Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 4, 587–599
DOI: 10.34768/amcs-2021-0040

FORENSIC DRIVER IDENTIFICATION CONSIDERING AN UNKNOWN SUSPECT

KLARA DOLOŠ [a,*], CONRAD MEYER [a], ANDREAS ATTENBERGER [a], JESSICA STEINBERGER [a]

[a]Central Office for Information Technology in the Security Sector
Zamdorfer Str. 88, 81677 Munich, Germany
e-mail: klara.dolos@zitis.bund.de

One major focus in forensics is the identification of individuals based on different kinds of evidence found at a crime scene and in the digital domain. Here, we assess the potential of using in-vehicle digital data to capture the natural driving behavior of individuals in order to identify them. We formulate a forensic scenario of a hit-and-run car accident with a known and an unknown suspect being the actual driver during the accident. Specific aims of this study are (i) to further develop a workflow for driver identification in digital forensics considering a scenario with an unknown suspect, and (ii) to assess the potential of one-class compared to multi-class classification for this task. The developed workflow demonstrates that in the application of machine learning in digital forensics it is important to decide on the statistical application, data mining or hypothesis testing in advance. Further, multi-class classification is superior to one-class classification in terms of statistical model quality. Using multi-class classification it is possible to contribute to the identification of the driver in the hit-and-run accident in both types of application, data mining and hypothesis testing. Model quality is in the range of already employed methods for forensic identification of individuals.

Keywords: natural driving behavior, digital biometry, OCC, CAN-BUS data, validation.

1. Introduction

A number of methods such as forensic dactyloscopy, anthropometry and odontology are major tools of criminal identification that have been existing for more than a century. These methods use physical characteristics for the identification of persons. Dactyloscopy (usage of fingerprints) for identification has been used to link suspected criminals to a crime scene or to identify victims. Large automated fingerprint identification systems (AFIS) are used to match unknown fingerprints against a database. This approach is used worldwide to identify criminals as well as victims. As digitalization increases in all areas of life, also digital traces and their use are becoming more available such as spectrographic voice and gait identification and keystroke biometrics (Benzaoui *et al*., 2014; Bouchrika *et al*., 2011; Müller, 2007).

The development of methods for digital biometry for other goals than forensics has come a long way. In the beginning, algorithms such as the Gaussian mixture density with rather few parameters were applied (Reynolds, 1994). With increasing computational power and larger data sets, more complex machine learning approaches such as feed-forward neural networks were developed and tested for, e.g., spectrographic voice identification (Ge *et al*., 2017). Lately, there has been an increasing interest in applying biometric methods on digital traces of driving behavior (Bernardi *et al*., 2018; Martinelli *et al*., 2020; Remeli *et al*., 2019; Dološ *et al*., 2020; Turunen and Dološ, 2021).

Although we focus on the identification of individuals using digital behavior data in the context of forensics, the topic can be related to a broader context. One early research field is how to use keystroke dynamics to estimate if a passcode was used by the person before (D'Lima and Mittal, 2015). If so, the passcode is considered not secure and could be rejected by the software. Another way of using digital biometry is the increase in authentication security using passcodes together with keystroke dynamics in the fixed-text context (Acien *et al*., 2020). This is similar to using the owner's driving style and, e.g., body weight for securing access control together with physical or electronic keys. Also

[*]Corresponding author

touch point data during swipe actions on touch screens could be used together with voice data for authentication for mobility services (Gupta *et al*., 2019). Other emerging fields are, e.g., authentication in virtual reality (Kupin *et al*., 2019) and smart cities (Ross *et al*., 2020) and, last but not least, pay-how-you-drive insurance contracts (Wakita *et al*., 2005; Carfora *et al*., 2019).

Research in the field of identification of individuals using digital biometric data share mostly two work steps: the feature extraction from the original data and the model parametrization (Mashao and Skosan, 2006; Tirumala *et al*., 2017; Gupta *et al*., 2019). Existing and novel approaches concerning feature extraction and model algorithms, especially for time series data, need to be assessed for their suitability on different digital traces. Understanding their effects and functioning will improve our ability to built robust classifiers. The application of machine learning methods on digital traces such as in-vehicle network data (for example, CAN bus data) is a promising approach (Dološ *et al*., 2020; Turunen and Dološ, 2021). For digital forensics this would be a great possibility, but these methods need a sound scientific foundation, especially regarding credibility and false conviction rates (Page *et al*., 2011; Thompson, 2006). This is supported by the finding that in 86 DNA exoneration cases, forensic science testing errors were the second leading cause of wrongful convictions (found in 63% of cases), only falling behind eyewitness misidentifications (71% of the cases) (Lieberman *et al*., 2008; Saks and Koehler, 2005).

Additionally to an explainable modeling procedure, communication of results, especially the reliability of the evidence, needs to be adapted to the needs in forensics (Christensen *et al*., 2014; Helm and Hagendorff, 2021) and the meaning of different measures of errors needs to be considered. To give an example, the interpretation of the false detection rate (FDR, as used by Dološ *et al.* (2020)) is pointing to the false conviction rate. The number of persons being imprisoned erroneously is a core value to be discussed in ethics and politics. The false positive rate (FPR, for example, used by Houck and Budowle (2002) as well as White *et al.* (2015)) corresponds to the individual probability of a suspect of being accused guilty when actually innocent. Such differences among model quality measures need to be considered in the discourse on credibility of forensic evidence used for identification.

In a prior study we showed how the actual driver could be identified out of a known and finite group of suspects based on digital vehicle data (Dološ *et al*., 2020). This however is rarely the situation in the real world. There, one (or more) known suspects are available, but it is likely that there was an unknown person involved. For this case the modeling approach and the forensic process needs to be adjusted. Two different approaches are commonly used for such tasks: one-class classification and multi-class classification. Since there are inconsistencies in the use of the term one-class classification, there is need to explain what is meant in this study with one-class and multi-class classification, respectively (Khan and Madden, 2010).

