Example-Based Query To Identify Causes of Driving Anomaly with Few Labeled Samples

Yuning Qiu
The University of Texas at Dallas
Richardson, TX 75080, USA
yxq180000@utdallas.edu

Teruhisa Misu
Honda Research Institute, Inc. USA
Mountain View, CA 95134, USA
tmisu@honda-ri.com

Carlos Busso
The University of Texas at Dallas
Richardson, TX 75080, USA
busso@utdallas.edu

***Abstract*—Driving anomaly detection is important for *advanced driver assistance systems* (ADAS) to increase driving safety and avoid traffic accidents. However, driving anomaly detection faces many challenges such as numerous and uncertain abnormal patterns observed on the road, sparsity of real anomaly cases documented with accurate labels, and rigid existing systems that rely on manually set thresholds and rules. Previous studies have proposed unsupervised methods for driving anomaly detection in the driver's behaviors or the road condition by identifying deviations from normal driving conditions. A challenge with unsupervised models is the lack of interpretability, where the cause of the anomaly is not always clear. We address this problem with an example-based query method that combines unsupervised anomaly detection methods with the *multi-label k-nearest neighbors* (ML-KNN) algorithm to interpret the detected driving anomalies by identifying their possible causes (e.g., surrounding objects or driver's errors). Our approach relies on a few manually labeled driving segments that are efficiently used as anchors to retrieve the causes of driving anomalies in a given driving segment. These anchors are projected into the embedding created by unsupervised driving anomaly detection systems. The experimental results show that this method can effectively identify the causes of driving anomalies, even for abnormal driving segments triggered by multiple causes. The evaluation shows the flexibility of our proposed solution, where we successfully implement the ML-KNN approach with three alternative feature representations.**

***Index Terms*—Unsupervised learning, anomaly detection and classification, model interpretability**

I. Introduction

The identification of abnormal driving conditions plays an important role in improving road safety. Various rule-based and pattern-based driving anomaly detection methods have been proposed, such as identifying aggressive or dangerous driving patterns [1]–[10], detecting drunk driving styles [11], quantifying driver distractions [12], [13], monitoring driver's fatigue [14], [15], or determining abnormal road conditions [16]–[18]. These driving anomaly detection methods face several important challenges:

- Abnormal patterns are uncertain. The variability across driving scenarios makes it difficult to maintain consistent criteria for determining abnormal driving events under different driving situations. This issue makes rule-based methods unreliable for some scenarios.
- Lack of real driving anomaly examples with accurate labels. Most driving scenarios do not include anomaly events. Therefore, it is hard to obtain a labeled database with enough abnormal driving events to train supervised models. The number of positive and negative samples is extremely unbalanced, challenging conventional machine learning formulations.
- Manually labeling is very difficult and time-consuming. The data collected in real driving environments usually do not have accurate *normal* or *abnormal* annotations. Creating these labels is very difficult given the potential size of data collected on the roads (e.g., SHRP2 data [19]).

As an appealing alternative formulation, studies have proposed unsupervised methods and contrastive learning approaches to identify driving anomalies [20]–[25]. The unsupervised methods formulate driving anomaly detection as a binary classification task (i.e., abnormal versus normal) by quantifying the deviations from normal patterns (e.g., outlier detection). A drawback of these approaches is that they cannot easily classify the detected abnormal driving events into specific anomaly classes. This drawback limits their interpretability and applicability.

This study explores the use of a few labeled examples to increase the interpretability of unsupervised driving anomaly detection methods. Our approach consists of an example-based query method that identifies possible causes of driving anomalies (e.g., surrounding objects or driver's errors). The formulation combines the embedding created by unsupervised driving anomaly models with the *multi-label k-nearest neighbors* (ML-KNN) algorithm. The assumption in our formulation is that similar anomalies are represented close to each other in the feature embedding of the unsupervised driving anomaly models. Therefore, we can project examples of target anomalies into the embedding, and use these samples as anchors. When a driving anomaly event is detected during the evaluation, we can directly compare its projection into the embeddings with the anchors. While this formulation is general and can be used with different unsupervised algorithms, we demonstrate the potential of this approach with two start-of-the-art multimodal unsupervised driving anomaly detection methods proposed by Qiu et al. [25] and Zhou et al. [26]. These models are trained with physiological data, the *controller area network bus* (CAN-Bus) data, and distances to

This work was supported by Honda Research Institute USA, Inc.

nearby vehicles, pedestrians, and bicycles. Overall, the main contributions of this work can be summarized as:

- We introduce an example-based query method for interpreting causes of driving anomalies that are detected by an unsupervised method that can only identify anomalies from normal events. This example-based query method makes the results of the unsupervised driving anomaly detection model more interpretable.
- We interpret the driving anomalies by retrieving the possible causes (e.g., pedestrians, bicyclists, and vehicles), rather than defining particular abnormal driving styles. This method can respond to the anomalies that are not seen by the system during training using a few labeled samples as anchors.
- We present experimental results to show that the proposed example-based query model manages to identify the causes of driving anomalies with only a few labeled samples, which saves significant labor efforts.

