A Comparative Study on Unsupervised Machine Learning Models for Detecting Sudden Lane Changes

T&P, CiTG

CIE5050-09

Research Report

Zhang Lanxin

5119863

Committee Members

Chair Supervisor	Dr. ir. H. Farah
Daily Supervisor	Y. Dong

Abstract

Lane-changing behaviour detection is a critical aspect of driving safety and traffic management. This study focuses on detecting sudden lane changes as a subset of abnormal driving behaviours. By analyzing the characteristics of abrupt lane changes, the aim is to develop effective data-driven unsupervised machine learning (ML) methods for their detection and classification. Three unsupervised ML models, namely Isolation Forest, Local Outlier Factor, and Robust Covariance are evaluated and compared using a dataset of lane-change events. The results show that the Isolation Forest and Local Outlier Factor models outperform the Robust Covariance model, with the Local Outlier Factor model excelling in precision and overall accuracy, achieving the best overall detection rate. Both Robust Covariance and Isolation Forest deliver satisfactory results. Conversely, the Robust Covariance model exhibits poor performance. The findings verify the capability of data-driven ML methods for enhancing road safety and driving experiences through effective detection of sudden lane changes using vehicle motion information data. Future work involves further improving the accuracy and reliability of the ML models, validating their generalizability on larger datasets, incorporating contextual information, and exploring their real-time implementation in driving assistance systems.

Table of content

1 Introduction

In the field of intelligent transportation, traffic safety is an important topic. As one of the most serious hazards worldwide, traffic accidents cause great casualties and property losses (Dept. for Transp., 2013; Peden et al., 2004). According to the WHO (2022), over 1.25 million people die, and up to 50 million suffer from injuries due to road crashes, which makes road accidents a leading cause of death.

Road crashes can not be completely prevented, but certain measures could be taken to try and reduce their occurrence. Most road accidents are caused by human factors such as driving behaviours (Saiprasert & Pattara-Atikom, 2013). According to the official reports in Beijing (Su et al., 2023), the causes of car accidents are generally attributed to drivers, vehicles, road conditions, and weather, among which the driver is the dominant factor and counts for about 95% of car accidents. Therefore, it is important to detect abnormal driving to improve traffic safety effectively.

Based on the discussion on the predominantly human-caused nature of traffic accidents, it is essential to highlight the significant association between abnormal driving behaviours and a substantial number of traffic incidents. In particular, within these abnormal driving behaviours, lane-changing actions driven by human factors are found to be closely linked to a significant portion of these accidents.

Lane-changing manoeuvres are associated with a substantial number of road traffic crashes. For instance, in 1999, about 539,000 two-vehicle lane-change crashes were reported in the U.S. (Basav Sen et al., 2003). In New South Wales, Australia, 3438 rear-end and 1171 lane-changing crashes were reported in 2017 (NSW-Transport., 2018). In Queensland, Australia, lane-change-related crashes represent about 4% of total crashes, whilst rear-end and side-swipe crashes each represent 4% of total crashes, respectively (Manager, 2009).

A successful lane change requires a driver to simultaneously allocate his/her mental attention to several decision-making factors, such as looking for an appropriate gap size on the adjacent lane, checking for a blind spot, maintaining the correct lane position and the distance to the leader on the current lane, and adjusting the driving speed (Ali et al., 2019). Given lane-changing manoeuvres' complexity and critical nature, this study focuses on detecting sudden lane changes as a subset of abnormal driving behaviours. Examining and analyzing the characteristics of these abrupt lane changes, this study aims to develop effective data-driven methods for detecting and classifying such behaviours, ultimately contributing to enhanced road safety and driving assistance systems.

Several algorithms and techniques can be used to recognize driving behaviour, among which the data-driven machine learning (ML) method possesses high potential. The ML

method is roughly divided into supervised and unsupervised learning approaches. Supervised ML needs labelled data. For example, Ly et al. used a Support Vector Machine (SVM) as a supervised method to explore the possibility of using the labelled vehicle's inertial sensors from the Controller Area Network (CAN) of a bus to build a profile of the drivers (Ly et al., n.d.). Assigning unknown data into categories by mining the underlying sources of unlabeled data is called the Unsupervised Machine Learning method. For example, clustering and Principal Component Analysis (PCA) from exploratory statistics were used to identify and explain driver groupings according to their driving behaviour (Constantinescu et al., 2010).

Unsupervised learning techniques are commonly employed when data is unlabeled, as is often the case with multi-variate anomalies, where it is uncertain whether the system behaviour is anomalous at any given point in time apart from obvious system failures. These techniques include Robust Covariance, Local Outlier Factors, and Isolation Forests (Nikita Butakov, 2020).

However, it is important to note that the absence of labelled data in unsupervised models makes it challenging to measure their accuracy precisely. It is easier to ascertain the exact accuracy of the models in identifying abnormal lane changes with the presence of labelled data for validation. Since there is no ground truth or reference for comparison, evaluating the performance of unsupervised ML models sometimes becomes more subjective.

In this study, the primary focus is on investigating sudden lane-changing behaviours in the open-sourced CitySim dataset. The data was first labelled and categorized according to rule-based criteria and human experts' experience to establish a baseline and ground truth for comparison. Subsequently, three unsupervised ML models, i.e., Isolation Forest, Local Outlier Factor, and Robust Covariance, were customized and implemented, and the performances of three unsupervised models were evaluated and compared regarding various metrics (including accuracy, precision, and recall ratio). Results demonstrated that Local Outlier Factor achieved the best overall performance outperforming the Isolation Forest and Robust Covariance.

2 Related Work

2.1 Lane-changing Behaviour Detection

The lane-changing (LC) behaviours are investigated for various types of roads, such as urban arterials, freeways, and expressways.

Decision models in lane-changing play a crucial role in predicting drivers' intentions. Drivers' LC intention could be predicted by operation inputs and vehicle states (Ng et al., 2020). Drivers' merging and give-way decisions were researched by a quantal response equilibrium framework and the Nash equilibrium solutions in lane changing (Arbis & Dixit, 2019). Most decision models can be probabilistic or deterministic-based (Li et al., 2019; Peng et al., 2020).

