Local and Global Interpretability Using Mutual Information in Explainable Artificial Intelligence

Conference Paper · November 2021

DOI: 10.1109/ISCMI53840.2021.9654898

3 authors:

Mir Riyanul Islam
Mälardalen University

Mobyen Uddin Ahmed
Mälardalen University

Shahina Begum
Mälardalen University

2021 8th Intl. Conference on Soft Computing & Machine Intelligence

Local and Global Interpretability Using Mutual Information in Explainable Artificial Intelligence

Mir Riyanul Islam*, Mobyen Uddin Ahmed† and Shahina Begum‡
School of Innovation, Design and Engineering
Mälardalen University
Västerås, Sweden
*mir.riyanul.islam@mdh.se, †mobyen.ahmed@mdh.se, ‡shahina.begum@mdh.se,
*Corresponding Author

***Abstract*—Numerous studies have exploited the potential of Artificial Intelligence (AI) and Machine Learning (ML) models to develop intelligent systems in diverse domains for complex tasks, such as analysing data, extracting features, prediction, recommendation etc. However, presently these systems embrace acceptability issues from the end-users. The models deployed at the back of the systems mostly analyse the correlations or dependencies between the input and output to uncover the important characteristics of the input features, but they lack explainability and interpretability that causing the acceptability issues of intelligent systems and raising the research domain of eXplainable Artificial Intelligence (XAI). In this study, to overcome these shortcomings, a hybrid XAI approach is developed to explain an AI/ML model's inference mechanism as well as the final outcome. The overall approach comprises of 1) a convolutional encoder that extracts deep features from the data and computes their relevancy with features extracted using domain knowledge, 2) a model for classifying data points using the features from autoencoder, and 3) a process of explaining the model's working procedure and decisions using mutual information to provide global and local interpretability. To demonstrate and validate the proposed approach, experimentation was performed using an electroencephalography dataset from road safety to classify drivers' in-vehicle mental workload. The outcome of the experiment was found to be promising that produced a Support Vector Machine classifier for mental workload with approximately 89% performance accuracy. Moreover, the proposed approach can also provide an explanation for the classifier model's behaviour and decisions with the combined illustration of Shapely values and mutual information.**

***Index Terms*—autoencoder, electroencephalography, explainability, feature extraction, mental workload, mutual information**

I. Introduction

Recent developments of Artificial Intelligence (AI) and Machine Leaning (ML) have been embraced in almost every domain in the form of automated and semi-automated systems. However, with the growing popularity of these systems, the AI/ML algorithms which act behind the systems, still endure acceptability issues due to the lack of explanations on the algorithms' inference mechanism and decisions. Realising the dire need of explaining or interpreting AI/ML model-based intelligent systems, the research domain of eXplainable Artificial Intelligence (XAI) emerged. Currently, XAI research is immensely spreading to develop methods of generating explanations to enhance the local and global interpretability of AI/ML models. Global interpretability refers to interpreting any model's inference mechanism, whereas local interpretability indicates the understandability of a specific decision from an AI/ML model [1]. Several tools are already proposed by researchers to generate explanations and interpretability of AI/ML models, such as Local Interpretable Model Agnostic Explanations (LIME) [2] and SHapley Additive exPlanations (SHAP) [3]. However, the understandability of the explanations from these tools are highly dependent on domain expertise.

Many fields from diverse domains have already been facilitated by XAI research, such as, image processing [4], anomaly detection [5], predictive maintenance [6] etc. On the contrary, safety-critical domains concerning human life, e.g., road safety has received less attention from the XAI researchers. Very few evidences are found in the literature like explaining motorbike riding pattern [7], whereas the depth of research in XAI is still shallow for drivers. However, AI/ML approaches had been well investigated for in-vehicle road safety features such as, drivers' drowsiness detection and intelligent speed assistance through utilising vehicular signals, neurophysiological signals, etc. Specifically, neurophysiological signals, e.g. electroencephalography (EEG) and electrocardiography (ECG), are one of the major tools for assessing a driver's in-vehicle performance [8]. The major challenge of utilising EEG signals in an AI/ML approach is the feature extraction procedure that demands high involvement of experts and manual computation. Automatic approaches are already proved to be efficient in extracting features from EEG leveraging the computation strength of convolutional neural network (CNN) based autoencoder [9] but lacks in explainability of the extracted features.

