

Forensic Activity Classification Using Digital Traces from iPhones: A Machine Learning-based Approach

Conor McCarthy*, Jan Peter van Zandwijk†‡, Marcel Worring*, Zeno Geradts*†

*University of Amsterdam, Amsterdam, The Netherlands

Email: {c.t.mccarthy, m.worring}@uva.nl

†Netherlands Forensic Institute (NFI), The Hague, The Netherlands

Email: {j.p.van.zandwijk, z.geradts}@nfi.nl

‡Amsterdam University of Applied Sciences, Faculty of Technology, Amsterdam, The Netherlands

Abstract—

Smartphones and smartwatches are ever-present in daily life, and provide a rich source of information on their users' behaviour. In particular, digital traces derived from the phone's embedded movement sensors present an opportunity for a forensic investigator to gain insight into a person's physical activities. In this work, we present a machine learning-based approach to translate digital traces into likelihood ratios (LRs) for different types of physical activities. Evaluating on a new dataset, NFI_FARED, which contains digital traces from four different types of iPhones labelled with 19 activities, it was found that our approach could produce useful LR systems to distinguish 167 out of a possible 171 activity pairings. The same approach was extended to analyse likelihoods for multiple activities (or groups of activities) simultaneously and create activity timelines to aid in both the early and latter stages of forensic investigations. The dataset¹ and all code² required to replicate the results have also been made public to encourage further research on this topic.

The potential of such digital traces from iPhones to inform forensic investigators about the user's physical activities has been studied, specifically using information on registrations of steps, distances and floors [32], [35], [33], to estimate the user's activities in a certain window of time. While such registrations can be easily interpreted, many other available registrations are less self-explanatory but nonetheless contain additional information that could help to identify physical activities. Machine learning techniques are well suited to this case, as these can take advantage of any patterns present in all available variables in the digital traces to classify activities, and consequently could increase the number of discernible activities.

For use in forensic investigations, and especially if being used as evidence in court, it is beneficial if the evidential strength of activity classifications made by machine learning algorithms can be quantified. An appropriate approach to achieve this is through the construction of a Likelihood Ratio (LR) system. The benefit of producing an LR in place of a typical classification is that it can be placed in context alongside other evidence to create a more complete picture of events. An LR also is a more transparent measure of evidential strength which aids reliability from a legal perspective. Therefore, our approach produces outputs in the form of LRs.

In this work, we present a machine learning approach for the interpretation of digital traces related to physical activities to estimate the LR of certain activities being performed in a specific time interval. The method uses timestamped digital traces extracted from smart devices and outputs an LR. The model can output LRs for two specific activities (the binary case), stating how much more likely the data was generated by the user performing activity A rather than activity B. For example, “The traces are fifty times more likely if the user was driving a car than sitting at home”. This form of LR is typical when presenting findings as evidence. Output can also be in the multiclass case, comparing the likelihood of multiple possible activities at once, loosening the necessary assumptions required for producing an LR of only two classes. This also enables the construction of a timeline of most likely activities at each moment, which can be informative to investigators.

INTRODUCTION

Smartphones and smartwatches have become a central component of modern life, integrated into almost all parts of daily routines. They are essential not only for communication, but also internet access, navigation, payments, and translation. Smart devices have consequently become indispensable, and people typically carry at least one on their person at all times. From a digital forensic perspective, the strong integration of digital devices into everyday life offers great investigative opportunities. Given the wide range of capabilities of modern devices, they contain a diverse set of data and could provide unique information that is not obtainable from other types of forensic evidence. We concentrate on the embedded sensors like accelerometers and gyroscopes, the raw data of which is processed into detailed information about physical activities, and stored on the device in the form of digital traces [34]. Significant research exists on the availability and extraction of digital traces from iPhones, whose traces will be the focus of this work.

¹<https://huggingface.co/NetherlandsForensicInstitute/datasets>

²https://github.com/Con-or-McCarthy/Data2Activity_1

RELATED WORK

LRs are a useful measure of evidential strength. They present the degree of support for one hypothesis versus another and provide a logically correct method to assist the investigator in assessing their uncertainty [1], [29], [11], [12], [24]. Usage of LR in forensic applications is well developed, with established methods for evaluation [7], [6], [3] and improving stability [36], and has been encouraged by institutions such as ENFSI as a suitable way to report evidence [2]. As such, forensic likelihood ratios have been utilised in domains such as speaker analysis [38], [25], face recognition [17], DNA [9], drug comparison [4], and glass analysis [30], among others.

Previous work on producing evidence from digital traces focuses on a single variable such as distance [37] or location [28], and assessing the LR of variable values being generated from different ground truth scenarios. To the best of the authors' knowledge, there is no existing work using digital traces to estimate the likelihood of quantities outside of those reported directly by the variables.

Using body-worn sensor data to classify activities is a large field of research in the machine learning community [27], [18], [13], [20], however the data is almost exclusively raw data in the form of high-frequency time series, and techniques do not easily translate to digital trace data.

COLLECTION OF DIGITAL TRACE DATA

Participants: Fourteen test subjects (6 females, 8 males, mean age 26.6y, standard deviation 8.8y) participated in the data collection experiments for this study. Each participant signed an informed consent form, agreeing to the collection and spreading of their data anonymously. For each subject, data was collected during multiple experimental sessions spread over several days. Due to scheduling conflicts, three participants could not participate in all sessions, and three additional subjects were recruited to complete those sessions in their place. This effectively provides a complete experimental dataset for eleven participants.

Smartphones: In the experiments, four different iPhones were carried simultaneously by the test subjects. The models and iOS versions of these iPhones are: 6+ (11.4.1), 7 (14.7.1), 11 (13.1.1), and XR (15.4.1).

Experimental protocol: In total each subject performed 19 different activities. These activities were carried out in four separate measurement sessions. Session 1 included movement activities, session 2 contained elevation change activities, session 3 forensically relevant dynamic activities, and session 4 consisted of transport activities. The order in which the different activities were performed within each session was randomised for each subject.

During sessions, subjects wore the iPhones at the following locations: front trouser pocket, back trouser pocket, breast pocket, backpack, and the hand. Phone placement was randomised across subjects, but was for each subject kept the same during all four experimental sessions.

Sessions 1, 2 and 3 were performed under controlled conditions. Between subsequent activities, the subject would

pause for one minute by either sitting or standing. To encourage diversity in their execution, while carrying out the activities, the subjects were given as few instructions as possible. At the end of each session, the subject completed a "free-living" section, in which they would perform the activities from that session in any order and for however long they wished without breaks. For each of the activities carried out by the subject, the start and end time of that activity was recorded by the experimenter. Details of each of the sessions is described in more detail below.

