Comparative Analysis of SMOTE and ROSE Oversampling Techniques for kNN-Based Autonomous Vehicle Behavior Modeling

Celine Serbouh Touazi, Iness Ahriz, Ndeye Niang, and Alain Piperno

***Abstract*—In this paper, we present a comparative analysis of Synthetic Minority Oversampling TEchnique (SMOTE) and Random OverSampling Examples (ROSE) oversampling techniques for K Nearest Neigbhors KNN-based autonomous vehicle behavior modeling. We address the challenges posed by imbalanced and mixed datasets in the context of autonomous vehicle testing, where the majority of test outcomes are classified as "OK" (safe) and fewer as "KO" (unsafe). We propose an enhanced approach that extends our previous work by incorporating ROSE as an alternative to SMOTE for generating synthetic samples. We integrate these resampling techniques with Leave-One-Out Cross-Validation (LOO-CV), applying resampling at each iteration to ensure data balancing is tailored to each training set. Additionally, we investigate the impact of different encoding strategies for categorical variables, including OneHot, binary encoding, and Factor Analysis of Mixed Data (FAMD). Our research aims to develop a robust classification model capable of accurately predicting autonomous vehicle behavior while effectively managing class imbalance and mixed data types, despite the limited availability of data due to costly and time-consuming testing procedures.**

***Index Terms*—IA, Autonomous Vehicles, SMOTE, ROSE.**

I. Introduction

The integration of Artificial Intelligence (AI) in various domains, including smart cities, is prominently reflected in vehicles equipped with Advanced Driver Assistance Systems (ADAS) and autonomous vehicles. The AI embedded in autonomous vehicles, which spans a range of functionalities, including perception, localization, and more, faces two significant challenges: a lack of repeatability, leading to variable behaviors under identical conditions, and a lack of robustness due to overfitting, which impairs its ability to generalize to new situations [1]. Given that vehicles are critical safety systems, even a minor failure rate is unacceptable. Therefore, before being marketed in Europe, vehicles must undergo a series of laboratory and open road tests that simulate real world conditions. Track tests on vehicles with proprietary models are notably costly and time-consuming due to the extensive setup of testing resources, resulting in limited available data. The tests are carried out with different values of the variables that define the scenarios. The variables encompass both qualitative (categorical) and quantitative (numerical), resulting in a mixed data set.

Furthermore, during track testing, it was observed that there were more successful tests, labeled OK, than unsuccessful ones, labeled KO. Since the Autonomous Driving system (AD) is designed, validated, and tested primarily within its Operational Design Domain (ODD), the test results are predominantly OK, with occasional KO results. This leads to imbalanced data, as the predominance of successful outcomes complicates the accurate prediction and identification of potential failures.

Thus, our primary objective is to construct a robust classification model capable of accurately predicting the behavior of autonomous vehicles. This includes addressing the challenges posed by unbalanced and mixed datasets, where input variables encompass both qualitative and quantitative data. Additionally, due to the limited amount of data resulting from costly and time-consuming tests, our focus is on developing a model that can effectively manage the uneven distribution of data classes, ensuring reliable predictions across various real-world scenarios encountered on the road.

In our previous work [2], we successfully addressed our objective by implementing three data resampling techniques: SMOTE, SMOTE-NC, and SMOTE-ENC. These methods were applied to tackle class imbalance and improve the performance of classification models. Building on this foundation, the present paper offers several key contributions to the field of autonomous vehicle behavior modeling:

- Extended resampling techniques: We introduce ROSE (Random OverSampling Examples) as an alternative to SMOTE for generating synthetic samples in imbalanced datasets. This expansion provides a new perspective on addressing class imbalance in the context of autonomous vehicle testing.
- Integration with Leave-One-Out Cross-Validation: We propose a novel approach of applying resampling techniques at each iteration of the LOO-CV process. This ensures that data rebalancing is tailored to each training set while maintaining the integrity of the test data, leading to a more robust evaluation of model performance.

Manuscript received January 20, 2025; revised February 6, 2025. Date of publication April 24, 2025. Date of current version April 24, 2025.

C. S. Touazi, I. Ahriz and N. Niang are with the Conservatoire National des Arts et Metiers, France (e-mails: celine.serbouh@lecnam.net, iness.ahriz@lecnam.net, n-deye.niang_keita@lecnam.net). A. Piperno is with the Union technique de l'automobile, du motocycle et du cycle (UTAC), France (e-mail: alain.piperno@utac.com).

The paper was presented in part at the International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 2024.

Digital Object Identifier (DOI): 10.24138/jcomss-2024-0121

- Comprehensive encoding strategy comparison: We investigate the impact of different encoding strategies for categorical variables, including OneHot, binary encoding, and Factor Analysis of Mixed Data (FAMD). This comparison aims to determine the most effective method for preserving information and enhancing resampling technique performance in mixed-type datasets.
- Application to autonomous vehicle behavior modeling: We apply these advanced techniques to the critical domain of autonomous vehicle testing, addressing the challenges of limited data availability due to costly and time-consuming testing procedures.

The rest of this article is organized as follows: in Section II, we present the related works, highlighting the existing approaches for data sampling and their applications. Section III describes in detail the adopted methodology, including the SMOTE and ROSE oversampling techniques and their integration with the kNN algorithm. Section IV focuses on the experiments carried out, the configurations used and the evaluation metrics, and the results are discussed in this section. Finally, Section V concludes the article by summarizing the main contributions and suggesting avenues for future research.

II. Related Works

In the field of machine learning, the quality and balance of training data are crucial for ensuring both model performance and generalizability. When datasets exhibit a significant degree of class imbalance—where certain classes are considerably underrepresented—this imbalance can introduce substantial bias into the learning process. Such bias often causes models to favor the majority class, leading to suboptimal performance, especially in terms of precision and recall for minority classes. This challenge becomes particularly critical in high-stakes applications such as fraud detection, medical diagnosis, and, pertinent to this study, the modeling of autonomous vehicle behavior. In these cases, the inability to effectively identify and account for minority classes can not only degrade the model's overall effectiveness but also lead to costly misclassifications with serious implications for safety in sensitive contexts like autonomous driving.

Class imbalance also affects traditional classification methods by skewing their focus toward the prevalent class and diminishing their performance with respect to rare or minority classes [3] [4]. For instance, logistic regression tends to underestimate the probabilities of the minority class in skewed datasets, while linear discriminant analysis can exhibit bias due to unequal covariance matrices that favor the dominant class. Even nonparametric methods, designed to optimize classification accuracy, may fail when accuracy measures are calculated without regard to class distribution, often resulting in high performance for the majority class at the expense of the minority class.