Microscopic models, on the other hand, focus on capturing the finer details of lane-changing behaviour. Variables such as the LC rate, velocity motivation, target lane choice, gap acceptance, target lane choice, and the direction of steering wheels are used to construct the microscope model (Guo et al., 2018). The insertion angle of the target vehicle is considered for a better understanding of how vehicle lane changing increases the probability of traffic accidents in the road segment (Yang et al., 2020).

Alongside decision and micro motion models, it is vital to consider the motivation factors behind lane-changing. Mathematic models such as regression, hybrid, and microscopic models were proposed to describe motivation factors and mechanisms of LC behaviour (Farooq & Juhasz, 2019). The regression model predicted the frequency of the LC in low illumination well based on factors such as gender, weekly driving time, and lane-change and risky driving behaviour factors (Kusuma et al., 2020).

2.2 Machine Learning Model

Recent studies (Eftekhari & Ghatee, 2018) have explored abnormal driving behaviour using clustering and shallow learning algorithms. Previous research on abnormal driving behaviour has employed supervised and unsupervised methods.

The training phase of supervised machine learning relies on labelled inputs and outputs. Once the model has learned the patterns and relationships in the labelled data, it can classify new and unseen datasets and make predictions. Jia et al. (2020) developed an LSTM-CNN model that combines the strengths of Long Short-Term Memory (LSTM) for processing time series data and Convolutional Neural Network (CNN) for processing matrix data. They detected extreme acceleration and deceleration points using statistical analysis of actual vehicle driving data and established a dataset for driving behaviour recognition. Training the LSTM-CNN model on this dataset yielded improved results. Shahverdy et al. (2021) proposed a lightweight 1D-CNN with high

efficiency and low computational complexity for classifying driver behaviour, specifically focusing on high-speed movement, braking, rapid speed changes, and quick steering. Ryan et al. (2021) simulated an end-to-end model of autonomous vehicles (AV) using Convolutional Neural Networks (CNN) to compare human and AV driving behaviour.

While as the name implies, unsupervised machine learning is a less guided approach than supervised machine learning. Using unlabeled training data, unsupervised machine learning models are trained. It allows the model to identify patterns, structures, and relationships within the data without predefined labels or human intervention. Mohammadnazar et al. (2021) introduced a framework that utilizes unsupervised machine learning methods to quantify instantaneous driving behaviour and classify driving styles in different spatial contexts. K-means and K-medoid methods were applied to group drivers into aggressive, normal, and calm clusters.

Similarly, Feng et al. (2018) proposed a novel technique for robustly classifying driving styles using the Support Vector Clustering approach. Their method aimed to differentiate variations in individual driving patterns and provide an objective driver classification. The authors identified four input signals (vehicle speed, engine speed, pedal position, and headway distance) and four statistical features (mean, standard deviation, maximum, and minimum values) as the parameters for feature extraction.

3 Methodology

3.1 Unsupervised ML models

Unsupervised machine learning models do not possess example input-output pairs that allow them to learn a function that maps the input features to outputs. Instead, they learn by finding structure within the input features. In unsupervised ML, "structure" refers to patterns, relationships, or regularities within the input features of the data. It involves identifying inherent dependencies or similarities among the data points without using labelled output information. By identifying structure, unsupervised learning models can uncover hidden patterns or groupings in the data, providing a deeper understanding of its inherent properties. In this study, Isolation Forest, Local outlier factor (LOF), and Robust Covariance are selected as the unsupervised ML models. In this section, the following paragraphs briefly introduce the three selected unsupervised ML models.

Isolation Forest

To look for anomalies, Isolation Forest (Lesouple et al., 2021) generates random isolation trees to isolate each data point. The number of branches required to isolate each point is computed for each tree. The mean of this number of branches defines the expected path length, which is used to isolate a point of interest. The expected path length is generally small for anomalies (contrary to nominal data) since anomalies are far from the majority of nominal data.

In statistics, the deviation can be assessed by the Z-score. The generalization of the Z-score for a point x_i in the case of a p-dimensional multi-variate probability distribution with some mean μ and covariance matrix Σ is known as Mahalanobis distance d_i, which is given by:

$$d_i = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}$$

Here is a simplified example to illustrate the Isolation Forest algorithm for anomaly detection:

- Data Preparation: The dataset comprises various features such as velocity, acceleration, coordinate and distance.
- Isolation Forest Construction:

  a. Random Subsampling: Randomly select a subset of instances from the dataset.

  b. Random Feature Split: Randomly select a feature and a split value to partition the selected instances. The split value can be any value within the range of the selected feature.

  c. Recursive Partitioning: Recursively split the instances based on the selected feature and split value. Continue this process until each instance is isolated in a

separate leaf node or a predefined stopping criterion is met (e.g., maximum tree depth).

- Path Length Calculation: Measure the average path length required to isolate each instance in an isolation tree. The path length is the number of edges traversed from the root to reach a particular instance. Anomalies are expected to have shorter average path lengths compared to normal instances.
- Anomaly Score Calculation: Calculate an anomaly score for each instance based on the average path length across all isolation trees. Instances with shorter average path lengths (fewer splits required for isolation) are assigned higher anomaly scores, indicating a higher likelihood of being an anomaly.
- Anomaly Detection: Identify instances with anomaly scores above the threshold as anomalies. These instances represent abnormal driving behaviours.

