Towards Classifying Human Circadian Rhythm Using Multiple Modalities

Kais Riani
Computer and Information Science
University of Michigan, Dearborn
kriani@umich.edu

Salem Sharak
Computer and Information Science
University of Michigan, Dearborn
sharak@umich.edu

Kapotaksha Das
Computer and Information Science
University of Michigan, Dearborn
takposha@umich.edu

Mohamed Abouelenien
Computer and Information Science
University of Michigan, Dearborn
zmohamed@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan, Flint
mburzo@umich.edu

Rada Mihalcea
Computer Science & Engineering
University of Michigan
mihalcea@umich.edu

John Elson
Ford Motor Company
jelson3@ford.com

Clay Maranville
Ford Motor Company
cmaranvi@ford.com

Kwaku Prakah-Asante
Ford Motor Company
kprakaha@ford.com

Waqas Manzoor
Ford Motor Company
wmanzoor@ford.com

2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-0019-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACII52823.2021.9597391

***Abstract*—Autonomous vehicles represent one of the most active technologies currently being developed, with research areas addressing, among others, the modeling of the states and behavioral elements of the occupants. This paper contributes to this line of research by studying the circadian rhythm of individuals using a novel multimodal dataset of 36 subjects consisting of five information channels. These channels include visual, thermal, physiological, linguistic, and background data. Moreover, we propose a framework to explore whether the circadian rhythm can be modeled without continuous monitoring and investigate the hypothesis that multimodal features have a greater propensity for improved performance using data points specific to certain times during the day. Our analysis shows that multimodal fusion can lead to an accuracy of up to 77% on identifying energized and enervated states of the participants. Our findings highlight the validity of our hypothesis and present a novel approach for future research.**

***Index Terms*—Multimodal dataset, circadian rhythm, action units, classification, machine learning, thermal imaging**

I. Introduction

With a total global investment in autonomous-vehicle technology exceeding $100 billion, a figure that will only rise with increased competition [1], this nascent technology is ripe for exploration and research. While there are many research elements related to autonomy which can be explored, an area which requires attention is the study of the state and behaviors of the occupants of an autonomous car. Understanding and analyzing the state of an occupant allow for the development of vehicles that can better cater to the needs of the travelers, enhancing the wellbeing and comfort of all involved parties.

One such state that can be modeled is the occupants' circadian rhythm in order to address the adverse effects of being enervated [2], which is an important attribute for an autonomous vehicle to monitor. As stated by [3], the Circadian rhythm refers to biological variations or rhythms with a cycle of approximately 24 hours that will persist even when the organism is placed in an environment devoid of time cues. For comparison, Diurnal rhythm is similar to Circadian rhythm in that it varies according to the time of day; however, it may or may not persist without external time cues. Practically speaking, there is little reason to distinguish between the two, as almost all diurnal rhythms are found to be circadian [3]. Therefore, we will be using the term Circadian rhythm.

By better understanding an occupant's state, the vehicle could rectify its occupants' enervated state by engaging in a conversation with them, by providing recommendations, by adjusting ride comfort, lighting, and audio settings to accommodate for a sleeping child, or the automatic adjustment of the cabin environment to maximize the rest of its passengers over long trips. These applications could benefit from a better understanding of the subjects' circadian rhythm, with an understanding of relevant visual and physiological cues. Studies such as [4], [5], and [6] have explored the detection of circadian state, but have mostly done so in a lab setting under continuous observation, and mostly in the domains of medicine, psychology, and the study of sleep disorders.

In this paper, we present a multimodal system that observes the circadian rhythm of a participating subject and identifies energetic or enervated periods through a series of short recordings. Detection of circadian rhythm using two data points has been performed by [7], where the circadian rhythm was estimated as the rate of change per day in the timing of the peak metabolite aMT6s between two urine collections. Others, such as [8], collected hair follicles at three times without sleep interruption and evaluated the circadian phase based on clock gene expression.

Our main contributions are three fold. Firstly, we introduce a novel multimodal dataset of 36 subjects consisting of surveys

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

collected at every recording session, thermal video recordings, RGB videos, audio recordings, and a four-sensor physiological feed including heart rate, skin conductance, respiration rate, and skin temperature. Secondly, this is the first study to our knowledge to model circadian rhythm using the thermal, linguistic, physiological and visual modalities, and the first to our knowledge to model the fusion of said modalities, which provides insights on the specific features, modalities, and fusions that are most capable of detecting an individual's energetic and enervated states. Finally, we analyze and evaluate our approach using demographic and behavioral features produced via a series of surveys.

II. Related Work

Among human behaviors, researchers focused on analyzing fatigue and sleep detection using both invasive and non-invasive modalities. The method proposed by Koichi Fujiwara et al. [9] is based on heart rate variability (HRV) analysis, which was validated by comparing with the (EEG)-based sleep scoring. RR interval (RRI) measurements were taken from 34 participants, and their sleep onsets were determined by a sleep specialist using the EEG data.

Classical non-invasive approaches used one or several of the following visual features for sleepiness detection: eye position and blinking, eye closure duration, and yawning frequency [10]. Garcia et al. [11] proposed a Deep Learning framework, where Computer Vision techniques were used to classify the cropped face from each frame into two classes: ”rested” or ”sleep deprived”.

