevertheless, it is strongly assumed that near perfect results of this study owe little to true generalization performance of an algorithm with its strong reliance on specific patterns encoded within a synthetic environment. As for current results, it must be admitted that these results confirm more of an existence proof rather than an expert judgment of performance of an algorithm for predicting road accidents.

Keywords: Accident Prediction, Driving Behavior Analysis, Machine Learning, Synthetic Dataset, Class Imbalance, Random Forest

1. Introduction
Road traffic accidents have always been a major concern for global safety, making a significant contribution to injuries, deaths, and economic loss on a global level [1]. Despite the development of safe vehicles and road systems, accident generation is still impacted by a multifaceted combination of driving patterns, environmental, and situational awareness factors. Notably, traffic accidents do not happen immediately but before have warning signs or micro-level driving patterns like high speed, low alertness, mobile phone usage, poor road conditions, and poor visual conditions [2, 3].
Recent works have proved that analysis of the above-listed microscopic driving variables could be helpful for gaining insight into accident risk and road safety for drivers [6, 7]. These microscopic variables are typically nonlinear in nature and are difficult to analyze using statistical analysis models, which is where machine learning models could be of use in this scenario related to accidents on roads [5, 9].

There have been attempts to use a variety of machine learning approaches for estimating the risk of accidents, including Logistic Regression, Support Vector Machines, and ensemble methods [2, 8]. Of these, the use of ensemble methods like Random Forests has gained considerable attention for its ability to handle non-linear interactions between variables, besides being more robust against over-fitting [4]. Nevertheless, the problem of dealing with a skewed class distribution, where examples of accidents are much fewer compared to non-accidents, has remained a challenge in accident prediction studies [10, 11].

Another factor that needs to be considered is the type of dataset being used for developing and testing the model. Although realistic datasets are practical in nature, they are also hard to access, noised, and have limitations in terms of being private. For this purpose, synthetic datasets are commonly being used in initial phase research for facilitating controlled research for reproducibility purposes [5]. The datasets are very helpful for examining the possibility of accident patterns being extracted from micro-driving factors, although limitations in terms of generalizability are also to be recognized.

With these thoughts in mind, this paper proposes a machine learning framework for accident prediction on a publicly available synthetic micro-level driving dataset of 2,000 examples with a total of five driving-related attributes [21]. Several machine learning algorithms are trained and tested in a fair way on an equal footing on identical experimental settings, with class imbalances handled using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm solely on training examples [10]. Several criteria are used for a comprehensive assessment of model results: accuracy, minority class F1 measure, analysis of a matrix of confusion, and application of a fivefold stratified validation to warrant reliability of processing [12, 13]. Through a consequent discussion on results, this paper can contribute to a clear transparent result on the feasibility of machine learning models in micro-level accident prediction.

2. Related Work

Road accident forecasting is an area that gained considerable attention from researchers owing to its practical applicability to transport safety systems. The early works on road accident forecasting usually employed statistical and rule-based methodologies to investigate accident causes. For instance, researchers concentrated on individual aspects such as violating speeds on roads and demographic characteristics of drivers. Nevertheless, such methodologies were usually inadequate to understand complex nonlinear relationships among different variables associated with road transport that affect road accident possibility to great extent [2, 3].

With the growing capabilities offered by advancements in machine learning methods, scientists have been increasingly working on data-driven models for modeling driving behavior and accident risk prediction. Antoniou *et al.* [3] discussed the correlation between driving behavior and accident involvement and stressed the relevance of behavior-based indicators for analysis purposes. Along similar lines, [6] showed the potential for classifying driver behavior based on data derived by sensors and emphasized the importance of micro-level indicators for risk analysis.

Several classic machine learning techniques have been utilized for the prediction of accidents, such as Logistic Regression, Support Vector Machines, and decision tree models [2, 7, 8]. Although linear models like Logistic Regression provide interpretability, they can be restrictive when dealing with nonlinear relations, as found in the context of driving, within the data. Support Vector Machines can improve the robustness of prediction by identifying complex boundaries but might face issues on a larger number of data samples [8]. These shortcomings led to the use of ensemble learning techniques, where a combination of models is used to make predictions.

Ensembles-based methods, specifically RF, have presented high performance in traffic safety and prediction models because they can efficiently capture the interaction among high-dimensional feature space with inherent robustness to overfitting [4, 5]. Wickramasinghe *et al.*, in their study [7], observed that ensemble methods outperform single classifiers while making predictions in traffic accidents, particularly with scenarios characterized by heterogeneous driver and environmental attributes, thus suggesting the applicability of the method to detect interactions among micro-level attributes related to driving behavior.

One of the key challenges pointed out by various research works is class imbalance, where accident examples are significantly less compared to other samples of general driving activities. Moreover, it may result in biased models that prefer predicting the majority class, thereby resulting in weak accident class detection capabilities [10, 11]. To overcome such class imbalance problems, data-level solutions like the Synthetic Minority Over-sampling Technique (SMOTE) have been used by many researchers to balance class distributions [10].