Session 1: Movement activities: The movements walking, running and cycling were performed multiple times for known distances. For running and walking there are two distances: ~ 240m and ~ 90m, each performed twice. For cycling, the subjects cycled a set number of loops around the car park of the experimental location three times. The number of loops each were varied between two, three, and four. During pauses between trials, the subject sat on a bench without backrest.

Session 2: Elevation Changes: The movements stair, escalator and elevator were performed four times each going upstairs and downstairs. The stair climbing and escalator movements were performed at different velocities. For stair climbing the subject was instructed to walk or to run on the stairs. For the escalator, the subject was instructed to stand or walk on the escalator. After each movement the participant would either stand or sit for one minute at the end of the stairs/escalator. During pauses between trial, the subject sat on a bench with backrest.

Session 3: Forensically relevant dynamic movements : This session includes the activities kicking, throwing, punching and dragging a heavy object. These activities were performed for time intervals of 10, 20 and 30 seconds. Kicking and punching were performed on a punching bag and was alternated between the dominant and non-dominant arm or leg. For the throwing task, the subject threw a ball weighing 1 kg using their dominant arm down a hallway to the experimenter, who rolled it back to the subject to pick it up and throw again until the task was complete. For the dragging activity the subject dragged the punching bag weighing 34 kg through a hallway.

Session 4: Transportation : In this session, the transport modes riding the train, tram, and in a car were carried out by the subjects while travelling home from the experimental location. If the subject had not used one or more of the modes of transport in their trip home, arrangements were made to carry out the missing modes in an additional session. In session 4, the participants recorded the start and end times of each travel activity by themselves.

Data acquisition and processing: After an experimental session, a full-filesystem extraction of the data of each of the four iPhone was produced using commercial forensic equipment present at Netherlands Forensic Institute. From these extractions, the files `healthdb_secure.sqlite` and `cache_encryptedC.db` were exported. These files are known to contain information related to physical activities performed by the phone's user [32], [33], [35].

After export, post-processing of the data contained in the exported files was performed. In this post-processing, data from various Tables from the databases were combined into a single dataset. For the file `cache_encrypte dC.db`, variables were extracted from the tables `StepCountHistory`, `MotionStateHistory` and `NatalieHistory` [33]. For the file `healthdb_secure.sqlite`, variables were extracted from tables `samples` and `quantity_samples` [32].

Extracted variables could either be categorical or numerical, and numerical variables could be further separated between cumulative and non-cumulative, where cumulative variables have an always increasing count, and non-cumulative do not. An example of a categorical variable is `type` from the table `MotionStateHistory` in `cache_encryptedC.db`, which can take one of six different values. An example of a cumulative numerical variable is `f loorsAscended` from the table `StepCountHistory` in `c ac he_encryptedC.db`, whose value either remains the same or increments. The variable `mets` from the table `Natali eHistory` in `cache_encryptedC.db`, is an example of a non-cumulative variable, which takes on a new, continuous value at each reading.

After post-processing, data was further aggregated into one minute intervals. In the case there were multiple registrations of a variable being within the same one minute interval, we used the modal variable for categorical variables, the sum of all values for cumulative variables and the mean for non-cumulative variables. In the case of a variable having no registration in a one minute interval, a missing value (NA) was imputed. NA values are used instead of zero values because a registered zero value has a different interpretation to no registration whatsoever. Ground truth values of activities executed were added to the data before aggregation. During aggregation, if two or more unique activity labels were present in a single minute of data, the minute would be split, with the corresponding variable readings aggregated to the appropriate label. This means that a one minute timestamp could appear more than once in the final dataset, each one containing data for less than a minute of activity. This approach was chosen over simply labelling the minute with the majority class to maximise data availability, particularly for activities that spanned a shorter duration, such as the *dynamic* activities.

Henceforth, we will refer to this dataset as `NFI_FARED`. Our dataset is fully publicly available for download and scripts for processing are available on the project's GitHub.

METHOD

From the data described in Collection of Digital Trace Data above, a machine learning approach consisting of a scorer and calibrator is created which can return an LR for any inputted sample of data (e.g. one minute from `NFI_FARED`). The choice of scorer, as well as the procedure followed to create and evaluate the LR system are detailed in the following sections. It should be noted that while results are based on the `NFI_FARED` dataset,

the described approach can be applied to other datasets containing timestamped traces and also extended in the event of additional traces becoming available to forensic practitioners.

Scorer Description

A score-based LR system works by assigning a single number to each observation to distinguish between H_1 -true and H_2 -true samples, where a higher score is more likely to come from a H_1 -true sample, and a lower score from a H_2 -true sample, in the binary case. In the multiclass case, K numbers are outputted, one for each of the K distinct hypotheses. Various options are available to achieve this, both in the form of fixed score functions and statistical models. Machine learning approaches learn parameters from the training data rather than relying on expert knowledge to find a good fit to a probability distribution. This increases flexibility in the features which can be used as input, since interpretability of inputs is not a requirement, unlike when crafting specialised scoring rules. Given a sufficiently representative training set, an adequately expressive model, and effective optimization, a machine learning model can learn parameters that approximate the optimal solution with respect to the underlying data distribution.

The digital trace data also exhibits certain characteristics that require consideration when selecting an appropriate model, namely the presence of many missing values, and features being a mix of categorical and numerical. The large amount of missing values arise as a consequence of the iPhone having irregular and inconsistent sampling rates between variables. Certain variables, such as `floorsAscended`, log a value only when a floor change is detected, resulting in a missing value in all samples without a floor change detected. Data imputation or removal is a standard method of handling missing values, however the missing values can still be informative to the investigator. For example, the lack of a `floorsAscended` registration could increase the likelihood of activity *driving*, while decreasing the likelihood of *upstairs*. Therefore, a model that can natively handle missing values is needed to leverage the maximum information. Similarly, the data containing both numerical and categorical features necessitates a model which can process both data types, to avoid potential bias from data transformations.

To satisfy these conditions, the tree-based machine learning model CatBoost [22] was selected. CatBoost uses gradient boosting, in which many decision trees are constructed iteratively and incorporated into the ensemble model. Each tree is fit on the residuals of the model in the last step to improve results. CatBoost natively deals with categorical variables without the need for preprocessing by using ordered encoding, in which target statistics from all the rows prior to a data point are considered to calculate a value to replace the categorical feature. CatBoost is faster than other gradient boosting methods such as XGBoost [8] and gradient boosting approaches are state-of-the-art

