Towards Non-Intrusive Camera-Based Heart Rate Variability Estimation in the Car under Naturalistic Condition

Shu Liu, Kevin Koch, Zimu Zhou, *Member, IEEE*, Martin Maritsch, Xiaoxi He, Elgar Fleisch, and Felix Wortmann

***Abstract*—Driver status monitoring systems are a vital component of smart cars in the future, especially in the era when an increasing amount of time is spent in the vehicle. The heart rate (HR) is one of the most important physiological signals of driver status. To infer HR of drivers, the mainstream of existing research focused on capturing subtle heartbeat-induced vibration of the torso or leveraged photoplethysmography (PPG) that detects cardiac cycle-related blood volume changes in the microvascular. However, existing approaches rely on dedicated sensors that are expensive and cumbersome to be integrated or are vulnerable to ambient noise. Moreover, their performance on the detection of HR does not guarantee a reliable computation of heart rate variability (HRV) measure, which is a more applicable metric for inferring mental and physiological status. The accurate computation HRV measure is based on the precise measurement of the beat-to-beat interval, which can only be accomplished by medical-grade devices that attach electrodes to the body. Considering these existing challenges, we proposed a facial expression based HRV estimation solution. The rationale is to establish a link between facial expression and heartbeat since both are controlled by the autonomic nervous system. To solve this problem, we developed a tree-based probabilistic fusion neural network approach, which significantly improved HRV estimation performance compared to conventional random forest or neural network methods and the measurements from smartwatches. The proposed solution relies only on commodity camera with a light-weighted algorithm, facilitating its ubiquitous deployment in current and future vehicles. Our experiments are based on 3,400 km of driving data from nine drivers collected in a naturalistic field study.**

***Index Terms*—heart rate variability, vital sign monitoring, non-intrusive measurement, in-vehicle environment, car data**

I. Introduction

Daily driving is an integral part of the day for many people, a fact that is frequently demonstrated by statistics. For example, in Germany 68% of the working population uses their car for commuting and more than 25% of them commute daily more than 30 minutes per direction [1]. However, in the U.S., about 90% of all citizens (aged 16 or older) drove 2.5 trips daily from 2019–2020 on average, which corresponds to about 1 hour of driving time or 30 miles ($\approx 48.3\ km$) of distance [2] per day. Moreover, the industry imagines the vehicle as the 3rd living space (after home and workplace) of people [3], which has a tremendous impact on people's lives. Nevertheless, driving is still a cognitively demanding task [4]. The prolonged driving time induces excessive stress [5], [6], which has the potential to impair mental and physiological health [4]. Furthermore, inattentiveness, drowsiness, and fatigue constitute one of the main factors of traffic accidents [7]. The timely recognition of the driver's status can be of significant benefit to improve driver states with just-in-time intervention (JITI) [8]–[11]. The recognition and regulation of driver status are particularly meaningful in the era of (semi-)automated vehicles. During the transition to fully automated vehicles (L2, L3 automation), drivers need to be mentally and physically prepared to take over the driving task at any given moment [12]. Therefore, the vision of future intelligent cars extends the idea of being a simple means of transportation toward a dedicated space where drivers' mental and physiological states are taken care of. Ultimately, identifying the status of drivers in vehicles is one important step toward safer driving and better life quality. From a broader perspective, the enhanced in-vehicle experience under the concept of ambient intelligence facilitates Internet-of-Things (IoT) enabled the transformation of a vehicle into a well-being and safety platform, where the driving performance, mental and physiological status are improved by restoring driver status in an optimal range, as illustrated in Figure 1 [13], [14].

Heart rate variability (HRV) and its measures are the most promising physiological signals to recognise driver status. Various studies have demonstrated their relevance to infer stats like stress, drowsiness, or inattentiveness. HRV is the variation in the time interval between heartbeats (inter-beat interval, IBI), and it can be characterised by HRV measures in time and frequency domains. In the context of our work, there are three most relevant and fundamental HRV measures in existing literature. In time domain, the root mean square of successive differences between IBIs (RMSSD) is a widely used measure. Increased RMSSD is associated with fatigue or drowsiness states, whereas stress can cause a decrease in HRV [15]–[17]. Furthermore, Taelman *et al.* observed that mental tasks

Manuscript received July xx, 20xx.

This research was funded and supported by the Bosch IoT Lab at the ETH Zürich and University of St. Gallen.

Shu Liu, Martin Maritsch, Xiaoxi He, and Elgar Fleisch are with the ETH Zurich, 8092 Zürich, Switzerland (e-mail: liush@ethz.ch; mmaritsch@ethz.ch; hex@ethz.ch; efleisch@ethz.ch).

Kevin Koch and Felix Wortmann are with the University of St. Gallen, 9000 St. Gallen, Switzerland (e-mail: kevin.koch@unisg.ch; felix.wortmann@unisg.ch).

Zimu Zhou is with the Singapore Management University, 178902 Singapore, Singapore (e-mail: zimuzhou@smu.edu.sg).

Corresponding author: Shu Liu

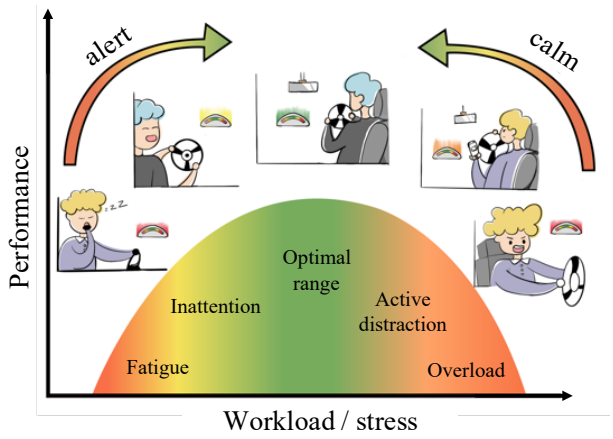


Fig. 1. Driving performance vs. cognitive load according to Yerkes-Dodson law, adapted from [14]. In future cars, intelligent vehicle systems are envisioned to be able to regulate excessive fatigue or stressful states of drivers, in order to further improve driving experience and safety [14].

can significantly reduce the proportion of successive normal beat to normal beat intervals (NN-interval) with a difference greater than 50 ms (pNN50) [18]. In addition to the signal in the time domain, HRV measures in the frequency domain are also powerful indicators. Patel *et al.* and Vicente *et al.* found statistically significant evidence that low frequency (LF) and high frequency (HF) ratio (LF/HF ratio) is in alert status higher than fatigue status while driving [19], [20]. To sum it up, excessively low or high states of the depicted HRV measures (RMSSD, LF/HF ratio, and pNN50) are strongly associated with the drivers' cognitive load and psychological status.

As a consequence, researchers and automobile manufacturers have taken pioneer efforts in driver heart rate detection. For example, BMW built a skin-resistance sensor into the steering wheel for heart rate monitoring [7], [21]. Similarly, Toyota and Denso monitored electrocardiography (ECG) and photoplethysmogram (PPG) using a steering wheel equipped with different electrodes and green light LEDs ($525nm$) [7], [22], [23]. In contrast, Ford and Denso utilised the driver seat [24], [25]. Although these methods seem to promise advanced and widely validated technology (as PPG used in today's smartwatches), researchers agree that these approaches cannot yet provide reliable measurements. For instance, [26] evaluated the performance of steering wheel integrated sensor under lab condition and concluded that there was an average error of 6% and the maximal error could escalate to 20%. Such error is far greater than commodity smartwatches, as we evaluated in this study (see Figure 9a). In [25], Wartzek *et al.* found that seat-integrated sensors could not reliably detect heart rate from drivers in all situations because seat integrated sensors are, for example, vulnerable to the thickness and the material of outer clothing as well as the weight of drivers.

