IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 October 2022, accepted 5 November 2022, date of publication 9 November 2022, date of current version 16 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221145

RESEARCH ARTICLE

Efficient Driver Drunk Detection by Sensors: A Manifold Learning-Based Anomaly Detector

ABDELKADER DAIRI[1], FOUZI HARROU[2], (Senior Member, IEEE), AND YING SUN[2]

[1]Computer Science Department, University of Science and Technology of Oran-Mohamed Boudiaf (USTO-MB), Bir El Djir 31000, Algeria
[2]Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

Corresponding author: Fouzi Harrou (fouzi.harrou@kaust.edu.sa)

This work was supported by the King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR), under Award OSR-2019-CRG7-3800.

ABSTRACT This study presents an effective data-driven anomaly detection scheme for drunk driving detection. Specifically, the proposed anomaly detection approach amalgamates the desirable features of the t-distributed stochastic neighbor embedding (t-SNE) as a feature extractor with the Isolation Forest (iF) scheme to detect drivers' drunkenness status. We used the t-SNE model to exploit its capacity in reducing the dimensionality of nonlinear data by preserving the local and global structures of the input data in the feature space to obtain good detection. At the same time, the iF scheme is an effective and unsupervised tree-based approach to achieving good detection of anomalies in multivariate data. This approach only employs normal events data to train the detection model, making them more attractive for detecting drunk drivers in practice. To verify the detection capacity of the proposed t-SNE-iF approach in reliably detecting drivers with excess alcohol, we used publically available data collected using a gas sensor, temperature sensor, and a digital camera. The overall detection system proved a high detection performance with AUC around 95%, demonstrating the proposed approach's robustness and reliability. Furthermore, compared to the Principal Component Analysis (PCA), Incremental PCA (IPCA), Independent component analysis (ICA), Kernel PCA (kPCA), and Multi-dimensional scaling (MDS)-based iForest, EE, and LOF detection schemes, the proposed t-SNE-based iF scheme offers superior detection performance of drunk driver status.

INDEX TERMS Anomaly detection, driver drunk detection, t-distributed stochastic neighbor embedding, isolation forest.

I. INTRODUCTION

The number of traffic accidents keeps increasing and causing more damage to society even with the advanced intelligent transportation systems. As reported by the World Health Organization, since 2016, traffic accidents are becoming among the top 10 causes of death [1]. Moreover, according to the WHO, about 1.3 million deaths each year are due to car crashes [2]. The risk of traffic accidents could be significantly increased when driving under the impact of alcohol and any psychoactive substance or drug. The WHO declared that approximately 40% of road traffic accidents are mainly caused by driving under the influence of alcohol [3], the fifth most common on-the-roads death cause [4]. In addition, driving drinking not only causes road traffic injuries but also causes financial losses of up to 500 million $ per year worldwide [5]. Therefore, accurate detection of drunk drivers is vital to mitigate road traffic accidents.

Automatically and accurately detecting car drivers under excess alcohol is essential for reducing road traffic accidents. Over the last decade, increasing interest in developing advanced technologies for detecting driving drinking. Generally speaking, there are two categories of driver alcohol detection: obtrusive-based and unobtrusive-based detectors [6]. Detecting drunk driving via the obtrusive-based techniques is carried out using physiological state changes of a driver, including blood alcohol concentration (BAC), breath alcohol concentration [7], electroencephalogram (EEG) signals [8],

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

and electrocardiogram (ECG) signals changes [9]. However, acquiring these data types (e.g., EEG signals and heart rate) is not accessible, particularly in driving. In addition, drivers may be troubled because of the surrounding environment with intrusive equipment. On the other hand, the unobtrusive techniques for detecting drunk driving are based on vehicle-based features and driving behavior. Different vehicle-based measures are generally used to detect drunk driving, including vehicle speed, acceleration, steering wheel movements, and lateral position. Other unobtrusive techniques employed image-based features to monitor the driver's face and state [10], [11]. For instance, authors in [12] introduced a breath-based alcohol detection system to control the ignition of the engine alcohol if the driver is drinking. This embedded system can be employed to prevent drunk driving and thus enhance traffic safety by reducing traffic accidents due to drunk driving. For instance, in [13], an Internet of Things (IoT)-based drunk detection strategy is introduced to prevent traffic accidents due to drunken driving. To this end, this IoT system is equipped with a set of sensors, including Heartbeat rate, Facial recognition, and alcohol concentration detection sensor.

Driving with excess alcohol may result in severe traffic accidents and serious injury, even deaths for the drivers and the public. Accurately detecting drunken driving is vital to improving traffic safety and helping avoid traffic incidents. Most of the developed detection approaches for drinking driving detection are generally designed using shallow supervised methods that require labeled data in training [14], [15], [16]. However, getting labeled data is not obvious and time-consuming. Thus, this study aims to design a semi-supervised data-driven detector for driving drinking detection that does not require labeled data. Unlike supervised algorithms, semi-supervised anomaly detection algorithms only employ the data of normal events to train the detection model, making them more attractive for detecting drunk drivers since it is not always easy to get accurately labeled data. Of course, the contributions of this study are summarized as follows.

- This study introduces an innovative approach for driving drinking detection by combining the advantages of the t-distribution stochastic neighbor embedding (t-SNE) model and isolation forest (iF)-based anomaly detection scheme. We used the t-SNE model to exploit its capacity in reducing the dimensionality of nonlinear data by preserving the local and global structures of the input data in the feature space to obtain good detection [17], [18]. Essentially, the original data are projected into the optimal low- dimensional space via the t-SNE, and then the iF detector is applied to the extracted features to realize anomaly detection. The key characteristic of the iF-driven anomaly detection scheme is its capacity to uncover anomalies without considering any distance or density metrics, reducing computational costs [19]. At first, the t-SNE-based iF detector is constructed based on training data (normal driving behaviors) and then used to detect drunk and driving behaviors. We assessed the effectiveness of this approach by using experimental data provided in [15] for alcohol detection in drivers by sensors and computer vision (i.e., physiological, biological, and visual characteristics). Specifically, three sensors are used for driver data acquisition. An MQ-3 gas sensor, which is sensitive to different gases and rapid to integrate into the system, is employed to sense the presence of ethanol. An MLX90621 temperature sensor is used to determine the facial thermal change of the driver. Also, the Raspberry Pi Camera is employed to compute pupil ratio. Of course, the multivariate data contains alcohol concentration and temperature in the car environment, face temperature, and pupil ratio.
- Furthermore, we compared the detection performance of the proposed t-SNE-based iF scheme to that of the Principal Component Analysis (PCA), Incremental PCA (IPCA), Independent component analysis (ICA), Kernel PCA (kPCA), and Multi-dimensional scaling (MDS)-based iForest, EE, and LOF detection schemes. In addition, the comparison has been performed with the standalone anomaly detection methods (i.e., iF, EE, LOF). The considered anomaly detection methods do not require labeling to identify anomalies. Four statistical indices are employed to compare the discrimination accuracy of the considered methods: accuracy, precision, F1-score, and the Area Under the Curve (AUC). Results demonstrated the superior detection performance of drunk driver status using the proposed t-SNE-based iF approach.

The remainder of this paper is organized as follows. Section II highlights literature reviews on the related works. Section III briefly describes the preliminary materials, including the tNSE and the iF anomaly detector. Section IV presents the proposed drunk driving detection approach. In Section V, we present the used data and the obtained results. Finally, we offer conclusions in Section VI.