In the present study, we used the term one-class classification to describe a model aiming at finding one target class in some data. Following this reasoning, the training data contains labeled data for the target class (“positives”) and unlabeled data representing several other classes (“negatives”) or data representing the background distribution (including the target class). This differs from the binary classification problem including two classes not only conceptually, but also regarding variances in the data and the degree of overlapping distributions. This differs from a “real” one-class classification using, e.g., a one-class SVM where only the target class is represented in the training data (“positives”) and no further information is available (Khan and Madden, 2010; Antal and Szabo, 2015; Mack and Waske, 2017). Although this seems to be a cost efficient approach, the information content in such training data is low. Especially for data with overlapping distributions for the target class with the background distribution, this approach will yield worse results. In contrast, in a multi-class classification there can be a focus on one single target class, but the data are completely labeled (no merged classes as “negatives” or data representing the background distribution). These data hold most information but at the same time are most expensive and also questionable concerning data privacy. The described three approaches, mainly defined by the training data type used, can thus be sorted following the amount of information they make use of and the model quality that can be achieved (Mack and Waske, 2017; Antal and Szabo, 2015; Khan and Madden, 2010):

$$\begin{aligned}&\textit{real one-class classification}\\ &\quad < \textit{target}/\textit{background one-class classification}\\ &\qquad < \textit{multi-class classification}.\end{aligned}$$

Application of a one-class classification with labeled data for the suspect and unlabeled data for other drivers is straightforward and cost efficient. Unlabeled data representing the background distribution are easy to gather because there is no need to track the driver but only the electronic vehicle data. Such data are already available in the context of the development cycle of vehicles and motor optimization. Such data could be even simulated for a better coverage of possible driving styles and at the same time considering the actual road situation during the incident. Application of a multi-class classification with labeled data using a driver data base together with the suspect data is more informed. Such data hold much more information. It is similar to large DNA or fingerprint databases already used in forensics and paternity tests. Unfortunately, to our knowledge, a driver database for forensic applications does not exist yet.

In this study we assessed the potential of using in-vehicle digital data to capture the natural driving behavior of individuals in order to identify them (Kwak *et al*., 2017; Martinelli *et al*., 2020; Dološ *et al*., 2020). We formulated a forensic scenario of a hit-and-run car accident considering a known as well as an unknown suspect. Specific aims of this study were (i) to further develop a workflow for driver identification in digital forensics considering a scenario with an unknown suspect, and (ii) assess the potential of one-class classification compared with multi-class classification for this task.

2. Methods

2.1. Forensic scenario: A hit-and-run accident. In this section we describe a possible use case for the method we developed. In a hypothetical hit-and-run accident law enforcement was able to identify the vehicle, which was involved, but is unsure about its driver. There is one known suspect named "B" who denies having driven the car. In-vehicle digital data were available, for example, provided by the insurance company as a result of a pay-how-you-drive car insurance contract. Such data provide information at high sampling rate for vehicle speed, accelerator pedal positions, steering wheel positions, and changes in these variables. Thus, it could be possible to identify the actual driver by his/her natural driving behavior calculated from in-vehicle network data. In our scenario, the suspect B was asked for a driving sample of approximately 40 min. These data were used to create a driver profile (using modeling). Using this driver profile, the probability of class membership of the suspect to the evidence data could be calculated together with model quality measures.

Since we are omniscient in this scenario, we know that actually a driver named "F" caused the accident and was the source of the evidence data.

2.2. Electronic vehicle data. For this study we used freely available data[1] (Kwak *et al*., 2017; Martinelli *et al*., 2020). In total ten drivers traveled between Korea University and the SANGAM World Cup Stadium in the surroundings of Seoul (South Korea). The experiment was performed in July 2015. The time factor was controlled by performing experiments from 8 p.m. to 11 p.m. on weekdays. Ten drivers completed two round trips resulting in 46 km total length. The data were collected from the city, highway and parking space. In the city there are signal lamps and crosswalks; at parking spaces driving cautiously was required.

The number of features recorded was 51 in 1 Hz sampling rate. Total driving time per individual was between 121 and 184 minutes. Data records with zero Vehicle_speed were excluded prior to the analysis. The ten drivers were labeled from "A" to "J" with around 7312 records for each driver (A: 5461, B: 9634, C: 5508, D: 10353, E: 6696, F: 8764, G: 6087, H: 7744, I: 5782, J: 7087).

Original features related to human behavior considered in this study were accelerator pedal value, master cylinder pressure, vehicle speed, steering wheel speed and steering wheel angle. Additionally, roughness was calculated with rolling window size $k = 20$ s for all features using the R-function `roughness{seewave}`. This is a compromise between the claim to maintain high temporal resolution and the need for a certain period of time for the calculation of roughness. Roughness was implemented as total curvature for the specified window size, i.e., as the integrated squared second derivative defined as follows:

$$\mathrm{RNS}_{(f)} = \int_{f-k/2}^{f+k/2} (f''(t))^2 \,\mathrm{d}t. \tag{1}$$

where f denotes the respective feature, t is the point in time and k is the window size for which roughness is calculated. In total we remained with 14 features (Table 1).

2.3. One-class classification. The first approach was a one-class classification (OCC) with the aim to correctly classify one target class out of all data. In total 10 random forest models (RF) (Liaw and Wiener, 2002; Breiman, 2001) with 300 trees were trained with data consisting of the target class (e.g., driver A, "positives") and a random sample of all other data ("negatives"). In order to decide for the number of trees necessary, we plotted the error against the number of trees similarly to, e.g., Oshiro *et al*. (2012). Model quality converged and did not increase any more. Such a fitted model can be applied to any other data and will give a probability for the class membership of the target class.

For each of the ten drivers we created training data, validation and evidence data using random block splitting. The data for each driver ("positives") were split into 200 sequences from which 75% were used for training. The remaining 25% were used for validation (calculation of model quality). For each driver also "negatives" were needed. Therefore, a negative sample of size 20% was drawn from all other data and used as "negatives" during model calibration. The remaining data were used as validation data together with the positives from the random block splitting (i.e., the data belonging to the

[1] `http://ocslab.hksecurity.net/Datasets/driving-dataset`.

Table 1. Model quality in OCC with random forests for each driver for the validation data.

Driver	Number of negatives (all other drivers)	Number of positives (driver data)	ACC	Balanced ACC	FDR.N negatives	FDR.P positives	FDR total
A	67655	1323	0.79	0.79	0.01	0.93	0.21
B	63482	2352	0.70	0.71	0.01	0.92	0.30
C	67608	1346	0.75	0.75	0.01	0.94	0.25
D	62763	2515	0.72	0.75	0.01	0.90	0.28
E	66420	1637	0.70	0.74	0.01	0.94	0.30
F	64352	2120	0.80	0.63	0.02	0.93	0.20
G	67029	1481	0.80	0.72	0.01	0.93	0.20
H	65372	1903	0.62	0.62	0.02	0.96	0.38
I	67334	1386	0.67	0.64	0.01	0.96	0.33
J	66029	1727	0.70	0.79	0	0.93	0.30

driver who caused the accident). The data of the person who caused the accident were placed in the middle of the time series to avoid edge effects. Our decision to use the same number of sequences for each driver resulted in differences in the number of data records (Table 1).

Model quality measures such as accuracy (ACC), false detection rate (FDR), false positive rate (FPR) were used in this work. They are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}, \tag{2}$$

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}, \tag{3}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}, \tag{4}$$

where TP, TN, FP, FN, P and N denote true positives, true negatives, false positives, false negatives, positives and negatives, respectively.