To overcome such drawbacks, Zheng *et al.* recently designed a radio frequency device and leveraged ultra-wide band (UWB) impulse. The drivers' heart rate can be inferred by analysing of the Doppler frequency shift of UWB signal induced by heartbeat, respiration and ambient noise [27]. Although the method of [27] can accurately detect drivers' heart rate, inherent disadvantages exist. First, such a method requires a special purpose UWB device, which is not readily available. Second, due to ambient noise and physical constraint of sampling rate, IBI can only be measured with moderate accuracy (about 50% of IBI measurements have an error greater than $50\ ms$ [27]). Such an accuracy limitation can be tolerated, when only the average heart rate is detected since HR is computed as the inverse of the mean IBI in a certain interval. The noise in the IBI measurement is cancelled out by the mean operation. However, considering critical metrics, such as RMSSD or LF/HF ratio, the inaccuracy will be magnified because these measures are sensitive even to the small inaccuracies in the measurements. Recently, with the advancement of computer vision techniques, remote PPG (rPPG) [28], [29] has attracted prominent attention. The fundamental principle of rPPG is as follows. Heartbeat (hence blood volume in vessels) induces subtle colour variations on the human skin surface, which can be captured by an RGB camera. Signal processing techniques are then applied to analyse such variation; thus human cardiac activities can be monitored. Although rPPG technique is appealing, many efforts are still needed before it can be applied to the real world scenario. Remote PPG is sensitive to illumination and motion artefacts. More importantly, commodity cameras record video at 30 or 60 Hz, which by Nyquist–Shannon sampling theorem is insufficient for the accurate measurements of IBIs, of which the variation is at millisecond-level. Existing research on HR/HRV detection using rPPG was conducted in well-defined lab conditions; therefore their generalisability to real-world scenarios remains unclear [30]. In a nutshell, the existing contact-less monitoring methods do not guarantee a reliable measurement of HRV in real world scenarios.

In light of these existing challenges, we propose an alternative way to monitor driver status through HRV. As described above, drivers's cognitive load and mental status are strongly characterised by excessively low or high HRV measures (*i.e.*, RMSSD, LF/HF ratio, and pNN50). Therefore, instead of attempting to derive HRV measures from inherently noisy IBI measurements, we propose a facial expression-based approach to detect the onset of HRV outliers. On the basis of existing literature, we define HRV outlier as samples whose values are one standard deviation below or above the mean [31], [32].

Facial expressions are strongly connected and influenced by the autonomic nervous system (ANS). On the one hand, human cardiac activity is controlled by ANS. The sympathetic nervous system (SNS) accelerate the heart rate through the discharge of epinephrine and norepinephrine while the parasympathetic nervous system (PNS) releases acetylcholine to induce deceleration [33]. On the other hand, ANS also functions involuntarily and cope-with affective arousal in reaction to circumstance accordingly [34]; To estimate HRV from facial expression, we employed the state-of-the-art machine learning scheme and developed a novel tree-based probabilistic fusion neural network approach. Compared with existing contact-less and non-intrusive UWB or rPPG based methods [27], [29], the advantage of our facial expression-based method and our contribution can be summarised as follows.

- Our approach relies on commodity RGB camera working

at 30 FPS, which is very likely to be integrated in future vehicles as a part of driver monitoring systems [35], [36]. Thus, no additional UWB devices are needed.
- We verified our approach based on around $3,400$ km (68.6 hours) of driving data collected from a two-week field study, involving nine participants during uncontrolled daily driving activities on public roads.
- A novel tree-based probabilistic fusion neural network approach is developed to optimise HRV estimation performance. The proposed tree-based probabilistic fusion framework outperformed conventional convolutional or recurrent neural networks and classic tree based machine learning models by up to 6.9% in balanced accuracy.
- We benchmark our method against consumer smartwatch measurement. Smartwatch can be seen as a proxy of the upper bound of rPPG since its close contact with the skin mitigates a large portion of noise due to illumination and motion artefacts. Our evaluation shows that the proposed approach can even outperform high-end consumable smartwatches by a large margin.
- To the best of our knowledge, this is the first study that verifies the feasibility of facial expression-based HRV outlier detection based on driving data collected from public roads in real driving scenarios. Since the overall environment is challenging compared with laboratory conditions, our results are likely to be more reliable.

The remainder of the paper is organised as follows: We present our field study in Section II. We introduce our methods for HRV estimation in Section III. Section IV summarises the results of our methods. The implication of the method and discovery is discussed in Section V. Finally, Section VI presents the conclusion.

II. Experiment Settings

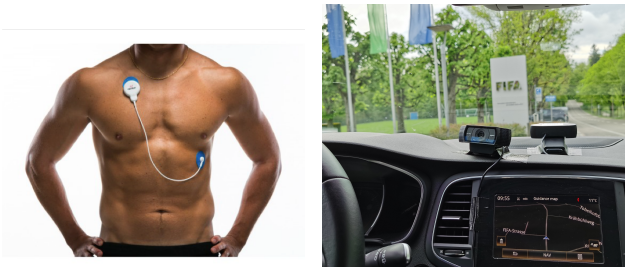
We conducted a two-week field study with nine daily commuters (originally then; one participant's data were removed due to corruption) during their normal driving routine on public roads. A variety of sensory data, including HRV, facial expression, and smartwatch records, is collected from the daily the driving activity of participants in naturalistic condition. The participants were supposed to use the vehicles for their daily drives, including business trips and vacations. Our university's ethics committee authorised the approval for the experiments prior to the study.

A. Subjects

The nine participants (four females and five males, mean age, 37 ± 8 years) were recruited from a large enterprise (more than 1,000 employees) through an internal call in their company social media forum. Our selection focused on ordinary daily commuters that are representative of a large variety of people.

B. Data Collection Equipment and Protocol

We mounted two webcams (Logitech HD Pro Webcam C920) on the dashboard of the vehicle to record the faces of the drivers and the traffic context. The traffic context information is not relevant to this study and will not be explored. HRV data were collected using a medical-grade heart monitoring device (Firstbeat Bodyguard 2), as shown in Figures 2a and 2b. The sampling rate of the heart monitoring device was 1000 Hz. For the sake of a more comprehensive comparison participants also wore a recent consumer smartwatch (Garmin vívoactive 3) during driving.



(a) Firstbeat measurement [37] (b) Webcam deployment

Fig. 2. Data collection equipment.

C. Characteristics of Driving Activity

It was crucial for our dataset to capture representative driving situations. This subsection presents some important statistics related to our dataset.

After data cleansing, we had about 68.6 hours of video data with associated HRV measurements during driving. The total driving distances of each participant are plotted in Figure 3. Most drivers drove for reasonably long distances (more than 300 km) during the field study. The GPS records of the vehicles are presented as a heatmap in Figure 4. As shown in this figure, most participants drove around the area of Stuttgart, Germany. Overall, our dataset covered a wide range of daily driving activities like commuting, shopping trips, and leisure activities at the weekend.

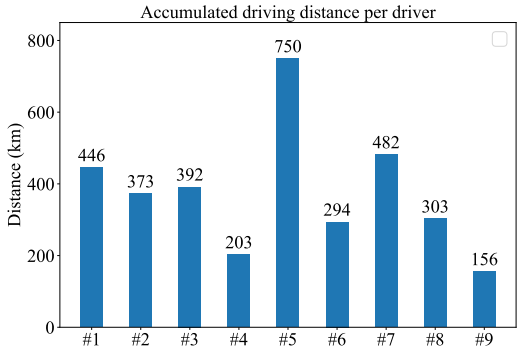


Fig. 3. Accumulated driving distance

D. Characteristics of Heart Rate Variability Measure

HRV is measured over a period of time. We applied an overlapping sliding window with a length of 5 min and a step size of 30 s to compute HRV measures (*i.e.*, RMSSD,

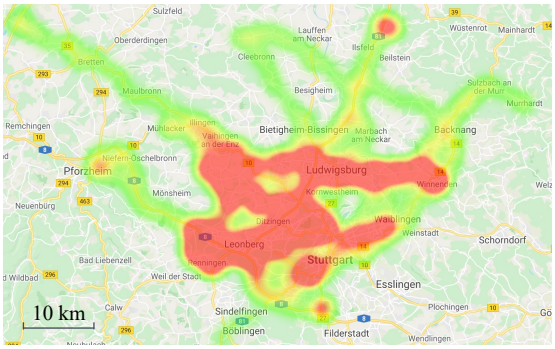
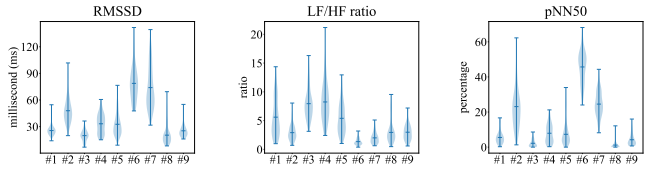


Fig. 4. GPS heatmap of the most active area



(a) RMSSD (b) LF/HF ratio (c) pNN50

Fig. 6. Distribution of RMSSD, LF/HF ratio, and pNN50 of the nine drivers

LF/HF ratio, and pNN50). The choice of 5 min follows the convention of short term HRV measurement [38], [39]. We refer to such 5-min segments as ***HRV segments***. The entire facial expression-based HRV detection framework is similar to the one used in [40] and is shown in Figure 5. We take facial expressions in ***HRV segments*** as input data to estimate the HRV measures associated with the corresponding segments. From the recording of Firstbeat Bodyguard 2, the ground truth of HRV measures is computed based on a standardised wearable data processing toolkit [41]. These HRV segments are randomly shuffled for training and testing. To avoid the intersection between training and test datasets caused by overlapping sliding windows, we discard HRV segments that intersect both datasets for each shuffle, as illustrated in Figure 16.