Local outlier factor (LOF)

LOF is another useful unsupervised ML algorithm that identifies outliers concerning the local neighbourhoods instead of using the entire data distribution (Breunig et al., 2000). LOF is a density-based technique that uses the nearest neighbour search to identify anomalous points. The advantage of using a LOF is identifying points that are outliers relative to a local cluster of points. For instance, when using the local outlier factor technique, neighbours of certain points are identified and compared against the density of the neighbouring points. The following steps can be applied when using a LOF model:

1) Calculate distance between P and all the given points using a distance function such as euclidean or Manhattan.
2) Find the k (k-nearest neighbor) closest point. For example, if K = 3, find the third nearest neighbor's distance.
3) Find the k closest points.
4) Find local reachability density using the following equation:

$$lrd_k(O) = \frac{\|N_k(O)\|}{\sum_{O' \in N_k(O)} reachdist_k(O' \leftarrow O)}$$

where reachable distance can be calculated as follows:

$$reachdist_k(O' \leftarrow O) = max\{dist_k(O), dist(O, O')\}$$

5) The last step is to calculate the local outlier factor as follows:

$$LOF_k(O) = \frac{\sum_{O' \in N_k(O)} \frac{lrd_k(O')}{lrd_k(O)}}{\|N_k(O)\|}$$

Robust Covariance

The Robust Covariance technique assumes that normal data points have a Gaussian distribution, and accordingly estimates the shape of the joint distribution (i.e., estimates the mean and covariance of the multivariate Gaussian distribution) (Nikita Butakov, 2020). It is based on the fact that outliers lead to an increase of the values (entries) in Σ , making the spread of the data apparently larger. Consequently, $|\Sigma|$(the determinant) will also be larger, which would theoretically decrease by removing extreme events.

Rousseeuw and Van Driessen (Peter J.Rousseeuw & Driessen Van Katrien, 1999) developed a computationally efficient algorithm that can yield robust covariance estimates. The method is based on the assumption that at least h out of the n samples are "normal" (h is a hyperparameter). The algorithm starts with k random samples with $(p+1)$ points. For each k sample, μ, Σ, and $|\Sigma|$ are estimated, the distances are calculated and sorted in increasing order, and the h smallest distances are used to update the estimates. In their original publication, the subroutine of computing distances and updating the estimates of μ, Σ, and $|\Sigma|$ is called a "C-step" and two such steps are sufficient to find good candidates (for μ and Σ) among the k random samples. In the next step, a subset of size m with the lowest $|\Sigma|$ (the best candidates) is considered for computation until convergence, and the one estimate whose $|\Sigma|$ is minimal is returned as output.

3.2 Evaluation metrics

Table 1 presents the confusion matrix, which serves as the basis for evaluating the discrimination performance of the selected model. A range of metrics will be employed to assess the overall effectiveness of the model in accurately classifying instances, ensuring a comprehensive evaluation of its performance (M & M.N, 2015).

Table 1 Confusion Matrix and the Corresponding Array Representation

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True-positive (TP)	False-negative (FN)
Predicted Negative Class	False positive (FP)	True-negative (TN)

In the context of binary classification, the model distinguishes between two classes: positive and negative. The positive class corresponds to the specific event or condition the model aims to identify, while the negative class represents the alternative possibility. For instance, in abnormal driving behaviour detection, the positive class may be labelled "abnormal," while the negative class represents the absence of abnormal behaviour. True Positive (TP) and True Negative (TN) indicate the number of instances accurately classified as positive and negative, respectively.

Within this study, True Positive (TP) signifies the accurate identification of anomalies, while True Negative (TN) denotes the correct identification of normal instances. Conversely, False Positive (FP) and False Negative (FN) indicate the misclassification of positive and negative instances. These terms represent the instances where anomalies

or normal cases were incorrectly identified. As a result, performance metrics such as accuracy, precision, and recall are computed utilizing these four measures to evaluate the model's effectiveness.

Accuracy measures the ratio of accurately classified instances to the total cases evaluated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision quantifies the proportion of correctly predicted positive patterns among the total predicted patterns in the positive class.

$$Precision = \frac{TP}{TP + FP}$$

Recall, as an additional valuable measure, tackles a different inquiry: it determines the proportion of actual Positives that are correctly classified.

$$Recall = \frac{TP}{TP + FN}$$

The F1-score combines the precision and recall of the model, defined as the harmonic mean of precision and recall.

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

Lastly, the Receiver Operating Characteristic-Area Under the Curve (ROC AUC) is employed as an evaluation metric. This metric assesses the model's performance by identifying the areas where it excels at classifying normal and anomaly situations. Plotting True Positive Rates against False Positive Rates produces ROC curves.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

4 Data Description

This study utilizes the CitySim dataset due to its inclusion of abnormal driving behaviour, making it suitable for calculating the necessary features for the thesis. CitySim (Zheng et al., 2022) is a collection of video-based trajectory data from drone recordings focusing on traffic safety in the United States. The dataset comprises vehicle trajectories extracted from 1140 minutes of drone videos captured at 12 different locations. It encompasses various road geometries such as primary freeway segments, weaving segments, expressway merge/diverge segments, signalized intersections, stop-controlled intersections, and intersections without sign/signal control.

The CitySim dataset is available in Comma Separated Value (CSV) files, where each row represents a waypoint belonging to a vehicle trajectory within a single frame. Each waypoint contains positional information for seven essential vehicle points, including the centre point, head, tail, and four bounding box vertices (refer to Figure 1). The dataset provides positional data in various formats, such as pixels, feet, and GPS coordinates. Moreover, it includes details on speed, heading (measured relative to both the global north and the image X-axis), and the lane number of the vehicle. Notably, the dataset exhibits high precision, with measurements accurate within a range of approximately 10 centimetres.