In addition to training challenges, evaluating model performance in the presence of rare classes presents its own set of complexities. In classification tasks, evaluating the accuracy of the classifier is as critical as model training, particularly in a class imbalance context. This is because both the selection of the best classification rule among alternatives and its applicability to real-world problems hinge on accurate performance measurements. The choice of evaluation metrics becomes pivotal, as common measures like overall error rate can be misleading in imbalanced settings. For example, in a dataset where a rare class represents only 1% of the data, a naive classifier that assigns all observations to the majority class could achieve a 99% accuracy, yet fail entirely in identifying the minority class [5] [6] [7]

To address this, class-independent metrics such as precision, recall, F-measure, and G-mean have been proposed, derived from confusion matrix observations. Precision measures the fraction of positive predictions that are correct, while recall focuses on the fraction of true positives identified. While precision is affected by class distribution, recall alone provides limited insight into false positive rates. These measures are often used together or combined into composite scores like the F-measure or G-mean. The Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) metric are also popular tools for evaluating classifiers in imbalanced contexts. ROC curves illustrate the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity), with steeper curves and larger AUC values indicating better performance. A completely random classifier would result in a diagonal ROC curve, while a perfect classifier yields a point in the top-left corner of the ROC space [8].

Despite the advances in performance metrics, challenges remain in estimating model accuracy reliably, especially for rare classes. Popular approaches, such as the apparent error (resubstitution) or holdout method, and more sophisticated methods like cross-validation or bootstrapping, are commonly employed [9]. However, the scarcity of rare class examples often leads to high-variance error estimates, undermining confidence in performance assessments. This limitation underscores the need for robust evaluation strategies that accurately reflect classifier effectiveness in real-world, imbalanced data scenarios. Without such strategies, even the most sophisticated learning methods risk yielding misleading conclusions about their performance, particularly when applied to sensitive domains with rare events.

To address the challenges of imbalanced data, researchers in [10] have organized existing methods into four principal categories. The first category, data-level approaches, aims to adjust the distribution of training data to reduce imbalance. These techniques typically involve either oversampling the minority class or undersampling the majority class, or in some cases, a hybrid approach combining both. By balancing the dataset at the data level, these methods enhance the model's exposure to minority class examples, thus improving its ability to learn from underrepresented instances without modifying the underlying algorithm. The second category consists of algorithm-level approaches, which adapt the learning algorithms themselves to mitigate the effects of data imbalance. Rather than modifying the dataset, these methods involve making adjustments within the algorithm's structure to improve sensitivity to minority classes. By altering aspects of the learning process such as adjusting decision thresholds, modifying loss functions, or introducing minority class-

focused constraints—these techniques allow models to handle imbalanced datasets more effectively, enhancing classification performance for all classes. The third category, cost-sensitive methods, incorporates class-dependent misclassification costs directly into the learning process. By assigning higher penalties to errors associated with the minority class, cost-sensitive approaches create an internal bias within the model that prioritizes accuracy for underrepresented classes. This strategy has proven valuable in applications where the cost of misclassifying minority class instances is significantly higher, allowing models to operate with greater sensitivity to the specific requirements of imbalanced data contexts. A fourth category has recently emerged: deep learning-based methods. Deep learning techniques have shown considerable promise in handling data imbalance due to their capacity for complex feature transformation and representation. By leveraging the high dimensionality and adaptability of deep neural networks, these methods create sophisticated features that can better capture the underlying patterns of minority classes, even in highly imbalanced datasets. This emerging class of techniques is particularly well-suited to applications involving large-scale, complex data, where traditional approaches may struggle to adequately represent minority class characteristics. Together, these four categories encompass a range of approaches aimed at mitigating the impact of data imbalance on model performance, providing researchers and practitioners with a diverse set of tools to enhance classification accuracy across imbalanced datasets.

One approach to tackle the challenge of imbalanced data involves employing algorithms that dynamically adjust the learning phase of classification models to accommodate the inherent disparities. Within scholarly research, a spectrum of techniques has emerged, typically falling into two overarching categories as outlined by [11]: single-class learning algorithms and ensemble learning methods.

Single-class learning algorithms operate by training models exclusively on data representing a single class, a strategy tailored to effectively address imbalanced datasets. Their primary objective is to identify instances belonging to the majority class and subsequently classify new samples using similarity metrics. This approach serves as a cornerstone in mitigating the impact of class imbalances on model performance.

For instance, in [12] authors proposed an innovative anomaly detection technique centered around a one-class Support Vector Machine (SVM), leveraging its capability to discern outliers in datasets characterized by skewed distributions. Similarly, authors in [13] leveraged a one-class SVM to discern deviations between normal and anomalous data, exemplifying the versatility of this approach in detecting and addressing imbalances within datasets.

Numerous investigations have delved into the realm of ensemble learning as a solution to the challenges posed by imbalanced data, with particular emphasis on two primary methodologies: bagging and boosting.

The classic bagging approach entails the utilization of bootstrap sampling. In case of classification multiple classifiers are trained on diverse subsets of the dataset, culminating in a collective decision through voting for the final prediction. To tailor bagging to address the intricacies of imbalanced datasets, various adaptations have been proposed. For instance, UnderBagging, introduced by [14], alleviates the impact of class imbalance by strategically under-sampling the majority class.

In contrast, boosting operates by iteratively training classifiers with a focus on rectifying misclassifications encountered during previous iterations. AdaBoost. M2 stands as a prominent example of this methodology [15].

In the pursuit of solutions for handling imbalanced data, another effective avenue lies in the realm of cost-sensitive methods. These algorithms take into account the costs associated with misclassification during their internal operations. Specifically, they assign a higher cost to the misclassification of minority class instances compared to majority class instances [10]. By incorporating this cost consideration into their decision-making process, these methods aim to optimize performance in scenarios where imbalanced classes pose significant challenges.

Moreover, the domain of deep learning has emerged as a powerful tool in various domains, demonstrating remarkable efficacy in tackling imbalanced datasets. Notably, [16] proposed a novel approach leveraging Deep Neural Networks (DNNs) to address imbalanced data scenarios. Their methodology involves utilizing DNNs to extract intricate features from samples belonging to the minority class, subsequently generating new pseudo-features to compensate for the scarcity of minority class samples. It's crucial to note that this approach doesn't generate entirely new data instances; instead, it focuses on enhancing the classification capabilities for unique and underrepresented samples through feature augmentation.

Finally, the simplest and most intuitive approach to address imbalanced data is to intervene at the data level, either by oversampling, adding new instances to the minority class, or by removing some elements from the majority class, or by employing a hybrid process.

In our study, due to constraints related to a limited dataset, we have chosen to use a combination of oversampling techniques to address class imbalance, specifically SMOTE (Synthetic Minority Over-sampling Technique), Random Oversampling, and ROSE (Random Over-Sampling Examples). SMOTE involves generating synthetic instances for the minority class by interpolating between existing examples of the same class within the feature space, which helps to reduce overfitting compared to simple replication [17].

Random Oversampling, a simpler approach, balances the dataset by randomly duplicating existing instances of the minority class. While effective in improving class representation, this method can sometimes lead to overfitting, as duplicated instances do not introduce new variability to the dataset.