Another newly explored area of research related to accident prediction deals with dataset availability and experimental reproducibility. Although real-world datasets are more practical in terms of representing actual driving conditions, there are also constraints like privacy issues, missing entries, and unregulated noises in real-world datasets. As a result, simulated datasets are utilized in initial experiments to enable experimentation under a controlled environment for systematic comparisons of models [5]. Nevertheless, it has also been supported in earlier research that simulated results require careful validation on real-world datasets before being put to practical use [7, 11].

In short, the state of existing research in the area shows that machine learning models and ensemble models offer efficacy in accident prediction tasks, although they also point out difficulties associated with class imbalance and real-world data. To build upon that research, the work conducted in this study investigates various machine learning models for use in accident prediction tasks using a micro-level driving data set.

3. Dataset Description

In this experiment analysis, a publicly available synthetic micro-level driving dataset is used for evaluation purposes by the authors with reference to Zenodo platform [21]. This dataset, named public_micro_dataset_2000.csv, entails 2,000 examples where each example uniquely describes a driving experience using a number of attributes related to driving. This experiment analysis aims to] create a simulated accident risk using a number of micro-level driving attributes.

Every observation in the data set is defined through the use of five numerical variables that describe vehicle speed, driver alertness level, mobile usage, road surface, and ambient sunlight intensity. Altogether, these variables encompass essential elements that tend to be directly or indirectly linked

with accidents during road safety analysis tasks [2, 6]. The target variable is binary and is referred to as "label_accident," and it takes a value of 0 for a non-accident observation and 1 for an observation related to an accident.

The class imbalance in the original dataset is quite evident, as it consists of around 84.4% of the non-accident class and the remaining 15.6% of the accident class. This kind of class imbalance is very common in accident prediction tasks in which the critical events occur much less often compared to normal driving patterns [10, 11]. However, in order to counter the effects of class imbalance and ensure that the model is not biased toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) method is used only on the training data [10]. The test data is left unaltered.

The data is then split using an 80:20 stratified Train-Test Split to ensure that the data represented in each split maintains its class distribution. The training data will then have a leveled class distribution after oversampling. The values of all features are encoded in numbers and normalized before training models.

However, it is necessary to highlight that these data have been generated synthetically or artificially, following deterministic rule-based patterns. Although these have proved beneficial for facilitating experimentation, they have limitations regarding capturing performances typical of real-world environments. Thus, these data may be considered suitable for conducting analyses related to feasibilities but cannot be applied directly for real-time operations. This is being insisted upon to maintain proper interpretations of outcomes, it is assumed, owing to references cited under [21].

4. Methodology

This section will discuss the methodological framework used for accident prediction based on micro-level driving factors. The proposed method includes a structured paradigm of feature representation, handling class imbalance, model selection, training procedure, and evaluation. This method will ensure a fair comparison of models, reproducibility of outcomes, and clarity of interpretation.

- **Feature Representation**

Every car incident is characterized with five indicators at the micro-level, which consist of car speed, driver alertness, use of the mobile device, road type, and surrounding sunlight strength. These indices collectively refer to both behavioral and environmental factors known to affect the possibility of accidents occurring [2, 6]. Continuous indicators are normalized to remove scale dominance, while binary attributes, which refer to the use of phones, remain unchanged. Every machine learns identically from every feature distribution because each feature has the same range.

- **Class Imbalance Handling**

Accident prediction is characterized by imbalanced data distribution owing to differences in the occurrence of accidental and normal samples by several orders of magnitude. In the used data set, samples of accidental data do not exceed 15.6%. However, training machines on imbalanced data increases the risk of biased prediction results that tend towards major classes rather than minorities [10, 11]. In this context, the problem of biased data distribution can be overcome by using SMOTE for synthesizing new samples of the minority class by interpolation of sample distribution based on Samples of accidental data that belong to the minority class of data samples [10].

- **Model Selection**

1. Four classical machine learning models are chosen for comparison and evaluation:
2. Logistic Regression, used as a linear probabilistic model for binary classification [9].
3. K-Nearest Neighbors (KNN), an instance-based learning algorithm that relies for classification on the proximity of points within the feature space;
4. Support Vector Machine (SVM), for which a maximum-margin decision boundary is built to distinguish among classes. The algorithm is applicable to
5. Random Forest. An ensemble learning algorithm involving many decision trees. Each decision tree is built using different subsets of data sampled with replacement and different random subsets of features. The general idea is to average out

Random Forest is a suitable technique for this particular task because of its capacity to handle non-linear correlations between the features, its resistance to overfitting, and stability when dealing with noise [4, 5]. Each of the models is trained on the same set of features.

