Fusion of Physiological Signals for Modeling Driver Awareness Levels in Conditional Autonomous Vehicles using Semi-Supervised Learning

Raul Fernandez-Matellan
UC3M
Madrid, Spain
Email: raulfern@pa.uc3m.es

David Puertas-Ramirez
UNED
Madrid, Spain
Email: dpuertasr@dia.uned.es

David Martin Gomez
UC3M
Madrid, Spain
Email: dmgomez@ing.uc3m.es

Jesus G. Boticario
UNED
Madrid, Spain
Email: jgb@dia.uned.es

***Abstract*—The evolution of autonomous vehicles (AVs) requires a paradigm shift towards the integration of human factors to improve safety and efficiency at levels 2, 3 and 4 of automation. This paper presents a comparison of three different fusion technologies (Low-Level fusion, Medium-Level fusion, and a hybrid fusion), highlighting the critical role of multimodal data integration and semi-supervised learning in predicting and adapting to levels of driver awareness. Our approach uses semi-supervised learning to deal with the data labelling problem, using unlabelled data to train an autoencoder and sparsely labelled data to train a 4-state classifier. Our model facilitates the fusion of data from different physiological signals, including skin electrodermal activity, heart rate, body temperature and acceleration. Using real driving data, the Medium-Level fusion approach gives the best performance, achieving 84% accuracy in predicting situations where the user may not be aware enough to take control of the vehicle. This research highlights the essential nature of fusion technologies to create adaptive and user-centred AV systems.**

I. Introduction

The journey towards fully autonomous vehicles represents one of the most ambitious endeavors in the realm of transportation technology. As vehicles evolve through the levels of automation defined by SAE International [1], the role of the human driver transitions from direct control to supervisory and, eventually, to a passenger [2]. However, this evolution is fraught with challenges, particularly in ensuring safety and reliability as vehicles assume greater control and a greater perceived safety/risk [3].

One of the paramount challenges lies in understanding and integrating human factors—those psychological, physiological, and behavioral elements that fundamentally influence how humans interact with AVs [4]. This paper addresses this challenge by presenting a comprehensive approach that leverages fusion technologies to create a symbiotic relationship between human drivers and autonomous systems, where Artificial Intelligence (AI) systems prioritise collaboration, cooperation, and mutual enhancement between humans and AI agents [5].

The integration of human factors into the development of AVs is not merely an option; it is a necessity. Human error has been identified as a contributing factor in a majority of vehicular accidents, where a Take Over Request (TOR) [6] in conditional automation needs to be promptly attended by the Fallback Ready User (FRU) [7]. However, as the vehicle assumes more driving tasks, understanding the state and readiness of the human to take over control in critical situations becomes crucial [7]. This requires a nuanced understanding of human factors, moving beyond simple behavioral models to incorporate real-time physiological data and environmental contexts. Our approach utilizes a multimodal fusion method, integrating various sources of data to predict and adapt to the driver's awareness levels, involving the comparison of three multimodal fusion approaches: a Low-Level , a Medium-Level, and a Hybrid combination. This methodology stands at the convergence of several critical areas of research: human factors engineering, machine learning, and autonomous vehicle technology.

Fusion technologies serve as the cornerstone of our approach. They allow for the integration of disparate data sources, including physiological signals such as skin electrodermal activity, heart rate, body temperature, and acceleration, alongside environmental and vehicular data [8]. This integration is accomplished through a semi-supervised learning model that trains an autoencoder and classifier to recognize and adapt to the subtleties of individual driver states. The fusion of these data sources requires advanced algorithms capable of extracting meaningful patterns from high-dimensional data, ensuring the model's predictive accuracy and adaptability.

Our methodology is based in a real-world context, using data collected from drivers in different driving scenarios to train and validate our model. This approach ensures that our system is not only theoretically sound but also practically applicable, as it achieves significant accuracy in predicting situations where the user may not be aware enough to take control of the vehicle. Our model demonstrates its potential to significantly improve vehicle safety and driver experience. It also highlights the essential role of fusion technologies in bridging the gap between human factors and autonomous driving systems.

In essence, our work aims to address a critical open issue in autonomous vehicle development: how to ensure that AVs can consider how the complexities of the road impact on

the complexities of human behavior. By integrating human factors into the essence of AV technology, we endeavor to create vehicles that are not only autonomous but also intuitive, responsive, and, above all, safe for all road users. This paper describes this goal, exploring the theoretical foundations, technological innovations, and practical applications of a fruitful fusion technology approach to the integration of human factors in autonomous vehicles.

The paper is structured into several sections, starting with a review of related work in this field, followed by a description of the proposed methodology. Next, the experimental results are discussed and finally, the paper concludes with future work.

II. Related Work

Considering users within the context of autonomous vehicles has recently emerged as a significant problem that needs to be addressed [9] [6] [7] [10]. Given the complexity of the problem and variations among individual users, it is argued that a multimodal approach is essential [9]. Relying on a single signal is insufficient to comprehensively model the user, leading to reduced overall model accuracy [11]. However, the selection of specific sensors remains a subject of debate, with options ranging from physiological, visual, or a combination of both [12] [13]. Regardless of the scenario, it is evident that sensor fusion must be performed. While numerous studies have explored different aspects of user modelling in autonomous vehicles, such as attention [14], drowsiness [15], or distraction [16], the topic of sensor fusion has often been overlooked, with limited literature addressing it [17]. To our knowledge, no studies have compared various fusion methods, which prompted the realization of this study.