To mitigate this limitation, we also employed ROSE, an advanced oversampling method that generates synthetic examples for both classes through a smoothed bootstrap sampling approach. This technique leverages kernel density estimation to create synthetic data points around the existing instances, introducing controlled variability while maintaining the underlying distribution of the dataset. ROSE not only addresses the imbalance but also helps improve the model's robustness

and generalization by avoiding the rigidity introduced by duplicated or overly-similar synthetic data [18].

Essentially, SMOTE increases the representation of the minority class by creating new data points that closely resemble existing minority instances. This method strategically interpolates between neighboring samples of the minority class, effectively addressing the imbalance issue. As a result, it enhances the model's capability to identify patterns and make accurate predictions across all classes.

Moreover, by leveraging SMOTE in the context of our study, we aim to bolster the robustness and generalization capabilities of our predictive model, enabling it to perform more reliably in real-world scenarios characterized by imbalanced class distributions.

Expanding on the groundwork laid by the original SMOTE technique, a plethora of adaptations and variations have since been introduced to further refine its efficacy. For instance, [19] documented approximately 100 variants of SMOTE by 2018, and the landscape has continued to evolve with the emergence of new approaches.

Among these variants, Borderline-SMOTE, introduced by [20], targets individuals belonging to the minority class residing at the borders, specifically those instances with neighboring points in the majority class. Similarly, [21] proposed two additional variants, namely SMOTE-ENN and SMOTE-Tomek, which take into account distribution overlaps and class boundaries when generating synthetic instances.

Furthermore, [22] introduced the Adaptive synthetic sampling approach for imbalanced learning (ADASYN) method , which focuses on synthesizing data points for challenging instances of minority classes by comprehensively understanding the underlying distribution of these instances. [23] proposed SMOTE-D, a deterministic approach that synthesizes artificial data points specifically for negative classes (majority class).

However, it's worth noting that while the original SMOTE technique and its variants are well-suited for numerical data, they face challenges when applied to datasets containing categorical variables. In such cases, these methods may fail to accurately identify the categories of qualitative variables, potentially leading to the creation of new and unintended categories.

To address this limitation, [17] introduced an extension known as SMOTE-NC (SMOTE-Nominal Continuous), which specifically caters to datasets with both nominal and continuous variables. Building upon this, [24] proposed a further extension named SMOTE-ENC, which enhances the capabilities of SMOTE-NC by refining its handling of categorical data.

In the field of road safety and autonomous vehicles (AVs), addressing the challenges posed by imbalanced datasets has become a focal point of research, particularly in predicting accident severity and associated risks. Authors of [25] proposed an innovative methodology to overcome these challenges by utilizing California DMV collision reports (2019-2021) enriched with environmental and road data from OpenStreetMap. Through the application of resampling techniques, such as SMOTE and ROSE, they successfully balanced data classes, enhancing the predictive accuracy of their models. Furthermore, their study integrated advanced feature selection methods, including Mutual Information, Random Forest, and XGBoost, to identify critical factors influencing accident severity, such as vehicle manufacturers, collision types, and points of interest (POIs). The combination of SMOTE and Random Forest demonstrated the highest predictive performance, highlighting the value of a well-balanced dataset before feature selection.

Similarly, authors of [26] focused on cybersecurity in the Internet of Autonomous Vehicles (IoVs) by developing an intelligent intrusion detection system (IDS) capable of addressing the imbalance inherent in car hacking datasets. Using rebalancing techniques such as NearMiss, Random Over-Sampling (ROS), and TomLinks in conjunction with machine learning models like k-Nearest Neighbors (k-NN), Logistic Regression, and Naive Bayes, their approach achieved exceptional performance. Notably, the k-NN model combined with ROS and TomLinks reached a 100% detection rate, surpassing existing methods and demonstrating the importance of tailored strategies for imbalanced datasets in enhancing system robustness.

Lane-changing behavior (LC), a critical component of traffic safety research, has also been extensively studied due to its complexity and its potential to cause accidents. Previous approaches have largely relied on machine learning models for risk prediction, yet these methods often face limitations due to class imbalances, where high-risk instances are significantly underrepresented. To address these challenges, oversampling techniques such as SMOTE and ADASYN, as well as generative adversarial networks (GANs), have been employed. However, their reliability in practical scenarios remains variable. Recent advancements, such as the CMSS (Control Method for Synthetic Samples) developed by [27], offer a promising alternative by integrating oversampling techniques with optimization algorithms like Particle Swarm Optimization (PSO). This method not only improves predictive accuracy but also enhances model generalizability and interpretability using tools like SHAP (Shapley Additive Explanations) to strategically adjust synthetic samples.

These studies collectively underscore the critical importance of developing robust methodologies for handling imbalanced datasets. They demonstrate significant progress in improving predictive models, enhancing cybersecurity in connected vehicle networks, and advancing risk detection systems, thereby contributing to safer and more reliable autonomous driving technologies.

In this paper, we aim to apply well-established methods for addressing unbalanced data to the task of predicting the behavior of AI-equipped vehicles. Specifically, we focus on scenarios that push the boundaries of their operational domains. By leveraging these advanced techniques, we seek to enhance the accuracy and reliability of behavioral predictions for autonomous vehicles operating under challenging and edge-case conditions. This approach not only addresses the inherent data imbalance but also contributes to the robustness of AI models in critical real-world applications where ensuring safety and performance is paramount.

III. System Model

We consider a real-world dataset $X_{l,p}$where $l \in [1, L]$ and $p \in [1, P]$. L is the number of individuals and P is the number of variables. We have a mixed dataset, i.e., M qualitative variables and N quantitative variables such that $P = M + N$.

The dataset used in this study is proprietary and originates from tests conducted by UTAC, the official organization responsible for vehicle approval in France. Due to the sensitive nature of vehicle testing data, accessing public datasets in this domain proves to be particularly challenging, as confidentiality and data protection are critical concerns.

Since we are dealing with binary classification, we will adopt the notation used in the literature. The positive class corresponds to the minority class, and the negative class corresponds to the majority class [28]. We consider Imbalance Ratio (IR) as the ratio of the number of instances in the majority class to the number of instances in the minority class [29]. Mathematically, it can be expressed as:

$$IR = \frac{C_{maj}}{C_{min}} \tag{1}$$

such as:

- C_{maj} represents the size of the majority class
- C_{min} represents the size of the minority class

IR = 1 represents a perfectly balanced dataset. The dataset is considered imbalanced when $IR > 1.5$, and extremely imbalanced when $IR \geq 9$ [28].

In the context of our study, a significant constraint is the limited size of our dataset, a situation frequently encountered in many real-world applications, particularly in behavioral prediction for autonomous vehicles. This small data problem increases the risks of overfitting and makes it challenging to accurately assess the model's performance. To address this limitation and maximize the use of each available instance, we have chosen to implement cross-validation. This approach allows us to evaluate the model's robustness while effectively using the available data.