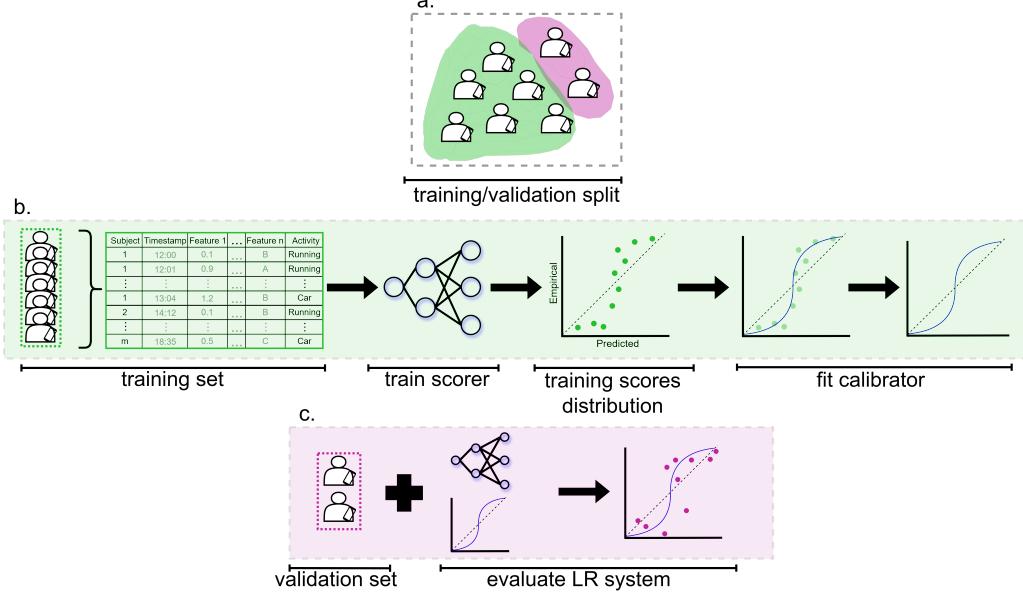


Fig. 1: Overview of method employed to train and evaluate an LR system using the proposed approach. a) Subjects are split between training (green) and validation (purple) to prevent data leakage. b) Data from training subjects is collected into the training set, containing the activity classes of interest, which is used to train the scorer (CatBoost[22]). After training, the scores for the entire training set are calculated to produce the training scores distribution, on which we fit the calibrator. The combination of scorer + calibrator is the LR system c) Validation data is taken from the validation subjects and fed through the LR system from b. The resulting likelihood ratios can then be evaluated using C_{llr} , PAV plots, Tippet plots, etc.

on tabular data, outperforming much more complex deep learning algorithms [5], [19].

Model Training

As illustrated in Figure 1a, the dataset is first split into a training and validation sets, with each subject assigned entirely to one set. This prevents data leakage resulting from subjects being shared between the training and validation sets. From the training subjects, all samples of the data relating to the relevant activities are placed into the training set. In Figure 1b, this is shown for the case of distinguishing activity *car* from activity *running*. The CatBoost model is then trained using the training set, producing a trained model which returns a single number for a given input sample. The multiclass case operates in the same fashion, but allows more than two classes to be included in the training set.

Likelihood Ratio Generation

After training, the entire training set is inputted to the model, producing a distribution of N scores (Figure 1b). To be used in crafting a reliable LR system, these scores must then be calibrated so they more closely align with empirical probabilities. Machine learning outputs can suffer from both under- and over-confidence, especially when calibration is not included in the training process. A range of calibrators are available for this purpose.

We employ logistic calibration for calibrating the scores, which works by defining a sigmoidal logistic curve:

$$c(s; w, m) = \frac{1}{1 + \exp(-w(s - m))}, \quad (1)$$

where s is the uncalibrated score produced by the model and c gives the logically calibrated score. m and w are parameters which are estimated from the training data to provide the line of best fit. The resulting logistic curve c is thus a calibration map that transforms uncalibrated scores s into calibrated scores $c(s)$. The process is visualised in Figure 1c, where the blue calibration curve c maps the green training scores onto the diagonal. Alternative calibration methods function similarly to produce more reliable probability estimates, and the best choice of calibrator depends on the nature of your dataset.

The calibrated score $c(s)$ is also the posterior probability p of H_1 being true for the sample. This is converted to posterior odds using the formula $p/(1 - p)$. The prior odds are calculated as the ratio of the number of H_1 -true samples and H_2 -samples in the training set. Bayes rules is then used to produce the LR as posterior odds / prior odds [15]. The values of the LRs are bounded using ELUB bounds to reduce sensitivity to extrapolation errors. ELUB bounds prevent LRs from becoming too extreme for edge values due to a lack of data points in these regions; further explanation can be found in [36]. In the multiclass case with more than two hypotheses, we do not convert scores to LRs and instead utilise the scores directly as likelihoods for our analysis. There are additional considerations to

be made when creating an LR for a multiclass system which can affect the precise interpretation of results, and discussion on this topic is provided in [26]. However, these considerations are not a focus of this work, and are left up to the practitioner.

Likelihood Ratio Evaluation

For an LR system to be useful, it must be effective in two dimensions: discrimination and calibration. Discrimination is how well the system distinguishes different classes. Discrimination is typically the primary goal for a classifier, and is measured with metrics such as accuracy. Such metrics are only concerned with the top predicted class from the classifier, and do not take level of confidence into account i.e. 51% and 100% probabilities are equally correct. For an LR system, level of certainty is a core part of an LR's utility, hence calibration is also assessed. Calibration measures how closely the predicted LRs reflect the observed frequencies of H_1 -true and H_2 -true samples from the validation set, i.e. does an LR of 20 reflect a H_1 -true sample occurring 20 times more than a H_2 -true sample for a given point in the validation set?

Best practice for evaluating LR systems is to assess discrimination and calibration in various ways [3], to verify that outputs are sensible and reasonable. Some examples of such methods are:

- Pool Adjacent Violators (PAV) plot: PAV transforms the scores to optimal (in terms of C_{llr}) LRs using isotonic regression. The PAV plot compares the outputted LRs to these “optimal” ones, with points ideally following the line $y=x$ [6], [37].
- Tippett plot: A Tippett plot shows the inverse cumulative densities of the log-likelihood ratio values given the two hypotheses H_1 and H_2 . Ideally the LRs given H_1 would have all high (above 0) log-likelihood ratios, and the LRs given H_2 low (below 0) log-likelihood ratios [23].
- Empirical Cross Entropy (ECE) plot: The ECE value is a single value for a set of LRs at a particular value of the prior odds, $P(H_1)/P(H_2)$. The ECE plot shows the ECE value as a function of these prior odds. The value is based on a penalty function that penalises misleading (supports wrong hypothesis) LRs. The ECE should always be lower than a completely non-informative system (outputs LR=1 for every sample). In the ECE plots included in this paper, the non-informative system is depicted as a dotted line [15].