Many studies have concentrated on studying drowsiness and distraction detection without considering the personal sleep patterns of different individuals, and in particular circadian rhythm detection. A limited number of studies investigated circadian rhythm while focusing mainly on analyzing physiological data. Masuda et al. [12] used a smart wear garment to estimate the value of time and heart rate (HR) to reach the lowest point in circadian rhythm by measuring electrocardiogram (ECG) during sleep. The system has shown a great potential in verifying the effects of jet lag on circadian rhythm. Another study by Kaduk et al. [13] provided a theoretical basis for integrating circadian rhythmicity studies into driver state monitoring. Other studies have worked on detecting circadian phase using different signals such as skin temperature, light exposure, activity and body position. [14]–[18]

III. Dataset

For our experiments, we gathered our data from 36 subjects of varying ethnicity, with each subject undergoing five recordings in a tobacco, alcohol and drug-free state to detect their circadian rhythm. The dataset consists of 24 males and 12 females with different demographic background, between 18 and 32 years old. In addition to thermal, audiovisual and physiological data, ten surveys were collected at different points in the study, as listed below. There were two main recording sessions, one in the morning and one in the evening, in addition to baseline recordings taken on an earlier day.

A. Behavior and Demographics

The following data was collected for each participant.

- Demographic data (baseline): An eight-question survey that collects background information, including race/ethnicity, sex at birth, gender, and age.
- Karolinska Sleep Questionnaire (baseline): The KSQ is a 31-question survey used to measure subjective sleep and sleepiness [19].
- Karolinska Sleepiness Scale (morning and evening): The KSS is a one-question survey that scores the level of sleepiness at the time of recording. The survey was taken twice: during the morning recording, and at a later hour, typically before going home. [20].
- Munich Chronotype Questionnaire (baseline): The MCQ has 14 questions that use midpoint between sleep on- and offset on non-workdays to assess chronotype [21].
- Big Five Inventory: The BFI (baseline) is a 44-item inventory that evaluates the subject using the Big Five factors of personality [22]–[24].
- Profile of Mood States 40 (morning and evening): The POMS40 Questionnaire is a 40-item psychological rating scale used to assess transient and distinct mood states. The survey is taken twice: during the morning and evening recordings [25].
- Pittsburgh Sleep Quality Index (evening): The PSQI is a 19-question self-rated questionnaire which assesses sleep quality and disturbances [26].
- Drug & Drinking Survey (baseline): A 17-question survey that identifies drug and drinking habits in the subjects, including caffeine usage, as well as their opinions on the harm caused by such drugs.
- Morningness-Eveningness Questionnaire (baseline): The MEQ, also known as the Owl and Lark Questionnaire is a 19-question survey that resolves the subject's chronotype by measuring times of peak alertness [27].
- Open Response (evening), which is a one-question survey asking subjects if they felt they had deviated from their normal drinking, sleeping, exercise, or smoking schedules over the previous week.

A small enclosed recording station was used to resemble to a certain extent a vehicle's environment. During each recording the following devices were used to collect visual, acoustic, thermal, physiological and linguistic data:

- Top-view Logitech's HD web camera, capturing 30 fps.
- Face close-up RGB Raspberry-pi camera at 25 fps.
- Low quality FLIR One thermal camera, recording the face of the subject at 7 fps.
- High-resolution FLIR SC6700 thermal camera, recording the subject's face at 100 fps, with a resolution of 640x512 pixels and 7.2M electrons.
- Four physiological sensors from Thought Technology Ltd.; Blood Volume Pulse (BVP) Sensor, Skin Temperature Sensor, Skin Conductance Flex/Pro Sensor, and Respiration Rate Sensor [28].
- A microphone recording speech during Active recordings.

B. Scenarios

The subjects were asked to schedule the first recording (baseline) at least three days prior to the remaining recordings with the intent of recording their baseline data. Before starting the two-minute baseline recording using our system of cameras and sensors, each participant took the KSQ, MCQ, MEQ Questionnaires, and the Drug and Drinking Survey. Following the recording, we asked the participants to sit still and breathe naturally for two minutes. Afterwards, they were asked to take the BFI and the Demographic Survey.

Following that, we held two sessions on a different day that was at least three days after the baseline recording. The subjects were asked to abstain from consuming any caffeine products the day of the recordings and the night before. One session took place in the morning, sometime between 8 AM to 11 AM, with a few cases around noon, as we asked the subjects to have their first session of the recordings within one hour after they woke up. The second session occurred later in the day, between 4 PM and 8 PM, typically before going home, with one case between 10 PM and 11 PM, based on when they woke up that day. The subjects were asked not to sleep between the two sessions.

Our hypothesis is that the subjects would be energized as soon as they wake up, while on the other hand, they start becoming enervated in the evening after a long active day, as supported by the control scenario in [29]. Each session lasted on average 20 minutes and consisted of two recordings: Silent and Active. At the end of each session, the subjects were asked again to take different surveys. At the end of the morning session, the participants took the KSS and POMS40 surveys, while in the evening session, the participants completed the PSQI, POMS40, and KSS surveys, and the open response. There was no sign of survey fatigue among the participants.