- **Learning and Validation Strategy**

The set of data is then split into a subset for training and testing, using an 80:20 stratified split. Models are trained independently based on a balanced version of the training set using SMOTE. To determine the ability of models to generalize, it is necessary to use cross-validation. To ensure accurate estimates of generalization, a 5-fold stratified cross-validation technique is used. There is a need to follow best practices in comparing classifiers, as discussed in best practices for comparing classifiers [12, 13].

- **Evaluation Metrics**

Model performance has been measured using a set of metrics that are complementary in nature. Accuracy has been measured to test its general classification ability. At the same time, F1-score for minority class has been highlighted to test the efficacy of accident detection. Further, confusion matrix analysis has been performed to provide insights to model classifications made along with distribution of errors. This combination of metrics has provided a fair assessment of both majority and minority class performance of models in an experiment, respectively [11].

5. Experimental Setup

This section details the practical implementation of the proposed methodology, including data preparation, baseline definition, and experimental configuration.

- **Dataset Preparation**

The experiments are performed on a publicly available synthetic micro-data dataset for driving, with 2,000 instances and five features related to driving, made available on the Zenodo platform by the authors in the paper [21]. The dataset is generated using synthetic rules that are deterministic, allowing for controlled experiments. All variables are converted to numerical values and standardized for training.

- **Baseline Models**

To benchmark the model's performances, two baseline predictors are introduced:

• Majority Class Predictor, where all instances are classified into the non-accident class;

• Random Predictor, where class labels are predicted randomly through uniform probability.

These baselines set the lower bound of the prediction performance and show the vulnerability of the prediction strategies in imbalanced accident prediction tasks [12].

- **Model Training Configuration**

Four machine learning models, namely Logistic Regression, KNN, SVM, and Random Forest, are trained on equal feature sets and data splits. The hyperparameters are set constant through various experiments. Random Forest is implemented with 100 decision trees, and default settings are employed for SVM and KNN.

Table 1: Model Hyperparameter Settings

Model	Key Parameters
Random Forest	n_estimators = 100
SVM	RBF kernel
KNN	k = 5
Logistic Regression	L2 regularization

- **Performance Evaluation Protocol**

Model performance metrics are measured using accuracy, minority class F1-score, and a study of the confusion matrix on an unseen test set. Moreover, cross-validation using a 5-fold stratified method has been used to test robustness and variance of model performance. The performance obtained through cross-validation has been visualized using box plots.

- **Interpretation Considerations**

Because a synthetic dataset's behavior can be deterministic, it has to be treated with some caution if near-perfect results are obtained with a specific model like Random Forest. Actually, what these results particularly show is a demonstration of feasibility in methodology, rather than guaranteed success in real-world generalization. This setup experimentally aims to provide a clear foundation for future validation in real-world driving datasets [5, 21].

6. Results and Discussion

- **Feature Distribution Analysis**

The distributions of thefive input features, which include speed, alertness of the driver, usage of the phone, road surface, and sunlight, are shown in Fig. 1 through Fig. 5 respectively. The graph shows a controlled spread of the data for each of the features to ensure realism in the data.

Phone use and other binary variables have very clear bimodal distributions, while continuous variables like speed, alertness, and sunlight have uniform distribution within their defined range. The distributions of target variables depicted in Fig.6 illustrate how there is an extreme imbalance of classes with cases of accidents being part of the minority classes.