II. Related Work

A. Driving Anomaly Detection and Classification

Identifying driving anomalies is important for traffic safety. Studies have proposed driving anomaly detection methods based on either predefined rules or patterns. Zhao et al. [27] detected aggressive driving events by setting thresholds on the vehicle's acceleration signal under different steering wheel angles (e.g., tuning lower thresholds for high steering wheel angles). While the model based on predefined rules can work well for cases like fast U-turn and swerving, the set of anomaly driving scenarios captured by the system is very limited. The pattern-based approaches detect anomalies with specific features or patterns, utilizing machine learning algorithms. Chen et al. [28] extracted statistical features of the acceleration and orientation of the vehicle from the *inertial measurement unit* (IMU) readings of a smartphone to train a *support vector machine* (SVM) classifier that identifies six types of abnormal behaviors: weaving, swerving, sideslipping, fast U-turn, turning with a wide radius and sudden brakes. However, due to the complexity and diversity of driving scenarios, it is quite difficult and time-consuming to exhaustively define all kinds of driving anomalies and design a pervasive and effective classification-based solution.

Studies have used contrastive learning and unsupervised learning algorithms to discriminate abnormal driving events, building more general driving anomaly detection methods [20]–[25]. Köpüklü et al. [24] proposed a contrastive learning-based neural network approach to differentiate between anomalous and normal driver behaviors. They collected a dataset using a driving simulator environment with normal and anomalous driving behaviors. They trained their model by drawing close pairs of normal driving video clips in an encoding embedding space, while pushing the anomalous clips away from the normal pairs. After training, they averaged the embedding of all the normal clips as a template. During inference, they computed the cosine similarity between the encoded embedding of the video and the normal template embedding, creating an anomaly score for that recording. Su et al. [29] proposed a *convolutional neural network* (CNN) and *bidirectional long short-term memory* (BLSTM) based model, utilizing infrared and depth videos facing the driver and the steering wheel to detect the driver's distraction as a binary classification task. They also conducted a multi-class action recognition task by classifying the detected driver's distraction into 16 classes, including writing messages using the right/left hand, talking to a passenger, and drinking using the right/left hand. Qiu et al. [21] proposed an unsupervised approach for driving anomaly detection based on conditional *generative adversarial networks* (GANs). They defined driving anomalies as driving events that deviate from expected driver behaviors. The generator of the GAN model is trained to generate a prediction for the upcoming data conditioned on the previously observed signals. An anomaly event is defined when the generator incorrectly predicts the event based on previous recordings. These approaches can detect anomalies even when the anomalies are unknown or unseen as long as they deviate from normal driving patterns. This approach is appealing given that anomaly driving detection is a long-tailed problem with infrequent and unexpected events that are not guaranteed to be represented in the data. We propose an example-based query system using the *multi-label k-nearest neighbors* (ML-KNN) algorithm to further broaden the interpretability of unsupervised driving anomaly detection models.

B. K-Nearest Neighbor Classifier

One of the components of our proposed approach is the *k-nearest neighbor* (KNN) algorithm. KNN is a supervised machine learning algorithm used mostly for classification tasks. It creates a feature representation where the training samples are projected. A sample in the test set is then projected into this feature representation. Then, its label is determined by the classes of the kth closest training samples.

Zhang et al. [30] proposed the ML-KNN algorithm, extending the KNN algorithm for multi-label classification tasks. For each test instance, they selected the k nearest neighbor samples. They used a Bayesian rule to calculate the probability of each class. The final label for the test sample is determined with the *maximum a posteriori* (MAP) principle. There may be more than one potential cause of driving anomaly due to the variety and complexity of real-life driving scenarios (e.g., a pedestrian and a bicycle crossing a street). Therefore, the ML-KNN formulation is suitable for identifying causes of driving anomalies that can be triggered by more than one factor.

III. Proposed Example-Based Approach

We formulate the driving anomaly interpretation problem as a multi-label classification task. We build an example-based query method for identifying causes of driving anomalies using the ML-KNN algorithm. The proposed approach is flexible and can be implemented with any deep learning unsupervised driving anomaly detection approach. We present alternative approaches in Section III-A. Then, Section III-B presents our proposed example-based approach to increase the interpretability of the unsupervised system.