In terms of modelling techniques, to fully exploit the capabilities of deep learning algorithms, a common approach is to convert 1D physiological data into 2D representations. This trend has been observed in analogous research efforts focused on the classification of physiological signals [18] [19]. Various techniques, including Markov transition field [20], Gramian angular field [21] and Recurrent Plots (RP) [22] can be used for this purpose. A series of empirical tests were conducted to evaluate the effectiveness of the three methods. The results showed that RP showed superior performance, thus justifying its selection. Given the scarcity of labelled data relative to the abundance of unlabelled data, our approach adopts a semi-supervised learning paradigm that integrates both labelled and unlabelled data to enhance model learning [23].

III. Methodology

A. Set-up for data acquisition

The research presented in this paper uses data collected from real-world autonomous driving environments, as described in our previous publication [11]. The experiments were conducted in two different scenarios: University of the West of Scotland (UWS) and Carlos III University of Madrid (UC3M). Each involved vehicles with different levels of automation. This approach allowed a comprehensive study of user-vehicle interactions. Studying drivers across different levels of automation and vehicle types improves the development of a robust and adaptable system. The first scenario (UWS) was configured and tested at the UWS using a Toyota Prius PHEV, with the SAE Level 2 autonomous [1] functionality enabled by OpenPilot software [24]. This software overrides the vehicle's original Controller Area Network (CAN) messages and sends its customised CAN messages over the vehicle's CAN bus to control the actuators and determine their precise behaviour. The second scenario, also a real-world outdoor scenario at UC3M, where a SAE Level 4 autonomous vehicle prototype has been deployed and configured to provide real-world experience on the university campus. This second vehicle is denominated iCab (Intelligent Campus AutomoBile) [25] and is an autonomous vehicle prototype based on an electric golf cart, where the steering wheel has been removed to give the user the real feeling of a SAE Level 4 autonomous vehicle. For the rest of the paper, we use only data from the UWS scenario. This dataset provides valuable insights into real-world driving conditions, including real road environments and different traffic situations.

To ensure continuous monitoring and capture of physiological data from the user without compromising driving comfort, we used the Empatica E4 wristband [26]. This device guarantees the non-intrusive nature of our methodology and its compatibility with driving conditions. Our experimental setup differs significantly from the methods previously used in this field [27] [28], which often rely on simulation and intrusive techniques to monitor driver physiological data.

All participants in our research were volunteers who were authorised to operate SAE Level 2 autonomous vehicles under real-world driving conditions. Data was collected from users in different age groups, from 25 to 50 years old. They also gave us permission to use their data for the purposes of this research project.

B. Physiological signals

The Empatica E4 wristband incorporates various sensors [26], including the Photoplethysmogram (PPG) sensor, which measures Blood Volume Pulse (BVP) and Heart Rate (HR). It also includes a 3-axis accelerometer to capture basic activity levels, a Galvanic Skin Response (GSR) sensor for continuous monitoring changes in skin properties, and an Infrared Thermopile for measuring skin temperature readings. The device has Bluetooth connectivity so it can connect to other devices to access the internet and upload data to the cloud. Recorded data is available for later download, with the following sampling frequencies: Blood Volume Pulse (BVP): 64Hz, Heart Rate (HR): 1Hz, Electrodermal Activity (EDA): 4Hz, Temperature (TEMP): 4Hz, 3-axis Accelerometer (ACC): 32Hz.

To standardize data processing, all signals were resampled to 4Hz using the Signal library from SciPy [29]. This resampling was undertaken for two primary reasons: firstly, our neural networks operate on a 120x120 image format, and with a 30-second sliding window, resulting in 120 data points (30x4), aligning seamlessly without requiring additional

rescaling. Secondly, our aim was to minimise signal manipulation wherever possible, considering the promising nature of the EDA signal in previous studies in other fields [30].

C. Dataset Generation

Several users participated in the data collection of the UWS scenario, three between the ages of 20 and 30 and one between 40 and 50. For the purposes of this article, we focus exclusively on data from a single 27-year-old user. We have empirically demonstrated the importance of tailoring the model to recognise the unique subtleties of each user and that the performance of the model improves significantly when trained on data specific to an individual user. The methodology used is an intra-subject approach, which involves the construction of individual models tailored to each user. This approach offers improved recognition accuracy by developing models that are specifically designed for each subject. However, the construction of these models requires a significant amount of training data for each individual. For this reason, the models are trained exclusively on data collected from a single user, with the aim of improving performance based on their specific physiological signals. The user who was selected for this work is a 27-year-old male. We recorded 26 driving sessions with this user, exploring a range of real-world driving scenarios to measure the driver's level of awareness. The data was classified into four distinct states of awareness: "Low-Low" (LL) indicates an excessive level of relaxation that can compromise driving ability. "Low" (L) indicates minimal attention given to manual driving tasks. "High" (H) indicates the minimal attention required during autonomous mode to react to unexpected events, while "High-High" (HH) indicates maximum attention during risky autonomous driving situations.

One challenge of our approach is the laborious, intricate, and costly process of labelling data. This is due to the complexity of assigning awareness levels to physiological data [31]. To mitigate these labelling complexities, we aim to implement our approach using semi-supervised learning techniques.

D. Semi-supervised learning approach

Semi-supervised learning techniques are employed to address the challenge of labelling data. While labelling is a significant bottleneck, we can generate a large amount of unlabelled data directly from the Empatica E4 device. Our approach consists of five distinct steps, where the labelled data is only used in the final step to train the final classifier.