Additionally, individual FDR and FPR were calculated. “Individual” FDR refers to the FDR only using classification results for the target class. For example, FDR for positives P was calculated as FDR.P and FDR for driver A was calculated as FDR.A defined by the following formulas:

$$\mathrm{FDR.P} = \frac{\mathrm{FP.P}}{\mathrm{FP.P} + \mathrm{TP.P}}, \tag{5}$$

$$\mathrm{FPR.A} = \frac{\mathrm{FP.A}}{\mathrm{TN.A} + \mathrm{FP.A}}, \tag{6}$$

FDR can be interpreted as the false conviction rate highly relevant in the field of forensics and the evaluation of the reliability of traces as evidence. FPR is the proportion of all negatives that still yield positive test outcomes and corresponds to a person’s probability of being accused guilty when actually innocent. FPR and ACC were calculated in order to compare the machine learning derived evidence quality with reports in the literature (Lieberman *et al.*, 2008; Saks and Koehler, 2005). Feature importance (Breiman, 2001, Ch. 10) was calculated as the mean value of all 10 one-class classification models.

In order to provide a visual interpretation, we calculated a prediction on the validation data for the target class (“positives”) and data for the other drivers (“negatives”). In more detail, the data consisted of a string of all sequences of driver data that were not used for training the model. Among those sequences also the target class (“positives”) was present. Applying the model to the validation data provided a probability for the class membership. In order to pronounce differences among predicted class probabilities p visually, they were squared p^2. Thereafter, curves were smoothed with a rolling mean (RM) with window size $k = 61$ s to reduce short-term fluctuations and to reveal the trend and at the same time remaining with an acceptable temporal resolution. For each data point the RM with window size k was calculated for feature values f as follows (valid for odd numbers of k):

$$\mathrm{RM}_k = \frac{1}{k} \cdot \sum_{i=1-(k-1)/2}^{(k-1)/2} f_i. \tag{7}$$

2.4. Multi-class classification. The second approach was a multi-class classification (MCC). The model output were probabilities for the class membership for all classes in the data, including the target class. A random forest with 300 trees was formed for all ten drivers. For validation, the data was split into test and training data using random block splitting. The data were split into 200 sequences from which 50% were used for training. Model quality measures were calculated using the validation data analogous to one-class classification. We additionally provided the no-information rate for ACC. The no-information rate is ACC for data for which the response and the predictors are independent. If the model ACC is not higher than the no-information rate, the model can be considered as useless.

In order to provide a visual interpretation, we calculated a prediction on the validation data. Curves were smoothed with rolling mean windows size 61 s in the same way as for the results of OCC. Additionally to the evaluation of the quality of class memberships of the original model predictions, predicted class memberships after smoothing were evaluated for the MCC approach.

2.5. Forensic scenario: Multi-class classification. To demonstrate the workflow and to make it easier to discuss different aspects of application of driver identification for either hypothesis testing or data mining, we chose two specific drivers for our use case. In the forensic scenario we stated that there was one know suspect B. B is one of the drivers for which classification worked well. We further defined F as the actual driver. F was chosen because it is also one of the well classified drivers and additionally not at the edges of the sequence to be plotted. Actually, any combination could be used or we could iterate over all possible combinations. For now, however, we provide a use case as a basis for the development of a workflow and the explanation of possible outcomes and pitfalls of forensic driver identification.

For this use case with the unknown suspect F, the model was not calculated with all the data for all ten drivers labeled with the alphabetic letters from A to J. Instead, the model was calculated using training data not containing driver F. This simulated the situation of having a driving sample of suspect B pooled together with a labeled driver data base for model training similar to automated fingerprint identification systems (AFISs). Predictions were made on data for all drivers A to J including driver B, who was the known suspect due to other reasons, and driver F, the evidence data derived from the vehicle or insurance company.

In more detail, first the data splitting was done as described above for the MCC model. After that, driver F was excluded from the training data and the model was parameterized for the remaining nine drivers. The overall model quality (ACC, FDR, FPR) was calculated based on test data also not containing F.

Model fitting and validation for the random forest algorithm could be calculated on an off-the-shelf computer within 1 to 2 hours.

3. Results

3.1. One-class classification. For each driver an OCC was calculated based on data with labels 1 (“positives”) for the target class and 0 (“negatives”) for data belonging to all other drivers (Table 1). For each driver the model provided a probability for class membership for each data record. Results showed that there was a higher predicted probability for the target class. Also the commonly applied model evaluation using ACC indicated that the models were statistically valid. Nevertheless, results were not satisfying for forensics. Especially the FDR for positives (FDR.P) showed clearly that this modeling approach needs further work (Table 1).

Time series prediction also resulted in no clear visual separability among positives and negatives (see Fig. 1) although a visible signal appeared which was clear, for example, for drivers C and J and also A. This is usually caused by a lower similarity of the data of these drivers to the other data. This corresponds to the highest balanced accuracy of 0.79 and 0.75 and FDR_total with 0.25 and 0.30 but not a low FDR.P, which is relevant for trusting the evidence in the conviction context.

The pattern with higher importance of roughness features for classification of positives is still remarkable (Table 2).

3.2. Multi-class classification. Multi-class classification (MCC) provided a better separability of drivers than one-class classification (OCC). The overall accuracy ACC of the model was 0.39. Compared with the no-information rate 0.17 given by the evaluation function, this is already a statistically valid model. But model quality measures relevant for forensics were still low. Especially, class specific FDRs (see Table 3, columns “original”) were too high baring in mind that this is the false conviction rate we would have to accept. Note that the additional information in the time series has not been used yet.

This is different when plotting the prediction for the time series data of the hold-out sample (Fig. 2). The model prediction resulted in a probability for class memberships for each point in time. Thus for each data record ten probabilities were predicted. The rolling mean was calculated as described in the methods. Squaring all values pronounced high values (for better visual interpretation). For most of the time series a rather clear attribution to one driver was possible. An exception was the time period around 25,000–35,000 s for drivers H, I J. ACC and FDRs became remarkably better (Table 3, columns “rolling mean”). FDR decreased from 0.61 to 0.45. ACC was 0.54 for this workflow. Together with the visual interpretation of the time series prediction on the hold-out sample, for most of the time a trustworthy attribution could be achieved (Fig. 2).

Feature importance also showed high relevance of roughness features together with vehicle speed and also steering wheel angle (Table 4).