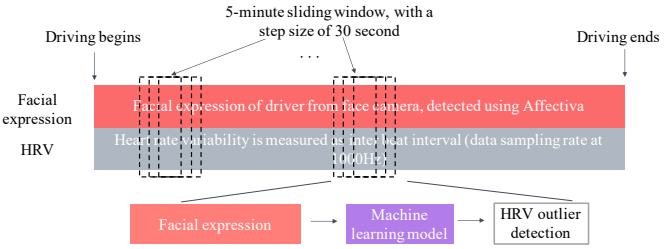


Fig. 5. HRV outlier detection framework

Next, we inspect the distribution of HRV, which is illustrated in Figure 6. Owing to the influence of age and gender, there is significant difference among participants in terms of the median and range [38]. To account for such individual factors, we define HRV outlier detection as a binary classification problem and predict whether a given driver's HRV measures are excessively low or high with respect to his/her personal empirical distribution. We distinguished between low and high HRV outlier detection, as formulated in Equation 1 and 2, respectively. Such definition is similar to [31], [32], in where the authors defined outliers for stress or mental status estimation as one standard deviation above or below the mean HRV. Consequently, HRV measures within one standard deviation of the personal mean are considered normal. It means that we develop two machine learning models, one for the detection of low outliers of HRV measures and the other for high outliers.

$$\underset{detector}{low} = \begin{cases} low\ outlier, & < per.\ mean - per.std \\ rest, & >= per.\ mean - per.std \end{cases} \quad (1)$$

$$\underset{detector}{high} = \begin{cases} rest, & <= per.\ mean + per.std \\ high\ outlier, & > per.\ mean + per.std \end{cases} \quad (2)$$

We performed data cleansing and removed IBI artefacts ($<$ 250 ms or $>$ 2000 ms). We removed HRV segments where the driver faces appeared in less than 70% of video frames as well as HRV segments with no valid IBI signal. Finally, we obtained in total 3876 HRV segments, the distribution of low and high outliers, and normal samples are given in Table I.

TABLE I
SAMPLE COUNTS FOR DIFFERENT CATEGORY

	RMSSD	LF/HF ratio	pNN50
low	513	584	423
normal	2840	2636	2858
high	523	656	595

III. METHODS

This section explains our approach to infer HRV outliers from facial expressions.

A. Data Preprocessing

Detection of Facial Action Units. The facial action units (AUs) is used in the facial action coding system (FACS) to describe the muscle movement currently active in the face, such as "nose wrinkle" or "cheek raise" [42] Based on the active level and the combination of AUs, facial expressions such as anger, fear, and joy can be quantitatively determined.

The manual coding process of FACS requires profound expert knowledge and is laborious. To alleviate this problem, we leveraged the automatic FACS coding algorithm from Affectiva, a spin-off of MIT's Media Lab. Affectiva's facial expression recognition technology uses computer vision and deep learning techniques to first detect the active level of AUs, based on which another mapping function is established between facial expression and AUs [43]. The Affectiva's major advantage is that it is built on a very large foundational dataset, consisting of more than 9.7 million facial images of people, with more than 5 billion facial frames [43]. Additionally, based on the in-vehicle data of more than 20,000 hours featuring more than 4,000 unique individuals, Affectiva is

well optimised to automotive in-cabin environment [43]. Given these features, Affectiva's solution can reliably capture facial movements. In this study, we used one of the latest stable versions (ics-2.2.1).

Feature Engineering. Affectiva detects AUs for each frame and presents the results as the activation level of each AU in the range of 0 to 100. The entire list of detected AU is the following:"browRaise","browFurrow", "noseWrinkle", "upperLipRaise", "mouthOpen", "eyeClosure", "cheekRaise", "eyeWiden", "innerBrowRaise", "yawn", "blink", "blinkRate", "lipCornerDepressor", "lidTighten" and "smile" (all 15 AUs provided by the Affectiva SDK). We build feature vectors (FVs) for AUs through a sliding window, with both the length and the step size equal to five seconds. In each sliding window, we compute mean, min, max, median, standard deviation, quantile-25%, quantile-75%, kurtosis and skewness for each AU. That is to say, for every 5 seconds, a 135-dimension (15 AUs $\times$ 9 features) FV is generated. From a 5-min HRV segment, a sequence of 60 FVs ($5min$ / $5s$ = 60) is generated.

In addition to AUs, HRV is heavily influenced by the time of day. We incorporate this prior information by including time features defined as current time (formatted in the 24 h-scale), day of the week, an indicator of driving at night, seconds before dawn, seconds after dusk, seconds before sunrise, and seconds after sunset. The last four features were set to zero if driving had occurred after or before the corresponding event. By merging the time features to each FV, the final input to the machine learning models has the shape of $60\,steps \times 142\,dim$.

B. Machine Learning Approaches

Standard Pipeline. We first verify the feasibility of the HRV outlier detection in the wild by exploring a random forest model. Despite the simplicity of tree-based models, they often outperform more complicated models such as neural networks or support vector machines (SVM) [44]. This is especially the case with a lack of prior insight about underlying data property or domain knowledge [45].

Our random forest based pipeline is depicted in Figure 7. In the training phase, we assign all FVs in an HRV segment the same label as the HRV segment, meaning that the input instance to the random forest is each FV. In the test phase, we perform prediction on all FVs in each HRV segment. The final prediction for one HRV segment is aggregated from prediction results of all FVs in that HRV segment. In this study, we use the majority vote as the aggregation function.

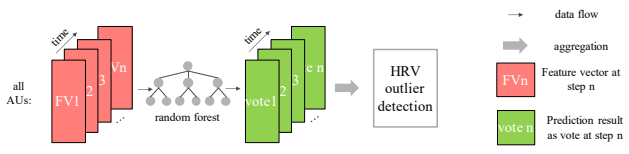


Fig. 7. HRV outlier detection using standard random forest pipeline

The input FVs are sequences of time-series data. Therefore, to further explore the possibility of other machine learning models, the choice of random forest can be replaced by prevalent (1D) convolutional neural network (CNN), recurrent neural network (RNN), and multilayer perceptron (MLP), etc. We used random forest as well as various neural networks as baseline method and the evaluation is presented in Section IV.

Tree-based Probabilistic Fusion Network (TPFN). As we will show in Table IV and V in Section IV, when tree-based models are applied in the standard pipeline, they usually outperform neutral network models. The tree-based model often perform better than other models in practise [44], [45]. The reason is that the hierarchy decision stage of tree-based models does not impose restrictions on the distribution of inputs; the merit of the ensemble mechanism of random forest makes it extremely robust to unseen data [44]–[46]. Unlike neural networks whose architecture is sensitive to specific data distribution and requires profound domain knowledge, recent work suggested that random forests can help discover the underlying structure of data [44], [45]. As such, we develop a hybrid model that uses a tree-based model to create a probabilistic embedding from data, which is further fused and processed by a neural network. The details of our proposed model are explained as follows.

We first compute the probability embedding of each AU. This is performed by building 15 random forests for the 15 feature subsets of all AUs. Each feature subset contains not only features of the corresponding AU, but also the prior mentioned time features. Therefore, each random forest takes FVs of 16 dim. (nine statistical features from AUs and seven time features) as input. We train these 15 random forests similar to those in Figure 7. After that, instead of aggregating the predictions of the random forests, we take the prediction probability (with closer to 0 being more likely to be class 0, and vice versa), which is again a time-series sequence of form 60 step $\times$ 15 AUs, as input to a neural network. The neural network take the fusion of the probability from the random forests and further predicts the HRV outlier for the entire sequence. In this study, we used a multilayer perceptron (two layers, each with 16 neurons and sigmoid as activation) to classify on every step the fused probability and then with a final classification that is aggregated (by majority vote) from the 60 votes.

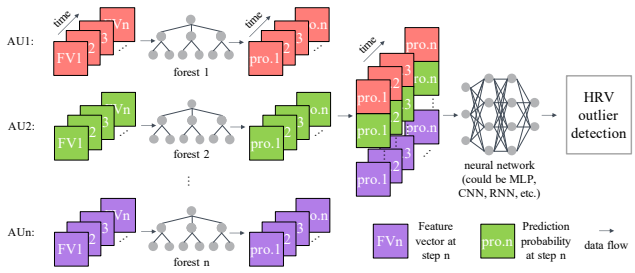
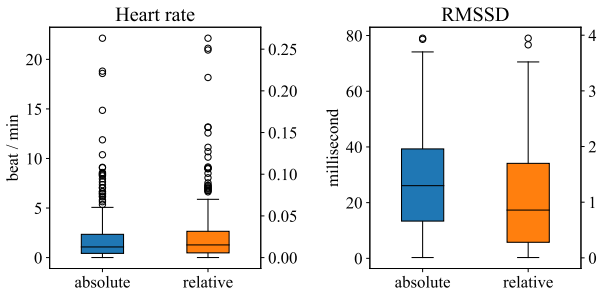


Fig. 8. Tree based probabilistic fusion model

IV. Evaluation

In this section, we evaluate the proposed method against state-of-the-art machine learning models. The evaluation is performed by constructing a *general model* for all drivers.