The initial step involves transforming one-dimensional data (HR, EDA, ACC, BVP, TEMP) into images using the Recurrence Plot technique. In the second step, key features are extracted from each image using an autoencoder that was previously trained on unlabeled data. If necessary, essential data from multiple images are fused in the third step. The fourth step involves applying Principal Component Analysis (PCA) to reduce feature dimensionality. Finally, the final classifier is trained using our dataset, which is divided into three subsets: training (70%), validation (15%), and test (15%). The model's accuracy is evaluated based on its performance on the test dataset.

E. Using Recurrent Plots to generate images

Recurrent Plot or Recurrence Plot (RP) is a tool used to measure the recurrence of a trajectory $\vec{x} \in \mathbb{R}^d$ in the phase space [22]. The equation 1 shows how a matrix is generated where N is the number of measured points, ϵ is a distance which is used as a threshold, Θ is the Heaviside function (being $\Theta(x) = 0$ if $x < 0,$ and $\Theta(x) = 1$ otherwise) , $||\cdot||$ is the norm used to measure the distance. The equation 2 is a simplified version of the previous equation to work on a continuous space. The dimension d can be used to combine various physiological signals into a single Recurrence Plot. When the state is defined as one-dimensional, the RP matrix is constructed by taking the term-by-term difference of the signal. Alternatively, if the user's state is represented by a vector containing measurements of each physiological signal along its dimensions, the RP can be generated by computing the distance between these two vectors.

$$R_{\mathrm{i,j}}(\epsilon) = \Theta(\epsilon - ||\vec{x_{\mathrm{i}}} - \vec{x_{\mathrm{j}}}||), \qquad i, j = 1, ..., N \tag{1}$$

$$R_{\mathrm{i,j}}(\epsilon) = ||\vec{x_{\mathrm{i}}} - \vec{x_{\mathrm{j}}}||, \qquad i, j = 1, ..., N \tag{2}$$

F. Autoencoder

The autoencoder consists of two main components: an encoder, which aims to represent information using fewer data, and a decoder, which attempts to reconstruct the original image. This method is commonly referred to as unsupervised because labels are unnecessary; the input image serves as the desired output image. Backpropagation is computed by calculating the difference (error) between the output of the decoder and the input image. The architecture designed for this autoencoder is depicted in Figure 1. It has been trained to reconstruct Recurrence Plots generated from non-labeled images recorded by the Empatica E4. In the following architectures, we exclusively utilize the encoder section, which provides us with a reduced vector representation.

G. Fusion Models

We present three different fusion architectures intended to combine information obtained from various physiological signals. The first architecture is described as Low-Level fusion, the second is Medium-Level fusion, and the third is a Hybrid model that combines both fusion techniques. We combine the seven signals extracted from the Empatica E4 (3-axis accelerometer, skin temperature, electrodermal activity, blood volume pulse, and heart rate) in various configurations.

1) Low-Level Fusion: Low-Level Fusion involves combining data before extracting its characteristics. The Recurrence Plot technique is utilized in this process. The vector 'x' is defined as a 7-dimensional vector, with each dimension representing one of the measurements from the physiological data. The L2 norm (Euclidean) is used to measure the distance

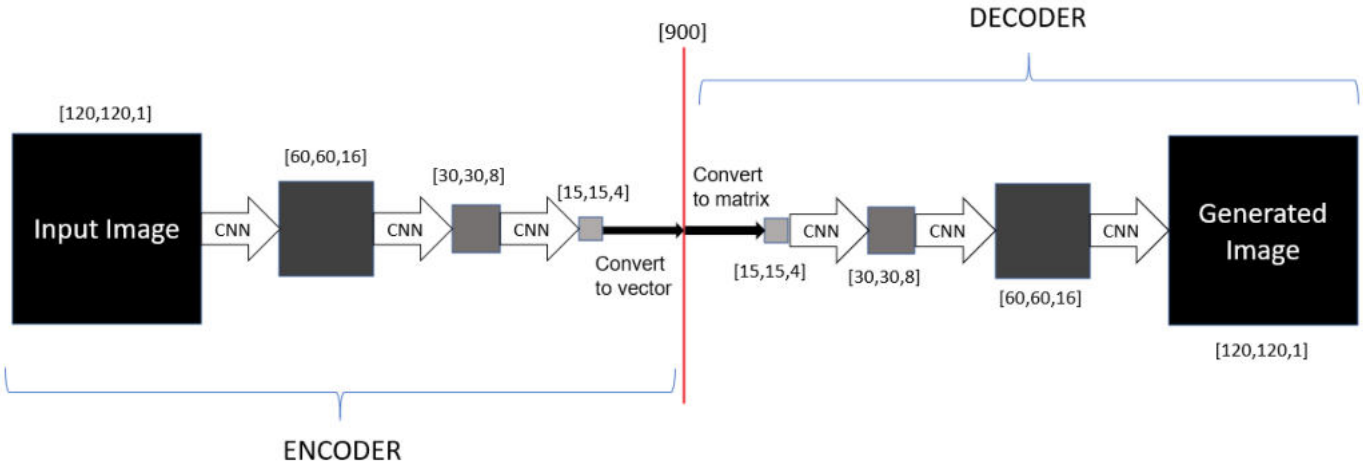

Fig. 1: Proposed autoencoder architecture: Encoder transforms [120,120,1] tensor to [15,14,4], flattened to 900 features. Decoder mirrors this process for image reconstruction.