3.3. Forensic scenario. An MCC was calculated for nine drivers in the same way as before. In the hold-out sample additionally the unknown suspect F was present. The data for F in the hold-out sample can be considered as evidence data belonging to an unknown suspect. When

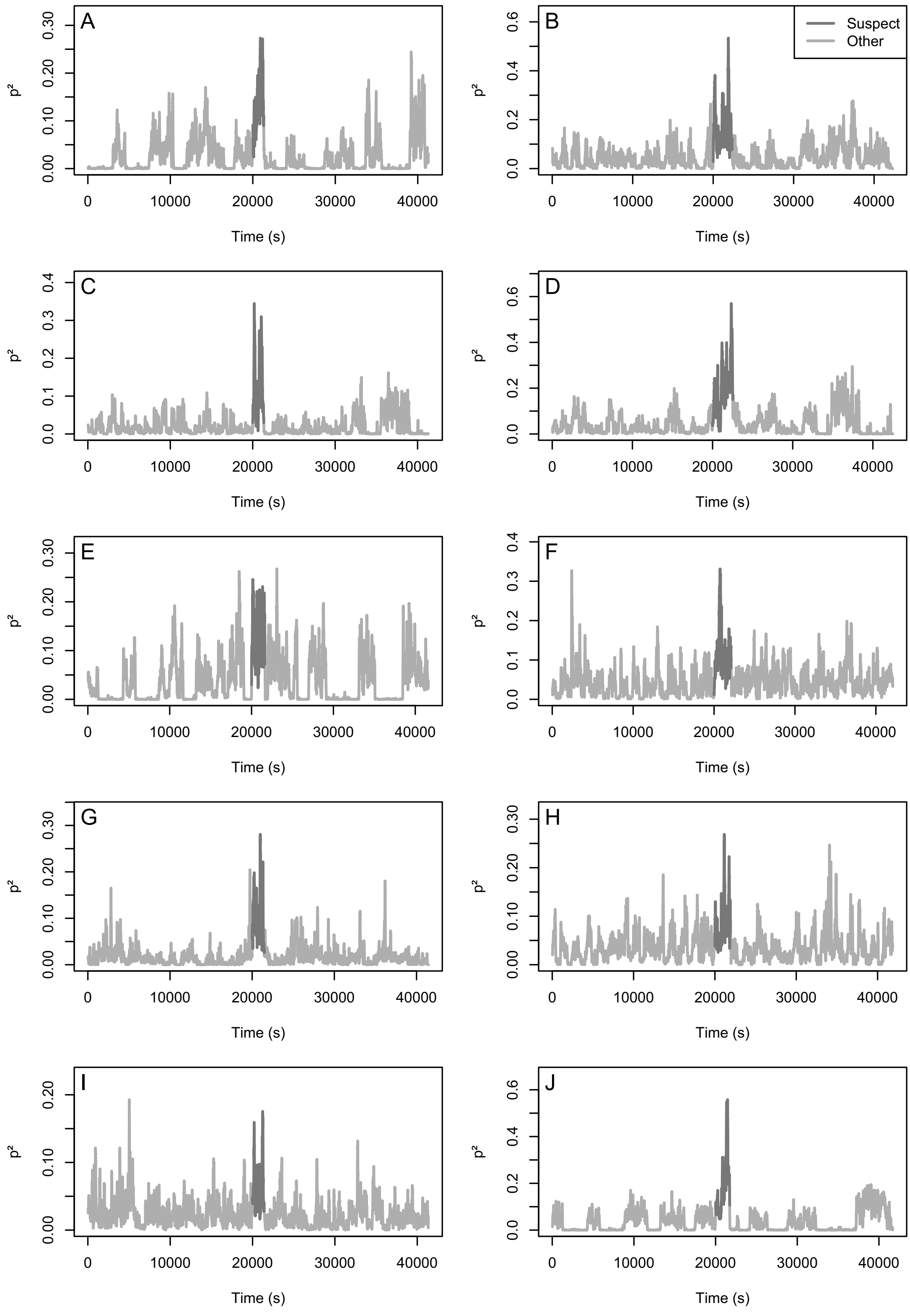

Fig. 1. Prediction of class probability (squared class probability p^2 for better visibility) for OCC on the validation data not used during model fitting. High values indicate a high probability of the data being attributed to the suspect.

Table 2. Mean feature importance for all ten OCC models. Features were sorted by the mean decrease in ACC for positives. Interpretation needs to be performed carefully since model quality was low.

Rank	Feature	Mean decrease ACC negatives (all other data)	Mean decrease ACC positives (driver)	Mean decrease ACC
1	Accelerator pedal value roughness	23.87	89.06	88.86
2	Vehicle speed	2.95	72.59	74.04
3	Master cylinder pressure roughness	2.41	67.79	67.54
4	Steering wheel speed roughness	14.03	62.5	65.64
5	Vehicle speed roughness	1.22	57.02	54.16
6	Steering wheel angle roughness	11.56	54.41	60.24
7	Path order	-6.29	51.99	47.73
8	Master cylinder pressure	3.83	49.52	41.08
9	Acceleration speed lateral	-7.69	31.05	20.79
10	Steering wheel angle	-7.94	30.18	17.41
11	Fuel consumption	10.5	27.1	31.86
12	Accelerator Pedal value	10.68	22.54	26.48
13	Acceleration speed longitudinal	10.7	20.97	26.19
14	Steering wheel speed	4.71	5.52	9.09

Table 3. FDR and FPR for MCC for the original predictions and rolling mean treatment. Overall model statistics (original/rolling mean): ACC $= 0.39/0.54$, FDR $= 0.61/0.45$, FPR $= 0.06/0.05$.

	FDR		FPR	
Driver	original	rolling mean	original	rolling mean
A	0.61	0.52	0.04	0.02
B	0.56	0.42	0.07	0.04
C	0.60	0.36	0.05	0.05
D	0.59	0.53	0.07	0.04
E	0.58	0.47	0.05	0.02
F	0.62	0.43	0.08	0.07
G	0.50	0.22	0.05	0.04
H	0.73	0.51	0.09	0.08
I	0.75	0.51	0.07	0.06
J	0.63	0.42	0.06	0.05

we follow a data mining workflow, we treat all drivers in the same way; thus, all of them were equally suspicious. For most of the time, class attributions were quite clear, however, predictions for class memberships for the F data were rather inconclusive (Fig. 3, Table 5). Note that also classifications for drivers H, I, J were inconclusive, C had some high values but they were not stable. It is thus not possible to conclude from this analysis that there has actually been an unknown suspect.

Nevertheless, one class needs to be the one with the highest predicted class probability. In our scenario driver D was most often miss-classified as being the source of the suspect data F (Fig. 4). This can be translated into D having the highest risk of false conviction. This points to the need of interpreting predicted probabilities for class memberships in relation to those for all other data. The probabilities for F data miss-classified as D were among the lowest and lower than for the true D data.

However, in the described hit-and-run case (see Section 2.1), there was a driver B who was identified as a suspect and therefore provided a driving sample to the police. This driving sample was hypothetically used together with data from a forensic driver data base. Clearly, the predicted class probability for the evidence data (i.e., driver F) was very low for B. Thus, the forensic driver identification workflow resulted in the exoneration of the known suspect B in our forensic scenario. For driver D there was no risk of false accusation, since these data belonged to a (hypothetical) forensic driver database.

4. Discussion

In this study we assessed the scenario of a hit-and-run car accident considering a known suspect B and an unknown suspect F (the actual driver during the accident). Our first approach was to use one-class classification. This is a well studied and successful approach in other research fields (Bergamini *et al.*, 2009; Mack *et al.*, 2014; Stenzel *et al.*, 2017). It allows us to directly model the probability of a driving sample or an evidence being a member of the target class. The second approach was a multi-class classification in which the suspect was missing in the training data. Both approaches provided us with new findings which will improve our abilities in applications of forensic driver identification and also other fields of digital biometry.