(a) Heart rate (b) RMSSD

Fig. 9. Absolute and relative errors of high-end smartwatch compared with Firstbeat [37]

In the following, we will provide an insight into the HRV measurement accuracy of the current high-end commodity smartwatch compared with medical-grade heart rate monitor (Firstbeat). Next, we comprehensively compare the proposed approach with various baseline methods.

A. Measurement Accuracy of Smartwatch

Smartwatches and other wearable devices are becoming popular in people's daily life. The low cost and ubiquitous property make them an ideal tool for health monitoring. Therefore, we should first inspect if their measurement accuracy meets the requirement of HRV detection in the wild.

For this purpose, we use the measurement of the Firstbeat as the gold standard to compute the errors of smartwatches. The absolute and relative errors of mean heart rate and RMSSD of HRV segments are illustrated in Figure 9. In this study, we have an overlap of 21.25 hours of Firstbeat data with smartwatch measurements. Due to the in-the-wild property of the experiment, drivers sometimes did not wear the smartwatch.

It it obvious in Figure 9a that the smartwatch can very accurately measure the average heart rate. The mean value of the absolute error is only around 1 beat per minute. This magnitude of error agrees with the latest systematic evaluation of smartwatches [15], [47]. However, the errors become significant when using the measured IBI from a smartwatch to compute RMSSD, as illustrated in Figure 9b. The mean of the relative errors is almost 100%. More comparisons between smartwatch and Firstbeat measurements are given in Figure 19. Although the sensors of smartwatches tremendously improved and will continue (e.g. ECG monitoring is now available in certain smartwatches, the prerequisite of its usage is that the users must sit still without arm movements; thus, limited applicability while driving [15]), the current high-end smartwatch that measures the accurate mean heart rate does not provide reliable HRV measurements while driving.

B. Comparison with Baseline Methods

In this subsection, we present the baseline methods to be compared and analyse the results quantitatively.

Baseline Methods. This part describes the baseline methods in detail. On the one hand, the chosen baselines, such as smartwatch and time models, are used to demonstrate that our proposed facial expression based approach is a good and necessary complement of currently prevalent heart rate monitoring methods; on the other hand, the comparison with the tree-based and neural network models can demonstrate that the proposed tree-based probabilistic fusion is an efficient way to learn data representation.

- **Smartwatch Model.** Although smartwatches exhibit unreliable measurements of HRV, as described in Section IV-A, it is still meaningful to evaluate whether the noise in the smartwatch is consistent. This means, for example, if the noise adds a consistent offset to the HRV measurements, HRV outlier detection can still be accurately performed since we are interested in whether HRV is lower or higher than the personal baseline level. We refer to the smartwatch model as **SM**.
- **Time Model.** It is well known that the time of day has a strong impact on HRV [48]. For example, HRV tends to be higher during working hours than at night because the body must react to the accumulated stress and cognitive load. We demonstrate the time-dependent variation of RMSSD in Figure 10. More examples of LF/HF ratio and pNN50 are given in Figure 17 and 18 in the Appendix.

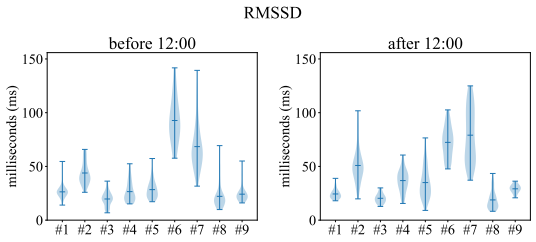


Fig. 10. RMSSD of the nine drivers in different time interval

Therefore, it is crucial to inspect the possibility of inferring HRV outlier purely based on time. To this end, instead of defining a rule based model, we build a Time Model (referred as **TM**) by constructing a random forest using only time features (7D). The TM resembles the settings in Figure 7 except for the input FVs.

- **Tree Based Models.** As described in Section III-B, random forest (refered as **RF**) can be used as the machine learning back-end in the pipeline. To explore the feasibility of other tree-based models, we further replace random forest with one of the latest tree-based models, the Deep Forest (referred as **DF**) [46]. For RF, DF and tree-part of TPFN, a grid-search of parameters is performed. The candidate parameters are described in Table VI. The optimal parameters for all tree-based models are determined as depth of tree = None (*i.e.*, unlimited depth), number of trees = 200, min samples split = 2, min samples leaf = 1.
- **Neural Network Models.** Over the last decade, neural network techniques have experienced tremendous improvement. Therefore, it is meaningful to benchmark our proposed tree-based probabilistic fusion approach with them. We implemented 1D convolutional neural network

(referred as **CNN**), Multilayer-perceptron (referred as **MLP**) as well as recurrent neural network (referred as **RNN**) to the time-series FVs. To be more precise, the **CNN**, as depicted in Figure 11, consists of two cascaded 1D convolutional filters (kernel size = 3, filter size = 64, dropout rate = 0.5, and activation = sigmoid) followed by a linear fully connected (FC) layer with 16 neurons and a Softmax operation that reduces the flatted convolutional output to two dimensions, corresponding to the binary classification.

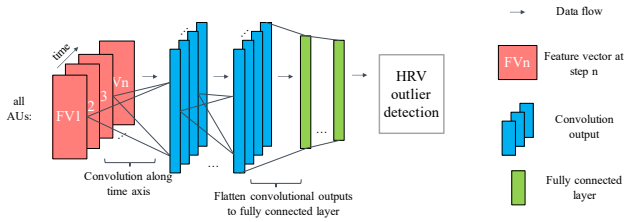


Fig. 11. CNN baseline model

RNN, as shown in Figure 12, uses two recurrent layers (dropout rate = 0.5) with 16 gated recurrent units (GRU) followed by a linear FC layer with 16 neurons and a Softmax operation that reduces the hidden states of GRUs to two dimensions, similar to CNN.

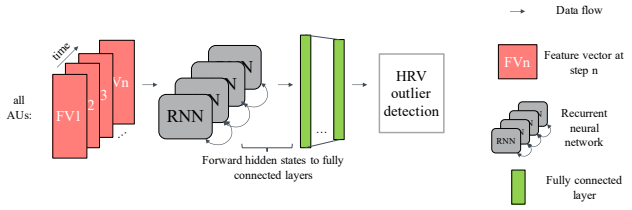


Fig. 12. RNN baseline model

MLP resembles the pipeline in Figure 7, where random forest is replaced by a two-layer multilayer perceptron (activation = sigmoid, dropout rate = 0.5) with 32 units in each layer. The classification is performed on each FV and the final prediction is the aggregation (majority vote) of all FVs in an HRV segment.

The chosen architectures for CNN, RNN and MLP are similar to the ones used in [49], [50], which have been proven to be effective in predicting various physiological and psychological status. Additionally, the optimal parameter settings of the above mentioned neural networks were determined using grid search. This is done by applying a 5-fold cross validation to the training dataset. Subsequently, each network is re-trained on a total training dataset with the optimal parameters. This procedure of parameter searching is similar to that of [50]. The candidate parameters for the grid search are described in Table VII - IX. Furthermore, for all networks, we further grid searched on optimiser (SGD and ADAM), normalisation schemes (Z-score normalisation, min-max normalisation, and logarithmic transformation[1]), and gradient clipping schemes (norm type = 2-norm, the option for max. norm was iterated over 1, 10, 100). Finally, all neural network models (including the neural network part of TPFN) are trained by ADAM with a learning rate of 0.005; Z-score normalisation and gradient clipping with max. norm = 10 are applied. The loss function is defined as cross-entropy.

Numeric Results. We perform the binary classification on an unbalanced dataset ($majority : minority \approx 82\% : 18\%$). Therefore *balanced accuracy* is used as the metric, which is an unweighted mean of accuracy over all classes. Thus, this metric is not biased towards the majority class and can provide a more accurate evaluation of the overall performance. As relevant HRV metrics, we selected RMSSD and evaluated LF/HF ratio and pNN50 since they are closely related to mental status, as explained in Section I.