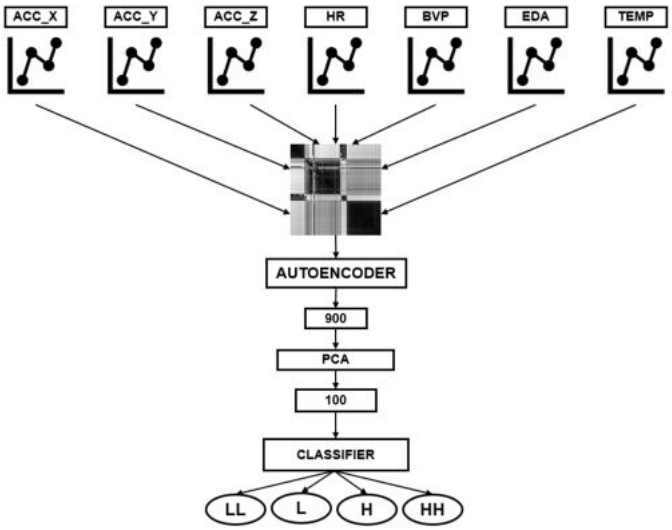

Fig. 2: Low-Level fusion architecture: Physiological signals merged via RP method, autoencoder extracts 900 features, processed by PCA, and fed to final classifier.

between two different vectors. Each variable was normalized using min-max normalization to bring them within the range of 0-1, as they each have their own measurement range. This normalization allows for the addition of the different components of each vector. The resulting image is then rescaled to fit within the 0-255 interval for display purposes. Finally, the image undergoes PCA for dimensionality reduction before being fed into the classifier. The architecture outlined is depicted in Figure 2, which visualizes the process of Low-Level fusion.

2) Medium-Level Fusion: Medium-Level Fusion involves extracting primary features before consolidating information. In this scenario, the vector 'x' is represented as a scalar number, allowing the RP equation to be applied. Seven Recurrence Plots are generated, one for each vector, and the autoencoder is then applied to each RP separately. The features are aggregated and the data is then run through PCA to reduce its dimensionality before being fed into a classifier. Figure 3 illustrates the architecture discussed.

3) A Hybrid Fusion Approach: The hybrid architecture incorporates a two-step process, starting with Low-Level fusion. In this process, data from the same sensor are combined to create a single Recurrence Plot. For example, accelerometer data generates an RP using a vector of three variables, each representing acceleration in one of the three axes (x, y, z). The PPG sensor combines BVP and HR signals to form its RP. The GSR sensor measures EDA, and the Temperature sensor records skin temperature independently, eliminating the need for Low-Level fusion. The structure described can be visualized in Figure 4.

Next, main features are extracted from each RP generated for the four sensors (accelerometer, PPG, GSR, and Temperature) using the autoencoder. These main vectors are then aggregated into one during a Medium-Level fusion process. Data are aggregated and then subjected to PCA for dimensionality reduction before being used as input for the classifier.

IV. Results

The study assesses three fusion architectures (Low-Level, Medium-Level, and Hybrid) using a dataset created from various real-world driving scenarios. The dataset includes data collected from a real driving environment where a driver of a SAE Level 2 autonomous vehicle experiences a variety of attention-demanding situations. RP is used to transform one-dimensional data into images for subsequent processing. The role of the autoencoder's bottleneck, where information is condensed into a one-dimensional vector, is crucial in the reconstruction process. Among the examined autoencoders with bottleneck sizes of 7200, 900, and 98 features, the 900-feature autoencoder is the most effective, striking a balance between image fidelity and the dimensionality of the bottleneck. Figure 5 presents a visual representation of the reconstruction