During the Silent recording, participants were asked to sit still, breathe naturally while staring at an image reflecting the time of the recording, for two minutes. Whereas the Active recording lasted for five minutes while allowing the subjects to speak freely on any topic they have in mind. The team handling the recordings left the lab during this time to allow the participant to speak freely.

IV. Methodology

A. Thermal Features

The thermal videos were analyzed following three main processing steps, based on the method proposed in [30]. First, we segmented the subjects' frame into five different regions, including the whole face, forehead, eyes, cheeks, and nose. Afterwards, we tracked these five regions throughout the recording by applying the tracking algorithm proposed in [31]. Finally, for each region of interest, we formed a thermal map by extracting statistical features.

We manually located the five ROIs in the first frame of each of the video recordings. Subsequently, we automatically captured points of interest in the detected ROIs, located where sharper changes in the colors existed, using a variation of the Shi-Tomasi corner detection algorithm [32].

A fast version of the Kanade–Lucas–Tomasi (KLT) tracking method [31] was applied on the detected points during the duration of the videos to stabilize the ROI bounding box. The points of interest were tracked by estimating the displacement between two successive frames. Afterwards, we applied geometric transformation [33] in order to map the interesting points between the frames by estimating their transformation based on similarity. As a precaution, we set a threshold of 95% of successfully mapped points between two consecutive frames to account for potential occlusion. In the presence of an occlusion, the frame was skipped and tracking continued at the point where the occlusion ended.

Finally, we created a thermal map that represented the thermal distribution in the ROI following these steps of ROI segmentation, Segment binarization, Image masking, and finally Thermal map cropping for each ROI. This process is illustrated in Fig. 1.

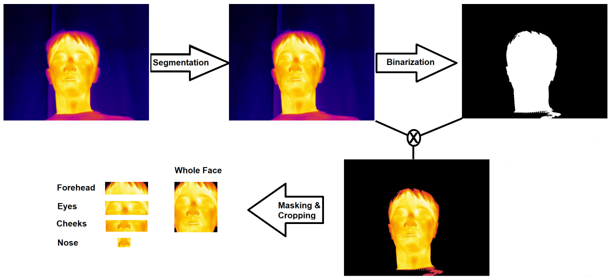


Fig. 1. Segmenting, binarizing, masking and cropping the thermal ROI.

The map consisted of the pixel mean, the minimum temperature, highest pixel value, the difference between maximum and minimum temperatures, and a 20-bin histogram over the pixels in the ROI that corresponds to the temperature distribution in the region, therefore, each region ended up with 24 thermal features. The final feature vector consisted of 120 features.

B. Visual Features

We extracted our visual features that represented the facial behavior of the subjects using the OpenFace library [34]. A Constrained Local Model was applied to extract facial and head positioning features based on a Point Distribution Model that is used to model the shape of a face, as well as local detectors that are used to detect landmark alignment at a specific pixel location [35]. The final feature vector covered facial landmark positions and head pose, eye gaze, and the presence and intensity values of Action Units (AUs).

After the detection of landmarks in the eyes region, OpenFace utilizes the intersection of the eyeball sphere with a ray passing through the pupil in order to estimate the position of the pupil, which is used to specify the eye gaze.

AUs were determined using the Facial Action Coding System (FACS). FACS is a framework that defines broad facial movements based on minute muscle movements. Built on FACS, AUs could be considered as a grouping of facial muscle movements. In that way, AUs present facial expressions with moderate complexity, offering broader understanding than what its individual constituents provide, while landmarks define details such as the layout and shape of the face [36].

OpenFace defined 18 distinct AUs along with their presence and intensity. These AUs include brow (AUs 1, 2, 4), lid (AUs 5, 7), cheek (AU 6), lip (AUs 10, 12, 15, 20, 23, 25, 28), jaw (AU 26), dimpler (AU 14), chin (AU 17), nose (AU 9), and blinking (AU 45). Presence and absence are coded as zero and one, respectively, and intensity is defined on a scale of zero (no presence) to five (highest intensity). The final feature vector for each recording consisted of a total of 709 features.

C. Physiological Features

Features from four sensors: Blood Volume Pulse (BVP), Skin Conductance, Skin Temperature, Respiration Rate were recorded using hardware from Thought Technology Ltd and then processed to extract statistical features. BVP was recorded at a sampling rate of 2048 Hz, while the other three modalities were sampled at a true rate of 256 Hz, which was upscaled to 2048 Hz to maintain consistency across the sensor feeds.

From these four sensors, 71 statistical features were derived in total, starting with 49 derived features from BVP, consisting of: Time domain statistical features such as mean, minimum, maximum and standard deviation which describe the overall behavior of the signal and also the relation between consecutive inter-beat interval (IBI), NN and pNN features that describe patterns in the interval between two normal heartbeats, and additional features which describe the spectral power statistics for very-low, low, and high frequency bands for which individual sets of statistical features were computed.

Six time-domain based statistical features for each of the three other sensors were also computed. Additionally, four features that describe the combined statistical patterns for BVP and Respiration Rate were computed. After feature extraction was completed, the average value for each feature for a given subject's recording was taken to reduce a time-series dataset into a single feature vector.