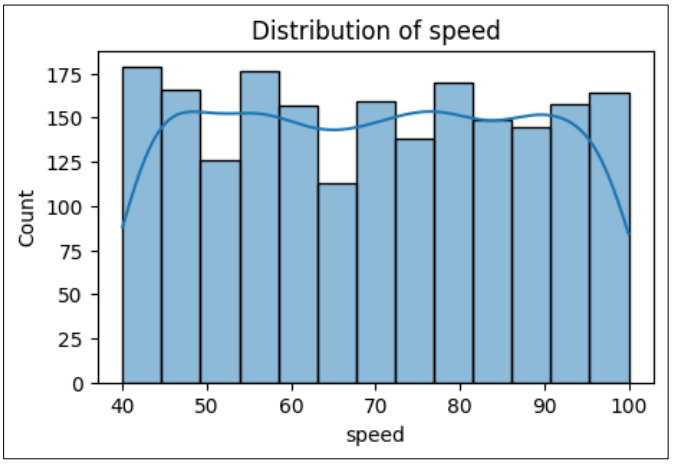

Fig 1: Distribution of speed

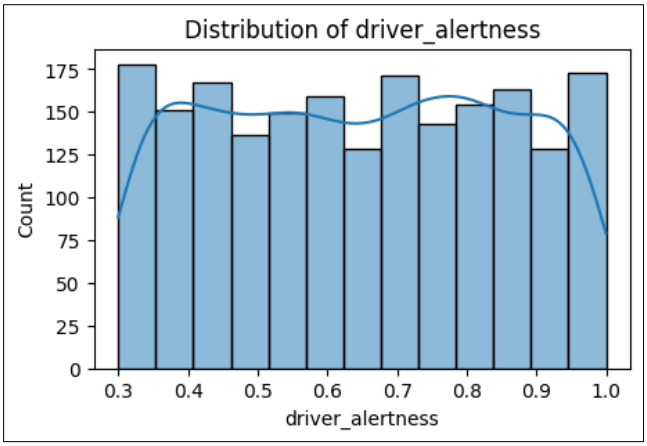

Fig 2: Distribution of driver alertness

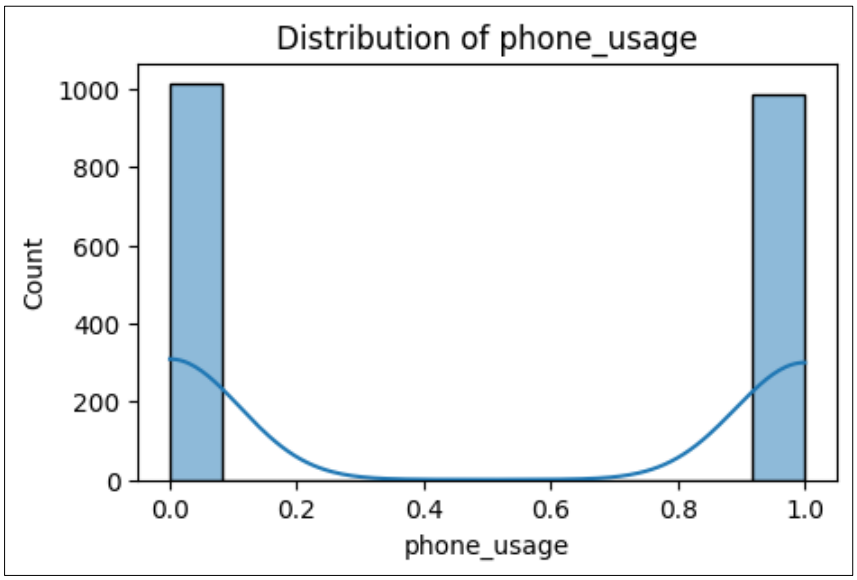

Fig 3: Distribution of phone usage

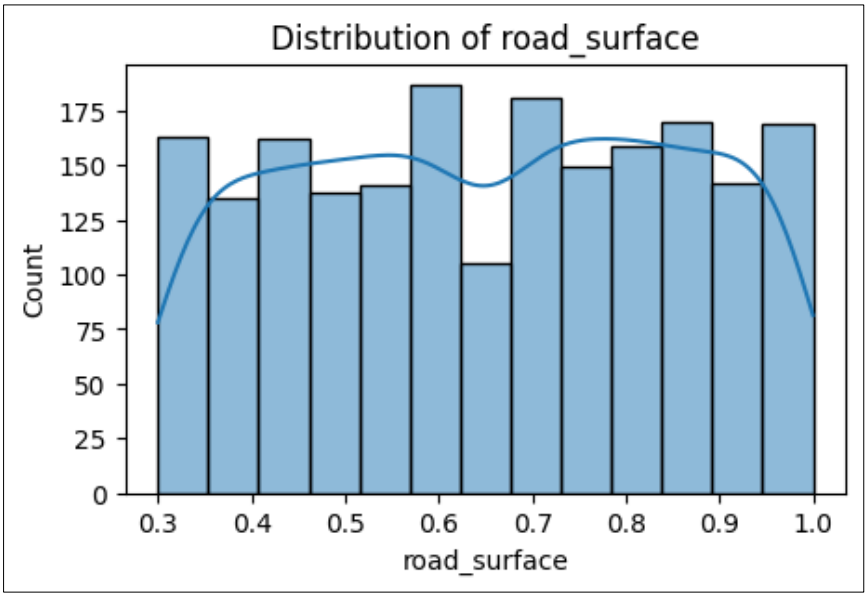

Fig 4: Distribution of road surface

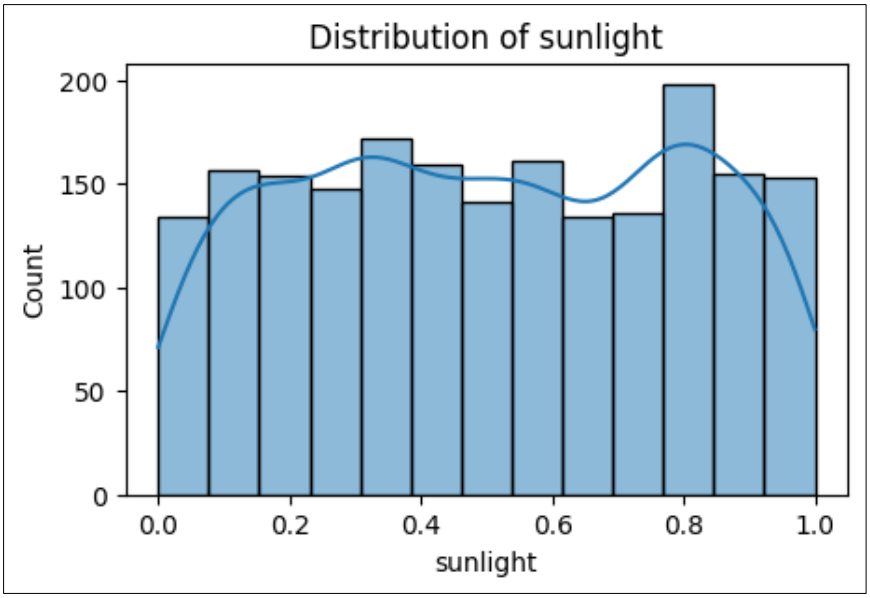

Fig 5: Distribution of sunlight

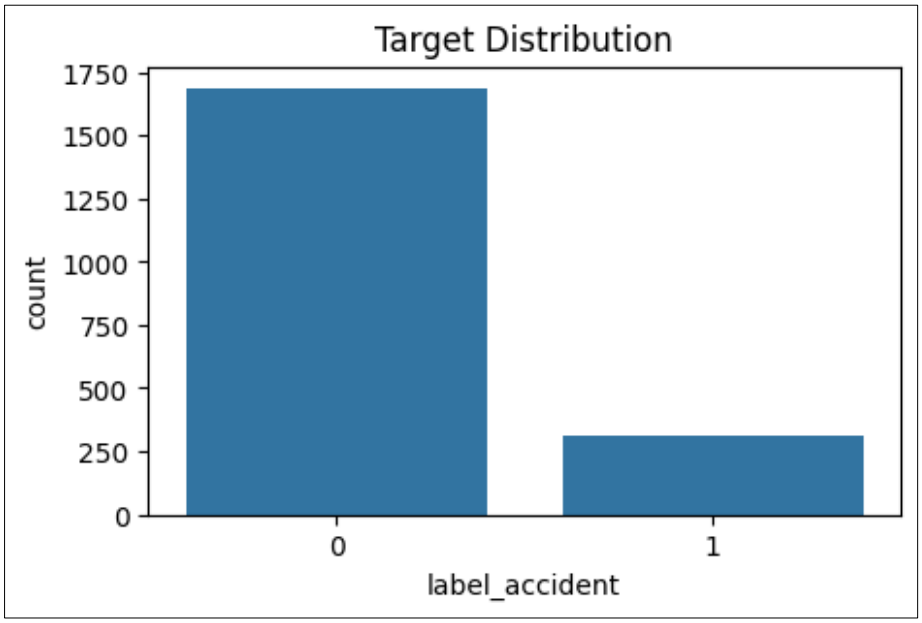

Fig 6: Target class distribution

Figures 1 through 6 show the distributions for the features that have been designed such that the variation is under control and the class distribution is balanced and realistic.

- **Baseline vs Machine Learning Models**

In order to provide context to the results of our models' performance, we tested some predictive models as a baseline before moving to more sophisticated machine learning models. In our first simple model, the Majority Class, our approach earned an accuracy of 84.5% but scored an F1 of 0 for minority class prediction because they were unable to identify any accident class instances. On the other hand, our Random Predictor earned varying results but did little to enhance our predictions of the minority class.

In contrast, the improvements shown by the machine learning models were quite significant in the minority class detection. The performance of the Logistic Regression model and the KNN model was moderate. The performance of the Random Forest classifier was better than all models.

Table 2: Performance Comparison of Baseline and Machine Learning Models

Model	Accuracy	F1 (Minority Class)
Majority Baseline	0.8450	0.0000
Random Predictor	0.7550	0.1404
Logistic Regression	0.9425	0.8414
KNN	0.8450	0.6220
SVM	0.4850	0.3268
Random Forest	0.9975	0.9920