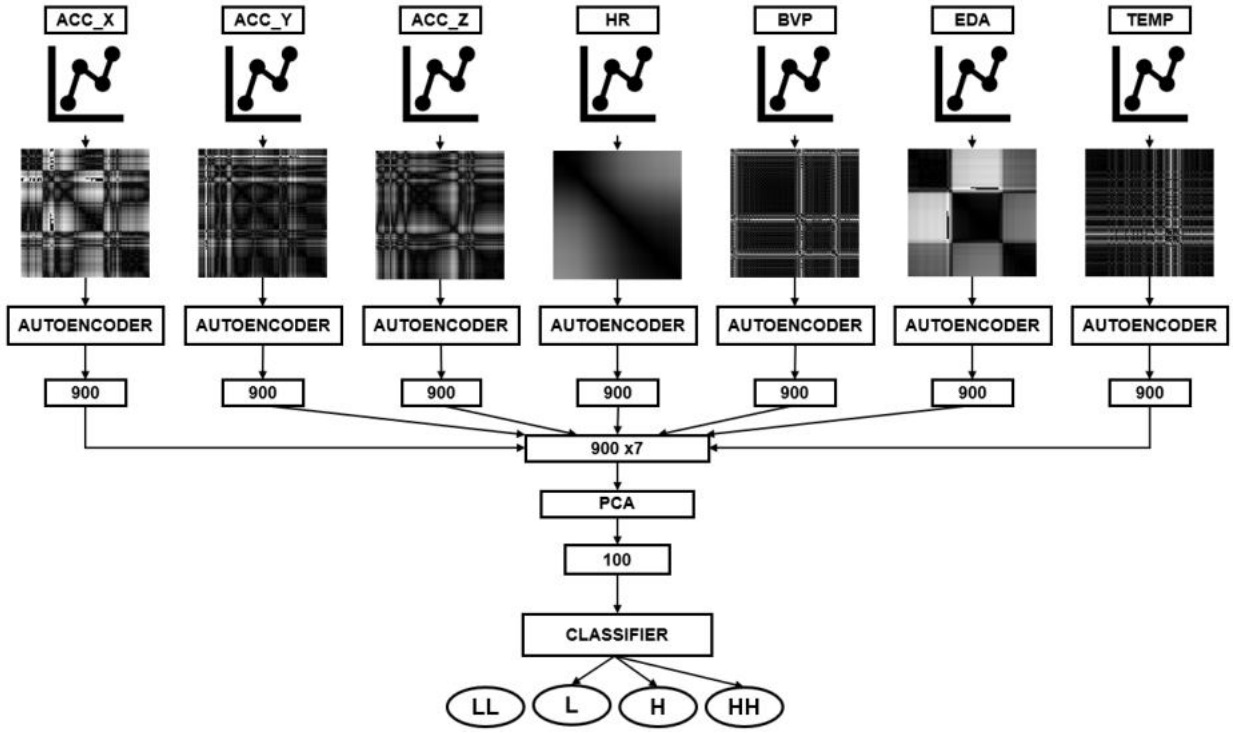

Fig. 3: Medium-Level fusion architecture: RP generates images for each physiological signal, fused output vectors undergo PCA for dimensionality reduction before classifier training.

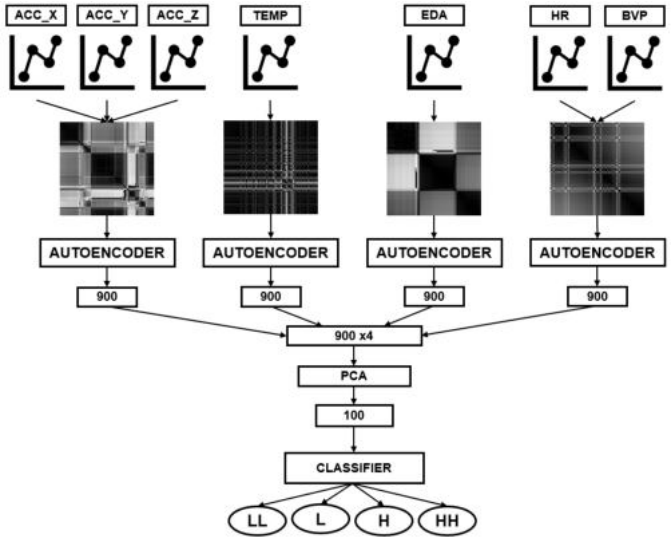

Fig. 4: Hybrid fusion architecture: RP is utilized for Low-Level fusion of signals from the same sensor. Subsequently, Medium-Level fusion is then used to combine data from different sensors, which are then merged into a vector, processed through PCA, and finally fed into the classifier.

capabilities. It demonstrates how the input image (120x120 pixels), is compressed through bottleneck sizes of 98, 900, and 7200 before being reconstructed to its original form. The autoencoder was trained using Recurrent Plots generated from unlabelled experiences. We used 152,120 unlabelled RP constructed from real driving experiences at both UC3M and UWS scenarios, containing data from different drivers in each scenario. This represents the unsupervised part of the semi-supervised learning approach that we are following.

The study primarily focuses on comparing the accuracy of three distinct architectures shown in Figures 2, 3, and 4 representing Low-Level fusion, Medium-Level fusion, and the Hybrid architecture, respectively. The RP technique is used to convert 1D signals into images. The autoencoder is then trained on the unsupervised RP images and PCA is applied to reduce dimensionality. The final classifier is then trained on the labelled dataset, which is common to the three different architectures. The classifier is a dense linear network consisting of three layers. Its purpose is to map 100 PCA components to four levels of driver awareness. Intermediate layers of 25 and 10 neurons, respectively, are used, each followed by a Rectified Linear Unit (ReLU) activation function. The dataset was divided into 70% training, 15% validation, and 15% test sets, giving a total of 2158 labelled samples. The class distribution includes 22% for HH, 23% for H, 34% for L, and 21% for LL. Due to the limited number of images, the models tended to overfit quickly, so the training had to be stopped after the sixth epoch. The results show that the Low-Level fusion model achieved 49% accuracy, the Medium-Level fusion model attained 70%, while the hybrid model dropped to 56% accuracy. Despite the detailed explanations in the discussion section, the Medium-Level fusion gave the best

Acceleration	EDA	HR	BVP	TEMP	Accuracy (%)
YES	YES	YES	YES	YES	**70.33**
NO	YES	YES	YES	YES	**54.76**
YES	NO	YES	YES	YES	**63.99**
YES	YES	NO	YES	YES	**67.86**
YES	YES	YES	NO	YES	**67.56**
YES	YES	YES	YES	NO	**69.34**
YES	NO	NO	NO	NO	**50.30**
NO	YES	NO	NO	NO	**40.35**
NO	NO	YES	NO	NO	**39.58**
NO	NO	NO	YES	NO	**53.57**
NO	NO	NO	NO	YES	**27.98**

TABLE I: Results of training the model using different combinations of physiological signals.

performance.