Numerically, the Log Likelihood Ratio Cost (C_{llr}) [7] captures many of the desirable components of the LR system in a single number. It is based on a strictly proper scoring rule, meaning the cost is minimised if and only if the forecast p equals the true distributions of outcomes q . This encourages honest reporting of probabilities and cannot be “gamed” with inaccurate likelihoods in the same way as metrics such as accuracy. The C_{llr} is also application independent, since it is integrated over the

priors and costs associated with different applications. It is computed as:

$$C_{llr} = \frac{1}{2\|S_1\|} \sum_{t \in S_1} \log_2(1 + \exp(-llr_t)) \\ + \frac{1}{2\|S_2\|} \sum_{t \in S_2} \log_2(1 + \exp(llr_t)) \quad (2)$$

Where S_1 is the set of H_1 -true samples, S_2 is the set of H_2 -true samples, and llr_t is the log-likelihood ratio under evaluation for trial t (log-likelihoods are used in place of likelihoods for practical reasons). Interpretation of the value of the C_{llr} is then as follows:

$$C_{llr} \begin{cases} = 0 : & \text{perfect} \\ \in (0, 1) : & \text{useful} \\ = 1 : & \text{not informative} \\ > 1 : & \text{not useful} \end{cases} \quad (3)$$

Thus, C_{llr} values below 1 are informative, down to a minimum of zero, which occurs for a “perfect” LR system outputting an LR of ∞ for all H_1 -true samples, and $-\infty$ for all H_2 -true samples. C_{llr} values of 1 or above are equivalent or worse than a completely non-informative system. It should be noted that beyond the above categorisation of C_{llr} values, precise interpretation is difficult and must be approached carefully [31].

C_{llr} captures loss from both the discrimination and calibration of the LR system. It is possible to separate the two components for a more detailed analysis. By recalculating C_{llr} on a set of perfectly calibrated likelihood ratios an estimate can be made of the discriminative loss C_{llr}^{min} of the system. The calibration loss is then the difference between the two: $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$.

C_{llr} can be generalised to the multiclass setting, termed Multiclass Cross-Entropy Cost (C_{mxe})[6]. The formula for C_{mxe} is:

$$C_{mxe} = \frac{1}{K} \sum_{k=1}^K \frac{1}{\|S_k\|} \sum_{t \in S_k} \log_2 \frac{\sum_{j=1}^K \exp(ll_{jt})}{\exp(ll_{kt})}. \quad (4)$$

Where K is the number of classes, S_k is the set of samples of class k , and ll_{kt} is the log-likelihood under evaluation for class k , given sample t ($ll_{kt} = \log_{10} P(\text{sample } t | \text{class } k)$). C_{mxe} is equivalent to empirical cross-entropy, and degenerates to C_{llr} when $K = 2$. Interpretation is similar to C_{llr} , aside from the reference value becoming $\log_2 K$. For convenience, we normalise $\hat{C}_{mxe} = C_{mxe}/\log_2 K$ so interpretation follows that of C_{llr} (eqn. 3).

RESULTS

The above described model is first evaluated in the binary setting, where the hypotheses are constructed in a one-versus-another setting. This application evaluates how well digital traces can be used in the LR system to generate LRs relating to say, the prosecution’s version of events versus the defense’s in a court setting, where clear cut hypotheses have already been established. The multiclass setting is evaluated next. This setting is more useful

in the investigatory phase, where less is known about the sequence of events, and it is convenient to explore multiple possible events simultaneously. For this purpose, the creation of an activity timeline from digital traces is also demonstrated, to illustrate potential applications of the model.

Unless otherwise stated, subject-wise cross-validation was carried out to generate all results. From each subject, iPhone model, and carry location, multiple samples are included in the training set, with a potentially unbalanced proportion of activities. To counteract the imbalance of classes, they are weighted by their inverse frequency. The overlap in iPhone and carry location between training and validation sets also violates some of the assumptions of binomial-distribution-based analysis (according to Kruskal-Wallis tests). We therefore report results based on multilevel bootstrapping [14], [10], [21]. For a single bootstrap sample, carry locations and iPhone types are sampled with replacement. If there are n unique carry locations, n samples with replacement are drawn, and similarly for iPhones. All datapoints which match the iPhone -carry location combination are then used as the validation set (including resamples), and performance metrics are calculated. This is repeated 1,000 times and the mean metrics across all bootstrap samples are reported.

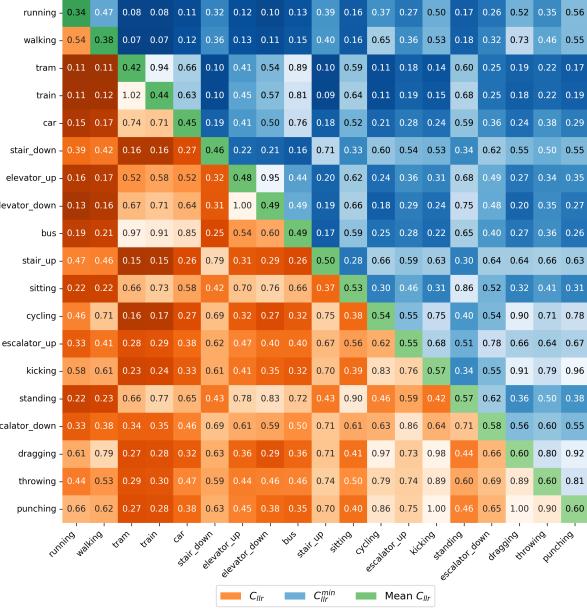


Fig. 2: C_{llr} and C_{llr}^{min} for LR systems produced from each combination of two activity classes. Darker colours indicate lower (better) values. Each LR system is produced with H_1 =row activity, H_2 =column activity. Diagonal values in green are the mean C_{llr} for an activity across all LR systems it is included.

Activity Pairs

NFI_FARED contains traces relating to 19 distinct activities. One of the most interesting points for a forensic investigator is to know which of these activities can be

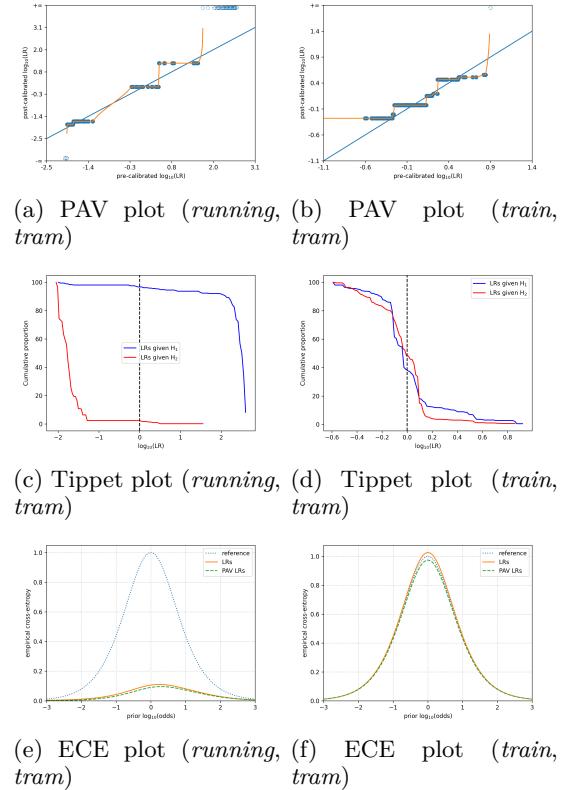


Fig. 3: Further analysis of the LR systems with the lowest (left column: (*running*, *tram*)) and highest (right column: (*train*, *tram*)) C_{llr} values.