In this context, we have opted to implement the Leave-One-Out Cross-Validation (LOO-CV) method. Unlike traditional k-fold cross-validation, where the data is divided into multiple subsets, LOO-CV involves using a single instance as the test set and the remaining data as the training set, iteratively taking a new observation as the test set each time. This technique is particularly suitable for small datasets as it enables the use of all available data for training in each iteration, thus maximizing the exploitation of rare examples and minimizing the risk of bias due to a small sample size.

An important aspect of our approach is the integration of oversampling techniques to manage class imbalance. However, it is crucial to note that oversampling techniques are applied exclusively to the training data at each step of the cross-validation process. Thus, during each iteration of the LOO process, the training data is resampled to adjust the class distribution, while the instance used as the test data remains an unmodified real example. This approach helps maintain the integrity of the test data and prevents model contamination, ensuring that oversampling does not influence the evaluation of the model's performance on unseen data. At each iteration, the imbalance ratio is kept constant and set to 1.5, which is the value recommended based on the results of our previous research. This rate was chosen because it provided optimal performance in that context. Consequently, we propose an adaptive approach that ensures the number of elements remains consistent across iterations. This approach also ensures that the testing is always performed on real examples from the original dataset, preserving the integrity of the evaluation process while maintaining the desired balance during training. This method is detailed in the algorithm 1, which illustrates how, at each step of the LOO, only the training data is oversampled.

Algorithm 1: Pseudo Algorithm of kNN-over-LOO

X : Initial dataset
L : Number of individuals
for $i = 1$ *to* L **do**
  X_{test} := $X[i]$
  X_{train} := $X[N - i]$
  $X_{train-resample}$:= RESAMPLE(X_{train})
  Train(kNN($X_{train-resample}$))
end

The RESAMPLE method represents one of the various oversampling techniques like SMOTE, SMOTE-NC, SMOTE-ENC, and ROSE.

A. Resampling Method based on SMOTE

In this section we will give more details about the application of SMOTE method and its variants to our problem.

1) SMOTE (Synthetic Minority Over-sampling Technique): This method is used to oversample individuals in the positive class. A randomly selected element x_l from the minority class is chosen, and the K-nearest neighbors (KNN) algorithm is applied to find its K nearest neighbors. Then a new element r_j is generated between$x_l$and one of its neighbors x_{lk}using a parameter α, according to the following equation.

$$\mathbf{r}_k = \mathbf{x}_l + \alpha.(\mathbf{x}_{lk} - \mathbf{x}_l) \tag{2}$$

2) SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous): it is a variant of SMOTE adapted to mixed data [17] that does not require encoding categorical variables.

Its operation follows algorithm 2.

When identifying the K-nearest neighbors, the distance calculation is adjusted slightly to account for the presence of non-encoded qualitative variables in the dataset. For nominal features, the distance is defined as the median of the standard deviations of the quantitative values, ensuring consistency regardless of the labels.

While SMOTE-NC performs well with binary classes, it faces challenges when dealing with nominal features that have multiple labels. Specifically, it struggles to interpret the varying relationships between each label and the minority class target. Moreover, SMOTE-NC requires the dataset to include at least one continuous attribute to operate correctly. To address these limitations, an improved version of SMOTE, named SMOTE-ENC, was introduced in [24]

Algorithm 2: Pseudo algorithm of SMOTE-NC

for $\boldsymbol{x}_l$ *positive individuals* **do**
  Identify the K nearest neighbors of $\mathbf{x}_l$;
  for *p=1:P* **do**
    if $x_{l,p}$ *is numerical* **then**
      Randomly choose one of the K nearest neighbors, denoted as $\mathbf{x}_{lk}$;
      $r_{k,p} = \alpha \cdot x_{l,p} + (1-\alpha) \cdot x_{lk,p}$;
    else
      $r_{k,p}$ = mode of the most frequent category among the K nearest neighbors;
    end
  end
end

3) SMOTE-ENC (Synthetic Minority Over-sampling Technique Encoded Nominal Continuous): It is an improved version of the SMOTE-NC variant, specifically developed for mixed datasets, which encodes nominal variables as numerical values. In this method, the numerical differences represent the importance of the change in relation to the minority classes.

The encoding of qualitative variables relies on χ^2 distance, which determines if two categorical variables are associated. However, the objective here is not to analyze the relationship between the target variable and the categorical variables, but rather to measure the distance between two points with different labels [24].

Algorithm 3: Encoding with SMOTE-ENC [24]

N: the number of continuous variables;
M: the number of categorical variables;
v: median of the standard deviations of continuous variables;
C_{min}: size of the minority class;
S: size of the training dataset;
$IR_{ENC} = \frac{C_{min}}{S}$;
for $m = 1 : M$ **do**
  for *each label E* **do**
    E_m = total number of label of variable m in the training dataset;
    $E'_m = E_m \times IR_{ENC}$;
    E_m^{min} number of labels in the minority class;
    $\chi = \frac{E_m^{min} - E'_m}{E'_m}$;
    if $N > 0$ **then**
      $E = \chi \times v$;
    else
      $E = \chi$;
    end
  end
end

After all categorical variables have been transformed into quantitative variables using the algorithm 3, a new dataset consisting solely of quantitative variables is created. At this stage, the SMOTE (Synthetic Minority Over-sampling Technique) method can be applied to balance the minority classes within the dataset.

B. Resampling Method based on ROSE

The principle behind ROSE is based on the idea that any additional data we collect follows the probability distribution of the underlying population of data belonging to the minority class. Therefore, one approach is to approximate this probability distribution and then sample from it to simulate the collection of real examples [18]. This is exactly what the ROSE algorithm does, as shown in algorithm 4.

Algorithm 4: Pseudo algorithm of ROSE

X: input data;
y: output labels;
k: minority class;
N_k: number of synthetic points to generate;
s: control factor for the kernel width ($s = 1$ by default);
Step 1: Estimate the conditional distribution $P(x|y = k)$
Compute the smoothing matrix h using Silverman's rule:

$$h = s \cdot N_k^{-1/(d+4)} \cdot D_\sigma$$

where D_σ is the diagonal matrix of feature standard deviations, and d is the dimensionality of the data.
Step 2: Generate synthetic points
for $i = 1$ ***to*** N_k **do**
  Randomly select a point x_i from X_k;
  Place a Gaussian kernel centered on x_i;
  Sample a new point x_{new} from this kernel;
  Add x_{new} to $X_{\text{synthetic}}$;
end

1) Estimating the conditional distribution $P(x|y = k)$ ROSE estimates the probability distribution $P(x|y = k)$ for each class k using kernel density estimation (KDE), a method for estimating the underlying probability distribution from observed data. such that

$$p(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

KDE operates by selecting a kernel function $K(x)$ (commonly a Gaussian) and positioning this function over each data point. The kernel function is then scaled and summed to produce a smooth estimate of the probability distribution. The scale of the kernel can be adjusted to improve the accuracy of the estimate.
The kernel function is a hyperparameter, and while there are various options for kernels, a Gaussian kernel with a scale parameter σ is commonly used, as it satisfies basic properties like smoothness and symmetry. In the case of ROSE, this Gaussian kernel is used to estimate the distribution of each class and generate synthetic samples. Where:

- $p(x)$ is the estimated density at x.
- n is the total number of data points.
- x_l are the data points.
- $K(\cdot)$ is the kernel function (for example, a Gaussian function).
- h is the bandwidth parameter, which controls the width of the kernel.