D. Linguistic Features

Audio recordings during the two Active segments per subject of approximately five minutes each were taken and transcribed to provide a total of 72 text transcripts, two per subject. The transcriptions were ensured to be consistent amongst each other to improve the quality of features that can be extracted. A variety of features were extracted from each transcript after tokenizing the text. A set of 58 features was derived in total, consisting of:

- Word density and sentence density measures, which were calculated by finding features, such as the number of words spoken, the number of sentences in the recording and the average number of words per sentence,
- Measures of the number and ratio of fillers used, split into filler sounds such as 'Um', 'Uh' and so on, filler words such as 'like', 'basically' and so on, and filler phrases such as 'kind of', 'I mean' and so on,
- Different 'end of sentence' punctuation, such as '.', ',', '?' and so on,
- Frequency counts for Part-Of-Speech Tags for the transcripts, and
- Polarity and subjectivity of the transcripts.

These features were extracted using NLTK and TextBlob libraries, to obtain a total of 58 features per transcript. In contrast to other modalities, there is no Silent recording from which linguistic data can be obtained. Hence, linguistic features extracted here should be associated with Active recordings of other modalities during multimodal classification.

V. Classification

There are two primary targets we used for defining the labels against which the features were classified and evaluated. The first one is based on the Time hypothesis (hereon referred to as the Time label), and the other one is based on the KSS set of labels (hereon referred to as the KSS labels) obtained from the results of the KSS survey. The Time label is derived based on the assumption that subjects were energized in the morning, within an hour after they woke up, and enervated in the evening, as discussed previously in Section III. For the KSS Label, we have a KSS score per session in the range of one to nine, where one indicates that the subject was most energized, and nine indicates the greatest enervation. We modeled this range into a binary or trinary problem, as seen in Fig. 2. This gave us two binary selections, and two trinary selections. In the '4-Drop 1-4' selection, recordings that received a score of five were discarded from this selection, on the basis that they were neither energized nor enervated recordings. The reasoning for having a third class was based on the fact that the scores in the middle of the scale represented very slightly energized/enervated individuals and hence can be categorized as a neutral third class with potentially distinct classification patterns as well. Hence, we have a total of five target label types that can be assigned to each recording, one binary label based on the Time hypothesis, and two binary and two trinary labels derived from different selections made from the KSS score range.

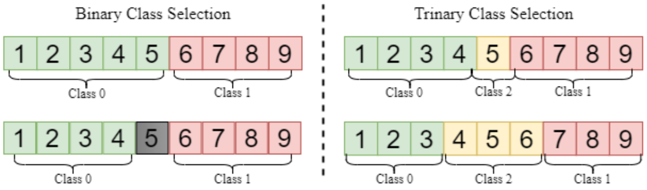


Fig. 2. Schematic representation for splitting KSS scores for Binary and Trinary classification.

Certain recordings were absent in the dataset as follows: One subject was missing a morning (Silent) recording in the physiological modality, two morning recordings missing in the visual modality, and all recordings missing in the thermal modality. Another subject's morning recordings, as well as two more subject's evening recordings were missing in the thermal modality as well.

Classification was performed using two supervised machine learning classifiers: Random Forest Classifier (RFC),

TABLE I
CLASS DISTRIBUTION OF RECORDINGS PER MODALITY.

Modality Type	Number of Subjects	Recordings per Session	Number of recordings	Class Distribution of Recordings
Visual	36*	Active, Silent	142	70 Morning, 72 Evening
Thermal	35*	Active, Silent	134	68 Morning, 66 Evening
Physiological	36*	Active, Silent	143	72 Morning, 72 Evening
Linguistic	36	1 only per session	72	36 Morning, 36 Evening

and Extreme Gradient Boosted Machine (XGB). Performance was evaluated using Leave-One-Subject-Out Cross Validation, which means that the training set excluded one subject's recording set at a time, with that subject's recording used for testing for the given fold. A baseline classification using random guessing was used to establish baseline metrics.

Classification was carried out using two approaches, unimodal classification and multimodal fusion. In unimodal classification, features from one modality at a time were used separately for classification. Given the nature of the demographic modality, being the same for a subject regardless of the time of recording, it was not used in unimodal classification. Multimodal fusion used features from more than one modality at a time. However, the methodology for how these features were combined differ, which can be further specified into early fusion, late fusion, and a combination of early and late fusion being dubbed as Meta fusion hereon.

Early fusion involved merging all features from all modalities before training a classifier. In this context, for every unique recording, features from each modality will be fused to form a larger feature vector per recording. Late fusion used multiple classifiers rather than using a single classifier. In this case, data from each modality is trained separately using distinct classifiers. The decisions of the classifiers across different modalities were then used in a simple majority voting system to give a final prediction for a given test recording.

Meta fusion, as shown in Fig. 3, is a combination of the early and late fusion approaches discussed above. The fusion occurs in two stages. Firstly, from the distinct individual modalities, a new set of distinct fusions were generated. For example, by taking four modalities V (visual), P (physiological), T (thermal and L (linguistic) to start with. The first stage then involved generating new fusions of the modalities, hereby generating VPTL, a fusion of features from all four modalities, VPL, a fusion involving three modalities, and TL, a fusion with two modalities.

In the first stage, a set of predictions were generated for each fusion using Leave-One-Subject-Out Cross Validation. In the second stage, a set of these fusions were selected, and by using majority voting on their decisions, final predictions were generated. The selection was based on the best performing final prediction set, generated initially by going through all possible combinations of the first stage fusions.