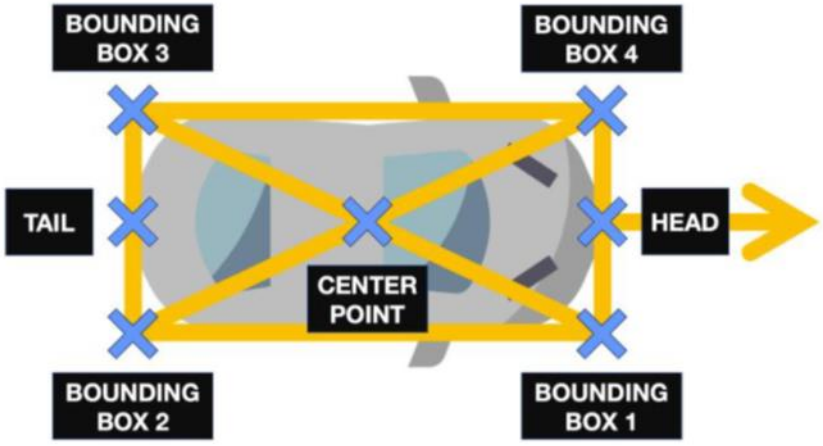

Figure 1 Vehicle bounding box feature description

Expressway A (Figure 2) in CitySim is chosen as a case study for analysing abnormal lane-changing behaviours. The weaving segment of Expressway A is particularly significant due to its actual occurrence of critical safety events, including cut-ins, merges, and other lane-changing behaviours.

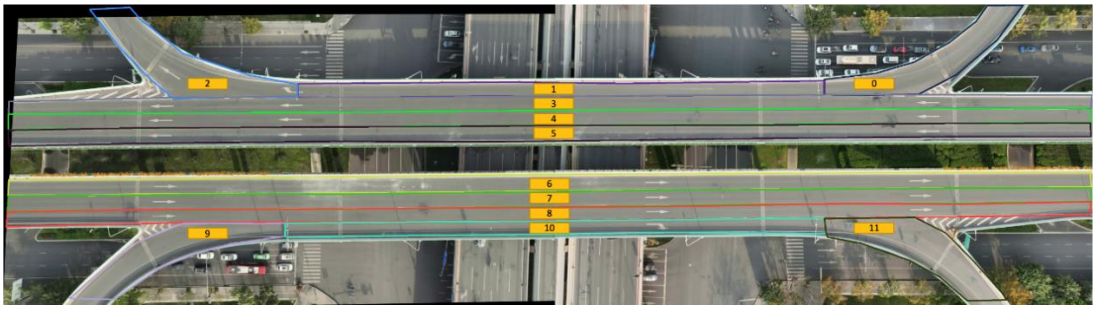

Figure 2 Expressway A

When a vehicle changes lanes by cutting into a target lane from a source lane, it can conflict considerably with the following vehicle on the target lane. This lane-changing behaviour is known to be a contributing factor in rear-end collisions. Given that the study focuses on the weaving segments of Expressway A, it is expected to observe various lane change behaviours in this area. (See in Table 2 and Figure 3).

Table 2 Lane changing behaviour at Expressway A

ExpresswayA-01	No lane changing	Once	Twice	Three times	Four times	Total
Lane-changing	325	97	118	47	5	592

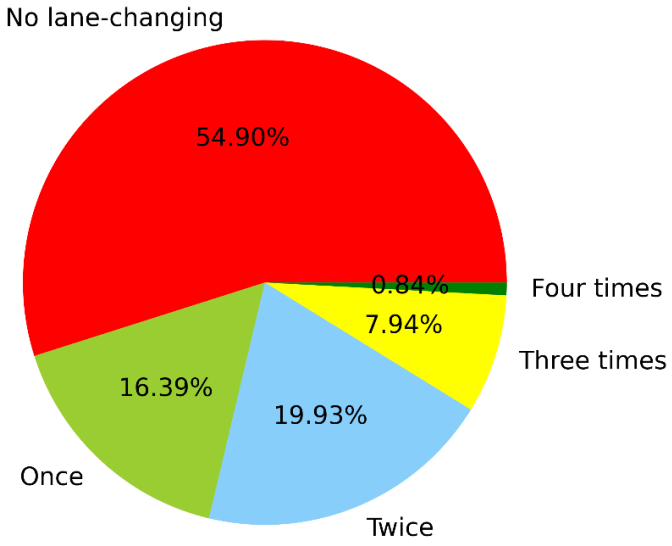

Figure 3 Proportional distribution of lane changing times

The dataset provides insights into the distribution of lane-changing behaviours, allowing for an analysis of their proportions. Notably, in approximately 54.9% of instances, a significant proportion did not involve lane changes, indicating that a substantial portion of observed driving behaviours did not include lane changes.

Among the instances that did involve lane changes, approximately 16.4% consisted of single-lane changes. This suggests that many driving instances included a single-lane change during the recorded period.

Instances with two-lane changes accounted for approximately 19.9% of the dataset, indicating a relatively higher frequency of instances where two-lane changes occurred. On the other hand, instances with three-lane changes were less common, comprising roughly 7.9% of the dataset. Instances with four-lane changes were the least frequent, representing only around 0.8% of the dataset.

The analysis reveals that most instances did not involve lane changes, followed by instances with a single lane change. Instances with multiple lane changes (two or three) were less frequent, while instances with four-lane changes were extremely rare. This distribution is visually depicted in Figure 3 through a pie chart.

In order to assess whether vehicles exhibit sudden lane-changing behaviour, the lateral acceleration of these vehicles is analyzed. Figure 4 illustrates the distinct characteristics of rapid lane-changing behaviour observed in real-life situations.