II. RELATED WORKS

Driving with excess alcohol can result in severe road traffic crashes to drivers and the public. Over the last decade, many researchers and engineers have developed data-driven methods to improve drunk driving detection for intelligent transportation systems [14], [20]. For instance, the authors in [16] introduced an approach for drunk driving detection using support vector machines (SVM) classifier. The SVM is applied to the extracted driving characteristics (i.e., lateral position and steering angle) to decide the state of the driver state (normal or drunk). Driving with excess alcohol could influence the slopes of steering angle and the slopes of vehicle lateral position. This study is conducted using a fixed-base driving simulator. Results showed that the SVM classifier obtained an overall accuracy of 80% in discriminating drunk driving. In [21], principal component analysis (PCA) has been employed for features selection, and SVM is applied to distinguish normal driving from drunk driving. The results showed that the SVM classifier achieved an accuracy of 70%, which still needs more improvement. In [22], Random Forest (RF) is employed to detect drunk driving based on driving behavior data collected from a driving simulator. After selecting the important features using the RF algorithm, SVM, AdaBoost, linear discriminant analysis (LDA), and RF have been applied to detect drunk driving under different road conditions. Results showed that RF and AdaBoost achieved the best classification performance based on seven features. Specifically, the classification accuracy reached by the RF and AdaBoost is slightly greater than 80%; while, the LDA and SVM achieved an accuracy of 75.93% and 74.07%, respectively.

The authors in [23] focused on developing driver behavior states detection strategy to discriminate three driver states: normal, drowsy, and drunk driving using vehicle-based measures. This study is conducted using a simulator, which enables obtaining data difficult to collect under real driving conditions, such as drowsy or drunk driving. Importantly, three models are constructed to discriminate the three behavior states: normal, drowsy, and drunk driving. An experiment with free-road driving is performed to get information about the drowsy and normal state, and another experiment is implemented under road driving to obtain information about drunk driving and normal driving. The data used for the detection is based on acceleration, velocity, yaw rate, and steering. Essentially, the first model aims to separate drowsy behavior from the normal one; the second model is used to discriminate drunk from drowsy states using features from the free-road data, and the last one, constructed using event-road driving data, focus on detecting abnormal events. Of course, each model is used to separate two states. The states identification is treated as a supervised classification using a machine learning model, namely Random Forest. In [24], a two-stage data-driven approach based on Markov models together with Recurrent Neural Networks is presented to detect drunk driving using onboard vehicle sensors. Specifically, several sensory data are collected and processed by Recurrent Neural Networks to predict the longitudinal acceleration in a supervised manner. This approach achieved an overall detection performance of 79%, which makes it very promising to prevent drunk drivers from driving.

Recently in [25], a two-stage deep learning approach is proposed to detect drunk driving using a Convolutional Neural Network (CNN). At first, the simplified VGG (Visual Geometry Group) network, a standard CNN, is applied to estimate the driver's age, and then the simplified Dense-Net for identifying the facial features of drunk driving for alcohol test discrimination. An accuracy of about 86.36% is achieved in the age discrimination step. The overall accuracy of 88.53% is obtained for the drunk driving detection stage. Authors in [26] address the abnormal driving detection using a stacked sparse autoencoders approach (SdsAEs) to model driving behavior features, specifically a softmax layer is considered for a classification task. Results showed the superior performance of the SdsAEs approach in detecting abnormal driving behavior compared to softmax regression, SVM, and a back-propagation neural network. Authors in [15] and [27] proposed a strategy for in-driver drunk status detection based on two inputs, a visual via image processing and sensors data. Specifically, the following input variables are used to classify normal driving from drinking dring status: the facial temperature of the driver, the pupil width, and the concentration of alcohol in the car environment. The problem of drunk detection is addressed via supervised classification techniques combined with a features selection, using machine learning models, such as SVM, k-nearest neighbors (kNN), Decision Tree, and Neural Network. The authors in [28] introduced an approach to identify the driver state by using physiological sensors and a capacitive hand detection sensor. They use cellular neural networks for monitoring the driver's stress level. Results showed promising performance of this approach in recognizing the driver states (i.e., stress or no stress) by providing detection accuracy of 92%.

III. MATERIALS AND METHODS

This section presents the materials needed to design the proposed drunk driving detection approach: the t-SNE and the isolation forest methods.

A. T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

The t-SNE is a nonlinear dimensionality reduction technique originally introduced by van der Matten and Hinton in 2008 to visualize high dimensional data in lower-dimensional space [17]. It is characterized by its capacity to capture much of the local structure in the high-dimensional data while also retaining global structure. More explicitly, if the original data contain numerous clusters, the t-SNE enables revealing the presence of these clusters in the low dimensional space. In recent years, the t-SNE has been widely employed in many research fields for visualizing high dimensional features [29], [30], [31], [32], [33], [34], [35].

Lets denote $\mathcal{D} = d_1, d_2, \ldots, d_l$ a high dimensional datasets, and $\mathcal{S} = s_1, s_2, \ldots, s_l$ the corresponding visual space. At first, the t-SNE calculates the dissimilarity separating the observation in the input space. To this end, the similarity between sample data points d_i and d_j is quantified using the Gaussian distribution in Equation (1), P_{ij}, with σ_i denotes the standard deviation of the Gaussian distribution centered on d_i,

$$P(j|i) = \frac{exp(-\|d_j - d_i\|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\|d_k - d_i\|^2/2\sigma_i^2)}, P(i|i) = 0. \quad (1)$$

It is worth pointing that in t-SNE, we set $P(i|i) = 0$ because only pairwise closenesses within data points are of interest. The joint probabilities of the high-dimensional points, which is a symmetrized version of the conditional similarity because it has the property that $P_{ij} = P_{ji}$ for $\forall i, j$, is expressed as:

$$P_{ij} = \frac{P(i|j) + P(j|i)}{2l}. \quad (2)$$

Using conditional or joint probabilities results in similar results, but optimizing the joint model is less computationally expensive [17].

For the lower space, the student-t probability distribution with one degree of freedom has been employed to compute the similarity between sample data points s_i and s_j, as in Equation (3).

$$\mathrm{q_{ij}} = \frac{(1 + ||\mathrm{s_i} - \mathrm{s_j}||^2)^{-1}}{\sum_{\mathrm{k=1,k\neq 1}}^{\mathrm{N}}(1 + ||\mathrm{s_k} - \mathrm{s_1}||^2)^{-1}} \tag{3}$$

Indeed, the student-t distribution has heavy tails than the Gaussian distribution, making it more suitable for discriminating crowded points in the inputs. Crucially, Student-t distribution is appropriate for representing dissimilar points in the input space by a larger distance in low-dimensional space. Then, the Kullback-Leibler divergence (KL) is applied to quantify the distance between distributions of data in original space and low-dimensional space. The KL distance is minimized to get coordinates of the data points in lower-dimensional space. The objective function $\mathcal{L}$ is defined as follows [36]:

$$\mathcal{L} = \sum_i KL(Pi\|Qi) = \sum_i \sum_j P(j|i)log(\frac{P(j|i)}{Q(j|i)}). \tag{4}$$

$P(j|i)$ represents the similarity between d_i and d_j while $Q(j|i)$ is used for y_i and y_j. Indeed, P (data distribution of the input data of higher dimension) equation (3), while Q (data distribution of the output data of low dimension).

The cost function $\mathcal{L}$ is minimized based on a gradient descent algorithm; the t-SNE stochastic gradient descent is achieved as follows:

$$\frac{\delta \mathcal{L}}{\delta s_i} = 4\sum (P_{ij} - Q_{ij})(s_i - s_j)\big(1 + (\|s_i - s_j\|^2)\big)^{-1} \tag{5}$$

After that, s_i is updated by the following equation:

$$s_i^t = s_i^{t-1} + \eta \frac{\partial \mathcal{L}}{\partial s_i} + \alpha_t (s_i^{t-1} - s_i^{t-2}), \tag{6}$$

where s_i^t represents the solution at iteration t, η denote the learning rate and α refers to momentum at iteration t. The learning rate decides the step size used at each iteration to optimize the objective function $\mathcal{L}$, while a relatively large momentum term could be introduced for accelerating the optimization procedure and avoiding poor local minimums.