Autoencoders of different architectures have been exploited in several studies to explain diverse tasks, like forecasting energy demand [10], classifying time series [11], detecting anomalies [5] and changes in temporal images [12], etc. Moreover, autoencoder has been used to enhance the quality of explanations from different explainability tools [13]. All of these works contribute to explain decisions or enhance explanations but no evidence was found on explaining the deep features that can be extracted using autoencoder.

One of the major challenges of explaining model and/or decision is to extract the underlying relation between the input and output. Recently, the concept of mutual information has drawn attention of XAI researchers due to its naive nature of quantifying relevancy between two random variables [14]. Upon realising the potential of mutual information and the urge of explaining features to induce global interpretability in AI/ML models, this study proposes a hybrid approach of feature explanation using mutual information associating the explanation generated by popular explainability tool SHAP. The idea solely depends on the fact that the mutual information is a proper mean of domain knowledge as demonstrated in several recent studies on recommender systems [15], automated fault diagnosis [16], feature extraction [17] etc.

Summarising, to expand the research domain of XAI and contribute to road safety, this study aims at utilising the EEG signals recorded from car drivers' to demonstrate the proposed concept of explaining autoencoder extracted features using mutual information, followed by explaining mental workload classification to achieve local and global interpretability. To achieve the aim of this study, two major objectives are set and stated below:

- To propose a novel approach of using mutual information to explain autoencoder extracted EEG features.
- To demonstrate a hybrid methodology of explaining mental workload classification from the autoencoder extracted features using SHAP and mutual information to induce local and global interpretability.

The remaining parts of this article are arranged as follows, Section II contains description of the materials and methods. In Section III, obtained results are presented discussed thoroughly. Finally, conclusion of this study and possible research directions are stated in Section IV.

II. Materials and Methods

A. Data Acquisition and Preprocessing

The data, specifically the EEG signals, that was analysed in this study was collected under the framework of the project BrainSafeDrive[1], through an natural driving experiment in a route around the urban areas of the periphery of Bologna, Italy. During the experiment, 20 male participants drove along a $2.5km$ long circuit route twice in normal and rush hour randomly for three laps. Moreover, each circuit consisted of easy and hard segments containing road through busy industrial area and comparatively quite residential area, respectively. The experimental road, hour and segments were selected to induce the drivers with different levels of workload. Additional description of the experimental protocol are available in the articles- [18] and [17].

To record EEG signals, the digital monitoring BEmicro system (EBNeuro, Italy) was used with active 15 EEG channels (FPz, $AF3$, $AF4$, $F3$, Fz, $F4$, $P5$, $P3$, Pz, $P4$, $P6$, POz, $O1$, Oz and $O2$) placed according to the $10-20$ International System. The sampling frequency was $256Hz$ and the channel impedance was kept below $20k\Omega$. During the experiments raw EEG signals were recorded and the processing was applied offline. In particular, each EEG signal has been firstly band-pass filtered with a fourth-order Butterworth infinite impulse response (IIR) filter (high-pass filter cut-off frequency: $1Hz$, low-pass filter cut-off frequency: $30Hz$). Afterwards, ARTE (*Automated aRTifacts handling in EEG*) algorithm [19] was deployed to remove various artefacts such as, drivers' movements and environmental noises, from the recorded EEG signals. Finally, the EEG signals were sliced into epochs of $2s$ ($0.5Hz$ of the frequency resolution) length using sliding window technique with a stride of $0.125s$ keeping an overlap of $0.825s$ between two continuous epochs. The windowing technique was performed to obtain higher number of observations in comparison with the number of variable and to contain the stationarity condition of the EEG signals [20].