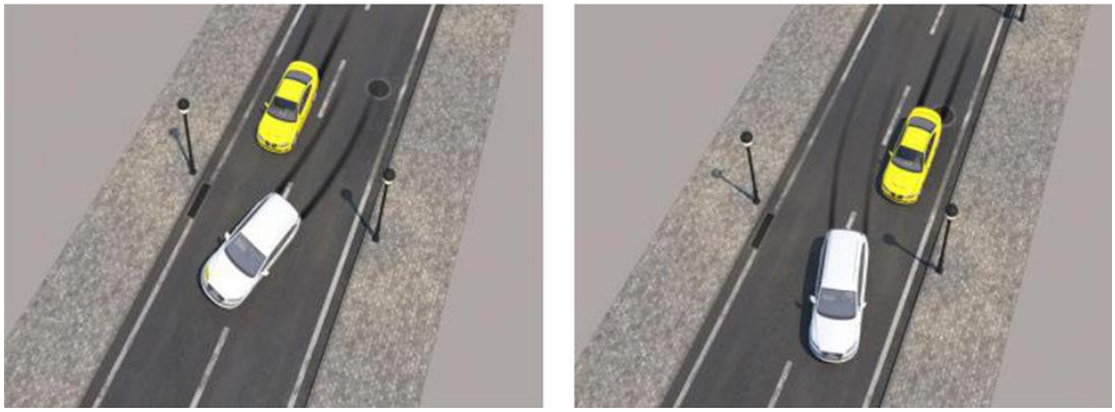

Figure 4 Rapid lane changing(Jia et al., 2020)

Based on the findings from the literature review, it is clear that unsupervised models are commonly used without the need for pre-labelling the data. However, to gain a more comprehensive understanding of the results obtained from the unsupervised models employed in this study (Isolation Forest, Local Outlier Factor, and Robust Covariance), it becomes necessary to label the lane-changing behaviours present in the dataset. Specifically, our focus will be on identifying sudden lane-changing behaviour and classifying them based on the metric of lateral acceleration. This initial data labelling process allows us to establish a baseline for evaluating the outcomes of the unsupervised models.

The following Table 3 and Figure 5 are the analysis of lateral acceleration for the vehicles that have lane-changing behaviour.

Table 3 Lateral acceleration

Expressway A	Mean	Std	Min	Max
Lateral Acceleration(m/s^2)	0.00	1.30	-6.59	6.42

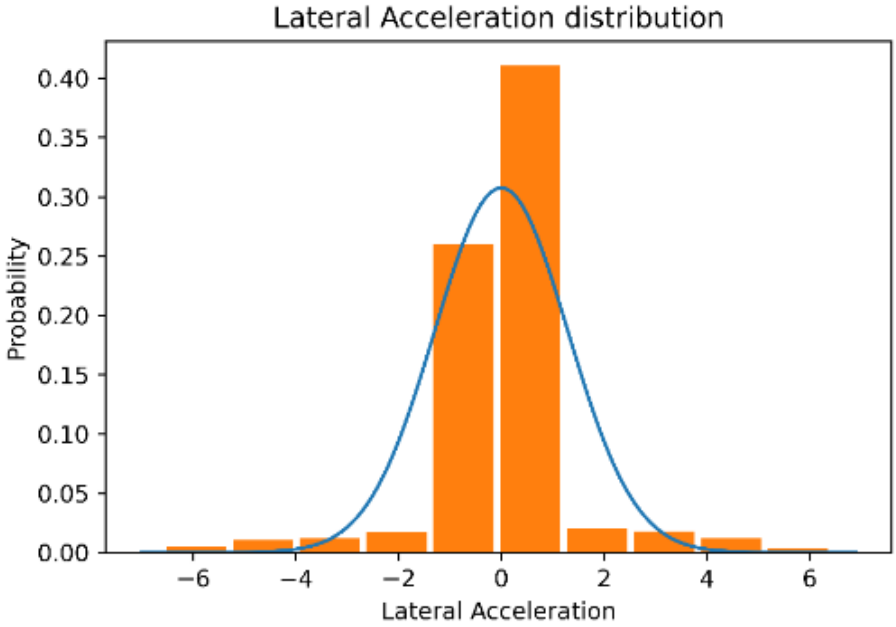

Figure 5 Lateral Acceleration Distribution

Most vehicles exhibiting lane change behaviour demonstrate a relatively constant acceleration of around 0 m/s², indicating lane changes performed at a consistent speed. However, there are outliers in the dataset where the acceleration of certain vehicles deviates from this pattern, as illustrated in Figure 6. Based on the normal distribution, we define outlier values as those exceeding 1.3 m/s² or falling below -1.3 m/s², which serve as the filtering criteria for identifying such cases.

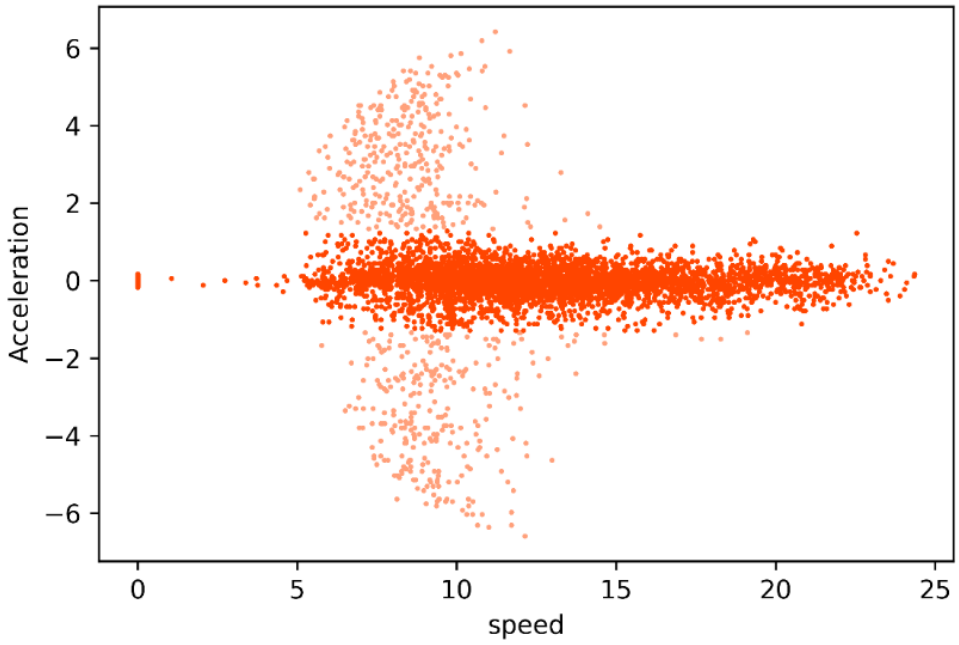

Figure 6 Extreme lateral acceleration and deceleration points distribution at different speeds (red=normal/orange=abnormal)

The scatter plot exhibits a distinct pattern resembling a bird, which is remarkably well-suited for analyzing lane-changing behaviours. Firstly, points on the bird's wings suggest a significant occurrence of abnormal lane changes involving high speeds. This indicates a potential risk of unsafe driving behaviour associated with sudden lane changes. On the contrary, the concentrated points forming the bird's body indicate a predominant occurrence of normal lane-changing behaviours with consistent lateral acceleration, indicating a more stable driving pattern.