Table 4. Feature importance for MCC. Features are ordered by mean decrease in ACC.

Rank	Feature	Mean decrease ACC	A	B	C	D	E	F	G	H	I
1	Vehicle speed	213	116	203	166	190	116	181	127	159	182
2	Steering wheel angle	198	67	74	70	102	82	110	91	83	88
3	Master cylinder pressure roughness	194	103	171	141	141	123	158	142	156	159
4	Accelerator pedal value roughness	192	165	164	160	209	107	153	149	178	199
5	Vehicle speed roughness	177	103	116	126	133	109	134	124	125	140
6	Steering wheel speed roughness	171	107	113	112	178	113	157	155	114	132
7	Acceleration speed lateral	163	65	86	73	102	85	102	77	92	83
8	Path order	158	95	108	110	135	95	133	103	131	122
9	Steering wheel angle roughness	151	91	142	109	138	94	108	105	132	132
10	Master cylinder pressure	129	87	72	67	90	123	93	98	98	108
11	Fuel consumption	78	54	48	52	63	61	73	64	64	65
12	Acceleration speed longitudinal	67	43	40	42	48	49	60	56	46	57
13	Accelerator pedal value	25	31	20	21	22	31	26	29	22	25
14	Steering wheel speed	22	12	30	20	20	14	19	16	19	17

Table 5. False detection rates (FDR) and false positive rates (FPR) for multi-class classification (MCC) for the forensic scenario, i.e., the model was calculated on training data without F. Overall model statistics: ACC $= 0.56$ (no information rate $= 0.20$), FDR $= 0.44$, FPR $= 0.05$.

Driver	FDR rolling mean	FPR rolling mean
A	0.46	0.03
B	0.47	0.05
C	0.18	0.04
D	0.45	0.03
E	0.40	0.04
G	0.14	0.04
H	0.54	0.09
I	0.39	0.08
J	0.58	0.06

4.1. Reliability of one-class and multi-class classification. The failure of OCC in this study was a surprise to us, because in other research fields much better results could be achieved (Märkel and Dološ, 2017; Stenzel *et al.*, 2017; Antal and Szabo, 2015). The main reason is likely a large within-subject variability due to strong differences in road conditions and traffic during data capture. When all other drivers were pooled in one group (the “negatives”) they covered the whole feature space and the target class could not be separated clearly.

MCC yielded much better results. There, labeled data was used for each driver separately. After the prediction was treated by rolling mean smoothing, the results were acceptable for most drivers. For those drivers, for which model predictions were inconclusive in the time series prediction, model quality measures and the time series were consistent, i.e., all indicated that the evidence was inconclusive and should not be used for these drivers.

The reason why applying a rolling mean was successful is that it reduced short-term fluctuations and highlighted longer-term trends. High probabilities for class memberships occurred only for few consecutive seconds. This explains an increase in the prediction accuracy for neighboring data. Another way to improve results also for OCC could be sophisticated data pre-processing. Special driving maneuvers could be used which are assumed to pronounce the characteristics in natural driving behavior. An approach worth being investigated is using not the time series and machine learning but to take a closer look at specific patterns in the data. Such patterns are value combinations which are rare, e.g., only a few seconds each hour, but occur with one person only (Turunen and Dološ, 2021). Such an analysis of driver specific patterns is complementary to the machine learning methodology. Especially efforts for improving the results of the OCC approach are needed, because it is data saving as well as close to forensic applications. For further research suitable data need to be created by well designed experiments considering findings of our study.

4.2. Forensic scenario. At the beginning of this section, we would like to highlight the importance recognizing fundamental statistical principles when using machine learning in forensics. Regardless of the specific machine learning approach, application of digital biometrics for forensic identification of persons can be applied in different ways (Dessimoz and Champod, 2008; Mordini, 2017). Here, two ways differing in their statistical interpretation were shown in the results and will

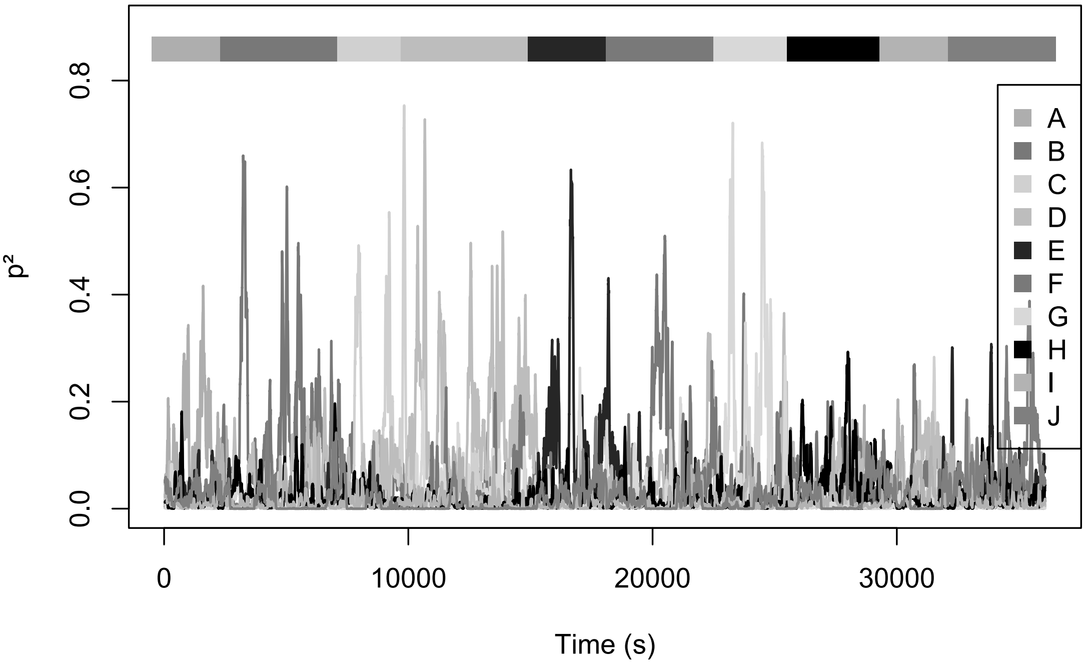

Fig. 2. Prediction of MCC on the hold-out sample: y-axis—squared class-probability for a better visual impression.

be discussed briefly. One way is to search for indicators which suspect out of all known suspects could be worth to be investigated first or with most resources. From the statistical perspective this is data mining and not hypothesis testing. This leads to the consequence, that findings provided by the model should have a low weight or no weight at all in decision making in a court hearing. However, it can be very useful to identify directions for investigations. In our scenario this would have resulted in a weak attribution of the evidence data to driver D and further investigation in that direction.