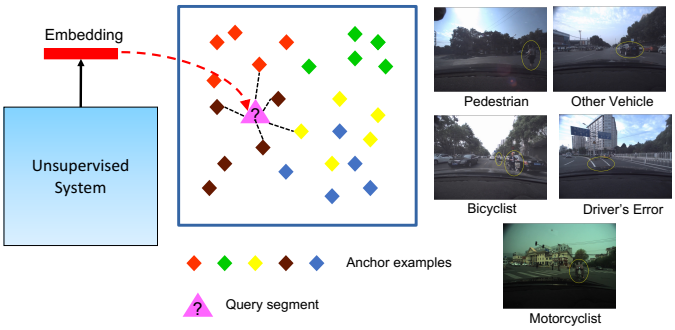
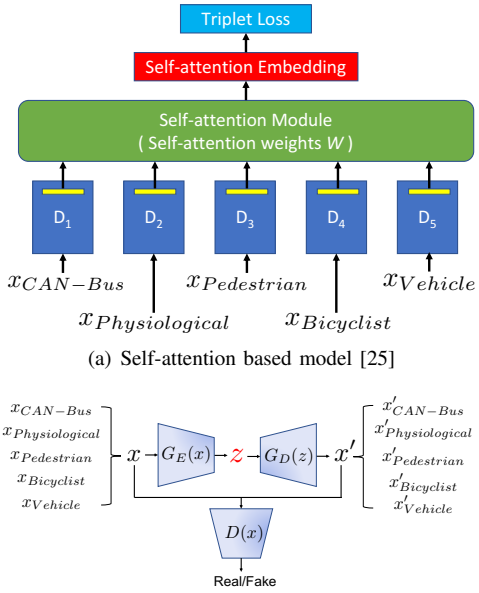


Fig. 1: Proposed formulation to increase the interpretability of the detected anomalous segments by an unsupervised multimodal model. The left side of the figure shows the unsupervised driving anomaly detection approach proposed by Qiu et al. [25]. The right side of the figure shows the implementation of the ML-KNN model that efficiently uses a few labeled samples as anchors, which are projected into the feature embedding provided by the unsupervised approach. The selected anchors are annotated with a single label.



(a) Self-attention based model [25]

(b) BeatGAN model [26]

Fig. 2: The unsupervised model considered in this work. Alternative feature representations are obtained from the self-attention embedding and self-attention weights of the model proposed by Qiu et al. [25], and the bottleneck representation z of BeatGAN [26].

A. *Unsupervised Driving Anomaly Detection System*

We present three alternative approaches to extract feature representation from unsupervised driving anomaly detection systems. The first two approaches use the scalable, multimodal framework proposed by Qiu et al. [25]. Figure 2(a) shows a diagram of the approach. The model consists of conditional *generative adversarial networks* (GANs) and attention mechanism, trained with the triplet loss function. The conditional GAN acts as a feature extractor and follows the principle presented in Qiu et al. [21]. The approach implements five different conditional GAN models using different modalities: CAN-Bus signals, physiological signals, distance to nearby pedestrians, distance to nearby vehicles, and distance to nearby bicycles (see description of the database in Sec. IV-A). By considering these five modalities, the unsupervised approach identifies anomalies not only from the maneuvers and reactions of the driver, but also from anomalies associated with the driving environment (e.g., a pedestrian crossing the street). The conditional GAN models are trained to generate upcoming data conditioned on previously observed signals. The discriminator determines if the signal is real or created by the generator. The generator and discriminator are implemented with CNNs and *long short-term memory* (LSTM) to extract discriminative feature representations that consider temporal relationships. The model extracts the intermediate layer of the discriminators as the embedding for each modality (highlighted in yellow in Fig. 2(a)). The attention mechanism [31] fuses the five modalities into a joint embedding, creating weights depending on their correlations. The embedding of each modality is first projected to the query and key vectors. The attention weights are obtained by calculating the dot product of the query vector of one modality and the key vectors of the other modalities. The attention weights are normalized using the Softmax function. Finally, the triplet loss function draws the joint embedding of the generated data close to the joint embedding of the corresponding real data, while pushing away the joint embedding of the generated data from a randomly selected driving segment. The difference between the joint embedding of real data and generated data is computed as the final anomaly score. We use two feature representations from this model. The first representation is the output embedding of the attention module, highlighted in a red box in Figure 2(a). This embedding produces a 64-dimensional vector. We refer to this method as the *embedding of the attention model*. The second representation is the self-attention weights. The attention module assigns the attention weights by calculating the dot product of the query and key vectors following the self-attention mechanism [26]. The attention weights indicate the correlation among different modalities. The dimension of the matrix with attention weights is 5×5, since we have five modalities. We are using five MHA, resulting in an embedding with a dimension of 125 (i.e., $25 \times 5 = 125$). We refer to this method as *weights of the attention model*.