distinguished from one another using the described approach, how well they can be distinguished, and equally as important, which ones cannot be distinguished. To assess this, an LR system was created for each possible pair of hypotheses involving a single activity. The C_{llr} of all systems was then computed, and results are presented in Figure 2. The figure consists of three sections: the lower triangle, in orange, shows the C_{llr} of the system where H_1 =activity on the row, and H_2 =activity on the column (H_1 and H_2 are interchangeable when calculating C_{llr} ; eqn. 2). The diagonal, coloured in green, shows the mean C_{llr} of an activity across all eighteen LR systems it was tested in, and the upper triangle, coloured in blue, shows the C_{llr}^{min} for each LR system.

For interpreting Figure 2, we will focus on the C_{llr} values. Of all 171 possible activity pairings, 167 of the associated LR systems have a C_{llr} below one, which means that these systems are at least more informative than an uninformative system, and could be of some value to an investigator. On average, *running* and *walking* have the lowest C_{llr} . We speculate that this is due to richer data availability for these activities, both from variables which measure steps directly, and others which only register when steps are detected e.g. **floorsAscended** [35].

Highest (worst) C_{llr} values occur for dynamic classes *punching*, *throwing*, and *dragging*. This brings into question how effectively digital traces can be applied to dis-

Scorer	Calibrator	Accuracy \uparrow	Mean $Ctr \downarrow$	% $Ctr \uparrow$ below \downarrow			
				1.00	0.75	0.50	0.25
CatBoost	LogReg	85.0	0.504	97.7	84.8	52.6	14.6
	KDE	85.0	0.511	97.1	82.5	52.6	14.0
	Gauss	85.0	0.546	97.1	80.1	47.4	10.5
Random Forest	LogReg	86.6	0.609	85.4	66.1	47.4	18.1
	KDE	86.7	0.617	84.2	71.3	53.2	18.1
	Gauss	86.7	0.612	91.8	74.9	40.9	10.5
XGBoost	LogReg	86.2	0.760	68.4	55.6	39.2	14.6
	KDE	86.2	0.871	64.3	52.6	31.6	9.9
	Gauss	86.2	0.787	67.3	55.6	34.5	8.2
Decision Tree	LogReg	83.3	1.305	42.7	24.6	8.2	0.6
	KDE	83.3	1.303	41.5	25.1	8.2	0.6
	Gauss	83.3	1.261	45.0	28.1	9.9	0.6

TABLE I: Ablation study results on all pairs of activities. LogReg=Logistic Regression, KDE=Kernel Density Estimation, Gauss=Gaussian calibrator.

tinguish these less typical activities. The erratic nature of the activities compared to others is likely a major source of difficulty for the systems. Although, even the worst mean C_{llr} for any class (*punching*) is still well within the useful range at 0.6.

The pairings with $C_{llr} > 1$ are (*train, tram*), (*elevator down, elevator up*), (*punching, kicking*), and (*punching, dragging*). The latter two pairs are pairs from the difficult to deal with dynamic category of classes (Table II), while the former two pairs of activities have particular trouble only when paired together. This is because of the similar nature of the activities in the pairs adding confusion, discussed further in Multiclass LR systems.

LR systems should be evaluated using a number of characteristics as well as C_{llr} . For practical reasons, extensive evaluation was not possible for all LR systems, but in Figure 3, PAV, Tippet, and ECE plots [15] are presented for the best (*running, tram*) and worst (*train, tram*) LR systems according to C_{llr} . The left column of the figure confirms that the (*running, tram*) system has strong discriminative performance (Figure 3c) and is well calibrated (Figures 3a and 3e). (*train, tram*) on the other hand, shows poor discrimination (Figure 3d), but actually has close to optimal calibration (Figures 3b and 3f). This also agrees with the values in Figure 2; $C_{llr} = 1.02$ is primarily from the discrimination loss $C_{llr}^{min} = 0.94$, with a calibration loss of only $C_{llr}^{cal} = 0.08$.

Figure 2 can be used as a reference in the early stages of analysis, to help assess whether digital traces could be helpful in a specific case.

Ablation

The described LR system consists of two components: a scorer and a calibrator. For each component, multiple options are available. In Table I, we compare the performance of twelve methods of producing LR systems by creating an LR system for each possible activity pairing and calculating the accuracy and C_{llr} . The mean values are presented, as well as the proportion of activity pairings with C_{llr} below the thresholds 1.00, 0.75, 0.50, and 0.25.

The combination of a CatBoost scorer and Logistic Regression calibrator performs the best. The mean C_{lir} of all systems is 0.504, and 97.7% of all systems are more informative than the reference, with over half of all systems having a C_{lir} below 0.5. The only category in which this

combination is not best overall is in percentage of C_{lrs} below 0.25, where Random Forest combined with Logistic Regression or Kernel Density Estimation is better (14.6% versus 18.1%). While this does indicate that Random Forest may have a higher ceiling for quality of LR systems, in the majority of cases CatBoost + Logistic Regression performs best.

Accuracy measures the percentage of correct classifications made by the scorer, which is the typical goal for a machine learning classifier. Interestingly, both Random Forest and XGBoost are better classifiers than CatBoost, but have worse C_{hr} . This demonstrates that when choosing a scorer, it is important to not just select the best classifier, since its performance in an LR system may be inferior to a weaker classifier with better calibration.

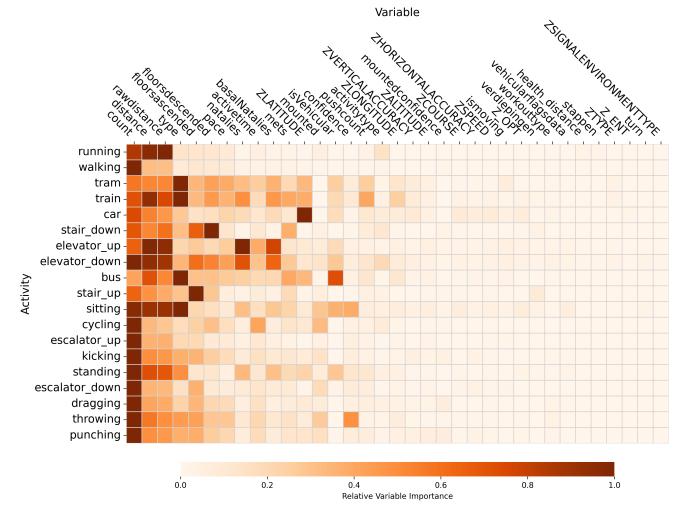


Fig. 4: Heatmap displaying average variable importances for each activity class.