We first compare the proposed approach with HRV outlier detection based on smartwatch measurements. That is to say, for SM we computed HRV measures from smartwatch-measured IBI and use the computed HRV measures to detect HRV outliers. Smartwatch measurements are available for 21.25 hours of 68.58 hours. To ensure a fair comparison, the proposed TPFN is trained for the remaining 47.3 hours. After that, TPFN and SM were validated on the same dataset where smartwatch measurements are available. The result is described in Table II and III. The proposed TPFN method outperforms SM in all cases, with an improvement ranging from 3.6% to 13.1%. The evaluation demonstrated that the IBI measured by smartwatches could not precisely compute HRV measures despite accurate heart rate measurement. The measurement noise does not constitute a constant offset, making HRV outlier detection based on smartwatches imprecise.

TABLE II
BALANCED ACCURACY OF LOW HRV OUTLIER DETECTION, SMARTWATCH (SM) VS. PROPOSED SOLUTION (TPFN)

Model	RMSSD	LF/HF ratio	pNN50
SM	68.2	52.7	55.1
TPFN	**73.3**	**60.3**	**68.2**

TABLE III
BALANCED ACCURACY OF HIGH HRV OUTLIER DETECTION, SMARTWATCH (SM) VS. PROPOSED SOLUTION (TPFN)

Model	RMSSD	LF/HF ratio	pNN50
SM	64.7	58.6	61.0
TPFN	**68.3**	**70.6**	**71.9**

Next, we compare the proposed TPFN with prevalent machine learning models. We randomly split the dataset into train (70%) and test (30%) sets. The final results are presented as the average of 10 repeated experiments using 10 different random seeds. The standard deviations of the 10 repetitions are indicated in brackets in corresponding tables.

[1]to avoid numerical issues, logarithmic transform is applied as $A = log(|A| + 1)$

The results are presented in Table IV and V. TM performs by distance the worst despite the strong correlation between HRV and time [38]. Neural network approaches (CNN, RNN and MLP), despite their higher complexity, achieve worse results than tree-based models (RF and DF). Finally, the best performance is achieved by the proposed hybrid TPFN model that combines the merits of both tree based model and neural networks. The TPFN model outperforms other best performing baseline models by an average of 3.4% and up to 6.9% in balanced accuracy.

TABLE IV
BALANCED ACCURACY OF LOW HRV OUTLIER DETECTION

Model	RMSSD	LF/HF ratio	pNN50
TM	60.3 (1.7)	58.0 (2.2)	61.5 (2.5)
RF	62.4 (2.1)	61.6 (2.7)	66.8 (2.2)
DF	62.8 (2.8)	61.8 (2.7)	66.6 (2.2)
CNN	59.1 (3.3)	57.9 (2.9)	55.3 (3.7)
RNN	58.4 (3.2)	56.1 (3.1)	57.3 (3.3)
MLP	60.5 (2.5)	55.5 (5.3)	56.3 (4.7)
TPFN	**69.7 (2.2)**	**65.3 (2.7)**	**71.4 (2.1)**

TABLE V
BALANCED ACCURACY OF HIGH HRV OUTLIER DETECTION

Model	RMSSD	LF/HF ratio	pNN50
TM	56.0 (2.4)	61.4 (2.8)	59.2 (2.6)
RF	65.5 (2.3)	64.5 (2.7)	65.5 (2.4)
DF	66.8 (2.4)	64.8 (2.7)	65.8 (2.3)
CNN	56.0 (3.5)	60.2 (3.4)	60.3 (4.7)
RNN	60.5 (2.3)	61.1 (2.7)	62.2 (4.2)
MLP	59.7 (3.8)	58.1 (3.8)	60.4 (4.5)
TPFN	**68.3 (2.1)**	**65.7 (2.7)**	**69.2 (2.2)**

To better visualise the performance of the outlier detection, the confusion matrices of TPFN are plotted in Figure 13 - 15. The confusion matrices show that the proposed TPFN is not particularly biased towards the majority class, except for the high outlier detection of LF/HF ratio in Figure 14. Meanwhile, the model maintains low false negative and false positive rates, as illustrated in the sub-diagonals of the confusion matrices.

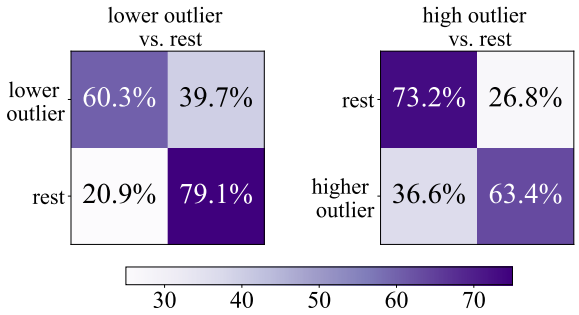


Fig. 13. Confusion matrix of RMSSD outlier detection

V. DISCUSSION

In this section, we discuss the proposed approach in terms of prediction usability, reliability and potential limitations.

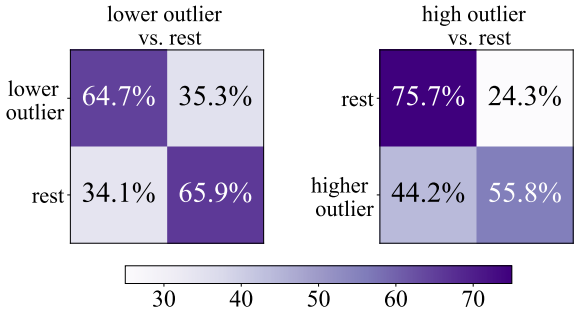


Fig. 14. Confusion matrix of LF/HF ratio outlier detection

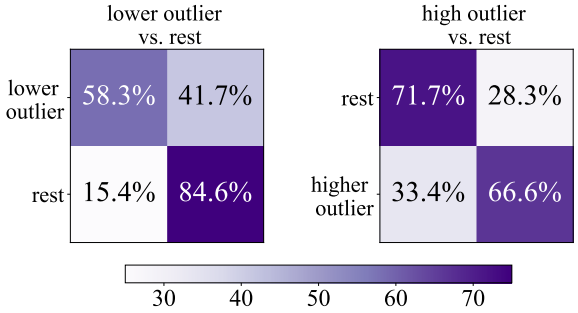


Fig. 15. Confusion matrix of pNN50 outlier detection

A. Usability

The proposed approach relies on a driver monitoring camera. Although such a camera is not widely installed in current vehicles, it is becoming an integral and essential component of future cars. The reason is that a driver monitoring camera is an essential safety feature that prevents inattention or drowsiness while driving. European Union (EU) is the pioneer in pushing forwards this safety feature. In 2019, a general safety regulation was passed by the EU Council of Ministers. The safety regulation requires that all new vehicles on the EU market must install advanced safety systems to prevent distraction and drowsiness. Such an advanced safety system is very likely to be implemented through a driver monitoring camera [35], [36], [51]. Starting in 2022, all new type-approved vehicles with a certain level of autonomy must fulfil this requirement. By 2026, this law will cover all newly produced cars regardless of their level of automation [51], [52]. In the United States, two safety-related traffic bills have been introduced or passed (H.R.2 - Moving Forward Act and S.4123 - SAFE Act of 2020). This may lead to the requirement that driver monitoring camera becomes mandatory in new vehicles [51]. In China, the regulations requiring long-distance trucks to use driver monitoring have already been implemented in certain regions, in particular for vehicles transporting hazardous goods. More similar regulations are expected to follow [51]. We can anticipate that driver monitoring cameras will become essential and mandatory in many regions of the world in the future. The proposed solution can be well integrated into future cars without any additional hardware cost.

The proposed solution plays an important complementary role to the emerging driver monitoring solutions such as

activity recognition and gaze detection [53], [54]. While driver activity recognition and gaze detection algorithms can infer whether a driver's behaviours are allowed during driving, these algorithms do not guarantee if a driver's mental status is favourable. Various studies suggested that incremental cognitive load impact drivers' visual behaviour and their gaze is therefore focused on the central road region [55], [56]. Even worse is that increased cognitive load reduces drivers' awareness of incidents occurring within the restricted visual field, namely "look but fail to see" [14]. Such shortcomings of driver activity or gaze monitoring techniques can be well compensated by our proposed algorithm, which assesses the cognitive load of drivers through HRV estimation.

Finally, the monitoring of HRV measures provides a significant derivative benefit from the well-being perspective. Driving is not the only source of stress and mental load in daily lives. Occupational burn-out, sentimental relation between couples, and mood disorders, etc. can lead to sub-optimal states and could manifest themselves in the changes of HRV measures [38], [57]. The HRV monitoring technique in combination with well-being interventions that regulate drivers' psychological status [9], [10], in essence, does not only reduces stress and cognitive load from driving, but also from other daily events [14]. The smart vehicles in the future should not only be a tool for transportation, but also an intelligent 3rd living space integrated with a wellness platform [14].