The third approach uses BeatGAN, which is an unsupervised anomaly detection method proposed by Zhou et al. [26]. Figure 2(b) shows the model, which has a generator and a discriminator. The generator of BeatGAN consists of an encoder ($G_E(\cdot)$) and a decoder ($G_D(\cdot)$) that reconstructs the input. The discriminator is trained to decide whether the reconstructed input is real or fake. We implement the BeatGAN model using fully-connected layers, as described in Zhou et al. [26]. The encoder of the generator ($G_E(\cdot)$) is implemented with five layers with 1024, 512, 256, 128, and 64 nodes, respectively. The decoder of the generator ($G_D(\cdot)$)

is implemented mirroring the structure of the encoder. The discriminator ($D(\cdot)$) is implemented with six layers with 1024, 512, 256, 128, 64, and 1 node, respectively. We concatenate the data from the five modalities (i.e., CAN-Bus signals, physiological signals, distance to nearby pedestrians, vehicles, and bicycles) as the input, and we extract the bottleneck embedding of the generator, z, as the feature embedding for the ML-KNN model. We refer to this representation as the *bottleneck of BeatGAN*.

B. Interpreting Causes of the Detected Anomalies

This section describes our approach to interpreting the cause of the detected anomalies with limited supervision. The approach leverages the embedding of the unsupervised driving anomaly detection model to estimate the distance between the projection of the input video into this space and the projection of anchors representing alternative types of anomalies. The approach relies on the assumption that similar driving anomalies are clustered together in the space provided by the embedding. We use a limited number of labeled samples as anchors (e.g., eight samples per concept), which provide improved interpretability of the detected driving anomalies. A strength of our approach is that it does not compromise the accuracy of the driving anomaly detection system to improve its interpretability, as our method is implemented after the unsupervised model is trained.

We implement our approach using the ML-KNN algorithm. Figure 1 illustrates the proposed approach. First, we select as anchors abnormal driving segments from a set of samples with the highest anomaly scores provided by the unsupervised driving anomaly detection system (Sec. III-A). The selected driving segments are labeled by human raters with the corresponding cause of the anomalies. As shown in Figure 1, for a given query driving segment, we find the *k* nearest anomaly anchors by calculating the Euclidean distance in the feature embedding.

We define the encoded embedding space of the driving segments as $\mathcal{X}$, and the label space of potential driving anomalies as $\mathcal{Y} = \{1, 2, 3, ..., Q\}$, where Q is the number of classes of driving anomalies. With these definitions, we can represent the segments as $S = \{(x_1, Y_1), (x_2, Y_2), \ldots, (x_m, Y_m)\}$, where $x_i \in \mathcal{X}$, $Y_i \in \mathbb{N}^Q$, and m is the number of driving segments. The element corresponding to class l, $Y_i(l)$ with $l \in \mathcal{Y}$, takes the value of 1 if $l \in Y_i$, and 0 otherwise. The system identifies the *k* nearest anchors $A(x_i)$ to x_i and counts the number of neighbors of x_i labeled with class l. We define the membership counting vector as:

$$C_{x_i}(l) = \sum_{i \in A(x_i)} Y_i(l), \quad l \in \mathcal{Y}. \tag{1}$$

For each test instance x_t, the system retrieves the *k* nearest anchors $A(x_t)$, and estimate $C_{x_t}(l)$. The label vector $\widehat{Y}_t$ is determined using the following MAP principle:

$$\widehat{Y}_t(l) = \underset{b \in \{0,1\}}{\operatorname{argmax}} P(H_b^l | E^l_{C_t(l)}), \quad l \in \mathcal{Y} \tag{2}$$

where H_b^l represents the event that x_t is related ($b = 1$) or unrelated ($b = 0$) to class l, and $E^l_{C_{x(l)}}$ denotes the event that there are exactly $C_x(l)$ anchors labeled with class l among the samples in $A(x_t)$. Using the Bayesian rule, Equation 2 can be rewritten as

$$\begin{aligned} \widehat{Y}_t(l) &= \underset{b \in \{0,1\}}{\operatorname{argmax}} \frac{P(H_b^l) P(E^l_{C_t(l)} | H_b^l)}{P\left(E^l_{C_t(l)}\right)} \\ &= \underset{b \in \{0,1\}}{\operatorname{argmax}} P(H_b^l) P(E^l_{C_t(l)} | H_b^l) \end{aligned} \tag{3}$$

Equation 3 is the final classification of the ML-KNN algorithm. The prior $P(H_b^l)$ and posterior $P(E_{C_t(l)}|H_b^l)$ probabilities can be obtained directly from the anchor set by counting the label frequency. Zhou et al. [30] introduced more details about the calculation steps.