The other way is to formulate a hypothesis based on other evidence, which suspect was most likely the actual driver. We stated that driver B was a suspect according to the police and based on other reasons than the driving sample. Then, an identification of the driver using the natural driving behavior is used as an additional confirmation or rejection. Statistically this is an important difference which leads to a stronger meaning of the predicted attribution of the evidence data to the suspect (Glymour *et al.*, 1997). Thus, if a hypothesis is formulated first and is confirmed or rejected by the model results, this has a stronger implication in decision making in a court hearing. That the evidence data (from driver F) was attributed to driver D is not relevant when using a hypothesis before model calculation. Further, in such application the data of driver D was part of the forensic driver database and thus very unlikely the source of the digital trace. This application considerably differed from a data mining approach in which all drivers would be considered as suspects. We recommend to use the presented workflow in an unknown suspect scenario with a prior hypothesis, thus to verify if, e.g., driver B was the source of the evidence data, in order to minimize the risk of false attribution.

Our best methods resulted in an ACC of 0.60 (with no information rate 0.19) and an FDR of 0.40 pointing to the false conviction rate. We do highlight that these values already indicate that the MCC model was statistically valid, i.e., significantly better than random. But in the application of machine learning methods in the field of forensics, not only the comparison against a null-hypothesis needs to be passed, but also the question which error rate in terms the false conviction rate (FDR) and the individual risk of false accusation (FPR) is accepted by society (Mordini, 2017, Ch. 16.6).

The values of the present study can be related to those of estimated false convictions of 4.1% (comparable to FDR in our study) for death-sentenced defendants in the US (Gross *et al.*, 2014) and also to other methods for identification in forensics. Spectrographic voice identification error rates were found to range from 31% to 63% (Koenig, 1986; Faigman, 2002; Saks and Koehler, 2005). Using face recognition software with human interaction miss-identification (FPR) ranged between (10–)30–40% (White *et al.*, 2015), microscopic hair comparisons were about 12% (Houck and Budowle, 2002), keystroke dynamics were about 7% (Antal and Szabo, 2015; Eude and Chang, 2018). Identification of latent fingerprints were reported to be better than

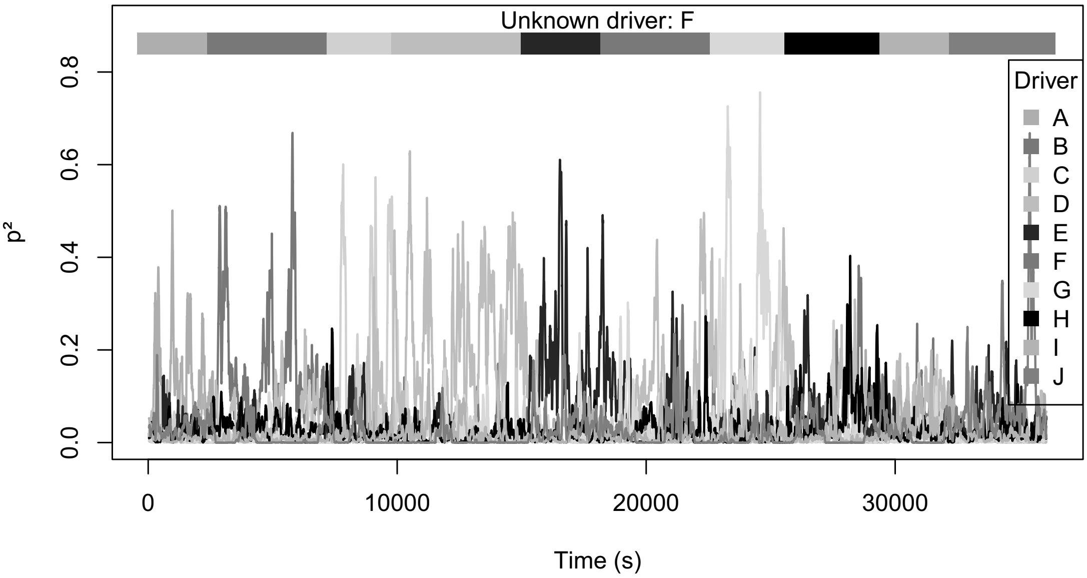

Fig. 3. Model prediction for the forensic scenario with the unknown suspect F using the multi-class approach: y-axis—squared class-probability for a better visual impression.

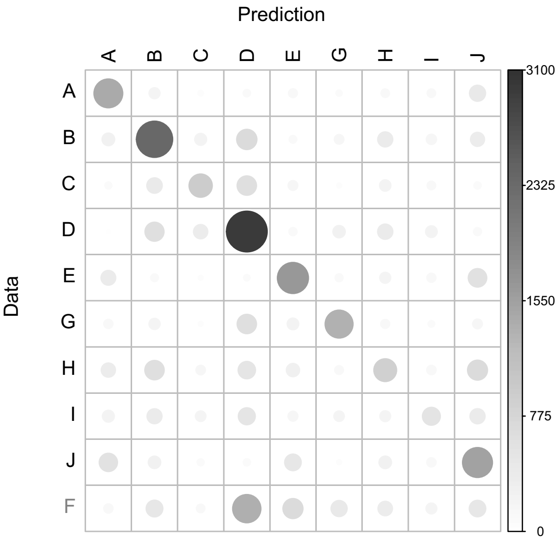

Fig. 4. Confusion matrix for the MCC forensic scenario with F being the unknown suspect. Note that this is only a binary application of classification without considering the actual value of the probability of the class membership (cf. Fig. 3).

that, but also a lack of proper data was reported (Haber and Haber, 2004). However, the larger the database to compare with, the higher the number of similar patterns in dactyloscopy and a FPR of 15.9% was reported (Koehler and Liu, 2021). Compared with an individual FPR of about 5% in our study, quality is in the range of already used forensic methods for identification.

However, it needs to be considered that the method of calculating FPR differed in the mentioned studies including ours so that there is a need for further investigation on the reliability of such identification methods. There is the need to develop best practice guidelines (Ikuesan and Venter, 2017; Champod and Tistarelli, 2017) for model evaluation in forensics including a discussion on relevant measures such as FPRs and FDRs. Likely, the error rates for digital behavior biometrics will never be as reliable as those of DNA evidence (Kloosterman *et al.*, 2014).

5. Conclusion

To the best of our knowledge, we were the first publishing research on forensic driver identification using driving behavior (Dološ *et al.*, 2020) (i) highlighting the importance of robust validation and the choice of a suitable model quality measure, (ii) pointing to differences in the application of multi-class and one-class approaches and (iii) the need to decide for hypothesis testing and data mining in advance. We conclude from this study that the two approaches used, one-class classification and multi-class classification, showed great potential for forensic driver identification. Besides the presentation of a machine learning workflow and its application to the forensic analysis of a hypothetical hit-and-run accident, our main conclusions are the following:

- OCC using unlabeled data could not be applied successfully, yet. We are positive that with further data and research especially on feature extraction, this approach could be sufficiently improved.