Note that in the t-SNE approach, the most important hyper-parameter is the perplexity, which defines the effective number of neighbors. In other words, the t-SNE output generated depends on the select values of its input, especially the *Perplexity* parameter. The value assigned to the *Perplexity* $\mathcal{P}$ is proportional to the σ_i^2, which means a small value will correspond to a small distance between to data points d_i and d_j. The perplexity is expressed as:

$$\mathcal{P}(P_i) = 2^{\mathcal{E}(P_i)}, \tag{7}$$

With $\mathcal{E}(P_i)$ denotes the Shannon's entropy of P_i [17]. There is no automatic way to choose the optimal perplexity value. Larger values of the perplexity leads could eliminate small-scale structures in the manifold; however, smaller perplexity values could falsely generate several sub-manifolds by using a small number of nearest neighbors. The optimal value of the Perplexity can be obtained by minimizing the cost function, $\mathcal{L}$, with respect to the Perplexity. The authors in [17] recommend choosing a perplexity value with the interval of [5, 50].

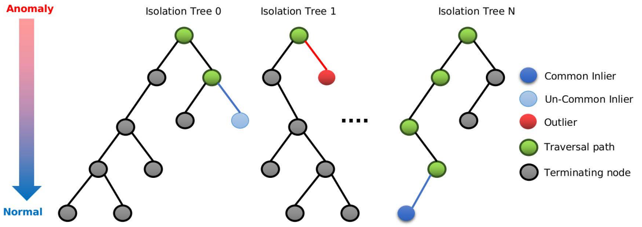


FIGURE 1. Isolation Forest for anomaly detection.

The time complexity of the t-SNE model is $O(N^2)$, where N denotes the number of data points [18]. In 2014, an improved t-SNE version, called Barnes Hut SNE, was developed to enhance time complexity and reduce it to $O(NlogN)$ [18]. More details about the t-SNE could be found [17], [37].

B. ISOLATION FOREST-BASED ANOMALY DETECTION

The Isolation Forest approach was primarily designed by Lui in 2008 [19] and improved later in 2011 [38] to deal with anomaly detection problems where only normal observations are available. Importantly, it is an unsupervised anomaly detection approach since it is designed without the need for labeled data. The essence of the approach is founded on the principle of the Decision Tree algorithm, and it identifies anomalies by isolating outliers from the data [38]. The iF is based on the well-known Random Forest, which consists of a set (ensemble) of decision trees constructed during the training phase [39]. Isolation Forest can be considered an ensemble learning approach to deal with classification and regression problems [40], [41]. For instance, in [40], a similarity-measured isolation forest is considered to detect anomalies in machine monitoring data. In [41], a combined approach using principal component analysis with the iF algorithm is introduced for partial discharge detection. Importantly, PCA is adopted to reduce the feature space to 2-D space, and the iF is applied to discriminate multi-source partial discharge signals.

Figure 1 illustrates the basic structure of the iF algorithm, which consists in building an ensemble of trees for a given data set. Essentially, the iF algorithm recursively splits the data by constructing an ensemble of trees until isolating all samples. Anomalies can be characterized by a short average path length on the trees. In other words, shorter paths are indicators for potential anomalies because a few numbers of anomalies lead to a smaller number of partitions [19].

Implementing the iF-based anomaly detection approach demands only two parameters specified: the number of trees and the size of sub-samples used for the splitting operations to build the forest. In [19], it has been shown that the detection performance of the iF approach can converge fast based on a small number of trees, and it only needs a small sub-sampling size to reach high detection accuracy. In the iF approach, anomalies in a dataset can be detected by analyzing the path lengths for the anomaly data points, with the splitting process being short, which mean that anomalies require few splits in isolation Trees to be isolated [42]. Furthermore, the anomaly score is computed from the mean path length across all the isolation trees in the forest.

In such an anomaly detection framework, anomalies are scored depending on the leaf depth and isolated after a few splits in a tree. Of course, anomalies are identified by fewer splits or shorter path lengths in the tree. A score is measured by assigning a score to detect anomalies using isolation susceptibilities of a given data point. Therefore, high susceptibilities (anomaly score) indicate potential anomalies, while data points with low anomaly scores are considered normal observations or inliers. Note that the iF approach is trained in an unsupervised manner, and it performs better for anomaly detection when the training dataset does not contain anomalies [38].

Lets denote $l(d)$ is the path length of a given data point d, and $\mathcal{D}$ a dataset composed of N data points. The minimum depth of a used decision tree is equals to $log(N)$ while the maximum depth is $N-1$. Essentially, the anomaly score is computed based on the path length of the trees within the forest. The anomaly score, $\mathcal{A}$, can be computed using the following formula [19]:

$$\mathcal{A}(d,N) = 2^{-\frac{E\left[l(d)\right]}{\alpha(N)}}, \tag{8}$$

where $E\left[l(d)\right]$ denotes the the expected path length of a given data point d from a collection of isolation trees, and $\alpha(N)$ is the average path length, expressed as [19]:

$$\alpha(N) \;= 2\lambda(N-1) - \frac{2(N-1)}{N}, \tag{9}$$

where $\lambda(i)$ is the harmonic number, which can be estimated as follows:

$$\lambda(y) \;= ln(y) + \epsilon, \tag{10}$$

With ϵ is the Euler Constant, i.e., $\epsilon = 0.5772156649$.

Overall, the anomaly score of d, $\mathcal{A}(d,N)$, is obtained by iTree from the training data of N samples, and the range of $\mathcal{A}(d,N)$ is within [0, 1]. It is worth pointing out that the anomaly score is oppositely proportional to the path length. The smaller the anomaly score, the higher the depth is, which indicates the higher the probability that the data point belongs to normal points. Finally, the anomaly detection is performed as follows.

$$\begin{cases} \text{an anomaly} & \text{if } \mathcal{A}(d,N) \text{ is close to } 1 \\ \text{Normal instance} & \text{if } \mathcal{A}(d,N) \text{ is close to } 0 \\ \text{Uncertain decision} & \text{if } \mathcal{A}(d,N) \text{ is close to } 0.5 \end{cases} \tag{11}$$

Noteworthy, an anomaly is flagged if $\mathcal{A}(d,N)$, while when $\mathcal{A}(d,N)$ is less than 0.5, then the data point is likely typical. In the final determination of drunk driving, when $\mathcal{A}(d,N)$ is close to 0.5, then a driver is considered under normal status.

The IF is intuitive, not time-consuming, and sensitive to an outlier in data, making it particularly suited for applications where low latency is necessary. The computational cost of IF in training and testing are is $O(t\ell\log\ell)$ and $O(nt\ell\log\ell)$, respectively. Here, ℓ refers to the subsampling size of the dataset [43], n denotes the size of the dataset, and t is the number of trees in the forest. Interestingly, ℓ needs to be small and constant across distinct datasets to reach a more satisfactory detection performance.

IV. THE T-SNE-BASED ISOLATION FOREST APPROACH

This study addresses the problem of drunk driving detection as an anomaly detection problem. Specifically, the goal is to identify the state of the monitored driver (normal or drunk) based on the collected multivariate time series data. A data-driven approach for drunk driving detection is presented by amalgamating the advantages of two unsupervised machine learning algorithms: manifold learning (i.e., t-SNE) and a decision-tree-based ensemble learning technique (i.e., Isolation Forest). The general framework of the proposed t-SNE-based iF detector is schematically illustrated in Figure 2.