IV. Experimental Settings

A. Driving Anomaly Dataset

This work relies on the *driving anomaly dataset* (DAD) [21], which consists of naturalistic driving recordings. The modalities collected in this database include the vehicle's CAN-Bus signals, the driver's physiological signals, the videos of the surrounding driving environment, and the distance to the nearby objects estimated with Mobileye technology (i.e., pedestrians, bicyclists, and vehicles). Our unsupervised approach uses CAN-Bus signals (speed, steering speed, steering angle, throttle angle, brake pressure, and yaw), physiological signals (electrocardiography, breath rate, and electrodermal activity), and distance to nearby pedestrians, bicycles, and vehicles. We use approximately 84 hours of urban driving recordings from 89 sessions. The data is split into the train (72 sessions, $\sim$70 hours), development (3 sessions, $\sim$4 hours), and test (14 sessions, $\sim$10 hours) sets. Qiu et al. [21] provides more details on the data collection.

B. Selection of Limited Labeled Samples

For our evaluation, we need to define (1) prototypical distractions to be used in our formulation, and (2) labeled examples for these anomalies to be used as anchors. Since the driving anomaly detection system considers the driver's physiological reaction and distance to the nearby objects, we set five possible causes of anomalies: pedestrians, bicyclists, motorcyclists, vehicles, and errors made by our driver. Notice that we could define other classes of anomalies, even if they are not observed in the training set. Our proposed approach should be able to identify similar cases during inferences as long as we can identify examples to be used as anchors.

While the DAD corpus has several annotations (e.g., driving maneuvers), it does not have labels for driving anomalies. From the test set, we selected 400 six-second segments, where 200 videos were from recordings without any annotation (e.g., normal driving conditions). The other 200 videos correspond to segments with the highest driving anomaly scores provided by the unsupervised approach proposed by Qiu et al. [25]. We asked three participants to watch the videos, answering two questions. First, we asked if they can see any driving anomaly

(a) Bicyclist, Pedestrian

(b) Motorcyclist, Other Vehicle

(c) Pedestrian, Vehicle

(d) Bicyclist, Other Vehicle

Fig. 3: Examples of abnormal driving events that are labeled as caused by multiple causes.

in the video (i.e., normal/abnormal). Second, we asked them to mark all the causes of anomalies that they can recognize in the video. The options were pedestrian, bicyclist, motorcyclist, other cars, and bad maneuvers by our driver, selecting all the options that apply. They were also able to indicate that there was no anomaly shown in the video.

For the estimation of the consensus labels from the evaluations, we assign a class and the possible causes of anomalies to a video if at least two evaluators agree on a given class. In total, we have 175 segments labeled as *abnormal*, and 225 segments labeled as *normal*. For the cause of anomalies, we have 59 segments for *pedestrian*, 49 segments for *bicyclist*, 38 segments for *motorcyclist*, 94 segments for *other vehicle*, 25 segments for *bad maneuvers by our driver* and 225 segments *without any anomaly*. From the 175 videos annotated as *abnormal*, 103 segments are evaluated to be exclusively triggered by one cause (i.e., 27 by *pedestrian*, 14 by *bicyclist*, 8 by *motorcyclist*, 46 by *other vehicle*, 8 by *bad maneuvers by our driver*), while 72 segments are labeled with multiple causes. Figure 3 shows some examples. We use this set of 175 videos for our evaluation.

C. Selection of Anchors

From the 175 abnormal segments, we manually select for each class eight typical abnormal driving segments with the highest anomaly scores as the anchor examples of driving anomalies (5 classes $\times$ 8 anchors = 40). The selected anchors are all annotated with a single label. These 40 labeled segments are the only supervision provided to train the ML-KNN model. This setting is ideal for the query-by-example formulation proposed in this study, where we only have a few labeled samples for an anomaly class or concept.

The other 135 abnormal segments are used as a test set to evaluate the model. These 135 segments were identified as anomalous by our unsupervised driving anomaly detection system. The model did not provide further information about what was anomalous in the videos. Our approach based on the ML-KNN algorithm uses 40 anchors to interpret the cause of the detected anomalies in these 135 videos.

TABLE I: Performance of the ML-KNN algorithm as a function of k using a single-label formulation.

Value of k	11	12	13	14	15	16	17
Attention Model (Embedding)	31.9%	31.9%	43.0%	**54.8%**	46.7%	41.5%	35.6%
Attention Model (Attention Weights)	53.3%	56.3%	59.3%	37.8%	64.4%	61.5%	**71.1%**
BeatGAN (z Embedding)	35.5%	43.0%	54.1%	**62.2%**	45.2%	36.3%	51.1%

V. Experiments and Results

The experimental evaluation consists of two part. First, we assess the performance of the proposed model to interpret the cause of the detected anomalies with limited supervision, contrasting it with alternative supervised approaches using a single class (Sec. V-A). Second, we evaluate the performance of the proposed system by considering this task as a multi-label problem (Sec. V-B).