Before the upcoming chapter, where an unsupervised model will be used for classifying abnormal lane-changing behaviours, this chapter focuses on the manual labelling of data using lateral acceleration as a metric. The subsequent chapter will then compare the manual data labelling and the unsupervised model approach. This comparison aims to evaluate the effectiveness and reliability of both methods in accurately identifying and classifying abnormal lane-changing behaviours.

5 Experimental Results and Comparison

5.1 Experimental Results

Isolation Forest

Figure 7 and Table 4 presents the results obtained from the Isolation Forest machine learning approach.

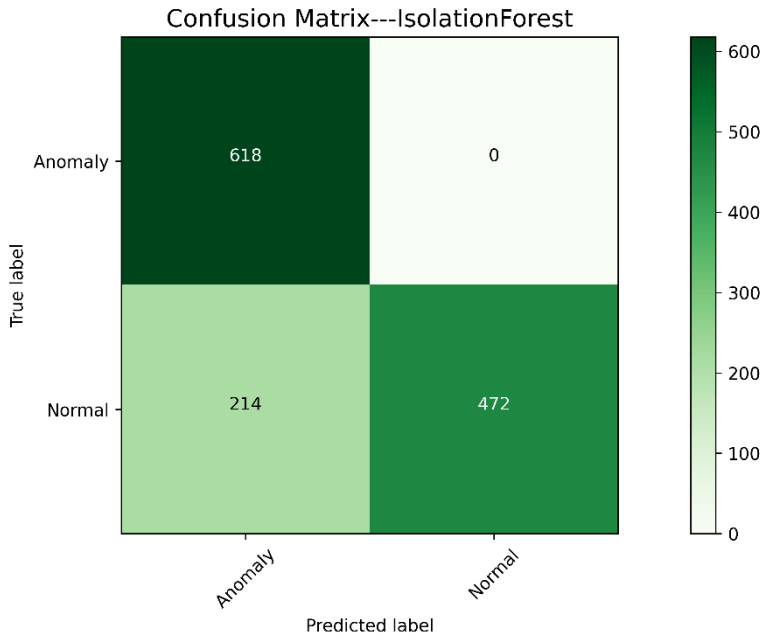

Figure 7 Confusion Matrix Isolation Forest

Table 4 Performance of Isolation Forest

	Precision	Recall	F1-score	Support
Abnormal	0.74	1.00	0.85	618
Normal	1.00	0.69	0.82	686
Accuracy			0.84	1304
Macro avg	0.87	0.84	0.83	1304
Weighted avg	0.88	0.84	0.83	1304
F1_score	0.8151986183074266			
FPR	0.3119533527696793			
TPR	1.0			
ACC	0.8358895705521472			

Local Outlier Factor

The findings from the Local Outlier Factor machine learning approach are visually depicted in Figure 8 and Table 5.

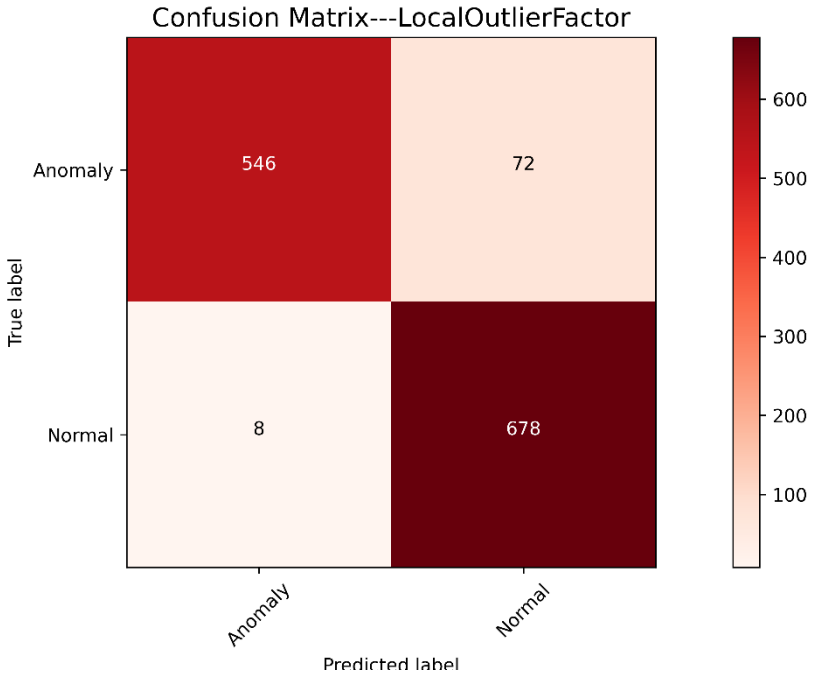

Figure 8 Confusion Matrix Local Outlier Factor

Table 5 Performance of Local Outlier Factors

	Precision	Recall	F1-score	Support
Abnormal	0.99	0.88	0.93	618
Normal	0.90	0.99	0.94	686
Accuracy			0.94	1304
Macro avg	0.94	0.94	0.94	1304
Weighted avg	0.94	0.94	0.94	1304
F1_score	0.9442896935933148			
FPR	0.011661807580174927			
TPR	0.883495145631068			
ACC	0.9386503067484663			

Robust Covariance

Figure 9 and Table 6 presents the comprehensive results obtained through the Robust Covariance machine learning method.