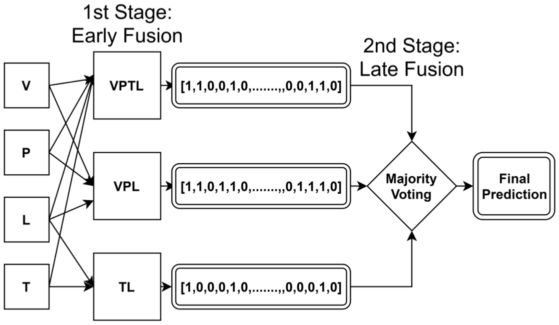


Fig. 3. Schematic representation for meta fusion.

VI. EXPERIMENTAL RESULTS

A. Classification Results

First we assessed the performance of each modality individually. Fig. 4 charts the performance of a single modality against one of the five classification tasks. The thermal modality has the best performance of approximately 65% accuracy using the Random Forest classifier with the Morning/Evening Time labels. However, for the KSS labels, the visual modality is the best performer with approximately 61% accuracy when using the same classifier. Note that the linguistic domain only used data from Active recordings.

We observed that the classifiers perform poorly when working with trinary labels and their performance was not significantly higher than the baseline performance. The exception to this is the physiological domain, where the the Gradient Boosted classifier performs with F1 scores of above approximately 42%, better by a margin of around 8% compared to the thermal, linguistic, and visual modalities. Interestingly, the physiological modality has the poorest performance for the Time labels amongst all the four modalities used, which seems to indicate that this modality is better suited towards survey-based labels for classification rather than time-based labels. Another observation is that using KSS for binary classification, where neither the energized nor the enervated class were dropped, achieved better performance than using all three classes.

Early fusion is used for the Active recordings only due to the absence of the linguistic modality for the Silent recordings. In early fusion, as seen in Fig. 5, we see a marked improvement in the evaluation metrics using the Time labels, reaching above approximately 71% in all the metrics measured. It can also be seen that trinary classification for early fusion performs poorly, which can be explained by the lack of a sufficient number of data points given the presence of more classes. Early fusion does not outperform the baseline metrics in several cases.

In late fusion, as seen in Fig. 5, similar improvement trends are seen using the Time labels. Moreover, late fusion results in an improved performance compared to early fusion using binary classification, with a relative improvement of approximately 20% for the 414bi label using the Gradient Boosted classifier. However, the visual modality unimodal classification

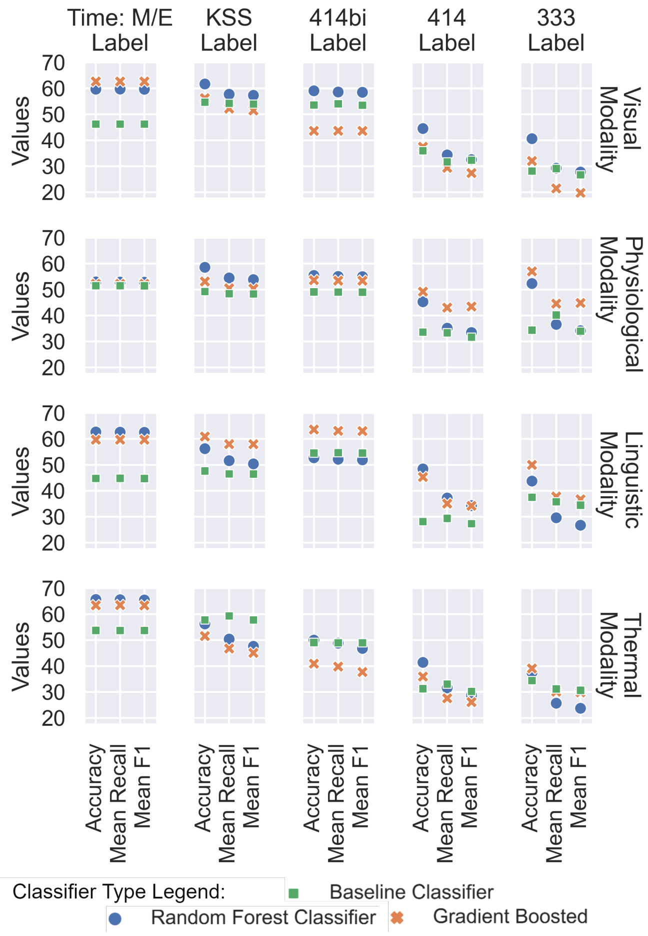


Fig. 4. Performance metrics for classification when using one modality at a time.

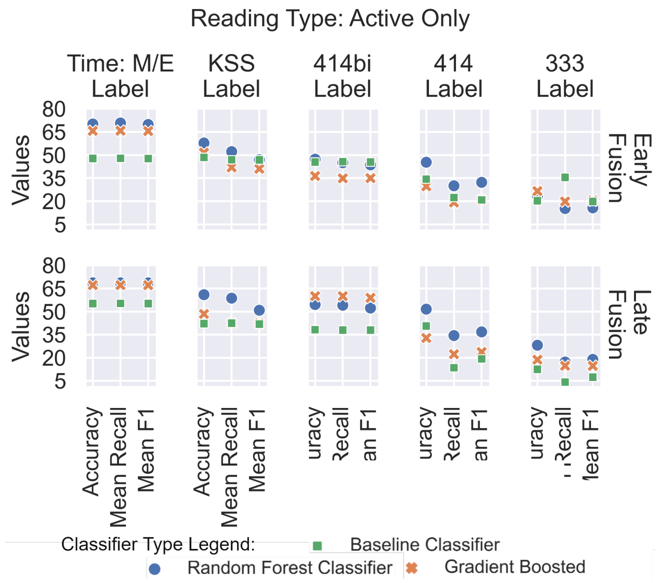


Fig. 5. Performance metrics for classification when using multimodal fusion with Active recordings.