A. Single-Label Evaluation

We formulate the single-label evaluation task as a classification problem to evaluate our proposed approach with standard supervised baselines. Given an unknown driving segment x_i, our model produces a multi-label output ($\widehat{Y}_i$), with the probability for each class to be relevant. We transform this formulation into a single-class problem by selecting the class with the highest probability of being relevant. We select as the evaluation metric the proportion of correct predictions referred to as *retrieval accuracy*. Table I reports the retrieval accuracy when the value of k varies from eleven to seventeen. We avoid considering higher values for k, since we only have 40 anchors for training (eight anchors per class).

Table I shows the results when our proposed approach is implemented with three feature representations: the embeddings of the attention model, the weights of the attention model, and the bottleneck of BeatGAN (Sec. III-A). For the embeddings of the attention model, we achieve the highest retrieval accuracy equal to 54.8% when the model considers the fourteen nearest anchors. Overall, the best performance is obtained using the weights of the attention model with a retrieval accuracy of 71.1% when considering the seventeen nearest anchors. This result demonstrates that the attention weights carry representative information among different driving segments. The proposed approach implemented with the bottleneck of the BeatGAN model achieves a retrieval accuracy of 62.2% when considering the nearest fourteen anchors.

We compare the proposed model with alternative supervised approaches. Since we also rely on 40 training samples, it is difficult and unreliable to build a classifier from scratch, without relying on the (unlabeled) data. A straightforward solution is to use features of the three discriminative representations considered in this study from the unsupervised driving anomaly detection systems. We implement three supervised systems, each of them implemented with features either from

TABLE II: Comparison of the retrieval accuracy of the proposed approach and alternative supervised methods.

	SVM	LR	RF	ML-KNN
Attention Model (Embedding)	33.3%	29.6%	26.9% (0.032)	**54.8%**
Attention Model (Attention Weights)	23.0%	27.4%	27.9% (0.029)	**71.1%**
BeatGAN (z Embedding)	31.1%	31.61%	30.71% (0.031)	**62.2%**

the embeddings of the attention model, the weights of the attention model, or the bottleneck of the BeatGAN model. The three supervised models are SVM with Radial Basis Function (RBF) kernel, *logistic regression* (LR), and *random forest* (RF). These classifiers are designed for a single-label task, but several examples have more than one label in our task. Therefore, we consider a success when the predicted label $\overline{y}_t$ for a driving segment $\overline{x}_t$ is in the true label set Y_t. We keep the same setting, training the classifiers with the selected 40 anchors and using the other 135 abnormal segments for inference. For the *random forest* model, we run the classification experiment 10 times and report the average and standard deviation value of the model performance. Table II shows the results, which indicate that the retrieval accuracies of the baseline methods are lower than 35%. The limited training set clearly affects these supervised models. In contrast, the proposed ML-KNN solution leads to clear improvements with accuracies as high as 71.1% when using the weights of the attention models.

B. Multi-Label Evaluations

We formulate the interpretation of driving anomaly detection as a multi-label problem, where several causes can be relevant. This section evaluates our approach using the ML-KNN framework with metrics that are usually used in multi-label classification problems: Hamming loss (↓), ranking loss (↓), and average precision (↑) [32]. We indicate with the symbol ↑ when a larger value leads to better performance and the symbol ↓ when a smaller value leads to better performance. The *Hamming loss* shows the mismatch between the relevant classes predicted by the model and the ground truth labels. The *ranking loss* indicates the average proportion of unrelated causes that are predicted to have higher probabilities than relevant causes. The *average precision* represents the average fraction of relevant causes ranked higher than a particular label $y \in Y_i$. Schapire et al. [32] provides the definitions of these metrics, which are more appropriate for this task than metrics used for single-label problems such as accuracy, precision, recall, and F-score since multiple classes can be relevant.

Table III shows the model performance when we vary from eleven to seventeen the number of k nearest neighbors considered by the ML-KNN algorithm. Overall, we obtain the best results considering the four multi-label metrics when k is either fourteen or fifteen using the three different feature representations. When using the weights of the attention model, we obtain an average precision equal to 0.648, indicating that the predicted top classes are often included in the ground truth. The best hamming loss is 0.283, suggesting that the class predicted by the ML-KNN algorithm is most of the time consistent with the relevant class in the ground truth labels. When using the bottleneck of the BeatGAN mode, our model achieves a retrieval accuracy of 0.622, and an average precision of 0.506. The performance of our proposed approach based on the ML-KNN algorithm demonstrates that the driving anomalies triggered by similar causes are located closer to each other in the feature embedding space generated by the unsupervised driving anomaly detection approaches. Even though the unsupervised approach was trained without any labeled sample, the models learn representative features that are characteristic of the types of driving anomalies considered in this study. Our proposed ML-KNN approach efficiently uses a few anchors per target class to increase the interpretability of the detected driving anomaly.