[1] http://brainsafedrive.brainsigns.com/

B. Feature Extraction

The feature extraction process was performed from two different perspectives. First, the features were extracted based on the *Power Spectral Density* (PSD) to incorporate domain knowledge in the feature set. In the second approach, convolutional autoencoder was developed to extract features from the EEG signals to contain deeper insights of the data and reduce human involvement. Both the approaches are briefly described below.

1) Features from Power Spectral Density: From the clean and segmented EEG signals, the PSD has been calculated for each EEG channel for each epoch using the *Fast Fourier Transformation* (FFT) and a Hanning window of equal epoch length, i.e., $2s$. Then, the EEG frequency bands of interest has been defined for each subject by estimating the *Individual Alpha Frequency* (IAF) value [21]. The IAF value was determined as the peak of the general alpha rhythm frequency ($8-12Hz$). Subsequently, average frequency of the theta band $[IAF-6, IAF-2]$, the alpha band $[IAF-2, IAF+2]$ and the beta band $[IAF+2, IAF+18]$, over all the EEG channels were calculated. Finally, a spectral feature vector containing 45 features (15 EEG channels $\times$ 3 Frequency bins) has been obtained from the frequency bands directly correlated to the mental workload, as manifested in the previous scientific literature [8]. In fact, one of the prime biomarkers of human mental workload is the ratio between Frontal Theta and Parietal Alpha spectral content [8].

2) Features from Convolutional Autoencoder: Traditionally, the convolutional autoencoder architecture consists of two segments, (i) encoder and (ii) decoder. A number of convolutional layers associated with pooling layers form the encoder segment to find the deep hidden features in the original signal. On the contrary, Decoder contains several deconvolutional layer to reconstruct the input signal from the features through minimising the residuals. The autoencoder trains through the process of encoding and reconstruction of predefined epochs and batch size. Here, several tweaking of the number of convolutional layers and associated parameters were performed and the encoder was finalised with three convolutional layers and three max-pooling layers followed by a flattening layer. Table I presents the summary of the layers of the encoder with a total of 732 parameters to train. The output shape of of the input layer is $(512, 16, 1)$ that contains 1 clean EEG signal epoch of length $2s$ (at $256Hz$ sampling frequency) from 15 channels and one channel was introduced with zeros to facilitate the design of the encoder. The decoder was designed in the inverse order of the structure of the encoder containing four convolutional layers and three upsampling layers facilitating the depooling mechanism. In each of the convolutional layers, batch normalisation with ReLU activation function was invoked with zero padding. The developed autoencoder utilised RMSprop optimisation with a learning rate of 0.002 and binary cross-entropy as the loss function. Finally, 32 features were extracted from the cleaned EEG epochs in accordance to the output shape of the flattening layer of the encoder.

TABLE I
SUMMARY OF THE DESINGED CONVOLUTIONAL ENCODER.

Layer Type	Output Shape	No. of Parameters
Input	(512, 16, 1)	0
Convolutional	(256, 8, 16)	80
MaxPooling	(128, 4, 16)	0
Convolutional	(64, 2, 8)	520
MaxPooling	(32, 1, 8)	0
Convolutional	(16, 1, 4)	132
MaxPooling	(8, 1, 4)	0
Flattening	(32)	0

After the preparation of feature sets, labels were added to the feature vectors according to the experimental road segment and time of driving based on the experimental design. Specifically, the feature vectors extracted from driving sessions on hard road segment during rush hour was labelled as $high$ mental workload. On the other hand, low mental workload labels were added to the features extracted from the data recorded during normal hour while driving on easier road segment as prescribed by the experts in the experimental protocol [17], [18].

C. Explanation of Extracted Features

The features extracted from the convolutional autoencoder are based on the underlying characteristics of the input data, in this study, the EEG signals. To understand and explain the features, mutual information was used to prove the relevance between the spectral features and the autoencoded features. The mutual information between two random variables is a metric to quantify the mutual dependence between the two variables. For measuring linear and nonlinear correlation, it is an ideal criteria. The mutual information has been considered as the base for many well-known methods, such as hidden Markov models and decision trees [16]. In fact, a recent study showed the use of mutual information in developing combined feature set from correlated features from different measurements [17].