- The reliability of MCC was in the range of known methods of forensic identification. However, it is crucial to discuss the importance of model quality measures such as FDR and FPR in an interdisciplinary manner. There still seems to be a lack of best practice guidelines for the application of machine learning in forensics.

- Using machine learning in forensics requires a decision on the purpose, data mining for criminal investigation or hypothesis testing for decision-making in court, prior to its application.

- An anonymized but labeled forensic database for the comparison of digital behavior such as the natural driving behavior and keystroke dynamics would improve the statistical strength of identification of individuals based on such traces.

Identification of persons using biometrics based on digital evidence is an automated way to establish the identity of a person on the basis of his or her digital behavioral characteristics. It is a promising way to gain additional information in casework and also useful for decision making in court.

Acknowledgment

The authors like to thank Byung Il Kwak, Ji Young Woo, and Huy Kang Ki for releasing the driving data for academic research (Kwak *et al.*, 2017). Furthermore, we would like to thank the organizers and members of the EU COST Action CA17124 *Digital Forensics: Evidence Analysis via Intelligent Systems and Practices (DigForASP)* and the project team DriVar for supporting this study.

References

Acien, A., Morales, A., Vera-Rodriguez, R., Fierrez, J. and Monaco, J.V. (2020). TypeNet: Scaling up keystroke biometrics, *International Joint Conference on Biometrics (IJCB), Houston, USA,* pp. 1–7, https://ieeexplore.ieee.org/document/9304908/.

Antal, M. and Szabo, L.Z. (2015). An evaluation of one-class and two-class classification algorithms for keystroke dynamics authentication on mobile devices, *20th International Conference on Control Systems and Computer Science, Bucharest, Romania*, pp. 343–350, http://ieeexplore.ieee.org/document/7168452/.

Benzaoui, A., Hadid, A. and Boukrouche, A. (2014). Ear biometric recognition using local texture descriptors, *Journal of Electronic Imaging* **23**(5): 053008, DOI:10.1117/1.JEI.23.5.053008.

Bergamini, C., Oliveira, L., Koerich, A. and Sabourin, R. (2009). Combining different biometric traits with one-class classification, *Signal Processing* **89**(11): 2117–2127.

Bernardi, M.L., Cimitile, M., Martinelli, F. and Mercaldo, F. (2018). Driver and path detection through time-series classification, *Journal of Advanced Transportation* (3): 1–20.

Bouchrika, I., Goffredo, M., Carter, J. and Nixon, M. (2011). On using gait in forensic biometrics, *Journal of Forensic Sciences* **56**(4): 882–889, DOI: 10.1111/j.1556-4029.2011.01793.x.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Carfora, M.F., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A. and Vaglini, G. (2019). A "pay-how-you-drive" car insurance approach through cluster analysis, *Soft Computing* **23**(9): 2863–2875, DOI: 10.1007/s00500-018-3274-y.

Champod, C. and Tistarelli, M. (Eds) (2017). *Handbook of Biometrics for Forensic Science*, Springer, Cham.

Christensen, A.M., Crowder, C.M., Ousley, S.D. and Houck, M.M. (2014). Error and its meaning in forensic science, *Journal of Forensic Sciences* **59**(1): 123–126, DOI: 10.1111/1556-4029.12275.

Dessimoz, D. and Champod, C. (2008). Linkages between Biometrics and Forensic Science, *in* A.K. Jain *et al.* (Eds), *Handbook of Biometrics*, Springer, New York, pp. 425–459.

D'Lima, N. and Mittal, J. (2015). Password authentication using Keystroke Biometrics, *2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India,* pp. 1–6, http://ieeexplore.ieee.org/document/7045681/.

Dološ, K., Meyer, C., Attenberger, A. and Steinberger, J. (2020). Driver identification using in-vehicle digital data in the forensic context of a hit and run accident, *Forensic Science International: Digital Investigation* **35**: 301090.

Eude, T. and Chang, C. (2018). One-class SVM for biometric authentication by keystroke dynamics for remote evaluation, *Computational Intelligence* **34**(1): 145–160, DOI:/10.1111/coin.12122.

Faigman, D.L. (Ed.) (2002). *Modern Scientific Evidence: The Law and Science of Expert Testimony*, 2nd Edn, West Group, St. Paul.

Ge, Z., Iyer, A.N., Cheluvaraja, S., Sundaram, R. and Ganapathiraju, A. (2017). Neural network based speaker classification and verification systems with enhanced features, *Intelligent Systems Conference (IntelliSys), London, UK*, pp. 1089–1094, http://ieeexplore.ieee.org/document/8324265/.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997). Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery* **1**(1): 11–28.

Gross, S.R., O'Brien, B., Hu, C. and Kennedy, E. H. (2014). Rate of false conviction of criminal defendants who are sentenced to death, *Proceedings of the National Academy of Sciences* **111**(20): 7230–7235, DOI: 10.1073/pnas.1306417111.

Gupta, S., Buriro, A. and Crispo, B. (2019). DriverAuth: A risk-based multi-modal biometric-based driver authentication scheme for ride-sharing platforms, *Computers & Security* **83**: 122–139.

Haber, L. and Haber, R.N. (2004). Error rates for human latent fingerprint examiners, *in* N. Ratha and R. Bolle (Eds), *Automatic Fingerprint Recognition Systems*, Springer, New York, pp. 339–360, DOI: 10.1007/0-387-21685-5_17.

Helm, P. and Hagendorff, T. (2021). Beyond the prediction paradigm: Challenges for machine learning in the struggle against organized crime, *Law & Contemporary Problems* **84**(3): 1–17.

Houck, M.M. and Budowle, B. (2002). Correlation of microscopic and mitochondrial DNA hair comparisons, *Journal of Forensic Sciences* **47**(5): 964–967.

Ikuesan, A.R. and Venter, H.S. (2017). Digital forensic readiness framework based on behavioral-biometrics for user attribution, *IEEE Conference on Application, Information and Network Security (AINS), Miri, Malaysia*, pp. 54–59, `http://ieeexplore.ieee.org/document/8270424/`.

Khan, S.S. and Madden, M.G. (2010). A survey of recent trends in one class classification, *in* L. Coyle and J. Freyne (Eds), *Artificial Intelligence and Cognitive Science*, Springer, Berlin, pp. 188–197, DOI: 10.1007/978-3-642-17080-5_21.

Kloosterman, A., Sjerps, M. and Quak, A. (2014). Error rates in forensic DNA analysis: Definition, numbers, impact and communication, *Forensic Science International: Genetics* **12**: 77–85.

Koehler, J.J. and Liu, S. (2021). Fingerprint error rate on close non-matches, *Journal of Forensic Sciences* **66**(1): 129–134, DOI: 10.1111/1556-4029.14580.

Koenig, B.E. (1986). Spectrographic voice identification: A forensic survey, *Journal of the Acoustical Society of America* **79**(6): 2088–2090, DOI: 10.1121/1.393170.