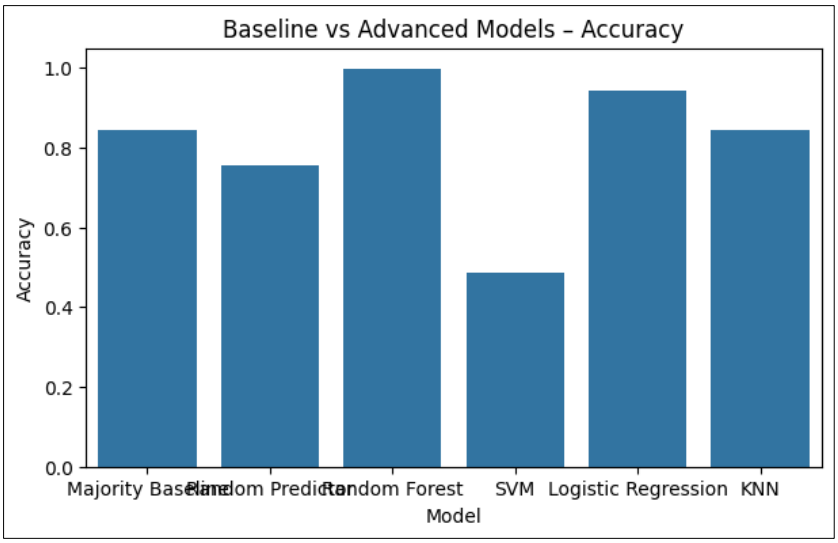

Fig 7: Baseline vs Advanced Models – Accuracy

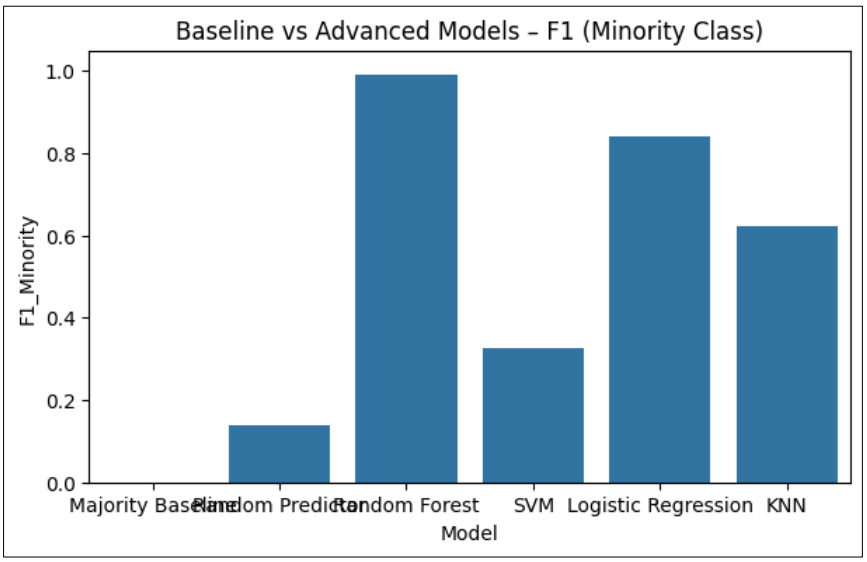

Fig 8: Baseline vs Advanced Models – Minority F1-score

As illustrated in Fig. 7 and Fig. 8, ensemble-based learning significantly improves minority-class detection compared to baseline predictors.

- **Confusion Matrix Analysis**

The confusion matrices of individual machine learning models are provided in Fig. 9 to Fig. 12. It has been observed that Logistic Regression, KNN models predict a level of accident classification, whereas a Random Forest Classifier differentiates between accident and non-accident images close to perfection.

The confusion matrix for the Random Forest classifier (Fig. 12) shows complete dominance on its diagonal, reflecting that there are no false positives and false negatives in the test data.