Variable Importance

An advantage of using Catboost is that it allows us to inspect the importance of each input variable when classifying, and how they vary between classes. Importance was calculated using `PredictionValuesChange` importance [22], which measures how predictions shift when variables are perturbed. The importance values for all eighteen pairings involving a particular class were aggregated and normalised, and the results are displayed in Figure 4. In this figure, we can inspect which variables contribute most to the classification of each activity class. Variables are arranged in descending order of average importance, and importance values are normalised row-wise. This analysis was also repeated weighting importance by the inverse C_{llr} of the pairing, to reduce contribution of non-informative classifiers, however this did not have any material impact on results.

Variable **count** is one of the top most important variables for most of the activities in NFI_FARED, which has been identified in previous work as the cumulative number of steps [33]. Alongside variables **distance** and **r**

awdistance, these variables have been previously studied for forensic applications [33]. **count** in particular is quite informative for activity classification, and merits further targeted research. Many variables contribute little to classifying any of the activities. 17 out of the 35 variables average less than 3% of the importance of the largest recorded value, most of which can be attributed to noise. These variables can be removed to reduce model size without a significant loss of prediction quality.

Distributions of variable importance is interesting to observe, with marked difference between activities. For example, *running* has variables **count**, **distance**, and **rawdistance** as its most important, none of which are the most important for *bus*, whose top variable is **type**, which has previously been identified as being related to travelling in a vehicle [33]. Such interactions between variables and activities can be instructive for future research further connecting these variables with real-world quantities.

Sensitivity Analysis

Above results are calculated using training and test sets without any overlap in subjects, to better assess generalisability to new subjects. However, there are two additional factors which can also affect generalisability: phone type and carry location. To investigate how results change when analysis is performed on a new phone/carry location, the evaluation procedure was repeated, with the alteration that after the training/validation set had been constructed for one subject fold of an activity pair, all samples from one of the factors (e.g. carry location: hand), were removed from the training set, and all samples **not** from that factor were removed from the test set. This produced a training and test set with no overlap in both subjects and that factor. This was repeated for every factor on each subject fold-activity pair combination, with the same LR system generation and evaluation being performed. The change in performance compared to training/test sets only keeping subjects distinct was then scrutinised. To remove the effect coming from reduced sample size, the procedure was also repeated with training/test set size reduced by removing samples at random to match the size of the modified sets, but without separation of the factors.

For carry location, a one-sided Wilcoxon signed-rank test showed no significant increase ($p\text{-value}=1.00$) in C_{llr} between separating by carry location and not. Performance was consistent when the tested data came from a carry location not present in the training set.

For phone type, the Wilcoxon signed-rank test showed a significant increase in C_{llr} with a $p\text{-value} \ll 0.01$. C_{llr} increased by a mean value of 0.04 when testing on a new phone type. Overall, 56% of C_{llrs} increased by 0.05 or less, 84% by 0.10 or less, and 96% by 0.15 or less. The largest single increase in C_{llr} was for the pairing (*car*, *elevator up*), increasing 0.19 from 0.50 to 0.69. Across all pairings, only two crossed the $C_{llr} = 1$ threshold to an uninformative system as a result of separating phone type, namely (*dragging*, *cycling*) and (*dragging*, *kicking*), both of which already had $C_{llr} > 0.95$.

Group	Activity Types
Movement	cycle, run, walk
Transport	bus, car, train, tram
Dynamic	drag, kick, punch, throw
Elevation	elevator, escalator, stair
Stationary	sit, stand

TABLE II: Semantic grouping of activity classes. Elevation activities include both up and down variations.

Overall, it can be stated that performance is expected to worsen if the data being tested is from a phone not present in the training set. However, no such consideration must be taken in regards to carry location.

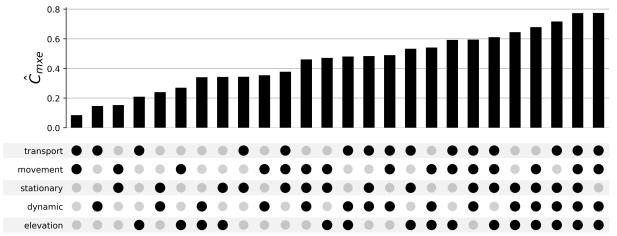


Fig. 5: UpSet plot[16] of \hat{C}_{mxe} for all unique combinations of activity groupings.

Multiclass LR systems

The proposed approach can be theoretically extended to as many hypotheses as there are classes in the training set. This may be of interest to an investigator in the earlier stages of an investigation, when not enough information is yet known to reduce the set of possible activities to only two options, and they wish to explore the likelihoods of multiple hypotheses simultaneously.

The naïve approach to a multiclass system is to simply include all activities in the training set and run the resulting system on your data. However, as seen in Figure 2, due to the overlapping nature of certain activities, some classes are mutually confused due to high similarities e.g. elevator up and elevator down. Including both of these activities as output options can cause the estimated likelihood of both to be suppressed, affecting performance. When naïvely using all nineteen of the NFI_FARED activities in a single LR system, the resulting \hat{C}_{mxe} is 0.96, close to uninformative. It therefore makes sense to group similar activities into distinct clusters, to avoid misleading deflation of likelihood values. This has the added benefit of forcing fewer restrictions on the investigator when selecting what activities are of interest.

It was found that grouping classes into the five categories which were used when creating the dataset performed as well or better than any tested unsupervised clustering method (see Figure 7 in the Appendix). Using these groupings also provides semantic meaning, aiding interpretation. The groups are listed in Table II. In this set up, an investigator may be interested in knowing whether a user was taking transportation, moving themselves, or stationary during a period of time. They could then select

these three groups in the multiclass system and inspect the resulting likelihoods.

Similarly to the binary case, some sets of groupings will be easier to distinguish than others. In Figure 5, the \hat{C}_{mxe} of the LR systems of all combinations of the activity groupings is displayed. Systems containing only two groups perform best, and \hat{C}_{mxe} increases with number of included groups. All systems are better than the reference, with the two worst (all groups included and all but stationary included) having a \hat{C}_{mxe} of 0.78.

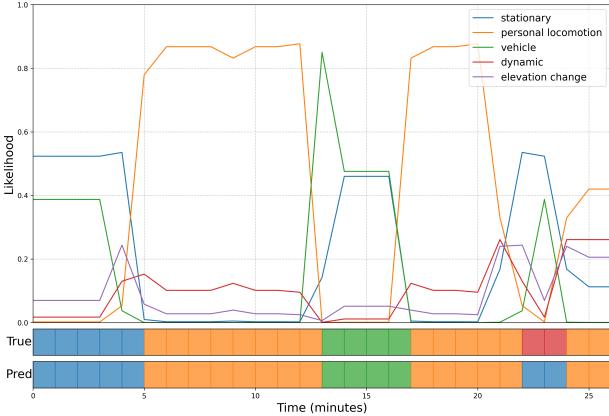


Fig. 6: Example timeline illustrating change in likelihood for each activity grouping as ground truth changes over time. Ground truth class and highest likelihood predicted class at each timestamp displayed at the bottom.