TABLE III: ML-KNN performance as a function of k under a multi-label formulation. (HL: Hamming loss; RL: Ranking loss; AP: Average precision; RA: Retrieval accuracy)

Feature	Metric	Value of k						
		11	12	13	14	15	16	17
Attention Model Embedding	HL (↓)	0.439	0.413	0.384	0.412	0.416	0.388	0.415
	RL (↓)	0.501	0.462	0.448	0.443	0.485	0.480	0.492
	AP (↑)	0.502	0.509	0.543	0.538	0.517	0.503	0.5
	RA (↑)	0.319	0.319	0.430	0.548	0.467	0.415	0.356
Attention Model Attention Weights	HL (↓)	0.350	0.339	0.302	0.382	0.283	0.287	0.370
	RL (↓)	0.339	0.313	0.299	0.362	0.330	0.314	0.329
	AP (↑)	0.582	0.608	0.683	0.604	0.684	0.690	0.542
	RA (↑)	0.533	0.563	0.593	0.378	0.644	0.615	0.711
BeatGAN z Embedding	HL (↓)	0.425	0.390	0.341	0.397	0.450	0.542	0.391
	RL (↓)	0.403	0.394	0.362	0.347	0.442	0.477	0.438
	AP (↑)	0.450	0.498	0.564	0.506	0.450	0.406	0.522
	RA (↑)	0.356	0.429	0.540	0.622	0.452	0.363	0.511

VI. Conclusions

The work proposed an example-based query system for retrieving possible causes of driving anomalies detected by unsupervised driving anomaly approaches, increasing the interpretability of their results. The approach uses an embedding from an unsupervised anomaly detection approach, which is combined with the ML-KNN algorithm to identify the potential cause of driving anomalies under the assumption that similar anomalies cluster close to each other in this embedding space. We show that the proposed approach is flexible, implementing the ML-KNN framework using three feature representations extracted from two different unsupervised anomaly detection models. The proposed solution effectively uses limited labeled samples (i.e., eight samples for each of the five classes considered in our evaluation), providing results that are better than supervised approaches trained with few samples. Our evaluation with single-label and multi-label formulations demonstrates the strengths of the proposed approach, increasing the interpretability of the anomalous segments detected by unsupervised approaches.

While the approach was trained with two particular unsupervised driving anomaly detection systems, our formulation is flexible and can be implemented with other methods as long as we can identify an appropriate feature embedding.

The increase in the interpretability of the results from the unsupervised driving anomaly detection model broadens the application of unsupervised methods for downstream tasks. Our proposed approach can be used to create actionable safety measures when a given type of anomaly is detected. The approach is also flexible. If a new type of anomaly is required, we only need to annotate a few examples to be used as anchors (e.g., using cellphone when driving). Furthermore, this method provides a possible solution to facilitate the automatic annotation of large databases, where our formulation can be used to retrieve potential candidate segments that are similar to a predefined type of anomaly. The choice of anchors can be important for the success of the k-NN classifier. Therefore, we will evaluate the sensitivity to the proposed approach on the selected anchors. We can randomly change or remove some anchors from each class and measure the impact on the results.

REFERENCES

[1] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *IEEE Intelligent Vehicles Symposium (IV 2012)*, Alcala de Henares, Spain, June 2012, pp. 234–239.

[2] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. González, "Safe driving using mobile phones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1462–1468, September 2012.

[3] T. Chakravarty, A. Ghose, C. Bhaumik, and A. Chowdhury, "MobiDriveScore - a system for mobile sensor based driving analysis: A risk assessment model for improving one's driving," in *International Conference on Sensing Technology (ICST 2013)*, Wellington, New Zealand, December 2013, pp. 338–344.

[4] J. Wahlström, I. Skog, and P. Händel, "Risk assessment of vehicle cornering events in GNSS data driven insurance telematics," in *IEEE Conference on Intelligent Transportation Systems (ITSC 2014)*, Qingdao, China, October 2014, pp. 3132–3137.

[5] J. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *SIGCHI Conference on Human Factors in Computing Systems*, Toronto, ON, Canada, April-May 2014, pp. 4047–4056.

[6] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "D3: Abnormal driving behaviors detection and identification using smartphone sensors," in *IEEE International Conference on Sensing, Communication, and Networking (SECON 2015)*, Seattle, WA, USA, June 2015, pp. 524–532.

[7] J. Wahlström, I. Skog, and P. Händel, "Detection of dangerous cornering in GNSS-data-driven insurance telematics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3073–3083, December 2015.