2) Generating synthetic points
   - Randomly select a point
   - Center a Gaussian distribution on it
   - Then sample a point from the Gaussian distribution.

C. Proposed Strategy

To apply the classification method to imbalanced data, we will follow four distinct strategies as shown in figure 1. These strategies are designed to enhance the model's performance by addressing the class imbalance issue, thereby ensuring better accuracy and robustness of the predictions.

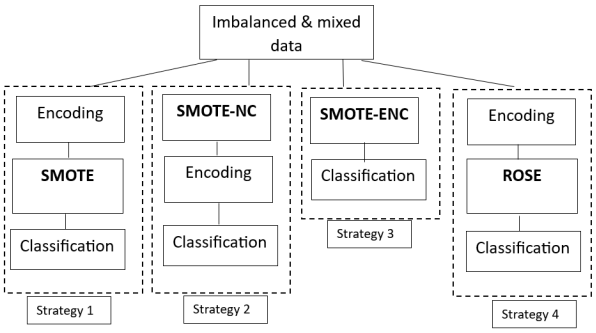

Fig. 1. Proposed strategies for classifying imbalanced data

- *Strategy 1*: In this strategy, we use the classic version of SMOTE, which applies to numerical data. Since our dataset contains mixed data, we consider two approaches:
  - Perform binary encoding for variables with two categories and OneHot encoding for variables with multiple categories.
  - Alternatively, use the FAMD (Factor Analysis of Mixed Data) method to directly process mixed data.

  Once all categorical variables have been handled (either encoded or transformed using FAMD), we will apply SMOTE to rebalance the minority classes. Finally, we will use a classifier on this rebalanced dataset to improve the accuracy and robustness of the predictions.
- *Strategy 2*: Since SMOTE-NC is suitable for mixed data, we apply it to our dataset without requiring prior encoding of variables, as mentioned in Strategy 1. SMOTE-NC directly handles both numerical and categorical variables, simplifying the class rebalancing process. After applying SMOTE-NC and obtaining a balanced dataset, we will consider binary/OneHot encoding or the FAMD approach to process categorical variables. Finally, we will use a classifier on this rebalanced dataset to improve the accuracy and robustness of the predictions.
- *Strategy 3*: This strategy relies on SMOTE-ENC, which integrates its own encoding technique, detailed in [24]. Unlike traditional approaches that require a separate step for encoding categorical variables, SMOTE-ENC directly incorporates this process into its resampling algorithm. Once the data is rebalanced using SMOTE-ENC, we proceed with classification.
- *Strategy 4*: This strategy is similar to Strategy 1 but replaces SMOTE with the ROSE algorithm. Like in Strategy 1, we handle mixed data using either binary/OneHot encoding or the FAMD method. ROSE will then be applied to generate synthetic samples by estimating the probability distribution $P(x|y = k)$ for the minority class and drawing samples from it. Finally, we will train a classifier on this rebalanced dataset to improve prediction accuracy and robustness.

IV. Simulation & Results

In this section, we present conducted simulations and analyze the results obtained, highlighting key insights and evaluating the performance of the proposed approach under various conditions

A. Simulation Setup

We consider in this study a dataset comprising $L = 142 \times a$ individuals and $P = 21 \times a$ variables, categorized as $N = 6 \times a$ quantitative variables and $M = 15 \times a$ qualitative variables. Among the qualitative variables, $14 \times a$ are binary, and hence, binary encoding will be applied to them. Additionally, a variables with multiple categories will undergo OneHot Encoding (OHE) as detailed in [30]. For the rest of the numerical application $a = 1$

To simplify the initial study, we opted to use kNN-based classification. Therefore, we first normalize the data using Standard Scaler, which involves centering and scaling each variable.

The output variable y is binary, with two categories: OK representing the negative class $C_{maj} = 127$ and KO representing the positive class $C_{min} = 15$. This results in an imbalance ratio $IR = \frac{127}{15} = 8,46 >> 1.5$, which indicating a scenario of imbalanced dataset.

As specified in algorithm 1 we will use the LeaveOneOut process for the traing and test of the model. Our assessment will utilize evaluation metrics (accuracy, precision, recall and F1 score) as specified in [8].

B. Results & discussion

As the primary goal is to implement a classifier, this section is dedicated to evaluating the performance of the k-Nearest Neighbors (kNN) algorithm combined with the Leave-One-Out (LOO) validation technique. The classification is implemented with original data by encoding the categorical variables by the mean of OneHot encoding. The kNN algorithm depends on a key hyperparameter, k, which specifies the number of neighbors taken into account. The value of k is determined empirically by varying it and analyzing the corresponding performance metrics. This process enables the identification of the optimal k value that maximizes accuracy.

TABLE I
KNN PERFORMANCE WITH IMBALANCED DATA

	Original data	*AFDM*
k	10	8
Confusion Matrix	Real Class \ Predicted Class: KO: KO 1, OK 14; OK: KO 0, OK 127	Real Class \ Predicted Class: KO: KO 1, OK 14; OK: KO 0, OK 127

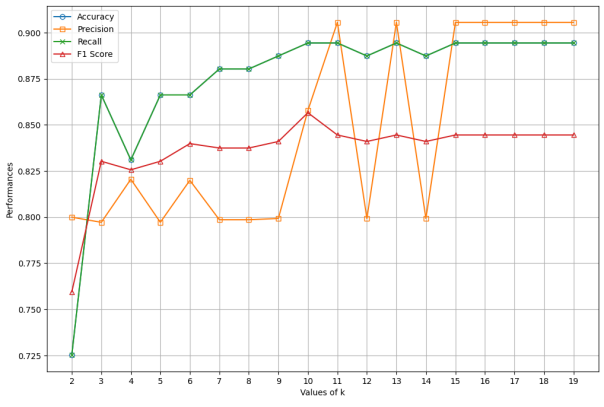

Fig. 2. Performance of kNN with imbalanced original data

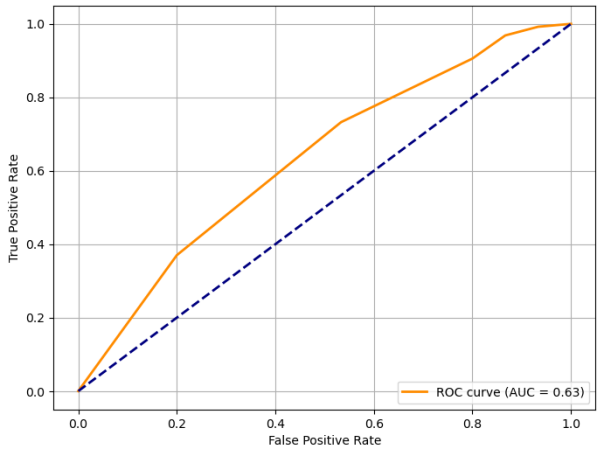

Fig. 3. ROC curve with imbalanced original data ($k = 10$)