Kupin, A., Moeller, B., Jiang, Y., Banerjee, N.K. and Banerjee, S. (2019). Task-driven biometric authentication of users in virtual reality (VR) environments, *in* I. Kompatsiaris *et al.* (Eds), *MultiMedia Modeling*, Springer, Cham, pp. 55–67, DOI: 10.1007/978-3-030-05710-7_5.

Kwak, B.I., Woo, J. and Kim, H.K. (2017). Know your master: Driver profiling-based anti-theft method, *arXiv* 1704.05223, `http://arxiv.org/abs/1704.05223`.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest, *R News* **2**(3): 18–22.

Lieberman, J.D., Carrell, C.A., Miethe, T.D. and Krauss, D.A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence?, *Psychology, Public Policy, and Law* **14**(1): 27–62, DOI: 10.1037/1076-8971.14.1.27.

Mack, B., Roscher, R. and Waske, B. (2014). Can I trust my one-class classification?, *Remote Sensing* **6**(9): 8779–8802.

Mack, B. and Waske, B. (2017). In-depth comparisons of MaxEnt, biased SVM and one-class SVM for one-class classification of remote sensing data, *Remote Sensing Letters* **8**(3): 290–299, DOI: 10.1080/2150704X.2016.1265689.

Martinelli, F., Mercaldo, F., Orlando, A., Nardone, V., Santone, A. and Sangaiah, A.K. (2020). Human behavior characterization for driving style recognition in vehicle system, *Computers & Electrical Engineering* **83**, Article 102504, DOI: 10.1016/j.compeleceng.2017.12.050.

Mashao, D.J. and Skosan, M. (2006). Combining classifier decisions for robust speaker identification, *Pattern Recognition* **39**(1): 147–155.

Mordini, E. (2017). Ethics and policy of forensic biometrics, *in* M. Tistarelli and C. Champod (Eds), *Handbook of Biometrics for Forensic Science*, Springer, Cham, pp. 353–365, DOI: 10.1007/978-3-319-50673-9_16.

Märkel, U. and Dološ, K. (2017). Tree species site suitability as a combination of occurrence probability and growth and derivation of priority regions for climate change adaptation, *Forests* **8**(6): 181.

Müller, C. (Ed.) (2007). *Speaker Classification I: Fundamentals, Features, and Methods*, Lecture Notes in Computer Science, Vol. 4343, Springer, Berlin, DOI: 10.1007/978-3-540-74200-5.

Oshiro, T.M., Perez, P.S. and Baranauskas, J.A. (2012). How many trees in a random forest?, *in* D. Hutchison *et al.* (Eds), *Machine Learning and Data Mining in Pattern Recognition*, Springer, Berlin, pp. 154–168, DOI: 10.1007/978-3-642-31537-4_13.

Page, M., Taylor, J. and Blenkin, M. (2011). Forensic identification science evidence since Daubert: Part II-Judicial reasoning in decisions to exclude forensic identification evidence on grounds of reliability: Identification evidence since Daubert (II), *Journal of Forensic Sciences* **56**(4): 913–917, DOI: 10.1111/j.1556-4029.2011.01776.x.

Remeli, M., Lestyan, S., Acs, G. and Biczok, G. (2019). Automatic driver identification from in-vehicle network logs, *arXiv* 1911.09508, `http://arxiv.org/abs/1911.09508`.

Reynolds, D. (1994). Experimental evaluation of features for robust speaker identification, *IEEE Transactions on Speech and Audio Processing* **2**(4): 639–643.

Ross, A., Banerjee, S. and Chowdhury, A. (2020). Security in smart cities: A brief review of digital forensic schemes for biometric data, *Pattern Recognition Letters* **138**: 346–354, DOI: 10.1016/j.patrec.2020.07.009.

Saks, M.J. and Koehler, J.J. (2005). The coming paradigm shift in forensic identification science, *Science* **309**(5736): 892–895, DOI: 10.1126/science.1111565.

Stenzel, S., Fassnacht, F.E., Mack, B. and Schmidtlein, S. (2017). Identification of high nature value grassland with remote sensing and minimal field data, *Ecological Indicators* **74**: 28–38.

Thompson, W. (2006). Tarnish on the “gold standard”: Recent problems in forensic DNA testing, *The Champion* **30**: 10–16.

Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S. and Wang, R. (2017). Speaker identification features extraction methods: A systematic review, *Expert Systems with Applications* **90**: 250–271.

Turunen, E. and Dološ, K. (2021). Revealing driver’s natural behavior—A GUHA data mining approach, *Mathematics* **9**(15): 1818.

Wakita, T., Ozawa, K., Miyajima, C., Igarashi, K., Itou, K., Takeda, K. and Itakura, F. (2005). Driver identification using driving behavior signals, *IEEE Conference on Intelligent Transportation Systems*, *Vienna, Austria*, pp. 907–912, http://ieeexplore.ieee.org/document/1520171/.

White, D., Dunn, J.D., Schmid, A.C. and Kemp, R.I. (2015). Error rates in users of automatic face recognition software, *PLOS ONE* **10**(10): e0139827, DOI: 10.1371/journal.pone.0139827.

Klara Dološ is a researcher at the Central Office for Information Technology in the Security Sector (ZITiS) in Munich, Germany. She received her MSc in geoecology at the Karlsruhe Institute of Technology (KIT) in 2009 and her PhD in ecological climate impact modeling at the University of Bayreuth in 2013. Her research interests include digital forensics and applications of statistics and machine learning in related fields.

Conrad Meyer received his MSc in electrical engineering from the University Erlangen, Germany, in 2011. Now he is working for the Digital Forensics Department at the Central Office for Information Technology in the Security Sector (ZITiS) in Munich, Germany. His research includes the reconstruction of events in the transportation sector.

Andreas Attenberger heads research activities in digital forensics at the Central Office for Information Technology in the Security Sector (ZITiS) in Munich, Germany. After finishing his diploma studies at LMU Munich in 2009, he was a research assistant with Bundeswehr University Munich (UniBW), Neubiberg, Germany, and participated in numerous projects with both military and corporate partners. He finished his PhD at UniBw in 2016 and took on a professorship with the University of Applied Sciences in Kufstein, Germany. His research focuses on the impact of increasing connectivity (connected cars, mobile devices, IoT solutions, etc.) on forensic investigations.

Jessica Steinberger is a researcher with the Digital Forensics Department of the Central Office for Information Technology in the Security Sector (ZITiS) in Munich, Germany. She received her BSc and MSc degrees in computer science from the University of Applied Sciences Bingen, Germany, and her PhD degree from the University of Twente, the Netherlands, in 2018. Her main topics of interest include network forensics in general and the combination of CEP with network forensics in particular.

Received: 9 April 2021
Revised: 25 August 2021
Re-revised: 10 September 2021
Accepted: 19 October 2021