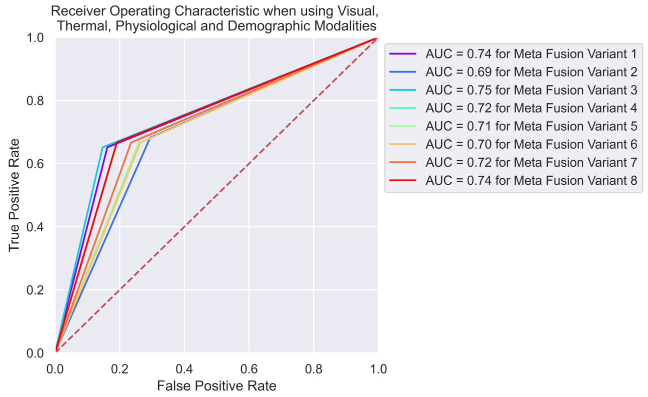


Fig. 6. AUC Curves for classification when using meta fusion with all recordings.

performs better on the KSS label when using Random Forest classification by approximately 8%. Most results indicate that multimodal fusion achieves the best improvement in performance using binary classification.

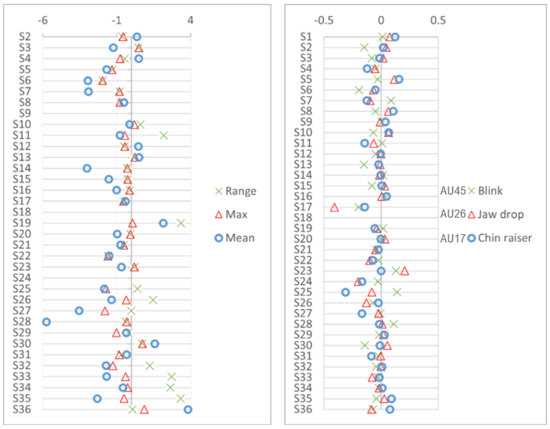
A meta fusion approach achieves the best performance compared to previous results, reaching an accuracy of approximately 77% and a recall of 79% for the best performing selection of modalities when using a Gradient Boosted classifier. The results for such different Fusion types are shown in Fig. 6, where each Fusion type represents a cascaded combination formed using multiple modalities, created using the procedure described in Section V. The fusions represent a multi-level combination of physiological, visual, thermal and demographic features. In particular, all Fusions types include a series of combinations, including the integration of all four modalities, except Fusion types 4 and 5. In addition, all fusions included a meta combination of three modalities combined, two modalities combined, and at least one single modality. It also achieves an improvement of approximately 8% over early fusion-based classification using the Time: M/E label for all four modalities combined, as seen earlier.

With all the results shown, we can see the measurable improvement in the performance metrics when using multimodal fusion, which is further enhanced using meta fusion.

B. Feature and Survey-based Analysis

Fig. 7.a shows an analysis of the mean, maximum, and range thermal features. Based on the Time labels, we calculated the graph bars by subtracting the enervated class average values from the energetic class values. Therefore, a positive value represents a high temperature in the morning session while a negative value represents a correlation between higher temperatures and an enervated subject. The resulting figure shows interesting patterns as the majority of subjects have higher temperatures in the enervated session, which agrees with Vaara et al. findings [37], where the evening temperature is higher despite the overall decrease over three days.

Fig. 7.b presents an analysis on how the AUs are associated with enervated and energetic subjects. The graph bars are constructed using the same approach in Fig. 7.a with a positive value indicating an association between an AU and an energetic state, and vice versa. The resulting graph shows that enervation is more associated with blinking (AU 45), which can be used as a valuable indicator when it comes to modeling the circadian rhythm. Furthermore, for the vast majority of the subjects, the chin raiser (AU 17) and the jaw drop (AU 26) AUs are more associated with enervation.



a.) Thermal modality: Face region feature analysis

b.) Visual modality: Raspberry Pi camera AU analysis

Fig. 7. Feature analysis for the Time label using Silent recordings

Fig. 8 presents the results of two surveys. The MEQ, shown left, resolves subject chronotype through a series of questions. Results of the survey indicate a predominantly evening or neutral subject body. The KSQ includes a series of questions, one of which asks subjects to self-report their perceived chronotype. The results of this question are shown right. Fig. 9 presents the coherence between the KSS and the KSQ across subjects. The KSS collects the subject's sleepiness levels at two times: once earlier with the first session and once with the later session. Here, we contrast the difference between the morning and evening sleepiness levels of the KSS, which implies the chronotype of the subject, against the self-reported chronotype of the KSQ. In both cases, a lower value indicates greater morning energization.