Timelines

Another application of the model is the creation of activity timelines from digital trace data. Establishing a timeline can be an instructive part of an investigation, and using our approach, one can be directly created in order to inspect the likelihood of different activities/activity groupings over time. Figure 6 shows a timeline for a 26 minute period in which a person was sitting, walked to the tram, walked for a few minutes after getting off the tram, kicked/punched someone for 2 minutes, then ran away, constructed by reordering samples from NFI_FARED. The timeline was constructed using the NFI_FARED dataset, extracting and reordering the datapoints from the validation set to follow the described pattern. The bottom row of squares shows the most likely class according to the model, with the correct group in the row above. It can be seen that the timeline is mostly accurate, correct for 24 of the 26 minutes, all correct for the groups *stationary*, *personal locomotion*, and *vehicle*. The *dynamic* group is wrongly identified and assigned a low likelihood. As in the binary case, identifying this group of activities is challenging.

DISCUSSION

The proposed approach to translate digital traces to LRs was tested and evaluated using iPhone traces from

the NFI_FARED dataset. iPhones were studied in this work because they have a consistent file storage system compared to the diverse ecosystem of Android OS's and are the most extensively studied, in particular in relation to physical activities. As more traces become available they can be inserted directly into our approach without adjustment. Currently, the results only pertain to digital traces from iPhones and their usefulness in an LR system.

The system was first analysed in the binary case, comparing the likelihood of the traces being generated by either one of two activities. It was found that according to the C_{llr} values, 167 pairs of activities could be used to produce an LR system which is more informative than a baseline reference system. Additional analysis was carried out on the best and worst systems, (*running, tram*), and (*train, tram*), respectively, which agreed with the initial conclusions from the C_{llr} and C_{llr}^{min} values. For use in actual casework, however, a fuller analysis would need to be carried out on any generated LR system, and its quality would have to be decided on a case-by-case basis. Such deep analysis was impractical to include given the volume of pairings, and the primary concern was assessing in which scenarios digital traces could be helpful to an investigator. The heatmap (Figure 2) is suggested to be used as an initial reference, to help inform whether it is worth looking into creating an LR system from digital traces, and expert judgement and case specific information should always be the deciding factors.

Investigations into variable importance using our machine learning approach can benefit future research into the meaning and application of digital traces. Additionally, the importance could lead researchers towards creating some rule-based systems for activity classification, which may be preferable in certain situations. For example, variable **type** is of high importance for classes *tram*, *train*, and *bus*, supporting previous findings linking the variable with transport activities. Finding a reliable link between this variable and these activities could prove fruitful.

An important point of consideration for using the proposed approach in investigations is the generalisability of results to the situation under study. It was shown that carry location does not need to be a concern when applying the approach, which is helpful since this information may be hard to discern in an actual case. However, the sensitivity analysis did reveal that phone type mismatches between the training and test set can adversely impact the quality of results. If the digital traces are taken from a phone not present in NFI_FARED, then the LR system trained on NFI_FARED is expected to have a C_{llr} 0.04 higher than what is presented here. NFI_FARED contains data from four iPhones that were relevant at the time of collection. As new phones and operating systems are continually being released, the dataset will require regular updating to provide maximum benefit and performance. Such continual updating is a considerable task, and therefore cooperation between agencies is required to keep valuable forensics datasets up to date. The outlined approach

can also work on an entirely new dataset of digital traces without overlap with NFI_FARED, relevant to the case at hand. If it is not possible to include the targeted phone and operating system in the training set, the investigator should take care when evaluating their LR system and be aware of the associated reduction in effectiveness.

The model was also extended to multiclass scenarios, where the likelihood that the traces were generated while one of multiple activities was being performed could be estimated. Multiclass is obviously a generalisation of the binary case; however, due to the drop in precision, it is recommended for use in the investigatory phase, and in practice, a binary setup would be presented for evidence. Multiclass is most useful to explore many possibilities at once in the early stages of an investigation, and further refinement is up to the expert. It was also found to be necessary to group activities into meaningful clusters, both to improve results and make analysis easier. The expert can begin by using all groups if desired, but as demonstrated in Figure 5, reducing the number of investigated groups improves the utility of the system. Reduction of groups could be in the form of outside knowledge, for example, if the period in question is too late for public transport, group *transport* could be ruled out, or if *dynamic* activities are not relevant to the case, they can also be omitted. It should be noted that multiclass results are presented using likelihoods, rather than LRs. The likelihoods can be transformed to LRs quite easily, such as using a one-versus-all approach, but in our analysis we found using likelihoods more instructive. The creation of activity timelines from digital trace data was also demonstrated as one such application of multiclass activity scenarios.

CONCLUSIONS AND FUTURE WORK

In this work we explored what benefit digital traces from smartphones could bring to forensic activity classification. Our outlined method provides a straightforward way to translate digital traces into LRs for activities, both in binary and multiclass settings. This is the first published use of these digital traces to classify activities, rather than validating the accuracy of the digital traces' proprietary labelling, and it was found that 167 possible pairs of activities could be distinguished. This analysis was facilitated by the collection of dataset NFI_FARED, containing a large and diverse set of digital traces labelled with nineteen distinct activities, which has also been made public.

The most important next step for improving how forensic practitioners can use digital activity traces is data collection. NFI_FARED is the first public dataset containing forensically relevant digital traces from iPhones, representing a large amount of experimental data and activities. However, as new models continue to be released, it is important for the dataset to be updated to maintain effectiveness. Additionally, data collected from new subjects and locations will further enhance the robustness of models trained on that data.

ACKNOWLEDGEMENTS

We would like to thank Loes Quirijnen for planning and collecting all the data for NFI_FARED. We also thank Marjan Sjerps for offering her time during discussions and providing valuable feedback on this paper, alongside Wauter Bosma and Pim Meulensteen.