At first, after the acquisition of driver data, the t-SNE is applied and projected the normalized data to feature space with a lower dimension than the input space, usually for 2D or 3D for visualization purposes. The input of t-SNE is the normalized dataset $\mathcal{X}$ is transformed in feature space as,

$$\mathcal{T} = tSNE(\mathcal{X}, Components, Perplexity).$$

The t-SNE features, $\mathcal{T}$, are used as input to the Isolation Forest detector to identify if the driver's drunk status. Note that the iF detector is trained based only on t-SNE features without anomaly (i.e., data from a driver under normal status). Then, it is used to decide if the new $\mathcal{T}$ is anomaly-free (no alcohol) or contains anomaly (driver under the impact of alcohol).

As mentioned above, the Isolation forest training is performed based on transformed data without anomaly (no alcohol), and all decision tree's depth is deeper than anomalies with a shorter path length accounting from the tree root. This structure of isolation trees is suitable for detecting alcohol cases (anomaly) from normal cases during the testing phase. The transformed testing data via the t-SNE are passed through the already built iF scheme in the testing stage. Specifically, the path depth is estimated to compute the anomaly score, then compared to a decision threshold for anomaly detection. If the computed anomaly score is greater than 0.5,

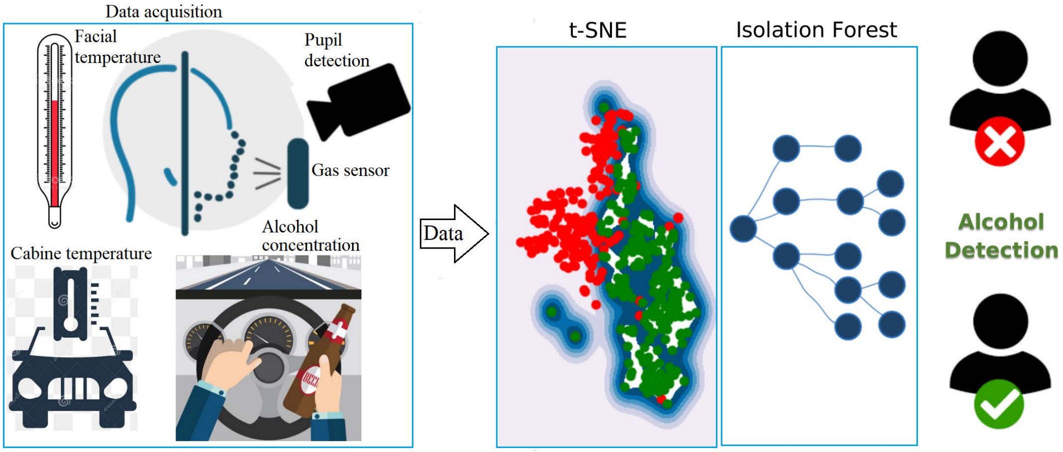


FIGURE 2. The proposed drunk driver detection framework.

an anomaly is declared (i.e., drinking driving); otherwise, the driver is under normal status (no alcohol). The proposed t-SNE-driven iF detection procedure is summarized in Algorithm 1.

```
Algorithm 1: The Proposed Approach Methodology
  Input: : Alcohol Detection Training dataset $\mathcal{X}$
  $\mathcal{X}$ = Normalization($\mathcal{X}$);
  $\mathcal{P} = p_1, p_2, \ldots, p_k$ : Set of Perplexities;
  $C = 2$: Components;
  $N = 150$: Number of Isolation Forest;
  for $Perp$ in $\mathcal{P}$ do
      $\mathcal{T} = tSNE(\mathcal{X}, C, Perp)$;
      $\mathcal{T}_{Anomaly}, \mathcal{T}_{Normal}$ = Split($\mathcal{T}$);
      IsolFor = IsolationForest(N,$\mathcal{T}_{Normal}$);
      prediction = IsolFor.predict($\mathcal{T}_{Anomaly}$);
      AUC = PerformanceEvaluation(prediction);
  end
  Choose the Perplexity that maximize AUC;
  End;
```

In this study, five statistical scores are employed to quantify the performance of the studied methods computed using a 2 × 2 confusion matrix: Accuracy, Precision, Recall, F1-score, and Area under curve (AUC) [44]. For a binary detection problem, the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are used to compute the evaluation metrics.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{14}$$

$$\text{F1} - \text{score} = 2\frac{\text{Precision.Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{15}$$

V. RESULTS AND DISCUSSION

A. DATA DESCRIPTION

This part is devoted to assessing the efficiency of the proposed approach in detecting drunk driving. The experiments are accomplished through actual data from a publicly available database provided in [15]. Three types of sensors are used to collect this data: a sensor of concentration of alcohol in the environment (physiological), a sensor that measure the temperature of the defined points on driver's face (biological) and another one that allows to identify and recognize the thickness of the pupil (visual characteristics). The dataset is relatively small with 390 data points (217 for no alcohol presence 173 for alcohol presence with different concentration). Five variables are collected to decide between drunk and normal driving behaviors: alcohol concentration in the car environment in ml/L, car environment temperature in degrees Celsius, face temperature min in degrees, face temperature max in degrees Celsius, and pupil ratio. Figure 3 illustrates the distribution of the five considered attributes, which indicates that these datasets are non-Gaussian distributed. Those empirical historical data would challenge traditional dimensionality reduction methods, such as PCA and MDS, that typically require linear and Gaussian distributions. Thus, nonlinear techniques designed without restricting the data distribution to be Gaussian, such as tNSE and KPCA, could be promising.

B. EXPERIMENTS AND SETTINGS

Three main experiments are conducted in this study:

1) At first, we evaluate the standalone anomaly detection schemes, including iF, EE, and LOF, in detecting drunk driving.
2) Then, we evaluate the performance of the t-SNE-based iF approach to detect drunk driving.
3) After that, we optimized the performance of the t-SNE-based iF approach detection performance based on different values of the perplexity parameter.
4) Finally, we compared the performance of the proposed approach with five commonly used dimensionality

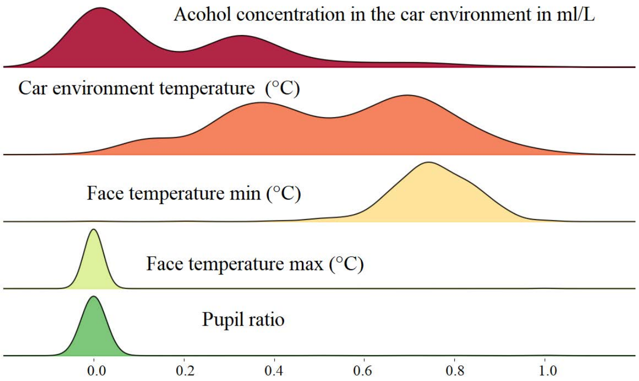


FIGURE 3. Distribution of the considered alcohol attributes.

reduction-based approaches: PCA, ICA, IPCA, KPCA and MDS-based anomaly detection.

In the first experiment, we applied three standalone anomaly detection methods, isolation Forest, Elliptical Envelope (EE) [45], and Local Outlier Factor (LOF) [46]. The parameters setting of these three detectors is listed in Table 1. We used the Grid Search approach to determine the optimal values of hyper-parameters. The three anomaly detectors are applied to the original data with dimensionality reduction. In the LOF detector, an anomaly score is computed for each observation by measuring the local divergence of the density of a given sample compared to its neighbors. In this study, the number of neighbors used in LOF is 20. In the EE detector, which aims to fit an ellipse around the data using a minimum covariance determinant (MCD), the proportion of points to be included in the support of the raw MCD estimate is 0.05.