After collating and scoring the survey results from all the allocated surveys, some interesting findings presented themselves. When taken together, as also can be seen in in Fig. 8 and Fig. 9, we note an incoherence between the subjects' perceived and resolved chronotype. This incoherence seems to be especially present for subjects who self-report a morning chronotype while resolving as evening-types. Given that the MEQ evaluates subject chronotype via a series of questions including time of sleep, and the KSS scores subject sleepiness at the time of recording, both surveys could be considered more objective than the referenced question from the KSQ, which only asks if the subject broadly believes they are a morning or evening-type person. The reason for the present incoherence, while interesting, is currently outside the scope of this study. It may be construed that subjects whose chronotypes are resolved as evening-type were less able to correctly identify their own chronotype; however, further focused study would be needed to verify such claims.

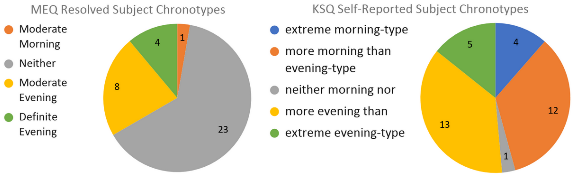


Fig. 8. MEQ and KSQ results

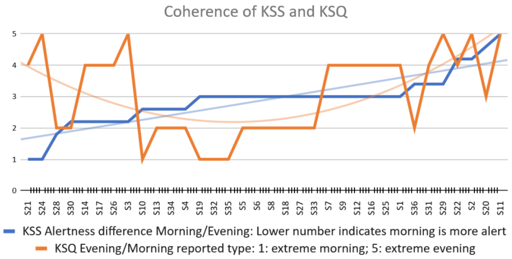


Fig. 9. KSS and KSQ coherence

VII. Conclusion

In this paper, we introduced a novel multimodal dataset for circadian rhythm detection and we proposed a data-driven, machine learning framework that presents the first attempt to model circadian rhythm using a multitude of data channels and a non-invasive approach. In addition, our framework analyzed the state and behavior of different individuals that could represent occupants of a vehicle.

Our research focused on the trade-offs that each of the explored modalities brought to the table. Our experimental results highlighted the advantages of using multimodal fusion, which resulted in relative improvements of up to 20% compared to unimodal classification. Further analysis demonstrated the improvements brought by the meta fusion approach, reaching an accuracy of approximately 77%. When considering specific features, we found that the thermal features provided the greatest indication of the circadian state. Our thermal-based analysis showed that the subjects had higher temperatures when they were enervated. Furthermore, it was observed that using the Time label based on our hypothesis achieved a significant improvement in performance compared to using self-reported indications of sleepiness levels. As part of a secondary analysis, we studied the subjects' survey responses and found indication of incoherence between certain subjects' professed chronotype and their resolved chronotype. We believe our work is a step toward understanding the state of a vehicle's occupants in order to enhance their comfort levels and well-being.

VIII. Acknowledgments

This material is based in part upon work supported by the Ford Motor Company. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Ford Motor Company or any other Ford entity.

References

[1] C. F. Kerry and J. Karsten, “Gauging investment in self-driving cars,” *Brookings Institution, October*, vol. 16, 2017.

[2] B. C. Tefft, “Acute sleep deprivation and culpable motor vehicle crash involvement,” *Sleep*, vol. 41, no. 10, 09 2018, zsy144. [Online]. Available: https://doi.org/10.1093/sleep/zsy144

[3] M. H. Vitaterna, J. S. Takahashi, and F. W. Turek, “Overview of circadian rhythms,” *Alcohol Research & Health*, vol. 25, no. 2, p. 85, 2001.

[4] R. R. Freedman, D. Norton, S. Woodward, and G. Cornélissen, “Core body temperature and circadian rhythm of hot flashes in menopausal women,” *The Journal of Clinical Endocrinology & Metabolism*, vol. 80, no. 8, pp. 2354–2358, 08 1995. [Online]. Available: https://doi.org/10.1210/jcem.80.8.7629229

[5] K. A. Thomas, R. L. Burr, and S. Spieker, “Maternal and infant activity: Analytic approaches for the study of circadian rhythm,” *Infant Behavior and Development*, vol. 41, pp. 80–87, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0163638314200403

[6] A. Rivera-Coll, X. Fuentes-Arderiu, and A. Díez-Noguera, “Circadian rhythms of serum concentrations of 12 enzymes of clinical interest,” *Chronobiology International*, vol. 10, no. 3, pp. 190–200, 1993. [Online]. Available: https://doi.org/10.3109/07420529309073887

[7] E. E. Flynn-Evans, H. Tabandeh, D. J. Skene, and S. W. Lockley, “Circadian rhythm disorders and melatonin production in 127 blind women with and without light perception,” *Journal of Biological Rhythms*, vol. 29, no. 3, pp. 215–224, 2014, pMID: 24916394. [Online]. Available: https://doi.org/10.1177/0748730414536852

[8] M. Akashi, R. Sogawa, R. Matsumura, A. Nishida, R. Nakamura, I. T. Tokuda, and K. Node, “A detection method for latent circadian rhythm sleep-wake disorder,” *EBioMedicine*, vol. 62, p. 103080, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352396420304564

[9] K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda *et al.*, “Heart rate variability-based driver drowsiness detection and its validation with eeg,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1769–1778, 2018.