These architectures offer versatility in the integration of different sensor types. Among them, the Medium-Level fusion model in Figure 3 is used due to its superior performance compared to the Low-Level fusion and hybrid counterparts. We trained the model with different combinations of physiological signals, with corresponding performance metrics detailed in Table I. Optimal performance, reaching 70.33%, was achieved when using all available signals. Interestingly, the accuracy of the model decreases significantly when certain signals are omitted. For example, if accelerometer data is excluded, the accuracy drops to 54.76%, while relying on this data alone results in a performance of 50.30%. Conversely, omitting skin temperature data has little effect on the model's performance, with an accuracy of 69.34%. However, training a model on temperature data alone yields only 27.98% accuracy, similar to random classification for four states (25%).

The model achieves an overall accuracy of 70.33%, calculated as the ratio of correctly predicted values to the total number of predictions. Further analysis shows that when the accuracy is broken down into the four different classes, the well-known F_1 metric yields the following values: 84.3% for the LL state, 68.8% for the L state, 64.6% for the H state and 58.6% for the HH state. The accuracy of the LL state is particularly noteworthy, with a precision of 92.3% and a recall of 77.2%. This indicates the model's ability to effectively predict instances where the driver is not alert, thus ensuring that control is not delegated in such situations.

V. Discussion

In this study, three fusion architectures were evaluated: Low-Level, Medium-Level and Hybrid, using a semi-supervised learning approach. Autoencoders were trained on large amounts of unlabelled data, followed by training of a supervised classifier to detect the driver's level of awareness. The Low-Level and Hybrid fusion methods showed poor performance, probably due to the loss of crucial signal features during fusion. The Mid-Level fusion approaches proved to be more effective as they allowed important information to be extracted from the signal prior to aggregation. The difference in performance can be attributed to the loss of information during the process of creating images from signals. Low-Level Fusion involves the merging of disparate raw signals into a unified composite signal prior to image generation, as illustrated in Equation 1. In contrast, Medium-Level Fusion generates images directly from the raw signals without any modification. With a Medium-Level fusion method, each individual signal can be treated separately. This makes it easily expandable with data from other sensors and adaptable according to specific needs, as long as data has been vectorised. Furthermore, even if different fusion methods are used, the feature extraction component can be adapted to any of the different architectures and data formats used by researchers [12] [13].

These results show that the methodology developed in this research can be applied through different levels of automation. At SAE Levels 0 and 1 it could act as a driver assistant, at Level 2 it could assist the driver by monitoring driving, and Levels 3 and 4 when the vehicle is prepared to take control when the driver is unable to take control in FRU [1]. Even at Level 5, our method can be adapted to increase the user comfort and detect medical emergencies. A proof of this versatile approach is that we work with two driving scenarios and vehicles, a Level 2 AV in UWS and Level 4 AV in UC3M [11]. Furthermore, the model can be adapted to each individual user which, according to related work [9] and supported by our own empirical results, it is crucial to achieve the highest level of accuracy.

Due to the modular design of the architecture, numerous combinations of methods could be explored in the future, which could improve accuracy. For example, the feature extractor, represented by the autoencoder in this case, could be replaced by more advanced and customized architectures to obtain a more refined representation of each signal. Furthermore, instead of employing PCA for the reduction of dimensionality after aggregation, an alternative approach could be to add another layer to the classifier. One of the advantages of using PCA is its ability to be initially optimized using unlabeled data, as the algorithm looks for the most effective means to represent the data with fewer variables. An alternative fusion technique worth exploring is to embed the generated Recurrent Plot into a channel and then allow it to the network [22].

Our comparison revealed that the various signals obtained from the Empatica E4 device [26] do not contribute equally to determining driver awareness. The skin temperature was found to have minimal impact, which could be attributed to the gradual nature of temperature fluctuations in the body or the accuracy limitations of the sensor. However, accelerometers were shown to significantly influence our system, since user movements were correlated with their state, which is consistent with findings from other studies [32]. Electrodermal activity (EDA) has also been found to be valuable in assessing user awareness and has also been found to be relevant in measuring other states, such as driver workload during takeover activities [27]. In our research, EDA provided less information than the accelerometer, but it turned out to be the most reliable physiological signal to discern the driver's state. Heart Rate (HR) and Blood Volume Pulse (BVP) data also contributed to the optimal performance of the model. We refrained from

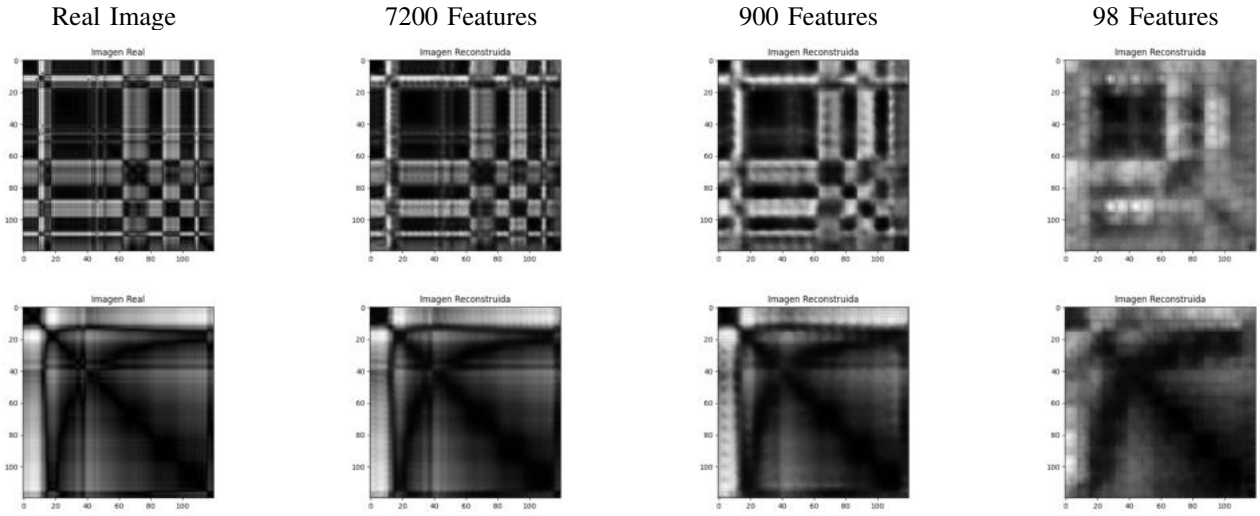

Fig. 5: Examples of the use of the trained autoencoder, from left to right: real image, image with 7200 features, image with 900 features and image with 98 features.