TABLE 1. Values of hyperparameters of the studied models.

Model	Parameters
t-SNE	Components=2, Perplexity ∈ [5,100]
MDS	Components=2 , eps=1e-3 , dissimilarity='euclidean'
PCA	Components=2
Isol-For	Estimator=150 , contamination=0.05
LOF	neighbors=20, algorithm='kd-tree', contamination=0.05
EE	contamination=0.05

The detection results of the three detectors (i.e., iF, EE and LOF) are listed in Table 2. Results reveal that the iF detector dominated the EE and LOF detectors by obtaining an AUC of 0.9452 and F1-score of 0.9448. It is followed by the EE detector, which showed a satisfactory detection accuracy with an F1-score of 0.9375 and an AUC of 0.9377. The LOF gives the lowest detection performance with an AUC of 0.64.

TABLE 2. Detection results of the three anomaly detectors.

Model	Accuracy	Precision	Recall	F1-score	AUC
iF	0.9410	0.9850	0.9078	0.9448	0.9452
LOF	0.6821	0.6476	0.9401	0.7669	0.6492
EE	0.9333	0.9799	0.8986	0.9375	0.9377

TABLE 3. Alcohol detection results using the t-SNE-based iF scheme under different perplexity values.

perplexity	Accuracy	Precision	Recall	F1-score	AUC
5	0.9049	0.8846	0.9539	0.9180	0.8985
10	0.9280	0.9200	0.9539	0.9367	0.9246
20	0.9486	0.9537	0.9537	0.9537	0.9480
30	0.9537	0.9626	0.9537	0.9581	**0.9537**
40	0.8869	0.8589	0.9539	0.9039	0.8781
50	0.8895	0.8625	0.9539	0.9059	0.8810
60	0.8869	0.8589	0.9539	0.9039	0.8781
70	0.9229	0.9119	0.9539	0.9324	0.9188
80	0.7121	0.6688	0.9537	0.7863	0.6821
90	0.8946	0.8697	0.9539	0.9099	0.8868
100	0.7378	0.6913	0.9537	0.8016	0.7110

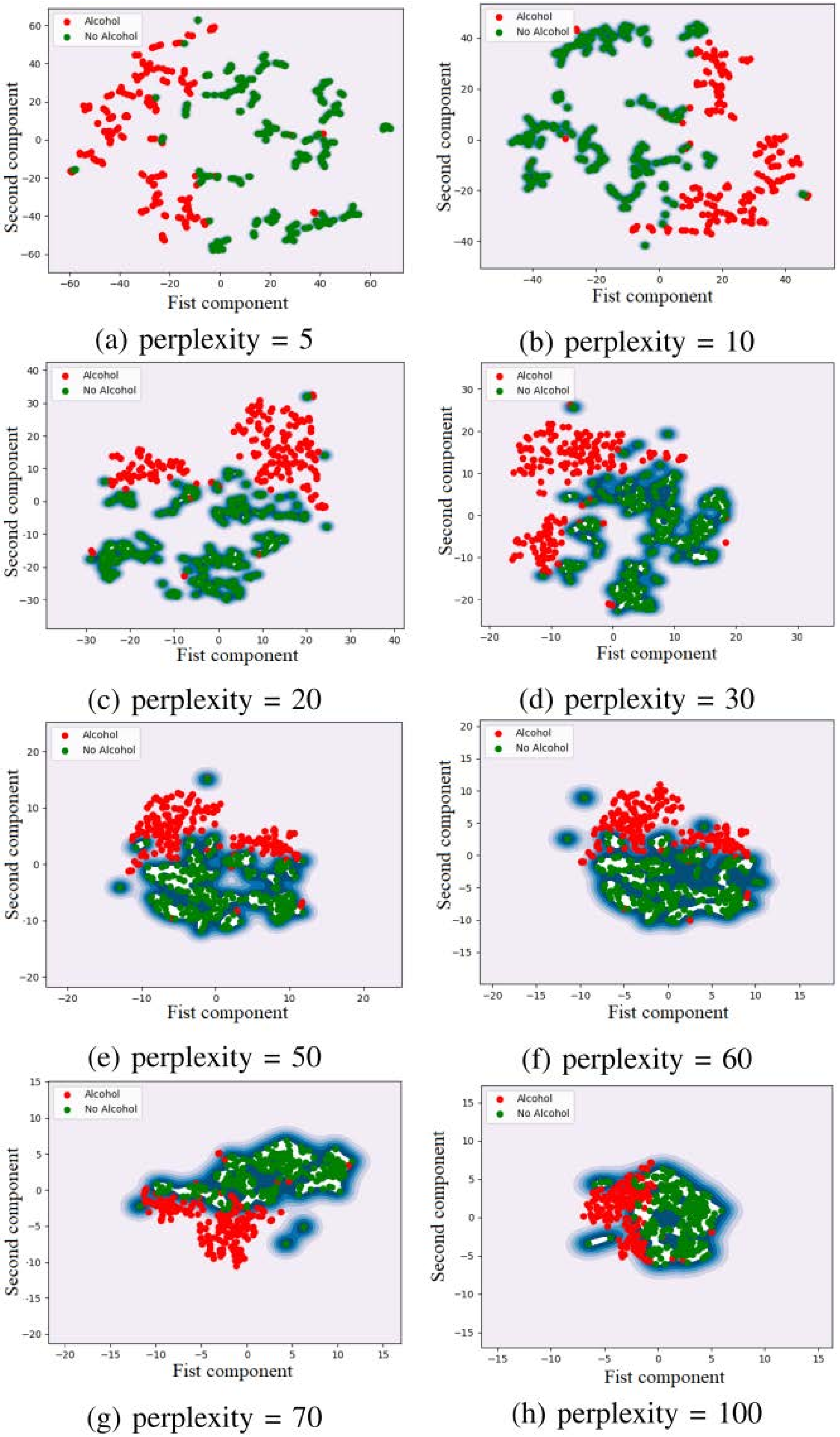
TABLE 4. t-SNE Alcohol detection results using LOF, with different perplexity.

perplexity	Accuracy	Precision	Recall	F1-score	AUC
5	0.8663	0.8287	0.9585	0.8889	0.8543
10	0.9023	0.8745	0.9631	0.9167	0.8944
20	0.9409	0.9327	0.9630	0.9476	**0.9381**
30	0.8920	0.8625	0.9583	0.9079	0.8838
40	0.8252	0.7790	0.9585	0.8595	0.8078
50	0.8946	0.8667	0.9585	0.9103	0.8862
60	0.9075	0.8851	0.9585	0.9204	0.9008
70	0.7301	0.6830	0.9631	0.7992	0.6996
80	0.8072	0.7564	0.9630	0.8473	0.7878
90	0.8509	0.8093	0.9585	0.8776	0.8368
100	0.8303	0.7841	0.9583	0.8625	0.8144

The second experiment is dedicated to verifying the performance of the proposed t-SNE-driven iF anomaly detection approach in detecting drunk driving. Detection results of the t-SNE-driven iF detector, under different perplexity values between 5 and 100, are listed in Table 3. To visually show the impact of the perplexity parameter on the final output of t-SNE, Figure 4 provides visual results of t-SNE applied to the alcohol dataset using different perplexity values. Results in Table 3 indicate that the t-SNE with a perplexity of 30 improves the alcohol detection using the iF detector by achieving a higher F1-score and AUC of 95.81 and 95.37% respectively. It can also be observed that perplexity 10 and 20 recorded AUC > 0.9, which is a good result.

Detection results based on t-SNE-based LOF and EE schemes under different perplexity values are reported in Table 4 and Table 5, respectively. The results show that t-SNE-based LOF and EE schemes with a perplexity of 20 can satisfactorily identify drunk driving from normal driving with an AUC of 93.81% and 93.99%, respectively. These two approaches provide almost comparable detection results.