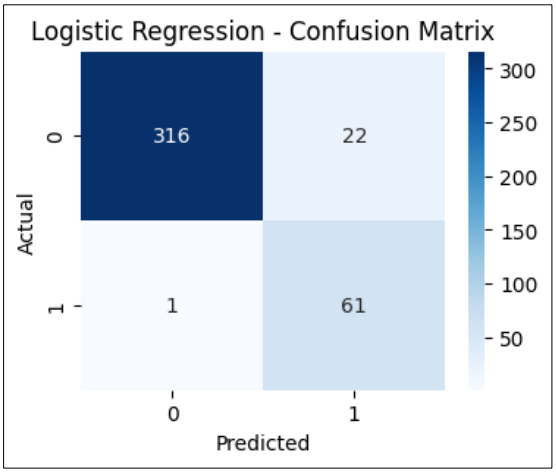

Fig 9: Confusion Matrix – Logistic Regression

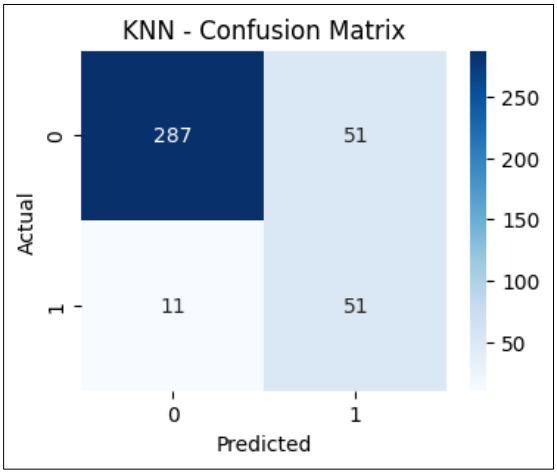

Fig 10: Confusion Matrix – KNN

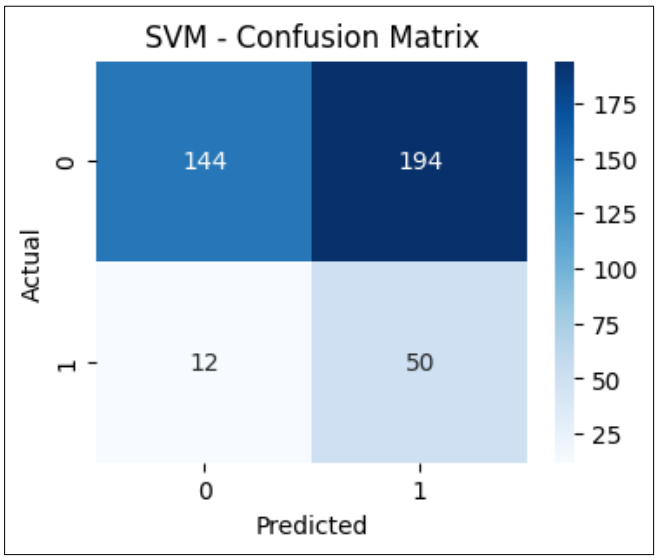

Fig 11: Confusion Matrix – SVM

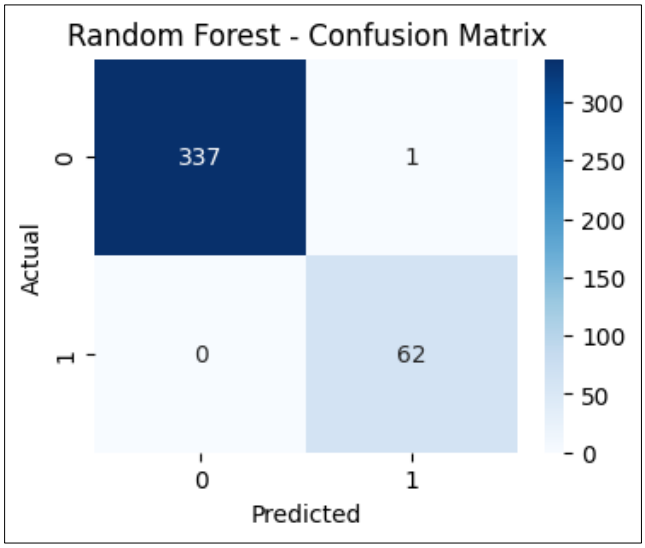

Fig 12: Confusion Matrix – Random Forest

The poor performance of SVM can be attributed to its sensitivity to scaling factors of features. This problem has been further aggravated due to a fixed kernel size.

The optimal classification obtained for Random Forest can be essentially attributed to the deterministic rule patterns that are inherent in the synthetic data.

- **Cross-Validation Performance**

Regarding robustness testing, 5-fold stratified cross-validation was performed. The accuracy achieved by the Random Forest classifier was 99.88% with low variance, as shown in Fig. 13.

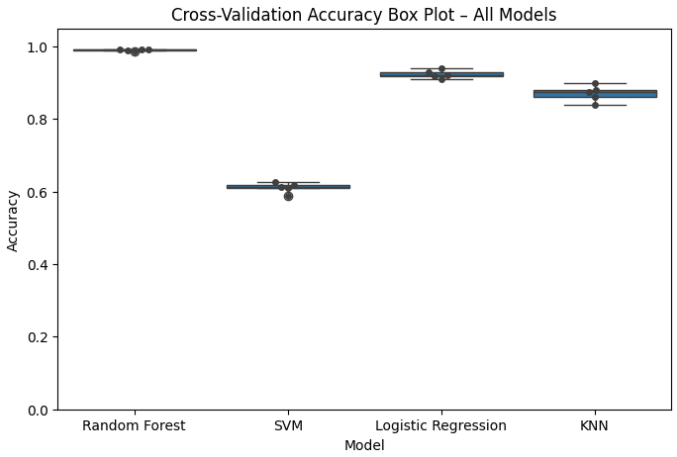

Fig 13: Cross-Validation Accuracy Box Plot

The low variance on each fold in Fig. 13 reflects good learning and robustness to data splitting.
Even though hypothesis testing was not a major concern, the good cross-validation performance with low variance suggests good learning capabilities.