REFERENCES

- [1] C. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Nov. 2004. Google-Books-ID: dqp2zioX7_EC.
- [2] C. C. G. Aitken, A. Barrett, C. E. H. Berger, A. Biedermann, C. Champod, T. N. Hicks, J. Lucena-Molina, L. Lunt, S. McDermott, L. McKenna, A. Nordgaard, G. O'Donnell, B. Rasmussen, M. J. Sjerps, F. Taroni, S. M. Willis, and G. Zadora. ENFSI guideline for evaluative reporting in forensic science. 2015.
- [3] D. L. Banks, K. Kafadar, D. H. Kaye, and M. Tackett, editors. *Handbook of Forensic Statistics*. Chapman and Hall/CRC, New York, Nov. 2020.
- [4] A. Bolck, H. Ni, and M. Lopatka. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):243–266, Sept. 2015.
- [5] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2022. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
- [6] N. Brummer and D. A. Van Leeuwen. On calibration of language recognition scores. In *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, pages 1–8, June 2006.
- [7] N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2):230–275, Apr. 2006.
- [8] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, Aug. 2016. ACM.
- [9] A. Collins and N. E. Morton. Likelihood ratios for DNA identification. *Proceedings of the National Academy of Sciences*, 91(13):6007–6011, June 1994. Publisher: Proceedings of the National Academy of Sciences.
- [10] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York, May 1994.
- [11] I. Evett. The logical foundations of forensic science: towards reliable knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1674):20140263, Aug. 2015. Publisher: Royal Society.
- [12] I. W. Evett. Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3):198–202, July 1998.
- [13] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu. A Survey on Deep Learning for Human Activity Recognition. *ACM Computing Surveys*, 54(8):1–34, Nov. 2022.
- [14] M. Guyll, S. Madon, Y. Yang, K. Burd, and G. Wells. Validity of forensic cartridge-case comparisons. *Proceedings of the National Academy of Sciences*, 120(20):e2210428120, May 2023. Publisher: Proceedings of the National Academy of Sciences.
- [15] A. J. Leegwater, P. Vergeer, I. Alberink, L. V. Van Der Ham, J. Van De Wetering, R. El Harchaoui, W. Bosma, R. J. Ypma, and M. J. Sjerps. From data to a validated score-based LR system: A practitioner's guide. *Forensic Science International*, 357:111994, Apr. 2024.
- [16] A. Lex, N. Gehlenborg, H. Strobelt, R. Villemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec. 2014.
- [17] A. Macarulla Rodriguez, Z. Geradts, and M. Worring. Likelihood Ratios for Deep Neural Networks in Face Comparison. *Journal of Forensic Sciences*, 65(4):1169–1183, 2020. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1556-4029.14324>.

- [18] C. McCarthy, L. Quirijnen, J. P. v. Zandwijk, Z. Geradts, and M. Worring. Hi-OSCAR: Hierarchical Open-set Classifier for Human Activity Recognition, Oct. 2025. arXiv:2510.08635 [cs].
- [19] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White. When Do Neural Nets Outperform Boosted Trees on Tabular Data? *Advances in Neural Information Processing Systems*, 36:76336–76369, Dec. 2023.
- [20] S. Mekruksavanich, A. Jitpattanakul, K. Sitthithakerngkiet, P. Youplao, and P. Yupapin. ResNet-SE: Channel Attention-Based Deep Residual Network for Complex Activity Recognition Using Wrist-Worn Wearable Sensors. *IEEE Access*, 10:51142–51154, 2022. Conference Name: IEEE Access.
- [21] K. Menzel. Bootstrap with Clustering in Two or More Dimensions, Dec. 2017. arXiv:1703.03043 [stat].
- [22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [23] D. Ramos, D. Meuwly, R. Haraksim, and C. E. H. Berger. Validation of Forensic Automatic Likelihood Ratio Methods. In *Handbook of Forensic Statistics*. Chapman and Hall/CRC, 2020. Num Pages: 20.
- [24] B. Robertson, G. A. Vignaux, and C. E. H. Berger. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Sept. 2016. Google-Books-ID: H5fCDAAAQBAJ.
- [25] E.-K. Sergidou, N. Scheijen, J. Leegwater, T. Cambier-Langenveld, and W. Bosma. Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150:1–8, May 2023.
- [26] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, Sept. 2023.
- [27] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, and S. Gupta. Deep ConvLSTM With Self-Attention for Human Activity Decoding Using Wearable Sensors. *IEEE Sensors Journal*, 21(6):8575–8582, Mar. 2021.
- [28] H. Spichiger. A likelihood ratio approach for the evaluation of single point device locations. *Forensic Science International: Digital Investigation*, 44:301512, Mar. 2023.
- [29] F. Taroni, S. Bozza, A. Biedermann, and C. Aitken. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, 15(1):1–16, Mar. 2016.
- [30] A. van Es, W. Wiarda, M. Hordijk, I. Alberink, and P. Vergeer. Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis. *Science & Justice*, 57(3):181–192, May 2017.
- [31] S. van Lierop, D. Ramos, M. Sjerps, and R. Ypma. An overview of log likelihood ratio cost in forensic science – Where is it used and what values can we expect? *Forensic Science International: Synergy*, 8:100466, Jan. 2024.
- [32] J. P. van Zandwijk and A. Boztas. The iPhone Health App from a forensic perspective: can steps and distances registered during walking and running be used as digital evidence? *Digital investigation*, 28:S126–S133, 2019. Publisher: Elsevier.
- [33] J. P. van Zandwijk and A. Boztas. The phone reveals your motion: Digital traces of walking, driving and other movements on iPhones. *Forensic Science International: Digital Investigation*, 37:301170, June 2021.
- [34] J. P. van Zandwijk and A. Boztas. Digital traces and physical activities: opportunities, challenges and pitfalls. *Science & Justice*, 63(3):369–375, May 2023.
- [35] J. P. van Zandwijk, K. Lensen, and A. Boztas. Have you been upstairs? On the accuracy of registrations of ascended and descended floors in iPhones. *Forensic Science International: Digital Investigation*, 47:301660, Dec. 2023.
- [36] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, and R. Stoel. Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, 56(6):482–491, Dec. 2016.
- [37] M. Vink, M. J. Sjerps, A. Boztas, and J. P. van Zandwijk. Likelihood ratio method for the interpretation of iPhone health app data in digital forensics. *Forensic Science International: Digital Investigation*, 41:301389, June 2022.
- [38] B. Xiao Wang, V. Hughes, and P. Foulkes. The effect of speaker sampling in likelihood ratio based forensic voice comparison. *The International Journal of Speech, Language and the Law*, 26(1):97–120, Sept. 2019.

APPENDIX

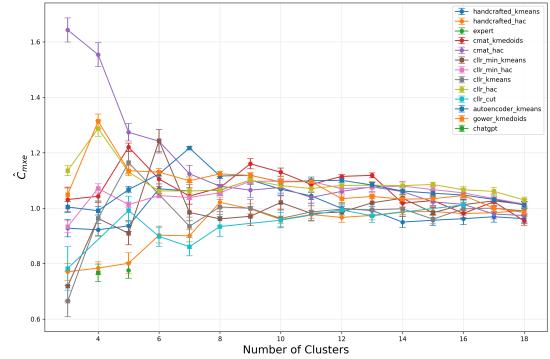


Fig. 7: \hat{C}_{mxe} values for different clustering methods as number of clusters is varied. Knowledge-based approaches **expert** and **chat-gpt** are in green.