1) Performances of Different Strategies with OneHot/Binary Encoding: Figure 2 depicts the performance of the kNN algorithm as a function of the hyperparameter k. The best performance is achieved at $k = 10$, where the model attains an accuracy of 88%. However, a closer analysis reveals an accuracy of 99% for the majority OK class and only 6% for the minority KO class. This indicates that the model struggles to correctly classify the minority class and is biased towards the majority class. Furthermore, an examination of the ROC curve and the AUC score in Figure 3 highlights the model's suboptimal behavior, with an AUC of 0.63, suggesting near-random performance. This poor outcome is primarily attributed to the imbalanced class distribution, which significantly hampers the model's ability to effectively differentiate between the two classes.

Table I presents the confusion matrix, providing a detailed view of the model's performance. The results indicate that the model incorrectly classified one OK instance as "KO" and misclassified 14 KO instances as OK. These findings highlight the significant challenges posed by the imbalanced class distribution, which causes the model to favor the majority OK class at the expense of accurately identifying the minority KO class.

In our study aimed at predicting whether a vehicle test outcome is OK or KO, we emphasize the importance of both false positives and false negatives. Misclassifying a KO as OK could lead to safety risks, while misclassifying an OK as KO results in unnecessary losses for the manufacturer due to disqualification. To mitigate these issues, we evaluated the model's accuracy for both OK and KO classifications.

As highlighted in Table II, Strategy 2 demonstrates the highest performance for OK tests, misclassifying only 3 OK instances as KO. However, for KO tests, Strategy 1 proves to be the most effective, achieving 60% accuracy, with Strategy 4 following closely. In stark contrast, Strategy 2 performs poorly on KO tests, failing to correctly classify any KO instances and yielding an AUC of 0.45, indicative of the kNN model's near-random behavior.

Notably, Strategies 1, 3, and 4 significantly improve the accuracy for the KO class compared to the standard kNN approach, which struggles with the challenges posed by completely imbalanced data.

2) Performances of Different Strategies with FAMD Encoding: The ROC curve shown in Figure 5, with an AUC of 0.53, indicates that the model's ability to differentiate between classes is nearly equivalent to random guessing. This is further corroborated by the confusion matrix, which shows that only 3 instances of the KO class are correctly identified, while 12 are misclassified. In contrast, the OK class is better handled, with 101 correctly classified instances versus 26 errors. These results underscore the limitations of Strategy 1 in addressing class imbalance.

When comparing the performance of kNN across datasets encoded using binary encoding, One-Hot encoding, and AFDM principal components, a similar overall behavior is observed. However, the model leveraging principal components distinguishes itself by accurately classifying all OK instances.

Table III shows a notable improvement in the accuracy for KO, which increases from 6% with standard kNN to 20% and 13% with Strategies 1 and 4, respectively. This underscores the importance of addressing class imbalance before developing classifiers.

TABLE II
COMPARISON OF DIFFERENT STRATEGIES BASED ON PERFORMANCE METRICS FOR *OneHot/Binary encoding* ($IR = 1.5$)

	Strategy 1	*Strategy 2*	*Strategy 3*	*Strategy 4*
$Accuracy_{OK}$	73%	98%	76%	83%
$Accuracy_{KO}$	60%	0%	33%	53%
AUC	0.68	0.45	0.65	0.67
Confusion Matrix	Real Class \ Predicted Class: KO: 9, 6; OK: 34, 93	Real Class \ Predicted Class: KO: 0, 15; OK: 3, 124	Real Class \ Predicted Class: KO: 5, 10; OK: 30, 97	Real Class \ Predicted Class: KO: 8, 7; OK: 22, 105

By comparing Tables II and III, it is evident that One-Hot/Binary encoding outperforms the other methods, particularly in classifying KO cases. This encoding technique provides the model with a more effective representation of the data, enabling better differentiation between the minority KO class and the majority OK class. The improved performance in classifying KO cases suggests that One-Hot/Binary encoding helps mitigate the effects of class imbalance, allowing the model to better manage the challenges posed by the under-represented KO class.

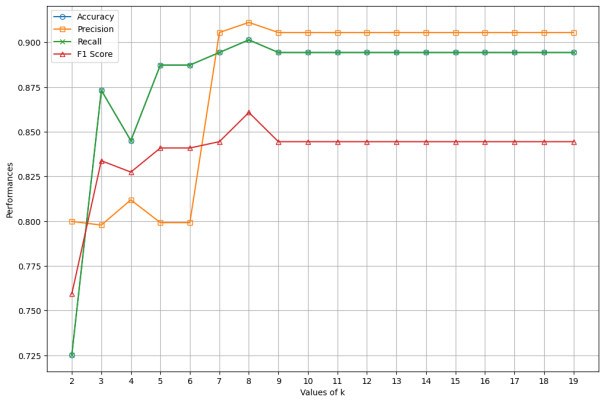

Fig. 4. Performance of kNN with imbalanced FAMD data

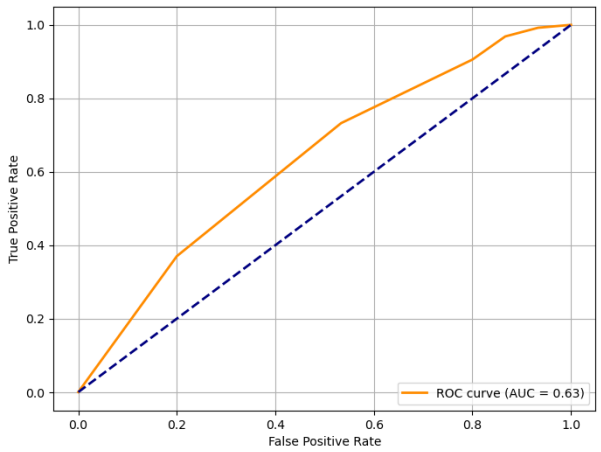

Fig. 5. ROC curve with imbalanced FAMD data ($k = 8$)

In a real industrial application, it is crucial to ensure that the generated synthetic data accurately reflects the distribution of actual data to maintain the model's reliability and effectiveness. To verify this, it is important to analyze the characteristics of both real and synthetic data, ensuring they are comparable and represent the same underlying patterns.