Theoretically, If X and Y are continuous random variables where $X, Y \in R^d$, the mutual information between X and Y is termed as $I(X, Y)$ and formulated as shown in equation 1 [14].

$$I(X, Y) = \int_y \int_x p(y, x) \; log_2 \frac{p(y, x)}{p(y) \; p(x)} \; dx \; dy \qquad (1)$$

In this study, F_s and F_a were considered for spectral and autoencoder extracted features, respectively depicting X and Y as stated in equation 1. Thus, computing the mutual information $I(F_s, F_a)$ generates the means of explaining the autoencoder extracted features by the spectral features as a substitute of domain knowledge. Afterwards, for better understanding of the explanation, the mutual information values are illustrated using Chord diagram [22] for the whole model or a single decision.

D. Mental Workload Classification

In order to classify drivers' mental workload from EEG features, Random Forest (RF) and Support Vector Machine (SVM) have been invoked leveraging the outcome of the previous studies- [17] and [9]. In the cited studies, authors compared the selected classifiers with several other AI/ML models such as, k-Nearest Neighbours (kNN), Multi-Layer Perceptron (MLP) and Logistic Regression and reported maximum accuracy by the selected ones in the binary classification of drivers' mental workload into $high$ and low. However, in this study, different kernel functions, e.g., Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid kernels for SVM have been deployed to investigate and report the change in performance metrics while classifying mental workload. Again, the varying number of estimators and depths were investigated while RF model was trained. After training and validating with 5-fold cross-validation of the aforementioned variants of the models, the classifiers resulting maximum performance accuracy were chosen to train the concluding model and generate explanation at global and local scope.

E. Explanation of Mental Workload Classification

To explain the SVM classifier trained to classify mental workload from autoencoder extracted features, open-source explainability tool SHAP was used. At first, explanations were generated at global and local scope by invoking the built-in functions. The main components of the explanations are the Shapley values associated with the autoencoder extracted features that control the behaviour of the model as a whole and for each individual classification tasks. Furthermore, following the method described in Section II-C, with pre-computed mutual information values the Chord diagrams were drawn that illustrates the relevance between the autoencoder extracted features and the grouped spectral features, i.e., *Theta*, *Alpha* and *Beta* on the basis of *Frontal* and *Parietal* scalp location. In Section III-B, Fig. 1 and 2 presents the global and local explanations, respectively.