- **Discussion and Interpretation**

Results show the ability of machine learning models to learn patterns pertinent to accidents based on micro-level driving variables. Learning methods based on ensemble methods, especially Random Forest, prove more effective due to their capability for recognizing nonlinear relationships between variables.
However, the extremely high level of accuracy and the ideal results of the classification should be considered carefully. This is because the dataset is governed by deterministic rules of generation and does not contain noise in the same way as a real-world setting. This makes the outcomes more a verification of the feasibility of the method than anything else. These are conditions of perfect classification, which in reality are not typically ever achieved in a real-world scenario of accident prediction, but are rather a result of controlled data generation.
Such perfect classification is rarely observed in real-world accident prediction tasks and is primarily a consequence of controlled synthetic data generation rather than model superiority.

7. Conclusion

In this study, the possibility of predicting risk related to road accidents using micro-indicators of driving and classical models of machine learning was explored. For evaluating the possibility of different classifiers on five driving-related attributes, a synthesized dataset relating to driving was prepared by considering 2,000 samples. From these results, it is revealed that classical models of machine learning can learn interactions among driving-related attributes like speed, alertness of the driver, phone usage while driving, surface of the road, and intensity of sunlight.
In terms of the models compared, it was found that Random Forest performed better than Logistic Regression, K-Nearest Neighbors, and Support Vector Machines. In particular, Random Forest performed nearly flawlessly in terms of accuracy and had the highest value of minority class F1-score. The results of cross-validation also showed that there was little variation in model performances.
Nevertheless, the exceptionally high level of performance in this study is primarily due to the deterministic rule-based structure of this synthetic dataset. Thus, the results presented here and elsewhere ought to be viewed from the perspective of proving feasibility rather than attempting to guarantee any kind of effective real-world results or generality. In any case, this work indicates the feasibility of learning methods based on ensemble approaches to model accident risk from micro-level driving behavior variables.

8. Future Work

Although the above analysis proves the feasibility of accident prediction through machine learning on artificial data, there are still some avenues that could be explored in future implementations. Firstly, the above model ought to be tested on real-world datasets involving noisy sensor reads, missing values, and uncertainty in behavioral patterns to verify generalization abilities in real-world settings. Secondly, models involving time-series patterns, like recurrent neural networks or transformers, could be employed to analyze time-series patterns of driving behaviors instead of focusing on tangible features.
Moreover, one area of further research could be adaptive learning approaches for accident risk assessment in real-time systems, followed by the application of explainable AI methods to enable the interpretation of the outcome of the proposed approach. Scaling up the proposed framework for large and diverse environments would be beneficial for ITS and safety on roads.

9. Acknowledgment

The authors sincerely thank Dr. Manoj Kumar Deshpande, Senior Director, and Dr. Piyush Choudhary, Head of the Department of Computer Science, for their guidance and support during this research.

References

1. World Health Organization. Global status report on road safety 2023. Geneva: World Health Organization; 2023.
2. Singh AP, Kathuria R. Analyzing driver behavior for accident prediction using machine learning. Accid Anal Prev. 2020;136:105406. doi:10.1016/j.aap.2019.105406
3. Antoniou C, Yannis G, Papadimitriou E. Driving behavior, accident involvement, and self-assessment. Transp Res Part F Traffic Psychol Behav. 2011;14(2):136–48. doi:10.1016/j.trf.2010.11.002
4. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324
5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009. doi:10.1007/978-0-387-84858-7
6. Elmitiny N, Yan X, Radwan E, Russo C, Nashar D. Classification of driver actions based on driving behavior. IEEE Trans Intell Transp Syst. 2013;14(2):820–9. doi:10.1109/TITS.2012.2236208
7. Wickramasinghe ML, Rathnayake N, Rathnayake U. Machine learning approaches for traffic accident prediction. IEEE Access. 2021;9:16233–47. doi:10.1109/ACCESS.2021.3053419
8. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. doi:10.1007/BF00994018
9. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Waltham, MA: Morgan Kaufmann; 2011. doi:10.1016/C2009-0-61819-5
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. doi:10.1613/jair.953
11. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84. doi:10.1109/TKDE.2008.239
12. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895–923. doi:10.1162/089976698300017197
13. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
14. Raschka S, Mirjalili V. Python machine learning. 3rd ed. Birmingham: Packt Publishing; 2019.
15. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39.

doi:10.1006/jcss.1997.1504
16. Lajevardi SM, Jafari S, Khosravi A. Driver risk assessment using machine learning techniques. Transp Res Part C Emerg Technol. 2021;128:103137. doi:10.1016/j.trc.2021.103137
17. Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.
18. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
19. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Sebastopol, CA: O'Reilly Media; 2019.
20. IEEE Standards Association. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. Piscataway, NJ: IEEE; 2019.
21. NexisparkX. Public micro-level driving behavior dataset for accident prediction [Data set]. Zenodo; 2025. doi:10.5281/zenodo.17603168

How to Cite This Article
Karma G, Bagwaiya H. Micro-Level Driving Behavior Analysis for Accident Prediction Using Machine Learning. International Journal of Multidisciplinary Research and Growth Evaluation. 2026;7(3):1145-1154.