In the last experiment, as benchmark methods, we assessed the performance of five dimensionality reduction techniques, namely MDS, PCA, ICA, IPCA, and KPCA in detecting drunk driving. These multivariate techniques are widely used in the literature by projecting multivariate data into a low-dimensional space, where most of the variability in data can



(a) perplexity = 5

(b) perplexity = 10

(c) perplexity = 20

(d) perplexity = 30

(e) perplexity = 50

(f) perplexity = 60

(g) perplexity = 70

(h) perplexity = 100

FIGURE 4. Ploting t-SNE with different perplexity values.

TABLE 5. t-SNE Alcohol detection results using Elliptic Envelope, with different perplexity.

perplexity	Accuracy	Precision	Recall	F1-score	AUC
5	0.8817	0.8548	0.9493	0.8996	0.8729
10	0.8509	0.8142	0.9493	0.8766	0.8380
20	0.9409	0.9447	0.9491	0.9469	**0.9399**
30	0.8997	0.8798	0.9491	0.9131	0.8936
40	0.8329	0.7923	0.9493	0.8637	0.8177
50	0.8792	0.8512	0.9493	0.8976	0.8700
60	0.9126	0.8996	0.9493	0.9238	0.9078
70	0.9280	0.9238	0.9493	0.9364	0.9252
80	0.7506	0.7045	0.9491	0.8087	0.7260
90	0.9177	0.9075	0.9493	0.9279	0.9136
100	0.9023	0.8836	0.9491	0.9152	0.8965

be maintained [47]. Generally speaking, linear techniques, including PCA, IPCA, MDS, and ICA, reduce data dimensionality by determining a linear combination of the original variables. They are suitable for handling data that is inherently linear. Nonlinear techniques, such as KPCA, permitted modeling and revealing of nonlinear relationships among multivariate data [47]. Similar to the t-SNE-based approach, we applied the considered linear and nonlinear dimensionality reduction techniques to the multivariate input data for feature extraction and applied the anomaly detection schemes (i.e., iF, EE, and LOF) to the extracted features for anomaly detection. These models are constructed using anomaly-free data and then used for anomaly detection. The values of the parameters of each model are listed in Table 1. Table 6 reports the detection performance achieved by PCA, IPCA, MDS, ICA, KPCA, and t-SNE-based iF, EE, and LOF detection methods when applied to detect drunk driving.

Drunk detection results using MDS, PCA, ICA, IPCA, and t-NSE-based iF, EE, and LOF methods are reported in Table 6. The proposed t-SNE-based iF detector offers superior driver drinking status discrimination performance by achieving an averaged accuracy of 0.9537, F1-Score of 0.9581, and an AUC value of 0.9537. This could be because the t-SNE preserves the local and global structures of the input data in the feature space. In addition, the t-NSE is an efficient nonlinear dimensionality reduction technique embedding multivariate data in a two-dimensional plane. Results in Table 6 indicate that the coupled t-SNE-based iF scheme provides better performance than that of the standalone detector (iF, EE, and LOF) for drunk driver detection. This confirms the benefit of using the t-NSE model in providing more relevant features. We observe that the KPCA-based EE detection scheme achieved the second-best result with an F1-score and AUC of 0.9466 and 0.9493, respectively. Linear dimensionality reduction-based detection schemes (PCA, MDS, ICA, and IPCA) follow it, as shown in Table 6.

Figure 5 displays the barplot of AUC values to visually aid the comparison of achieved results by the considered twenty-one detection schemes. Results show that the t-SNE-based iF detector obtains the most accurate drunk driving detection with an AUC = 95.37%. Overall, the detection accuracy is enhanced when using the t-SNE features compared with the original features. In other words, the t-SNE-based iF scheme outperformed the standalone iF, EE, and LOF anomaly detector in detecting drunk driving. Furthermore, as observed in Figure 5, using a nonlinear dimensionality technique (i.e, the t-SNE) for alcohol detection delivers improved detection performance with AUC = 95.37% compared to the approaches using linear dimensionality reduction techniques for features extraction; i.e., the PCA and KPCA-based EE achieved AUC = 94.93%, and the MDS-EE obtained an of AUC = 93.20%. It could be attributed to the capacity of the t-SNE in capturing nonlinear features in data and the sensitivity of the iF detector in uncovering abnormal observations. In short, the obtained results demonstrate and reveal the promising performance of the combined t-SNE with isolation forest in detecting drunk drivers detection.

TABLE 6. Drunk detection results using the considered schemes.

Model	Accuracy	Precision	Recall	F1-score	AUC
MDS-iF	0.9026	0.9590	0.8618	0.9078	0.9078
MDS-EE	0.9282	0.9701	0.8986	0.9330	0.9320
MDS-LOF	0.9026	0.9838	0.8387	0.9055	0.9107
PCA-iF	0.8821	0.9724	0.8111	0.8844	0.8911
PCA-EE	0.9436	1.0000	0.8986	0.9466	0.9493
PCA-LOF	0.9051	0.9839	0.8433	0.9082	0.9130
KPCA-iF	0.9487	0.9581	0.9493	0.9537	0.9486
KPCA-EE	0.9436	1.0000	0.8986	0.9466	0.9493
KPCA-LOF	0.9103	0.9840	0.8525	0.9136	0.9176
ICA-iF	0.8846	1.0000	0.7926	0.8843	0.8963
ICA-EE	0.9436	1.0000	0.8986	0.9466	0.9493
ICA-LOF	0.9103	0.9892	0.8479	0.9132	0.9182
IPCA-iF	0.8923	0.9834	0.8203	0.8945	0.9015
IPCA-EE	0.9385	0.9898	0.8986	0.9420	0.9435
IPCA-LOF	0.9103	0.9892	0.8479	0.9132	0.9182
tSNE-iF	0.9537	0.9626	0.9537	0.9581	**0.9537**
tSNE-EE	0.9409	0.9447	0.9491	0.9469	0.9399
tSNE-LOF	0.9409	0.9327	0.9630	0.9476	0.9381

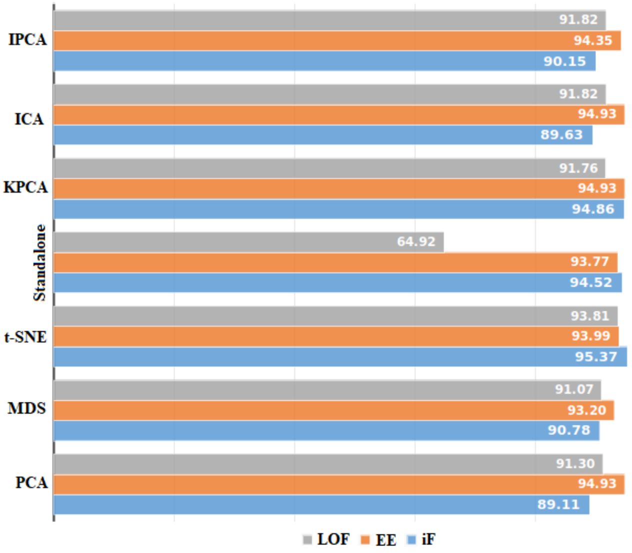


FIGURE 5. AUC values of the twenty-one investigated methods.

TABLE 7. Computation cost.