B. Reliability

The proposed HRV estimation solution provides supportive service to improve the user experience. Upon the detection of excessively low or high HRV measures, intelligent vehicle systems can deliver corresponding interventions to regulate the sub-optimal states of drivers. State-of-the-art driver stress or mental regulation strategies mainly consist of music or mindfulness intervention, breath exercise, control of ambient auditory, lighting or aero (wind) feedback, and odour stimulation, etc. [9], [10], [58]–[60].

Unlike obstacle avoidance or pedestrian detection systems that have almost zero tolerance for false detection, the HRV estimation in our context can engage in ambiguity when the system is uncertain about its estimation. This is in line with the Guidelines for Human-AI Interaction proposed by Amershi *et al.* [61]. There is a certain grey zone that tolerates ambiguous decisions. In the case of uncertainty, the reliability of the system can be further improved by adopting the interaction between system and users, for instance, through verbal communication [62], an inquiry of the necessity of intervention [9], or adjusting intensity/option of intervention [63] etc. On the other hand, interventions yield a stronger effect, especially if users are in a sub-optimal state (and hence a state of high "vulnerability"), because more potential for improvement exists. That being said, wrongly applied interventions (*i.e.*, user in the optimal state) to regulate low HRV measures are unlikely to move the user from optimal state to a state of high HRV measures [63]. If interventions are provided based on wrong HRV estimation, the consequence is not as dangerous as, for instance, miss-detection of lane marks or pedestrians.

C. Limitations

This research should be assessed considering its limitations. Even though our experiments were performed under naturalistic conditions, a very challenging setting, the proposed approach does not generalise to the leave-one-subject-out setting. This drawback could be attributed to the fact that we have only nine drivers in our dataset. The limited sample size is not diverse enough for a machine learning model to learn a generalisable pattern among different subjects. We expect that a large scale field study with a greater number of the subject could be conducted to further explore the generalisability of facial expression-based HRV estimation.

In addition, it is worth noting that the estimation of LF/HF ratio is generally inferior than the other HRV measures. We believe that this difference can be explained by the fact that subject respiration heavily influences the frequency components of HRV [64], [65]. More specifically, both respiration and autonomic nerve activities contribute to the deviation in LF/HF ratio, whereas only the latter factor can be reliably interpreted by the proposed facial expression-based inference model. For future work, we see potential in integrating a respiration detection module and thus fusing the information of breath to further improve the HRV estimation, which shall be one of our focuses in future research.

Furthermore, although the excessively low or high HRV measures are strong indicators of certain physiological and psychological status, an exact measurement of HRV measures could bring more insight into a user's health status (*e.g.*, monitoring of hypertension or other cardiovascular diseases), which is not accomplished in this study as no study subject reported any relevant complications. The exact measurement of HRV relies on precise capturing of IBI, which can be achieved by rPPG under well-defined lab condition. The fundamental mechanism of rPPG that detects the heartbeat induced peak of blood volume in a vessel is a more straightforward approach for measuring the exact value of HRV. However, rPPG is not as robust as our approach and is vulnerable to ambient noise due to illumination and motion artefacts [66]. With the positive results demonstrated in this study, researchers in the future could focus on a fusion approach that leverages the robustness of our approach to reduce noise in rPPG; thus, achieving a reliable HRV measurement in the wild. At the same time, future work could extend our work by investigating the feasibility of applying the proposed facial expression-based HRV estimation outside the vehicle. For example, a potential use case could be patients with cardiovascular diseases who need low-cost monitoring of their current condition. The mandatory step to validate our approach would be the collection of medical data from affected patients and subsequent experiments on this data.

VI. Conclusion and Outlook

Several studies and surveys pointed out that the sub-optimal state of drivers is the main cause of traffic accidents [7]. The National Highway Traffic Safety Administration (NHTSA) suggested that 94% of accidents resulted from human errors [67]. Therefore, a strategy for monitoring drivers' status and

driving performance becomes crucial in the reduction of the number of accidents. Such a driver monitoring system is particularly meaningful in the upcoming era of ever automated vehicles, where driver status needs to be maintained to ensure a seamless takeover of the control of cars. Although several HRV estimation approaches have been proposed, the mediocre accuracy, inconvenient deployment and the lack of ubiquity prevent them from becoming a practical and prevalent solution.

To address the existing challenges and embrace future technologies, we proposed a facial expression-based approach for HRV measure outlier detection. The reason is that empirical research showed that excessively low or high HRV measures are strongly correlated with various sub-optimal mental and psychological states of people [15], [16], [19], [20]. The merit of the proposed approach is three-fold. First, HRV estimation is a meaningful and even necessary complement to visual human activity recognition (HAR) based driver monitoring. While HAR captures drivers' physical behaviours, HRV estimation evaluates their mental status. Second, driver monitoring cameras will become a mandatory component of future vehicles in many regions. Therefore the proposed approach does not induce any extra hardware cost, providing a higher degree of ubiquity than smartwatches and UWB based technologies. Our evaluation demonstrates that the proposed tree-based probabilistic fusion network approach outperforms a consumer smartwatch in HRV measure outlier detection by up to 13.1% in terms of balanced accuracy. The positive results and the ubiquity of the proposed approach demonstrated its great potential in improving driving experience and safety. Finally, the proposed tree-based probabilistic fusion network approach outperforms other prevalent pure tree-based or neural network based methods by an average of 3.4% in balanced accuracy. The idea of the tree-based probabilistic embedding should inspire researchers to consider the possibility of hybrid models that leverages the merits of the tree-based models, especially when no rich prior domain knowledge is available.

The concept of facial expression-based estimation of HRV measures proposed in this work could further facilitate various IoT-based services and applications. For example, in mobile crowd sensing [68], [69], car ridesharing companies (Uber, Didi, etc.) could determine whether a driver is an optimal state based on the proposed HRV estimation approach. After that, task allocation can be optimised by assigning more demanding tasks to the drivers of better states or enforcing mandatory pause to the drivers who are temporally not fit for working. Thus, the quality of service will be improved. Another example is smartphone-based mobile sensing of user physiological and psychological states. One major limitation of smartphone-based sensing is the lack of accurate physiological data [31], [50], [70]. With the help of the proposed method, users' HRV estimation can be shared via data link between smartphones and the devices that capture facial expressions (*e.g.*, intelligent vehicles, webcam of laptops, and surveillance cameras). In this way, smartphone-based mobile sensing can achieve a more comprehensive understanding of users' status. The method proposed in this work, in essence, conceptualises a more robust and more accurate way of pervasive monitoring of users' mental states. The concept targets "IoT Data Analytical Services", one of the ten main challenges in developing an IoT service outlined by Bouguettaya *et al.* [71]. The purpose of IoT data analytics is to distil heterogeneous IoT data in order to provide domain-specific actionable knowledge of adequate quality [71]. In our vision, the facial expression-based HRV estimation of users should not only be limited to drivers, but can also be generalised to broader applications where users' mental state should be considered. As such, we expect to see interdisciplinary research from psychology, neuroscience, and computer science could benefit from our idea and further push forward the pervasive sensing of user status.

REFERENCES

[1] Statistisches Bundesamt, "Ergebnisse des Mikrozensus 2016. Report Report EVAS-Nr.12211," accessed: 2020-05-10. [Online]. Available: https://www.destatis.de

[2] American Automobile Association (AAA) Foundation for Traffic Safety, "New american driving survey: Updated methodology and results from july 2019 to june 2020," accessed: 2021-08-09. [Online]. Available: https://aaafoundation.org/new-american-driving-survey-updated-methodology-and-results-from-\#july-2019-to-june-2020/

[3] Robert Bosch GmbH, "The car as 3rd living space," accessed: 2021-06-04. [Online]. Available: https://www.bosch.com/stories/the-car-as-3rd-living-space/

[4] A. Stutzer and B. S. Frey, "Stress that doesn't pay: The commuting paradox," *The Scandinavian Journal of Economics*, vol. 110, no. 2, pp. 339–366, 2008.

[5] S. Zepf, M. Dittrich, J. Hernandez, and A. Schmitt, "Towards empathetic car interfaces: emotional triggers while driving," in *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2019.

[6] K. Chatterjee, S. Chng, B. Clark, A. Davis, J. D. Vos, D. Ettema, S. Handy, A. Martin, and L. Reardon, "Commuting and wellbeing: a critical overview of the literature with implications for policy and future research," *Transport Reviews*, vol. 40, no. 1, pp. 5–34, 2020.