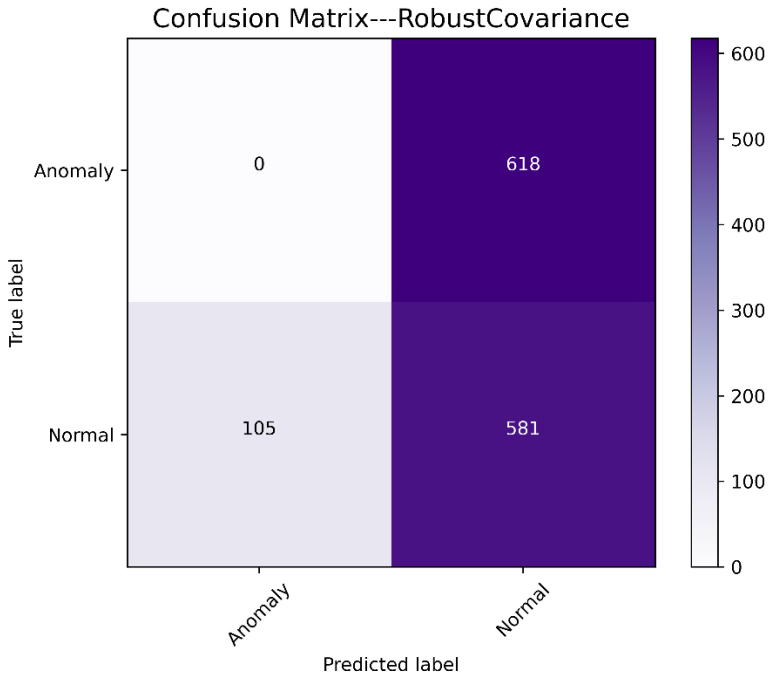

Figure 9 Confusion Matrix Robust Covarianc

Table 6 Performance of Robust Covariance

	Precision	Recall	F1-score	Support
Abnormal	0.00	0.00	0.00	618
Normal	0.48	0.85	0.62	686
Accuracy			0.45	1304
Macro avg	0.24	0.42	0.31	1304
Weighted avg	0.25	0.45	0.32	1304
F1_score	0.616445623342175			
FPR	0.15306122448979592			
TPR	0.0			
ACC	0.4455521472392638			

5.2 Comparison

The following table 7 compares the model performance metrics for Isolation Forest, Local Outlier Factor, and Robust Covariance, highlighting their accuracy, precision, recall, F1-Score, False Positive Rate (FPR), and True Positive Rate (TPR).

Table 7 Comparison of unsupervised machine learning

	Accuracy	**Precision**	**Recall**	**F1-Score**	**FPR**	**TPR**
Isolation Forest	0.84	0.74	1.00	0.82	0.31	1.0
Local Outlier Factor	0.94	0.99	0.88	0.93	0.01	0.88
Robust Covariance	0.45	0.00	0.00	0.00	0.15	0.00

The Isolation Forest model achieves relatively good accuracy, correctly classifying 84% of instances. It demonstrates a precision of 0.74, indicating that 74% of the instances predicted as abnormal are indeed true positives. The model also achieves a perfect recall of 1.00, correctly identifying all actual abnormal instances. The F1-score, a harmonic mean of precision and recall, is 0.82, representing a balanced performance. The FPR of 0.31 suggests that around 31% of normal instances are incorrectly identified as abnormal, while the TPR of 1.0 indicates that all abnormal instances are correctly detected.

The Local Outlier Factor model demonstrates high accuracy, correctly classifying 94% of instances. It achieves an impressive precision of 0.99, indicating that 99% of the instances predicted as abnormal are true positives. The model's recall is 0.88, suggesting it captures 88% of the actual abnormal instances. The F1-score of 0.93 indicates a well-balanced performance. The FPR is only 0.01, indicating a very low rate of misclassifying normal instances as abnormal. The TPR of 0.88 signifies the model's ability to identify a substantial portion of abnormal instances.

The Robust Covariance model exhibits poor accuracy, correctly classifying only 45% of instances. The precision and recall values are both 0.00, indicating that no instances are correctly identified as abnormal, resulting in no true positives. Consequently, the F1-score is also 0.00, indicating a complete failure in performance. The FPR of 0.15 signifies a relatively high rate of misclassifying normal instances as abnormal, while the TPR of 0.00 indicates a complete failure in detecting actual abnormal instances.

In summary, the Isolation Forest and Local Outlier Factor models perform better than the Robust Covariance model. The Isolation Forest model balances precision and recall, while the Local Outlier Factor model excels in precision and achieves high overall accuracy. On the other hand, the Robust Covariance model performs poorly, failing to detect any true positives and exhibiting high false positives.

6 Discussion

The results highlight the effectiveness of the Isolation Forest and Local Outlier Factor models in detecting sudden lane changes, outperforming the Robust Covariance model. The Isolation Forest model demonstrates a balanced performance in precision and recall, indicating its ability to identify true positives while minimizing false positives. On the other hand, the Local Outlier Factor model exhibits a high precision rate, indicating a low rate of false positives and high confidence in the detected anomalies.

The poor performance of the Robust Covariance model suggests that it may not be suitable for detecting sudden lane changes in this context. Further investigation and improvements are necessary to enhance its accuracy and reliability. Possible enhancements involve exploring different outlier detection algorithms or incorporating

additional features and data sources.

In terms of future work, it would be beneficial to evaluate the performance of these models on larger and more diverse datasets to validate their generalizability. Additionally, incorporating contextual information such as weather conditions, road types, and traffic density could provide a more comprehensive understanding of the factors influencing sudden lane changes and improve the accuracy of the detection models. Furthermore, investigating the real-time implementation of these models in driving assistance systems and exploring their integration with other advanced driver assistance technologies could be valuable avenues for future research.

Overall, this study contributes to abnormal driving behaviour detection by comparing and evaluating different unsupervised models for detecting sudden lane changes. The findings lay the foundation for further research and development of more robust and accurate detection systems, ultimately enhancing road safety and driving experiences.