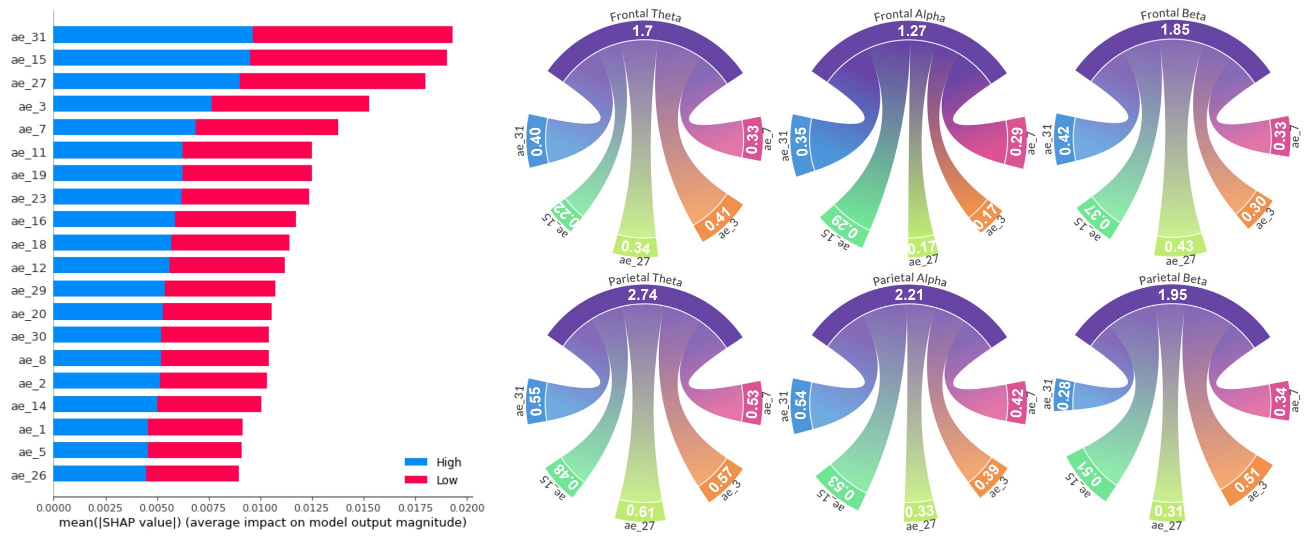

Fig. 1. Global explanation of mental workload classifier model with SHAP values with bar plot (left) and mutual information illustrated with Chord diagrams for six spectral feature groups (right).

III. Results and Discussion

The outcome of this study is presented in this section from two different aspects- mental workload classification and explaining the trained classifier model followed by explaining a single decision. For each of the aspects, the results are discussed in corresponding subsections.

A. Mental Workload Classification

For mental workload classification, the analysed dataset initially contained 65507 instances, where 36630 and 28877 instances were labelled as low and $high$ mental workload, respectively. As the complete dataset were substantially large, the instances of low were randomly down-sampled to match the number of instances labelled as $high$ that leaves the final dataset size to 57754. Both the RF and SVM classifiers were trained and cross validated with 5-fold cross-validation. After the training phase, the performance metrics were calculated with the hold-out dataset, where the total number of observations was 11550 and low mental workload was considered as the positive class. Table II presents the performance metrics of the mental workload classifiers. The performance metrics were selected depending on the balanced characteristics of the dataset. From the summary, it was observed that SVM with RBF kernel produced the highest classification accuracy. On the other hand, linear kernel came out to be most unsuccessful classifier in classifying mental workload. This signifies the non-linear characteristics of the EEG signals.

TABLE II
Performance Summary of Mental Workload Classification using RF and SVM Classifier Model on the Holdout Test Set.

Classifier	Accuracy	Precision	Recall	F_1 score
RF	88.59%	0.9995	0.7723	0.8713
SVM	89.45%	0.9831	0.7876	0.8746

B. Global and Local Explanation

Performing the approaches described in Section II-C and II-E, the explanations are generated using SHAP. At first a summary plot has been drawn using built-in functions that illustrate the prime features inferring the model's decision in terms of Shapley values. Furthermore, the spectral features associated mutual information values to the autoencoder extracted features are illustrated using Chord diagram to generate global explanation as illustrated in Fig. 1. For a single instance, local explanation is generated using SHAP values contributing to the decision (Fig. 2). Here, similar association between the autoencoder and spectral feature groups can also be shown as it is illustrated in global explanation.

Currently, several explainability tools are available to generate explanation for models' decisions, in other terms, at local scope. LIME is one of the methods to produce local explanation. However, here SHAP has been used for enhancing interpretability of mental workload classifier model since it has the capability to produce both local and global explanation that aligns with the objective of this study. But, the difficulty of understanding the Shapley values associated with the autoencoder extracted features for the end users had been overcome using spectral feature groups of EEG signals. Mutual information values, in naive term, relevancy between the spectral features and autoencoder features were calculated followed by the representation of Chord diagrams to facilitate the domain experts.