[7] Y. Choi, S. I. Han, S.-H. Kong, and H. Ko, "Driver status monitoring systems for smart vehicles using physiological sensors: A safety enhancement system from automobile manufacturers," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 22–34, 2016.

[8] H. Sarker, M. Sharmin, A. A. Ali, M. M. Rahman, R. Bari, S. M. Hossain, and S. Kumar, "Assessing the availability of users to engage in just-in-time intervention in the natural environment," in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 909–920.

[9] K. Koch, V. Tiefenbeck, S. Liu, T. Berger, E. Fleisch, and F. Wortmann, "Taking mental health & well-being to the streets: An exploratory evaluation of in-vehicle interventions in the wild," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2021.

[10] J. Lee, N. Elhaouij, and R. Picard, "Ambientbreath: Unobtrusive just-in-time breathing intervention using multi-sensory stimulation and its evaluation in a car simulator," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, 2021.

[11] K. Koch, V. Mishra, S. Liu, T. Berger, E. Fleisch, D. Kotz, and F. Wortmann, "When do drivers interact with in-vehicle well-being interventions? an exploratory analysis of a longitudinal study on public roads," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, 2021.

[12] National Highway Traffic Safety Administration, "Automated vehicles for safety," accessed: 2021-04-23. [Online]. Available: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety\#topic-road-self-driving

[13] E. Aarts and B. de Ruyter, "New research perspectives on ambient intelligence," *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 1, pp. 5–14, 2009.

[14] J. F. Coughlin, B. Reimer, and B. Mehler, "Monitoring, managing, and motivating driver safety and well-being," *IEEE Pervasive Computing*, vol. 10, no. 3, pp. 14–21, 2011.

[15] D. Hernando, S. Roca, J. Sancho, A. Alesanco, and R. Bailon, "Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects," *Sensors*, vol. 18, no. 8, 2018.

[16] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, no. 3, pp. 235–245, 2018.

[17] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Frontiers in Human Neuroscience*, vol. 13, 2019.

[18] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of mental stress on heart rate and heart rate variability," in *Proceedings of the European Conference of the International Federation for Medical and Biological Engineering*, 2009, pp. 1366–1369.

[19] M. Patel, S. Lal, D. Kavanagh, and P. Rossiter, "Applying neural network analysis on heart rate variability data to assess driver fatigue," *Expert systems with Applications*, vol. 38, no. 6, pp. 7235–7242, 2011.

[20] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Drowsiness detection using heart rate variability," *Medical & Biological Engineering & Computing*, vol. 54, no. 6, pp. 927–937, 2016.

[21] L. T. D'Angelo, J. Parlow, W. Spiessl, S. Hoch, and T. C. Lueth, "Unobtrusive in-car vital parameter acquisition and processing," *Ambient Assisted Living*, pp. 257–271, 2011.

[22] M. Osaka, "Customized heart check system by using integrated information of electrocardiogram and plethysmogram outside the driver's awareness from an automobile steering wheel," accessed: 2021-06-04. [Online]. Available: http://cdn.intechopen.com/pdfs-wm/27021.pdf

[23] M. Osaka, H. Murata, Y. Fuwamoto, S. Nanba, K. Sakai, and T. Katoh, "Application of heart rate variability analysis to electrocardiogram recorded outside the driver's awareness from an automobile steering wheel," *Circulation Journal*, vol. 72, pp. 1867–1873, 2008.

[24] M. Walter, B. Eilebrecht, T. Wartzek, and S. Leonhardt, "The smart car seat: personalized monitoring of vital signs in automotive applications," *Personal and Ubiquitous Computing*, vol. 15, no. 7, pp. 707–715, 2011.

[25] T. Wartzek, B. Eilebrecht, J. Lem, H.-J. Lindner, S. Leonhardt, and M. Walter, "Ecg on the road: Robust and unobtrusive estimation of heart rate," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 11, pp. 3112–3120, 2011.

[26] L. T. D'Angelo, J. Parlow, W. Spiessl, S. Hoch, and T. C. Lüth, "A system for unobtrusive in-car vital parameter acquisition and processing," in *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*, 2010, pp. 1–7.

[27] T. Zheng, Z. Chen, C. Cai, J. Luo, and X. Zhang, "V2ifi: In-vehicle vital sign monitoring via compact rf sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, 2020.

[28] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 2957–2960.

[29] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

[30] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 1056–1062.

[31] T. S. Buda, M. Khwaja, and A. Matic, "Outliers in smartphone sensor data reveal outliers in daily happiness," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–19, 2021.

[32] J. M. Smyth and K. E. Heron, "Is providing mobile interventions "just-in-time" helpful? an experimental proof of concept study of just-in-time intervention for stress management," in *Proceedings of the IEEE Wireless Health*, 2016, pp. 1–7.

[33] B. F. Robinson, S. E. Epstein, G. D. Beiser, and E. Braunwald, "Control of heart rate by the autonomic nervous system. studies in man on the interrelation between baroreceptor mechanisms and exercise," *Circulation Research*, vol. 19, no. 2, pp. 400–411, 1966.

[34] R. Isaacson, *The limbic system*. Springer Science & Business Media, 2013.

[35] Official Journal of the European Union, "REGULATION (EU) 2019/2144 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," accessed: 2021-08-09. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R2144&from=EN\#d1e32-24-1

[36] European Commission, "Report on advanced driver distraction warning systems," accessed: 2021-08-09. [Online]. Available: https://ec.europa.eu/docsroom/documents/45901?locale=en

[37] Firstbeat Technologies Oy, "Firstbeat bodyguard 2," accessed: 2020-05-10. [Online]. Available: https://international-shop.firstbeat.com/product/bodyguard-2/

[38] M. Malik, A. J. Camm, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger, F. Lombardi, A. Malliani, A. J. Moss, J. N. Rottman, G. Schmidt, P. J. Schwartz, and D. H. Singer, "Heart rate variability. standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[39] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, 2017.

[40] S. Liu, K. Koch, Z. Zhou, S. Föll, X. He, T. Menke, E. Fleisch, and F. Wortmann, "The empathetic car: Exploring emotion inference via driver behaviour and traffic context," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, 2021.

[41] S. Föll, M. Maritsch, F. Spinola, V. Mishra, F. Barata, T. Kowatsch, E. Fleisch, and F. Wortmann, "Flirt: A feature generation toolkit for wearable data," *Computer Methods and Programs in Biomedicine*, p. 106461, 2021.

[42] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press, USA, 1997.

[43] A. Mcmanus, "Affectiva automotive ai," 2020, accessed: 2020-10-28. [Online]. Available: https://go.affectiva.com/auto

[44] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[45] S. Wang, C. Aggarwal, and H. Liu, "Using a random forest to inspire a neural network and improving on it," in *Proceedings of the SIAM International Conference on Data Mining*, 2017, pp. 1–9.

[46] Z.-H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017, pp. 3553–3559.

[47] A. Shcherbina, C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler, and E. A. Ashley, "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," *Journal of Personalized Medicine*, vol. 7, no. 2, p. 3, 2017.

[48] H. Tsuji, F. J. Venditti Jr, E. S. Manders, J. C. Evans, M. G. Larson, C. L. Feldman, and D. Levy, "Determinants of heart rate variability," *Journal of the American College of Cardiology*, vol. 28, no. 6, pp. 1539–1546, 1996.

[49] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: An exploratory study," in *Proceedings of the ACM International Symposium on Wearable Computers*, 2019, pp. 211–219.

[50] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2017.

[51] F. Lyrheden, "Driver monitoring (dms) on its way to becoming mandatory in vehicles around the world," accessed: 2021-07-19. [Online]. Available: https://smarteye.se/blogs

[52] European New Car Assessment Programme, "Euro ncap 2025 roadmap," 2021, accessed: 2021-07-19. [Online]. Available: https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf

[53] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F.-Y. Wang, "Driver activity recognition for intelligent vehicles: A deep learning approach," *IEEE transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5379–5390, 2019.

[54] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014–2027, 2015.

[55] M. Sodhi, B. Reimer, and I. Llamazares, "Glance analysis of driver eye movements to evaluate distraction," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 529–538, 2002.

[56] T. W. Victor, J. L. Harbluk, and J. A. Engström, "Sensitivity of eye-movement measures to in-vehicle task difficulty," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 2, pp. 167–190, 2005.

[57] J. Holzman and D. Bridgett, "Heart rate variability indices as biomarkers of top-down self-regulatory mechanisms: a meta-analytic review," *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 235–255, 2017.

[58] D. Dmitrenko, E. Maggioni, G. Brianza, B. E. Holthausen, B. N. Walker, and M. Obrist, "Caroma therapy: pleasant scents promote safer driving, better mood, and improved well-being in angry drivers," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2020.