using more invasive methods of measuring physiological data, such as an EEG cap, despite their potential for greater accuracy [13], due to concerns about their intrusiveness and potential impact on participants' responses to the experiments.

We have demonstrated the adaptability of our model in extracting information from various sources. Although the work presented in this paper focuses on estimating driver alertness via signals from the Empatica E4 [26], our model has the flexibility to seamlessly incorporate data from additional sources in the future. For example, video capturing the user's face expressions, movements, gestures, blink rate, body position or environment [33], but also data from autonomous vehicles. The only requirement is to vectorise the information before aggregation. By extracting primary features from each raw source before combining them, Medium-Level fusion provides the best results in this context. Ultimately, our multimodal approach is highly adaptable and allows combining labelled and unlabelled data from different sources.

Our model shows good performance in classifying the Low-Low state of awareness but experiences reduced accuracy for other states. This decline can be attributed to the inherent strength of Low-Low states' labeling compared to other states. Despite our efforts to use consistent labelling methods, the complexity of the labelling process inevitably leads to inaccuracies, especially at higher levels of arousal. This is a common challenge when it comes to variables such as stress and concentration [31]. Furthermore, the choice of window size has a significant impact on the accuracy of the model [19]. For example, using a 10 second window gives an accuracy of 60.83% compared to 70.33% using a 30 second window. Using an excessively large sliding window has its drawbacks. For instance, using a 120-second window results in overlapping images, which leads to the generation of similar images. Therefore, if the dataset is randomly divided, the model will be trained and tested on comparable images, which could lead to inaccurate evaluation of the model's performance. For this reason, the results presented in this paper use a 30-second slider window to determine awareness levels.

VI. Conclusions and future work

In conclusion, our study demonstrates the effectiveness of a multimodal approach employing Medium-Level fusion techniques compared to Low-Level fusion using RP and a proposed hybrid architecture. The study was carried out using data collected from a real SAE Level 2 autonomous vehicle. Our best methods achieved an accuracy rate of 70.33% on a four-class classification problem, with the Low-Level state accuracy exceeding 84%. This study highlights the accuracy of the model in predicting situations in which the user may not be aware enough to regain control of the vehicle. Additionally, the model demonstrated flexibility in handling different inputs and using various techniques to extract data from raw images. The research highlights the importance of accelerometer data in predicting user states while driving.

In the future, we plan to expand our research to include other environments, specifically those with a conditional Level 4 autonomous vehicle that can drive around university campuses. This environment will allow us to subject drivers to risky situations in a controlled and secure environment, which will provide additional information about the driver's awareness and responsiveness. Furthermore, we intend to investigate various architectural designs that can improve the accuracy and resilience of our models, thus advancing the field of autonomous vehicle intelligence and safety.

Acknowledgment

This work was supported by the Spanish Government under Grants TED2021-129485B-C41 and TED2021-129485B-C44.

REFERENCES

[1] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," SAE International, Tech. Rep., 2021.

[2] J. Zhao, "HUMAN FACTORS IN AUTOMATED VEHICLES: A Bibliographic Study on Enhancing Trust and Situation Awareness in Human Machine Interface," Ph.D. dissertation, 12 2023.

[3] Y. Zou, X. Fu, D. Ye, and R. Ouyang, "The Roles of Perceived Risk and Expectancy Performance in Passenger's Acceptance of Automated Vehicles," in *International Conference on Human-Computer Interaction*. Springer, 2023, vol. 14023 LNCS, pp. 497–511.

[4] 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, no. March, pp. 1–27, 3 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2019.00057/full

[5] T. A. O'Neill, C. Flathmann, N. J. McNeese, and E. Salas, "Human-autonomy Teaming: Need for a guiding team-based framework?" *Computers in Human Behavior*, vol. 146, p. 107762, 9 2023.

[6] W. Morales-Alvarez, O. Sipele, R. Léberon, H. H. Tadjine, and C. Olaverri-Monreal, "Automated driving: A literature review of the take over request in conditional automation," *Electronics (Switzerland)*, vol. 9, no. 12, pp. 1–34, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/12/2087

[7] D. Puertas-Ramirez, A. Serrano-Mamolar, D. Martin Gomez, and J. G. Boticario, "Should Conditional Self-Driving Cars Consider the State of the Human Inside the Vehicle?" in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: ACM, 6 2021, pp. 137–141. [Online]. Available: https://dl.acm.org/doi/10.1145/3450614.3462243

[8] J. Oyekan, Y. Chen, C. Turner, and A. Tiwari, "Applying a fusion of wearable sensors and a cognitive inspired architecture to real-time ergonomics analysis of manual assembly tasks," *Journal of Manufacturing Systems*, vol. 61, no. October, pp. 391–405, 2021. [Online]. Available: https://doi.org/10.1016/j.jmsy.2021.09.015

[9] C. Collet and O. Musicant, "Associating vehicles automation with drivers functional state assessment systems: A challenge for road safety in the future," *Frontiers in Human Neuroscience*, vol. 13, p. 408476, 2 2019.