Reference

Constantinescu, Z., Marinoiu, C., Vladoiu, M., Marinoiu, C., & Vladoiu, M. (2010). Driving Style Analysis Using Data Mining Techniques. In *Communications & Control: Vol. V* (Issue 5).

Jia, S., Hui, F., Li, S., Zhao, X., & Khattak, A. J. (2020). Long short-term memory and convolutional neural network for abnormal driving behaviour recognition. *IET Intelligent Transport Systems*, *14*(5), 306–312. https://doi.org/10.1049/iet-its.2019.0200

Nikita Butakov. (2020). *How to build robust anomaly detectors with machine learning*.

Peter J.Rousseeuw, & Driessen Van Katrien. (1999). *A_fast_algorithm_for_the_minim*.

Ali, Y., Haque, M. M., Zheng, Z., Washington, S., & Yildirimoglu, M. (2019). A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C: Emerging Technologies*, *106*, 113–131. https://doi.org/10.1016/j.trc.2019.07.015

Arbis, D., & Dixit, V. V. (2019). Game theoretic model for lane changing: Incorporating conflict risks. *Accident Analysis and Prevention*, *125*, 158–164. https://doi.org/10.1016/j.aap.2019.02.007

Basav Sen, John D. Smith, & Wassim G. Najm. (2003). *Analysis of Lane Change Crashes*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: Identifying Density-Based Local Outliers*.

Constantinescu, Z., Marinoiu, C., Vladoiu, M., Marinoiu, C., & Vladoiu, M. (2010). Driving Style Analysis Using Data Mining Techniques. In *Communications & Control: Vol. V* (Issue 5).

Dept. for Transp., L. U. K. ,. (2013). *Reported Road Casualties in Great Britain: Quarterly Provisional Estimates Q1, 2011*.

Eftekhari, H. R., & Ghatee, M. (2018). Hybrid of discrete wavelet transform and adaptive neuro fuzzy inference system for overall driving behavior recognition. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 782–796. https://doi.org/10.1016/j.trf.2018.06.044

Farooq, D., & Juhasz, J. (2019). Simulation-based analysis of the effect of significant traffic parameters on lane changing for driving logic “cautious” on a freeway. *Sustainability (Switzerland)*, *11*(21). https://doi.org/10.3390/su11215976

Guo, M., Wu, Z., & Zhu, H. (2018). Empirical study of lane-changing behavior on three Chinese freeways. *PLoS ONE*, *13*(1). https://doi.org/10.1371/journal.pone.0191466

Jia, S., Hui, F., Li, S., Zhao, X., & Khattak, A. J. (2020). Long short-term memory and convolutional neural network for abnormal driving behaviour recognition. *IET Intelligent Transport Systems*, *14*(5), 306–312. https://doi.org/10.1049/iet-its.2019.0200

Kusuma, A., Liu, R., & Choudhury, C. (2020). Modelling lane-changing mechanisms on motorway weaving sections. *Transportmetrica B*, *8*(1), 1–21. https://doi.org/10.1080/21680566.2019.1703840

Lesouple, J., Baudoin, C., Spigai, M., & Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, *149*, 109–119. https://doi.org/10.1016/j.patrec.2021.05.022

Li, X., Wang, W., & Roetting, M. (2019). Estimating Driver's Lane-Change Intent Considering Driving Style and Contextual Traffic. *IEEE Transactions on Intelligent Transportation Systems*, *20*(9), 38–3271. https://doi.org/10.1109/TITS.2018.2873595

Ly, M. Van, Martin, S., & Trivedi, M. M. (n.d.). *Driver Classification and Driving Style Recognition using Inertial Sensors*.

Manager, B. (2009). *Road Traffic Crashes in Queensland 2009*. http://www.tmr.qld.gov.au

Mohammadnazar, A., Arvin, R., & Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. *Transportation Research Part C: Emerging Technologies*, *122*. https://doi.org/10.1016/j.trc.2020.102917

Ng, C., Susilawati, S., Kamal, M. A. S., & Chew, I. M. L. (2020). Development of a binary logistic lane change model and its validation using empirical freeway data. *Transportmetrica B*, *8*(1), 49–71. https://doi.org/10.1080/21680566.2020.1715309

Nikita Butakov. (2020). *How to build robust anomaly detectors with machine learning*.

Peden, M. M., World Health Organization., & World Bank. (2004). *World report on road traffic injury prevention*. World Health Organization.

Peng, J. S., Wang, C. W., Fu, R., & Yuan, W. (2020). Extraction of parameters for lane change intention based on driver's gaze transfer characteristics. *Safety Science*, *126*. https://doi.org/10.1016/j.ssci.2020.104647

Peter J.Rousseeuw, & Driessen Van Katrien. (1999). *A_fast_algorithm_for_the_minim*.

Saiprasert, C., & Pattara-Atikom, W. (2013). Smartphone enabled dangerous driving report system. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1231–1237. https://doi.org/10.1109/HICSS.2013.484

Su, Z., Woodman, R., Smyth, J., & Elliott, M. (2023). The relationship between aggressive driving and driver performance: A systematic review with meta-analysis. *Accident Analysis and Prevention*, *183*. https://doi.org/10.1016/j.aap.2023.106972

WHO. (2022, June 20). *Road traffic injuries*.

Yang, Q., Lu, F., Wang, J., Zhao, D., & Yu, L. (2020). Analysis of the insertion angle of lane-changing vehicles in nearly saturated fast road segments. *Sustainability (Switzerland)*, *12*(3). https://doi.org/10.3390/su12031013

Zheng, O., Abdel-Aty, M., Yue, L., Abdelraouf, A., Wang, Z., & Mahmoud, N. (2022). *CitySim: A Drone-Based Vehicle Trajectory Dataset for Safety Oriented Research and Digital Twins Figure 1. Post encroachment time conflicts in a single frame from the CitySim dataset Expressway A weaving segment location*. https://github.com/ozheng1993/UCF-SST-CitySim-Dataset.