[59] P. E. Paredes, Y. Zhou, N. A.-H. Hamdan, S. Balters, E. Murnane, W. Ju, and J. A. Landay, "Just breathe: In-car interventions for guided slow breathing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, 2018.

[60] S. Balters, M. L. Mauriello, S. Y. Park, J. A. Landay, and P. E. Paredes, "Calm commute: Guided slow breathing for daily stress management in drivers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, 2020.

[61] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-ai interaction," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019.

[62] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 217–226.

[63] I. Nahum-Shani, E. B. Hekler, and D. Spruijt-Metz, "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework." *Health Psychology*, vol. 34, no. S, 2015.

[64] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[65] T. E. Brown, L. A. Beightol, J. Koh, and D. L. Eckberg, "Important influence of respiration on human r-r interval power spectra is largely ignored," *Journal of Applied Physiology*, vol. 75, no. 5, pp. 2310–2317, 1993.

[66] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[67] National Highway Traffic Safety Administration, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," accessed: 2021-07-19. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506

[68] J. Wang, L. Wang, Y. Wang, D. Zhang, and L. Kong, "Task allocation in mobile crowd sensing: State-of-the-art and future opportunities," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3747–3757, 2018.

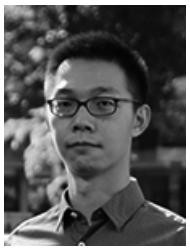
[69] J. Wang, F. Wang, Y. Wang, L. Wang, Z. Qiu, D. Zhang, B. Guo, and Q. Lv, "Hytasker: Hybrid task allocation in mobile crowd sensing," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 598–611, 2020.

[70] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, 2018.

[71] A. Bouguettaya, Q. Sheng, B. Benatallah, A. Ghar-Neiat, S. Mistry, A. Ghose, S. Nepal, and L. Yao, "An internet of things service roadmap," *Communications of the ACM*, 2021, to appear.

Shu Liu is a PhD candidate at the Bosch IoT Lab at ETH Zurich. His research interest includes the area of Machine Learning and Computer Vision in the context of Internet of Things / V2V and Autonomous Driving. He holds a M.Sc degree in Electrical Engineering from ETH Zurich and a B.Sc degree from TU Munich. Before joining the Bosch IoT Lab he worked as a Research Assistant at the Computer Vision and Geometry Group of ETH Zurich in cooperation with MIT.

Kevin Koch is a PhD candidate at the Bosch IoT Lab at the University of St.Gallen (HSG). His research focuses on the potential of the connected car for driver state analysis and driving performance. Kevin holds a Diploma (M.Sc.) in Information Systems from Technische Universität München (TUM) and a bachelor degree (B.Sc.) in Business Information Technology from Frankfurt School of Finance & Management. In addition, he completed several international stays, including a research stay at the National Institute of Informatics in Tokyo (NII).



Zimu Zhou received his B.E. in 2011 in the Department of Electronic Engineering, Tsinghua University and his Ph.D. in 2015 in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He is currently an assistant professor at the School of Information Systems, Singapore Management University. His research interests include mobile and ubiquitous computing.

Martin Maritsch is a PhD candidate at the Bosch IoT Lab at ETH Zurich. His research interests include in-vehicle sensor data analytics, physiological signal processing, and explainable/interpretable machine learning. He studied at Aalto University, Helsinki and Graz University of Technology and holds a diploma (MSc) in Software Engineering & Management. Prior to joining the Bosch IoT Lab he worked as a researcher in setting up a global, high-secure IoT platform for industrial equipment in the automotive sector.

Xiaoxi He received his B.Sc. degree in electrical engineering and information technology from Leibniz University Hannover in 2015, and the M.Sc degree from ETH Zurich in 2017. He has been a Ph.D. Candidate with the group of Prof. L. Thiele, Computer Engineering and Networks Laboratory, ETH Zurich, since 2017. His current research interests include deep learning theory and machine learning applications in mobile embedded system.

Elgar Fleisch is professor of Information and Technology Management at the ETH Zurich (MTEC) and the University of St. Gallen (HSG) and Director of the Institute of Technology Management (ITEM-HSG). His research interests focus on the current fusion of the physical and digital world into an Internet of Things. With his transdisciplinary team, his goal is to understand this fusion in the dimensions of technology, applications and social implications and, based on this, to develop new technologies and applications for the benefit of the economy and society. Elgar Fleisch is co-founder of several ETH and HSG spin-off companies and member of various academic steering committees and supervisory boards.

Felix Wortmann is professor, senior lecturer, and scientific director of the Bosch IoT Lab, a research collaboration between Bosch, the University of St. Gallen and ETH Zurich. His research interests include the Internet of Things, machine learning, blockchain, and digital innovation in mobility, energy, and health. Felix Wortmann received a BScIS and MScIS from the University of Münster, Germany, and a PhD in Management from the University of St. Gallen. He gained several years of industry experience in a German-based multinational software corporation.

APPENDIX

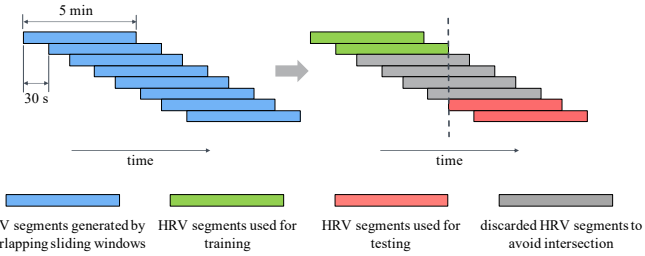


Fig. 16. Processing of HRV segments to avoid intersection between training and test data

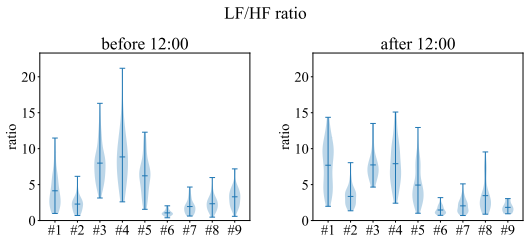


Fig. 17. LF/HF ratio of the nine drivers in different time interval

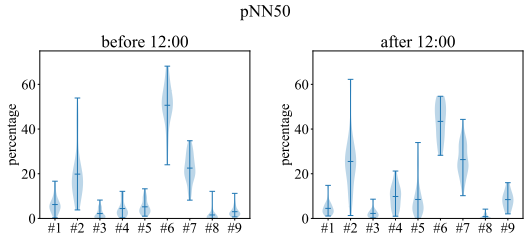
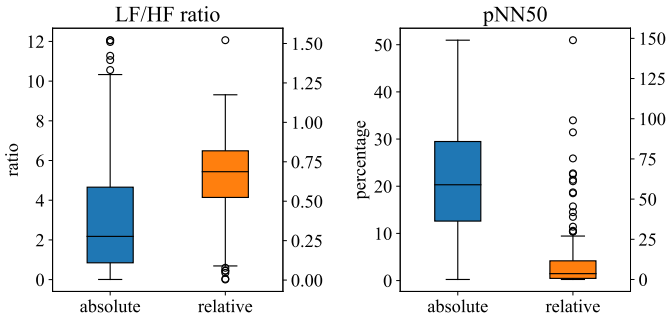


Fig. 18. pNN50 of the nine drivers in different time interval

TABLE VI
CANDIDATE PARAMETERS FOR GRID SEARCH FOR TREE-BASED MODELS

RF, DF, TFPN	
depth of tree	10, 20, 50, None
number of trees	50, 100, 200, 300
min samples split	2, 5, 10
min samples leaf	1, 2, 5



(a) LF/HF ratio (b) pNN50

Fig. 19. Absolute and relative errors of high-end smartwatch compared with Firstbeat [37]

TABLE VII
CANDIDATE PARAMETERS FOR GRID SEARCH FOR CNN

CNN	
# conv. layers	2, 4, 8
# filter per layer	8, 16, 32
kernel size	3, 5, 7
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	ReLU, sigmoid
FC layer settings (neurons in each layer)	[16], [32], [64], [16,16], [32,32], [64,64]

TABLE VIII
CANDIDATE PARAMETERS FOR GRID SEARCH FOR RNN

RNN	
# layers	1, 2, 4
# hidden units	8, 16, 32
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	ReLU, sigmoid
FC layer settings (neurons in each layer)	[16], [32], [64], [16,16], [32,32], [64,64]

TABLE IX
CANDIDATE PARAMETERS FOR GRID SEARCH FOR MLP

MLP	
# layers	2, 4, 8
# neurons in layer	8, 16, 32
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	ReLU, sigmoid