[10] J. Ma, Y. Wu, J. Rong, and X. Zhao, "A systematic review on the influence factors, measurement, and effect of driver workload," *Accident Analysis & Prevention*, vol. 192, p. 107289, 11 2023.

[11] D. Puertas-Ramirez, R. Fernandez-Matellan, D. Martin-Gomez, J. G. Boticario, and D. Tena-Gago, "Improving Autonomous Vehicle Automation Through Human-System Interaction," *The 37th annual European Simulation and Modelling Conference*, pp. 294–300, 2023.

[12] A. A. Lopez-Aguilar, S. A. Navarro-Tuch, L. M. Camacho-Bustamante, and M. R. Bustamante-Bello, "An Analysis of Monitoring Technologies for the Objective Evaluation of User Experience on Autonomous Vehicles," *2023 International Symposium on Electromobility, ISEM 2023*, 2023.

[13] E. J. C. Nacpil, Z. Wang, and K. Nakano, "Application of Physiological Sensors for Personalization in Semi-Autonomous Driving: A Review," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 19 662–19 674, 9 2021.

[14] M. Bonyani, M. Rahmanian, S. Jahangard, and M. Rezaei, "Dipnet: Driver Intention Prediction for a Safe Takeover Transition in Automated Vehicles," *SSRN Electronic Journal*, no. April, 12 2021. [Online]. Available: https://www.ssrn.com/abstract=3982515 https://papers.ssrn.com/abstract=3982515

[15] W. Kim, E. Jeon, G. Kim, D. Yeo, and S. Kim, "Take-Over Requests after Waking in Autonomous Vehicles," *Applied Sciences*, vol. 12, no. 3, p. 1438, 1 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/1438

[16] S. Yang, J. Kuo, and M. G. Lenné, "Patterns of Sequential Off-Road Glances Indicate Levels of Distraction in Automated Driving," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 2056–2060, 2019.

[17] Y. Du, A. W. Black, L. P. Morency, and M. Eskenazi, "Multimodal polynomial fusion for detecting driver distraction," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, pp. 611–615, 2018.

[18] R. Elalamy, M. Fanourakis, and G. Chanel, "Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals," *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021*, pp. 1–7, 2021.

[19] J. Lee, H. Lee, and M. Shin, "Driving stress detection using multimodal convolutional neural networks with nonlinear representation of short-term physiological signals," *Sensors*, vol. 21, no. 7, p. 2381, 2021.

[20] H. Xu, J. Li, H. Yuan, Q. Liu, S. Fan, T. Li, and X. Sun, "Human activity recognition based on gramian angular field and deep convolutional neural network," *IEEE Access*, vol. 8, pp. 199 393–199 405, 2020.

[21] J. Yan, J. Kan, and H. Luo, "Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network," *Sensors 2022, Vol. 22, Page 3936*, vol. 22, no. 10, p. 3936, 5 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/10/3936/htm https://www.mdpi.com/1424-8220/22/10/3936

[22] N. Marwan, M. Carmen Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports*, vol. 438, no. 5-6, pp. 237–329, 1 2007.

[23] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[24] Comma.ai, "openpilot — open source advanced driver assistance system," 2024. [Online]. Available: https://comma.ai/openpilot

[25] P. Marin-Plaza, A. Hussein, D. Martin, and A. de la Escalera, "icab use case for ros-based architecture," *Robotics and Autonomous Systems*, vol. 118, pp. 251–262, 2019.

[26] Empatica Inc., "Empatica E4," 2024. [Online]. Available: https://www.empatica.com/en-eu/research/e4/

[27] D. Min, A. Gluck, C. Menassa, V. Kamat, D. Li, and J. Brinkley, "Predicting Driver Takeover Performance in Conditional Automation (Level 3) through Physiological Sensing," 1 2024. [Online]. Available: http://deepblue.lib.umich.edu/handle/2027.42/191950

[28] A. Barisic, P. Sigrist, S. Oliver, A. Sciarra, and M. Winckler, "Driver Model for Take-Over-Request in Autonomous Vehicles," in *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: ACM, 6 2023, pp. 317–324. [Online]. Available: https://dl.acm.org/doi/10.1145/3563359.3596994

[29] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[30] A. Serrano-Mamolar, M. Arevalillo-Herráez, G. Chicote-Huete, and J. G. Boticario, "An intra-subject approach based on the application of hmm to predict concentration in educational contexts from nonintrusive physiological signals in real-world situations," *Sensors*, vol. 21, no. 5, p. 1777, 2021.

[31] M. Saneiro, O. C. Santos, S. Salmeron-Majadas, and J. G. Boticario, "Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches," *Scientific World Journal*, vol. 2014, 2014.

[32] G. Li and W.-Y. Chung, "Combined eeg-gyroscope-tdcs brain machine interface system for early management of driver drowsiness," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 1, pp. 50–62, 2018.

[33] A. Tavakoli, S. Kumar, X. Guo, V. Balali, M. Boukhechba, and A. Heydarian, "HARMONY: A Human-Centered Multimodal Driving Study in the Wild," *IEEE Access*, vol. 9, pp. 23 956–23 978, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9343252