[8] F. Li, H. Zhang, H. Che, and X. Qiu, "Dangerous driving behavior detection using smartphone sensors," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 1902–1907.

[9] P. Vavouranakis, S. Panagiotakis, G. Mastorakis, C. X. Mavromoustakis, and J. M. Batalla, "Recognizing driving behaviour using smartphones," in *Beyond the Internet of Things: Everything Interconnected*, J. Batalla, G. Mastorakis, C. Mavromoustakis, and E. Pallis, Eds. Cham, Switzerland: Springer International Publishing, January 2017, pp. 269–299.

[10] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong, and M. Li, "Fine-grained abnormal driving behaviors detection and identification with smartphones," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2198–2212, August 2017.

[11] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *International Conference on Pervasive Computing Technologies for Healthcare*, Munich, Germany, March 2010, pp. 1–8.

[12] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013, pp. 1–6.

[13] ——, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[14] Z. Zhu and Q. Ji, "Real time and non-intrusive driver fatigue monitoring," in *IEEE International Conference on Intelligent Transportation Systems*, Washington, DC, October 2004, pp. 657–662.

[15] R. Chai, G. R. Naik, T. N. Nguyen, S. H. Ling, Y. Tran, A. Craig, and H. T. Nguyen, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 715–724, May 2017.

[16] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *ACM Conference on Embedded Network Sensor Systems (SenSys 2008)*, Raleigh NC USA, November 2008, pp. 323–336.

[17] Z. Liu, M. Wu, K. Zhu, and L. Zhang, "SenSafe: A smartphone-based traffic safety framework by sensing vehicle and pedestrian behaviors," *Mobile Information Systems*, vol. 2016, pp. 1–13, October 2016.

[18] C. Yang, A. Renzaglia, A. Paigwar, C. Laugier, and D. Wang, "Driving behavior assessment and anomaly detection for intelligent vehicles," in *IEEE International Conference on Cybernetics and Intelligent Systems (CIS 2019) and IEEE Conference on Robotics, Automation and Mechatronics (RAM 2019)*, Bangkok, Thailand, November 2019, pp. 524–529.

[19] T. A. Dingus, J. Hankey, J. F. Antin, S. E. Lee, L. Eichelberger, K. Stulce, D. McGraw, M. Perez, and L. Stowe, "Naturalistic driving study: Technical coordination and quality control," Transportation Research Board of the National Academies, Washington, D.C., USA, Technical Report SHRP 2 Report S2-S06-RW-1, July 2014. [Online]. Available: https://trid.trb.org/view/1316354

[20] M. Zhang, C. Chen, T. Wo, T. Xie, M. Bhuiyan, and X. Lin, "SafeDrive: Online driving anomaly detection from large-scale vehicle data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2087–2096, August 2017.

[21] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and canbus data," in *ACM International Conference on Multimodal Interaction (ICMI 2019)*, Suzhou, Jiangsu, China, October 2019, pp. 164–173.

[22] ——, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Intelligent Transportation Systems Conference (ITSC 2020)*, Rhodes, Greece, September 2020, pp. 1–7.

[23] Z. Wang, G. Yuan, H. Pei, Y. Zhang, and X. Liu, "Unsupervised learning trajectory anomaly detection algorithm based on deep representation," *International Journal of Distributed Sensor Networks*, vol. 16, no. 12, pp. 1–21, December 2020.

[24] O. Köpüklü, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2020)*, Virtual Conference, January 2021, pp. 91–100.

[25] Y. Qiu, T. Misu, and C. Busso, "Unsupervised scalable multimodal driving anomaly detection," *IEEE Transactions on Intelligent Vehicles*, vol. to appear, 2022.

[26] B. Zhou, S. Liu, B. Hooi, X. Cheng, and J. Ye, "BeatGAN: Anomalous rhythm detection using adversarially generated time series," in *International Joint Conference on Artificial Intelligence (IJCAI 2019)*, Macao, China, August 2019, pp. 4433–4439.

[27] H. Zhao, H. Zhou, C. Chen, and J. Chen, "Join driving: A smart phone-based driving behavior evaluation system," in *IEEE Global Communications Conference (GLOBECOM 2013)*, Atlanta, GA, USA, December 2013, pp. 48–53.

[28] J. Chen and Q. Ji, "A probabilistic approach to online eye gaze tracking without explicit personal calibration," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1076–1086, March 2015.

[29] L. Su, C. Sun, D. Cao, and A. Khajepour, "Efficient driver anomaly detection via conditional temporal proposal and classification network," *IEEE Transactions on Computational Social Systems*, vol. Early Access, pp. 1–10, March 2022.

[30] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, July 2007.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[32] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, May 2000.