Model	Encoding time (S)	Detection time (S)
PCA	0.0019	0.1919
ICA	0.016	0.1891
IPCA	0.0385	0.3255
KPCA	0.0471	0.2327
MDS	7.1721	0.1826
t-SNE	1.4763	0.1754

Now, the computation cost time of the investigated methods is examined (Table 7). All experiments have been conducted via a laptop with CPU intel i3 under Ubuntu 20.04.4 LTS with 8GB of RAM (Random access memory) to guarantee a fair comparison. The considered methods have been implemented using Python 3.8 with Keras and Scikit-Learn 0.22, and the time costs for each method are recorded and compared. The time cost of each approach can be evaluated with regard to the encoding part of the used dimensionality reduction model (E.g., PCA, KPCA, and t-SNE) and anomaly detection using the iF scheme. The encoding and detection time of PCA, ICA, IPCA, KPCA, MDS, and t-SNE are (0.0019, 0.1919), (0.0160, 0.1891), (0.0385, 0.3255), (0.0471, 0.2327) (7.1721, 0.1826), and (1.4763, 0.1754), respectively.

We observe that the PCA-based approach requires a lower runtime requirement than the nonlinear dimensionality methods. But, its simple structure cannot capture non-Gaussian and nonlinear features. ICA-based iF scheme follows it, as it is a linear dimensionality reduction method without restricting the data distribution to be Gaussian. Both linear methods (PCA, ICA, and IPCA) achieved lower computational costs than nonlinear methods (KPCA and t-SNE), but they are unsuitable for nonlinear processes. MDS is computationally expensive.

In summary, this study showed that drunk driving detection using the t-SNE-driven iF anomaly detection approach is feasible and effective. It could be attributed to the ability of the t-SNE technique in preserving local geometry and global information of the multivariate data after dimensionality reduction, which is not the case with the linear dimensionality reduction techniques (i.e., PCA, MDS, ICA, IPCA) that may not capture the nonlinear structure in the data. Thus, the detection accuracy of drunk driving using the t-SNE method is better than the PCA, ICA, IPCA, and MDS-based methods. Also, this approach outperformed KPCA-based schemes in detecting drunk drivers. This is because the multivariate data collected to detect drunken driving is non-Gaussian and nonlinear. The t-SNE technique bypasses the data distribution problem by transforming the data distance problem into a probability distribution problem. Moreover, the use of the iF anomaly detector (a sensitive to uncover anomalies in multivariate data) based on the t-SNE features improved the drunk driving detection process. It is found from the results that the perplexity values within [5, 50] could provide good recognition performance, which is in concordance with the literature. The best detection performance is obtained with a perplexity of 30, so there is no need to take a large number of neighbors in the t-NSE. Furthermore, this study revealed the good detection capacity of the t-SNE-based iF approach to deal with a relatively small-sized dataset.

VI. CONCLUSION

Accurately detecting drunk driving is undoubtedly necessary for reducing traffic accidents and improving road safety. In this study, a data-driven methodology to detect drunk drivers is introduced. Importantly, to enhance drunk driving detection, this merges the extended capacity of the t-SNE nonlinear dimensionality reduction as a features extractor and the discrimination ability of the iF in anomaly detection. After normalizing the input data, the t-SNE is employed to extract the characteristics of collected multivariate data. Then, the iF detector is to t-SNE features to detect potential drunk driving. The major advantages of this approach are its assumption-free on data distribution and no need for labeled data in its design to perform anomaly detection. The detection effectiveness is assessed on actual public data collected by sensors and a digital camera. We compared the proposed t-SNE-iF approach with several semi-supervised detection approaches, t-SNE-based EE and LOF schemes, PCA, MDS, ICA, and IPCA-based iF, EE, and LOF methods, and the standalone anomaly detection schemes (i.e., iF, EE, and LOF). Results demonstrated the superior detection performance of drunk driver status based on the proposed approach. Thus, this study revealed the promising performance of the t-SNE-based anomaly detection approach for alcohol detection in drivers.

Despite the improved detection performance greater than 95%, future works will improve its capacity to discriminate drunk from normal driving by associating other sources of input like visual data (facial images) and driver behavior. The t-NSE-based model is relatively computationally demanding, hence parallel computing could provide possible solutions. Notably, more computational resources are needed when a more complex model structure is adopted. A more computationally-efficient t-SNE version, Barnes Hut SNE, has been developed in [18]. Another potential amelioration may rely on applying optimization techniques, such as Bayesian optimization, to select the optimal value of the perplexity during the training stage. Furthermore, another direction of improvement consists of using data augmentation techniques to generate large-sized data, which improves the construction of models and thus enhances the detection process. Also, it will be interesting to investigate the detection capability of this data-driven anomaly detection methodology in engineering applications, such as photovoltaic systems monitoring.

REFERENCES

[1] *The Top Ten Causes of Death*, World Health Org., Geneva, Switzerland, 2018. Accessed: Mar. 1, 2022.

[2] *Road Traffic Injuries*, World Health Org., Geneva, Switzerland, 2021. Accessed: Mar. 1, 2022.

[3] H. Haghpanahan, J. Lewsey, D. F. Mackay, E. McIntosh, J. Pell, A. Jones, N. Fitzgerald, and M. Robinson, "An evaluation of the effects of lowering blood alcohol concentration limits for drivers on the rates of road traffic accidents and alcohol consumption: A natural experiment," *Lancet*, vol. 393, no. 10169, pp. 321–329, Jan. 2019.

[4] *Global Status Report on Road Safety*, World Health Org., Geneva, Switzerland, 2015. Accessed: Mar. 1, 2022.

[5] A. L. Paredes-Doig, M. D. R. Sun-Kou, and G. Comina, "Alcohols detection based on Pd-doped SnO_2 sensors," in *Proc. IEEE 9th IberoAmerican Congr. Sensors*, Oct. 2014, pp. 1–3.

[6] Z. Li, H. Wang, Y. Zhang, and X. Zhao, "Random forest–based feature selection and detection method for drunk driving recognition," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 2, 2020, Art. no. 1550147720905234.

[7] M. Sakairi, "Water-cluster-detecting breath sensor and applications in cars for detecting drunk or drowsy driving," *IEEE Sensors J.*, vol. 12, no. 5, pp. 1078–1083, May 2012.

[8] S. Khardi and M. Vallet, "Drowsiness of the driver: Eeg (electroencephalogram) and vehicle parameters interaction," in *Proc. Int. Tech. Conf. Enhanced Saf. Vehicles*, 1995, pp. 443–461.

[9] C. Wu, K. Tsang, H. Chi, and F. Hung, "A precise drunk driving detection using weighted kernel based on electrocardiogram," *Sensors*, vol. 16, no. 5, p. 659, May 2016.

[10] Y.-S. Chen and C. Chia-Tseng, "Facial image recognition system for a driver of a vehicle," U.S. Patent 8 300 891, Oct. 30, 2012.

[11] J. Ljungblad, B. Hök, A. Allalou, and H. Pettersson, "Passive in-vehicle driver breath alcohol detection using advanced sensor signal acquisition and fusion," *Traffic Injury Prevention*, vol. 18, no. 1, pp. S31–S36, May 2017.

[12] G. Gasparesc, "Driver alcohol detection system based on virtual instrumentation," *IFAC-PapersOnLine*, vol. 51, no. 6, pp. 502–507, 2018.

[13] K. Sandeep, P. Ravikumar, and S. Ranjith, "Novel drunken driving detection and prevention models using Internet of Things," in *Proc. Int. Conf. Recent Trends Electr., Electron. Comput. Technol. (ICRTEECT)*, Jul. 2017, pp. 145–149.

[14] A. Halin, J. G. Verly, and M. V. Droogenbroeck, "Survey and synthesis of state of the art in driver monitoring," *Sensors*, vol. 21, no. 16, p. 5558, 2021.