Upon examining Figures 6 and 7, which present the box plots for both real and synthetic data, we observe that they are scaled similarly. This indicates that the synthetic data falls within the same range and distribution as the real data, which contradicts the conclusions of the article [31], which demonstrates that synthetic examples generated by these techniques, such as SMOTE, are often incorrectly attributed to the minority class, thereby misleading classification models. Moreover, the consistency in scale suggests that the synthetic data retains the key features and variability of the real-world data, which is crucial for ensuring accurate and robust model performance in real-world conditions.

To enhance the analysis, we can observe the table IV that shows a comparison between real and synthetic data using several similarity metrics: Wasserstein distance and mean feature variation. Wasserstein distance measures the dissimilarity between two probability distributions in terms of optimal transport, and it is particularly used in generative models, as shown by [32]. Finally, Mean Feature Variance evaluates the dispersion of data for each feature, providing a measure of the stability of synthetic data compared to real data. Overall, the synthetic data are close to the real data, with relatively low Wasserstein distances, indicating good overall correspondence. Lastly, the feature variance remains constant, showing similar dispersion across both datasets. These results indicate that, in general, the synthetic data resemble the real data, but there are notable differences in certain specific aspects, highlighting areas for improvement in the generative model

The fact that the synthetic data remains within the Operational Design Domain (ODD) of the vehicle further assures that the generated data can be safely used for testing and model training, reflecting realistic scenarios the vehicle may encounter. This alignment minimizes the risk of introducing biases or errors that could arise from synthetic data that deviates significantly from the real-world distribution.

V. CONCLUSION

In this study, we conducted a comparative analysis of SMOTE and ROSE oversampling techniques for kNN-based autonomous vehicle behavior modeling. Autonomous vehicles rely on proprietary artificial intelligence systems that are

TABLE III
COMPARISON OF DIFFERENT STRATEGIES BASED ON PERFORMANCE METRICS FOR *FAMD encoding* ($IR = 1.5$)

	Strategy 1	*Strategy 4*
$Accuracy_{OK}$	80%	98%
$Accuracy_{KO}$	20%	13%
AUC	0.53	0.67
Confusion Matrix	Predicted Class / Real Class: KO: 3, 12; OK: 26, 101	Predicted Class / Real Class: KO: 2, 13; OK: 2, 125

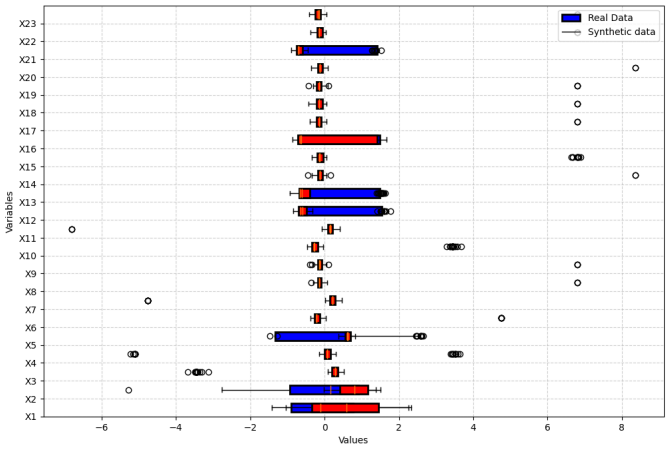

Fig. 6. Box plot on real & synthetic data with OneHot/binary encoding

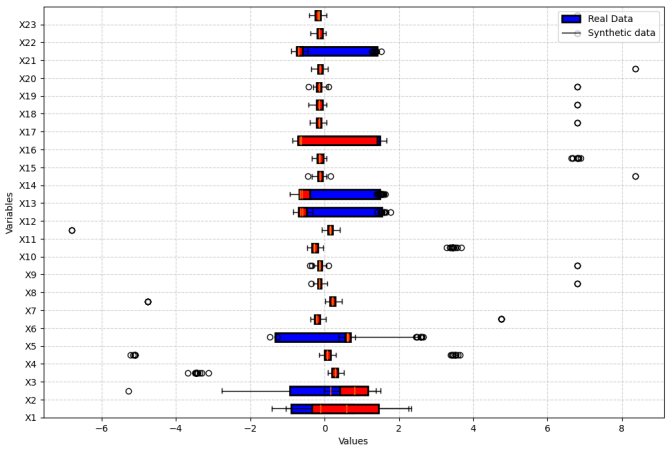

Fig. 7. Box plot on real & synthetic data with FAMD

TABLE IV
COMPARISON BETWEEN REAL AND SYNTHETIC DATA USING WASSERSTEIN DISTANCE AND MEAN FEATURE VARIANCE

Feature	*Wasserstein Distance*	*Mean Feature Variance*
X1	1.13	1.09
X2	0.85	
X3	0.23	
X4	0.39	
X5	0.71	
X6	0.48	
X7	0.49	
X8	0.36	
X9	0.35	
X10	0.23	
X11	0.08	
X12	0.16	
X13	0.39	
X14	0.31	
X15	0.77	
X16	0.45	
X17	0.07	
X18	0.36	
X19	0.36	
X20	0.31	
X21	0.09	
X22	0.08	
X23	0.09	

often inaccessible, necessitating the use of predictive models to ensure operational reliability. These models integrate various data sources and employ advanced machine learning techniques to enhance navigation, object recognition, and obstacle avoidance capabilities, ultimately ensuring safe and efficient operation. Before commercialization in Europe, these vehicles undergo rigorous testing in diverse scenarios, where outcomes are classified as either OK (safe) or KO (unsafe). A significant challenge arises from the inherent imbalance in the dataset, characterized by a predominance of OK cases relative to KO cases. Addressing this imbalance is crucial for developing robust classifiers capable of accurately predicting vehicle behavior.

Our primary objective was to construct a reliable classification model that effectively manages the uneven distribution of data classes while accommodating both qualitative and quantitative variables. Given the limited amount of data resulting from costly and time-consuming tests, we focused on enhancing model performance through resampling techniques. In our previous work, we successfully implemented three data resampling methods—SMOTE, SMOTE-NC, and SMOTE-ENC—to tackle class imbalance. Building on this foundation, we extended our approach by incorporating ROSE (Random OverSampling Examples) as an alternative method for generating synthetic samples. ROSE approximates the underlying probability distribution of the minority class and samples from it, providing a different perspective on data augmentation.

Additionally, we integrated these resampling techniques with the Leave-One-Out Cross-Validation (LOO-CV) method, applying resampling at each iteration to ensure that data rebalancing is tailored to each training set while preserving the integrity of the test data. This approach allows for a more robust evaluation of model performance on unseen examples. Furthermore, we explored various encoding strategies for categorical variables. In addition to conventional One-Hot and binary encoding techniques, we introduced Factor Analysis of Mixed Data (FAMD) as an alternative method for handling mixed-type data. By comparing these encoding strategies, we aimed to identify which method best preserves information and enhances the performance of resampling techniques in the context of class imbalance. Specifically, for the classification of the minority KO class, we observed that the accuracy for KO increased from 6% with imbalanced data to 60% with SMOTE and 53% with ROSE. This emphasizes the importance of building models on balanced data.