[10] G. Sikander and S. Anwar, “Driver fatigue detection systems: A review,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, 2018.

[11] M. García-García, A. Caplier, and M. Rombaut, “Sleep deprivation detection for real-time driver monitoring using deep learning,” in *International conference image analysis and recognition*. Springer, 2018, pp. 435–442.

[12] H. Masuda, S. Okada, N. Shiozawa, M. Makikawa, and D. Goto, “The estimation of circadian rhythm using smart wear,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4239–4242.

[13] S. I. Kaduk, A. P. Roberts, and N. A. Stanton, “The circadian effect on psychophysiological driver state monitoring,” *Theoretical Issues in Ergonomics Science*, pp. 1–25, 2020.

[14] V. Kolodyazhniy, J. Späti, S. Frey, T. Götz, A. Wirz-Justice, K. Kräuchi, C. Cajochen, and F. H. Wilhelm, “Estimation of human circadian phase via a multi-channel ambulatory monitoring system and a multiple regression model,” *Journal of biological rhythms*, vol. 26, no. 1, pp. 55–67, 2011.

[15] ——, “An improved method for estimating human circadian phase derived from multichannel ambulatory monitoring and artificial neural networks,” *Chronobiology International*, vol. 29, no. 8, pp. 1078–1097, 2012.

[16] M. A. Bonmati-Carrion, B. Middleton, V. Revell, D. J. Skene, M. Rol, and J. A. Madrid, “Circadian phase asessment by ambulatory monitoring in humans: Correlation with dim light melatonin onset,” *Chronobiology international*, vol. 31, no. 1, pp. 37–51, 2014.

[17] J. E. Stone, A. J. Phillips, S. Ftouni, M. Magee, M. Howard, S. W. Lockley, T. L. Sletten, C. Anderson, S. M. Rajaratnam, and S. Postnova, “Generalizability of a neural network model for circadian phase prediction in real-world conditions,” *Scientific reports*, vol. 9, no. 1, pp. 1–17, 2019.

[18] P. Cheng, O. Walch, Y. Huang, C. Mayer, C. Sagong, A. Cuamatzi Castelan, H. J. Burgess, T. Roth, D. B. Forger, and C. L. Drake, “Predicting circadian misalignment with wearable technology: validation of wrist-worn actigraphy and photometry in night shift workers,” *Sleep*, vol. 44, no. 2, p. zsaa180, 2021.

[19] G. Kecklund and T. Åkerstedt, “The psychometric properties of the karolinska sleep questionnaire,” *J Sleep Res*, vol. 1, no. Suppl 1, p. 113, 1992.

[20] T. Åkerstedt and M. Gillberg, “Subjective and objective sleepiness in the active individual,” *International journal of neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[21] T. Roenneberg, A. Wirz-Justice, and M. Merrow, “Life between clocks: daily temporal patterns of human chronotypes,” *Journal of biological rhythms*, vol. 18, no. 1, pp. 80–90, 2003.

[22] O. P. John, L. P. Naumann, and C. J. Soto, “Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.” 2008.

[23] O. John, E. Donahue, and R. Kentle, *The Big Five Inventory–Versions 4a and 54*. Berkeley, CA: University of California,Berkeley.

[24] V. Benet-Martínez and O. P. John, “Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english.” *Journal of personality and social psychology*, vol. 75, no. 3, p. 729, 1998.

[25] S. L. Curran, M. A. Andrykowski, and J. L. Studts, “Short form of the profile of mood states (poms-sf): psychometric information.” *Psychological assessment*, vol. 7, no. 1, p. 80, 1995.

[26] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, “The pittsburgh sleep quality index: a new instrument for psychiatric practice and research,” *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.

[27] J. A. Horne and O. Östberg, “A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms.” *International journal of chronobiology*, 1976.

[28] BioGraph Infiniti and FlexComp Infiniti User Manual, Thought Technology. [Online]. Available: www.thoughttechnology.com

[29] J. Arendt, A. Borbely, C. Franey, and J. Wright, “The effects of chronic, small doses of melatonin given in the late afternoon on fatigue in man: a preliminary study,” *Neuroscience letters*, vol. 45, no. 3, pp. 317–321, 1984.

[30] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, “Detecting deceptive behavior via integration of discriminative features from multiple modalities,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1042–1055, 2017.

[31] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc, “Gpu-based video feature tracking and matching,” in *EDGE, workshop on edge computing using new commodity architectures*, vol. 278, 2006, p. 4321.

[32] J. Shi *et al.*, “Good features to track,” in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[33] R. Hartley and A. Zisserman, “Multiple view geometry in computer vision (cambridge university, 2003),” *C1 C3*, vol. 2, 2013.

[34] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[35] J. M. Saragih, S. Lucey, and J. F. Cohn, “Deformable model fitting by regularized landmark mean-shift,” *International journal of computer vision*, vol. 91, no. 2, pp. 200–215, 2011.

[36] Y.-I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[37] J. Vaara, H. Kyröläinen, M. Koivu, M. Tulppo, and T. Finni, “The effect of 60-h sleep deprivation on cardiovascular regulation and body temperature,” *European journal of applied physiology*, vol. 105, no. 3, pp. 439–444, 2009.