[15] P. D. Rosero-Montalvo, V. F. López-Batista, and D. H. Peluffo-Ordonez, "Hybrid embedded-systems-based approach to in-driver drunk status detection using image processing and sensor networks," *IEEE Sensors J.*, vol. 21, no. 14, pp. 15729–15740, Jul. 2021.

[16] Z. Li, X. Jin, and X. Zhao, "Drunk driving detection based on classification of multivariate time series," *J. Saf. Res.*, vol. 54, pp. 61–64, Sep. 2015.

[17] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[18] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2015.

[19] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[20] M. Nasr Azadani and A. Boukerche, "Driving behavior analysis guidelines for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6027–6045, Jul. 2022.

[21] H. Chen and L. Chen, "Support vector machine classification of drunk driving behaviour," *Int. J. Environ. Res. Public Health*, vol. 14, no. 1, p. 108, Jan. 2017.

[22] Z. Li, H. Wang, Y. Zhang, and X. Zhao, "Random forest-based feature selection and detection method for drunk driving recognition," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 2, 2020, Art. no. 1550147720905234.

[23] K. H. Lee, K. H. Baek, S. B. Choi, N. T. Jeong, H. U. Moon, E. S. Lee, H. M. Kim, and M. W. Suh, "Development of three driver state detection models from driving information using vehicle simulator; normal, drowsy and drunk driving," *Int. J. Automot. Technol.*, vol. 20, no. 6, pp. 1205–1219, Dec. 2019.

[24] H. Harkous and H. Artail, "A two-stage machine learning method for highly-accurate drunk driving detection," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2019, pp. 1–6.

[25] R. C.-H. Chang, C.-Y. Wang, H.-H. Li, and C.-D. Chiu, "Drunk driving detection using two-stage deep neural network," *IEEE Access*, vol. 9, pp. 116564–116571, 2021.

[26] J. Hu, X. Zhang, and S. Maybank, "Abnormal driving detection with normalized driving behavior data: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 6943–6951, Jul. 2020.

[27] P. D. Rosero-Montalvo, V. F. López-Batista, D. H. Peluffo-Ordóñez, V. C. Erazo-Chamorro, and R. P. Arciniega-Rocha, "Multivariate approach to alcohol detection in drivers by sensors and artificial vision," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.* Cham, Switzerland: Springer, 2019, pp. 234–243.

[28] S. Muhlbacher-Karrer, A. H. Mosa, L.-M. Faller, M. Ali, R. Hamid, H. Zangl, and K. Kyamakya, "A driver state detection system—Combining a capacitive hand detection sensor with physiological sensors," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 4, pp. 624–636, Apr. 2017.

[29] A. Gisbrecht, A. Schulz, and B. Hammer, "Parametric nonlinear dimensionality reduction using kernel t-SNE," *Neurocomputing*, vol. 147, pp. 71–82, Jan. 2015.

[30] C.-Y. Lee and W.-C. Lin, "Induction motor fault classification based on ROC curve and t-SNE," *IEEE Access*, vol. 9, pp. 56330–56343, 2021.

[31] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, "T-distributed stochastic neighbor embedding (t-SNE): A tool for eco-physiological transcriptomic analysis," *Mar. Genomics*, vol. 51, Jun. 2020, Art. no. 100723.

[32] W. Lu and X. Yan, "Variable-weighted FDA combined with t-SNE and multiple extreme learning machines for visual industrial process monitoring," *ISA Trans.*, vol. 122, pp. 163–171, Mar. 2022.

[33] M. Pan, J. Jiang, Q. Kong, J. Shi, Q. Sheng, and T. Zhou, "Radar HRRP target recognition based on t-SNE segmentation and discriminant deep belief network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1609–1613, Sep. 2017.

[34] D. A. Senanayake, W. Wang, S. H. Naik, and S. Halgamuge, "Self-organizing nebulous growths for robust and incremental data visualization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4588–4602, Oct. 2021.

[35] H. Liu, J. Yang, M. Ye, S. C. James, Z. Tang, J. Dong, and T. Xing, "Using t-distributed stochastic neighbor embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data," *J. Hydrol.*, vol. 597, Jun. 2021, Art. no. 126146.

[36] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proc. NIPS*, vol. 15, 2002, pp. 833–840.

[37] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, Oct. 2016. [Online]. Available: https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600

[38] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery from Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[40] C. Li, L. Guo, H. Gao, and Y. Li, "Similarity-measured isolation forest: Anomaly detection method for machine monitoring data," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[41] Y.-B. Wang, D.-G. Chang, S.-R. Qin, Y.-H. Fan, H.-B. Mu, and G.-J. Zhang, "Separating multi-source partial discharge signals using linear prediction analysis and isolation forest algorithm," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 2734–2742, Jun. 2020.

[42] A. Mensi and M. Bicego, "Enhanced anomaly scores for isolation forests," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108115.

[43] Y. Chabchoub, M. U. Togbe, A. Boly, and R. Chiky, "An in-depth study and improvement of isolation forest," *IEEE Access*, vol. 10, pp. 10219–10237, 2022.

[44] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.

[45] P. J. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[46] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

[47] T. Cheng, A. Dairi, F. Harrou, Y. Sun, and T. Leiknes, "Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques," *IEEE Access*, vol. 7, pp. 108827–108837, 2019.

ABDELKADER DAIRI received the Engineering degree in computer science from the University of Oran 1 Ahmed Ben Bella, Algeria, the Magister degree in computer science from the National Polytechnic School of Oran, Algeria, and the Ph.D. degree in computer science from the University of Oran 1 Ahmed Ben Bella, in 2018. From 2007 to 2013, he was a Senior Oracle Database Administrator (DBA) and the Enterprise Resource Planning (ERP) Manager. He is currently an Assistant Professor in computer science at the University of Science and Technology of Oran-Mohamed Boudiaf. He has over 20 years of programming experience in different languages and environments. His research interests include programming languages, artificial intelligence, computer vision, machine learning, and deep learning.

FOUZI HARROU (Senior Member, IEEE) received the M.Sc. degree in telecommunications and networking from the University of Paris VI, France, and the Ph.D. degree in systems optimization and security from the University of Technology of Troyes (UTT), France. He was an Assistant Professor with UTT for one year and with the Institute of Automotive and Transport Engineering, Nevers, France, for one year. He was also a Postdoctoral Research Associate with the Systems Modeling and Dependability Laboratory, UTT, for one year. He was a Research Scientist with the Chemical Engineering Department, Texas A&M University at Qatar, Doha, Qatar, for three years. He is currently a Research Scientist with the Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology. He is the coauthor of two books *Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches: Theory and Practical Applications* (Elsevier, 2020) and *Advanced Road Traffic Modelling and Management Using Statistical and Deep Learning Methods* (Elsevier, 2021). He is the author of more than 180 refereed journals and conference publications and book chapters. His current research interests include statistical decision theory and its applications, fault detection and diagnosis, and deep learning. He received two IEEE ECBIOS 2021 Best Paper Awards. He is featured in Stanford University's list of the world's Top 2% Scientists for the year 2020.

YING SUN received the Ph.D. degree in statistics from Texas A&M, in 2011. She held a two-year postdoctoral research position at the Statistical and Applied Mathematical Sciences Institute and the University of Chicago. She was an Assistant Professor with The Ohio State University for a year before joining KAUST, in 2014. At KAUST, she established and leads the Environmental Statistics Research Group, which works on developing statistical models and methods for complex data to address important environmental problems. She has made original contributions to environmental statistics, in particular in the areas of spatiotemporal statistics, functional data analysis, visualization, computational statistics, with an exceptionally broad array of applications. She received two prestigious awards the Early Investigator Award in Environmental Statistics presented by the American Statistical Association and the Abdel El-Shaarawi Young Research Award from the International Environmetrics Society.