Our findings indicate that addressing class imbalance through effective resampling techniques and appropriate encoding strategies significantly improves classification performance in autonomous vehicle behavior modeling. This research contributes to the ongoing efforts to develop reliable predictive models that can adapt to real-world scenarios while ensuring safety in autonomous driving systems.

In conclusion, it is possible to improve various metrics, particularly AUC, which measures the model's ability to distinguish between the two classes (OK/KO), by exploring other classification algorithms such as decision trees and random forests. It would also be interesting to study the impact of kernel choice in the ROSE method on the resampling results.

REFERENCES

[1] C. Ebert and M. Weyrich, "Validation of automated and autonomous vehicles," *ATZelectronics worldwide*, vol. 14, no. 9, pp. 26–31, 2019.

[2] C. Serbouh, I. Ahriz, N. Niang, and A. Piperno, "Predictive modeling of autonomous vehicle behavior with imbalanced & mixed data," in *2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2024, pp. 1–6.

[3] D. J. Hand and V. Vinciotti, "Choosing k for two-class nearest neighbour classifiers with unbalanced classes," *Pattern recognition letters*, vol. 24, no. 9-10, pp. 1555–1562, 2003.

[4] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[5] D. Furundzic, S. Stankovic, S. Jovicic, S. Punisic, and M. Subotic, "Distance based resampling of imbalanced classes: With an application example of speech quality assessment," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 440–461, 2017.

[6] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of artificial intelligence research*, vol. 19, pp. 315–354, 2003.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[9] R. A. Schiavo and D. J. Hand, "Ten more years of error rate research," *International statistical review*, vol. 68, no. 3, pp. 295–310, 2000.

[10] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2012.

[11] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64 606–64 628, 2021.

[12] S. Yin, X. Zhu, and C. Jing, "Fault detection based on a robust one class support vector machine," *Neurocomputing*, vol. 145, pp. 263–268, 2014.

[13] S. Luca, D. A. Clifton, and B. Vanrumste, "One-class classification of point patterns of extremes," *Journal of Machine Learning Research*, vol. 17, no. 191, pp. 1–21, 2016.

[14] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," *Pattern Analysis & Applications*, vol. 6, pp. 245–256, 2003.

[15] W. Wang and D. Sun, "The improved adaboost algorithms for imbalanced data classification," *Information Sciences*, vol. 563, pp. 358–374, 2021.

[16] T. Konno and M. Iwazume, "Cavity filling: Pseudo-feature generation for multi-class imbalanced data problems in deep learning," *arXiv preprint arXiv:1807.06538*, 2018.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[18] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, pp. 92–122, 2014.

[19] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[20] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[21] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[22] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008, pp. 1322–1328.

[23] F. R. Torres, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Smote-d a deterministic version of smote," in *Pattern Recognition: 8th Mexican Conference, MCPR 2016, Guanajuato, Mexico, June 22-25, 2016. Proceedings 8*. Springer, 2016, pp. 177–188.

[24] M. Mukherjee and M. Khushi, "Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features," *Applied System Innovation*, vol. 4, no. 1, p. 18, 2021.

[25] P.-F. Kuo, W.-T. Hsu, D. Lord, and I. G. B. Putra, "Classification of autonomous vehicle crash severity: Solving the problems of imbalanced datasets and small sample size," *Accident Analysis & Prevention*, vol. 205, p. 107666, 2024.

[26] S. Alshathri, A. Sayed, and E. E.-D. Hemdan, "An intelligent attack detection framework for the internet of autonomous vehicles with imbalanced car hacking data," *World Electric Vehicle Journal*, vol. 15, no. 8, p. 356, 2024.

[27] Y. Li, L. Yu, F. Liu, D. Wu, and L. Xing, "Predicting lane-changing risk considering class imbalance problem: a control method for synthetic sample," *Transportation Safety and Environment*, p. tdae027, 2024.

[28] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu, and Y.-C. Tian, "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets," *Information Sciences*, vol. 422, pp. 242–256, 2018.

[29] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evolutionary computation*, vol. 17, no. 3, pp. 275–306, 2009.

[30] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," 2018.

[31] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEE Access*, vol. 10, pp. 47 643–47 660, 2022.

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

Celine Serbouh Touazi received her engineering degree and a master's degree in computer science from the Higher School of Computer Science in Algiers in 2022. Currently a Ph.D. student at CNAM Paris, within the Statistical Methods for Data Mining and Learning (MSMDA) team and the Signal Processing, Electronic Architectures, and Automation (LAETITIA) team, she also works at UTAC in the Expertise (ESA) department, where she applies artificial intelligence methods to analyze and understand the behavior of autonomous vehicles.

Iness Ahriz received the Engineering degree from University of Constantine, Algeria in 2005, a Master degree in Telecommunication from Sorbonne Paris Nord University, in 2007 and the Ph.D. degree from Sorbonne University, France, in 2010. Since 2011, she has been an Associate Professor with the CEDRIC Research Laboratory of Conservatoire National des Arts et Métiers, Paris. Her research interests include indoor localization, machine learning and joint communication and localization strategies.

Ndeye Niang is a professor of statistics, lecturer and researcher in data analysis at the CEDRIC laboratory of the CNAM Paris. Doctor in statistics (PhD thesis on multidimensional methods for statistical process control, University of Paris IX Dauphine), Ndèye Niang is a specialist in Data Analysis, Data Mining and Big Data Analytics. She has worked on the analysis of qualitative variables in data mining, in particular on correspondence analysis and several discrimination methods and on the clustering of variables prior to association rules mining in large databases. Through several master thesis and PhD supervision, she collaborates with many companies and research centres for the application of advanced statistical methods to real life problems in automotive industrie, indoor air quality, customer feedback management, drug side effects among others. She is currently working on the development of unsupervised and supervised methods for high dimensional data particulary for heterogeneous data, missing data and multi-block data. She is author or co-author of several publications.

Alain Piperno is Autonomous Vehicles Testing & Homologation senior expert, in UTAC, France He is Technical leader for innovations (level 3 Euro NCAP rating, new ADAS-AD testing solutions, shuttles & robot-taxi regulation, UTAC challenge, ADAS-AD driving training, AI evaluation, blockchains), R&D projects, RFQ , trainings. He was Research Leader & training doer in 2015-2019 in French research institute VEDECOM & in UTAC CERAM. He worked 25 years in RENAULT: autonomous vehicles safety, R&D, managed some RENAULT end to end innovations: tele-diagnosis, connected car services, stolen vehicle tracking, EV smart charging, smartphone software. Acquired a large automobile experience in RENAULT Engineering (body in white industrialization, design, costs and international management, architecture, electronic) & RENAULT Quality (customer incident, reliability, safety). 1988: graduated of ENSTA PARIS Engineer school and IAE PARIS business administration 1985: graduated of ECOLE POLYTECHNIQUE PARIS Engineer School.