IV. Conclusion

The contribution presented in this article is twofold: (1) proposal and illustration of a novel approach of using mutual information to explain EEG features extracted by convolutional autoencoder; this approach, to our knowledge, is the only procedure to explain the autoencoder extracted features; (2) demonstration of explaining drivers' mental workload

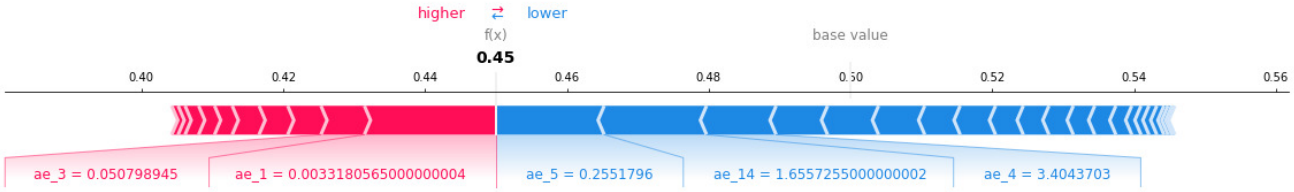

Fig. 2. Example of a local explanation with SHAP.

classification at local and global scope, based on autoencoder extracted EEG features using SHAP and mutual information. In a broader terms, explaining EEG signal classification that can be further adopted in other domains utilising the EEG signals.

The experimental results of this study have been encouraging, but there is space for improvements and further research. In terms of deep learning techniques, investigating other architectures, such as Recurrent Neural Network (RNN) as a combined alternative to the working sequence of autoencoder and RF or SVM classifier. As regards explainability, exploiting similar parameters to mutual information for explaining the features in more understandable form incorporating domain knowledge. Moreover, improving the quality and form of explanations both at feature and decision level through a validation phase that involves experts and end users.

Acknowledgment

This article is based on the study performed as a part of the project BrainSafeDrive, co-funded by the Vetenskapsrådet - The Swedish Research Council and the Ministero dell'Istruzione dell'Università e della Ricerca della Repubblica Italiana under the Italy-Sweden Cooperation Program.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[4] S.-L. Wu, H.-Y. Tung, and Y.-L. Hsu, "Deep learning for automatic quality grading of mangoes: methods and insights," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA: IEEE, Dec. 2020, pp. 446–453.

[5] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using shapley additive explanations," *Expert Systems with Applications*, vol. 186, p. 115736, Dec. 2021.

[6] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and explainable predictive maintenance: semi-supervised deep learning for anomaly detection and diagnosis in press machine data," *Applied Sciences*, vol. 11, no. 16, p. 7376, Aug. 2021.

[7] M. L. Abadi and A. Boubezoul, "Deep neural networks for classification of riding patterns: with a focus on explainability," in *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Oct. 2021.

[8] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, Jul. 2014.

[9] M. R. Islam, S. Barua, M. U. Ahmed, S. Begum, and G. Di Flumeri, "Deep learning for automatic EEG feature extraction: an application in drivers' mental workload classification," in *Human Mental Workload: Models and Applications*, L. Longo and M. C. Leva, Eds. Cham: Springer International Publishing, 2019, pp. 121–135.

[10] J.-Y. Kim and S.-B. Cho, "Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space," *Expert Systems with Applications*, vol. 186, p. 115842, Dec. 2021.

[11] G. Leonardi, S. Montani, and M. Striani, "Deep feature extraction for representing and classifying time series cases: towards an interpretable approach in haemodialysis," in *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference, FLAIRS 2020*, Florida, Miami - USA, 2020, p. 4.

[12] L. Bergamasco, S. Saha, F. Bovolo, and L. Bruzzone, "An explainable convolutional autoencoder model for unsupervised change detection," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2020, p. 1513, Aug. 2020.

[13] S. M. Shankaranarayana and D. Runje, "ALIME: autoencoder based approach for local interpretability," in *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, H. Yin, D. Camacho, P. Tino, A. J. Tallón-Ballesteros, R. Menezes, and R. Allmendinger, Eds. Cham: Springer International Publishing, 2019, pp. 454–463.

[14] T. M. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[15] Z. Noshad, A. Bouyer, and M. Noshad, "Mutual information-based recommender system using autoencoder," *Applied Soft Computing*, vol. 109, p. 107547, Jun. 2021.

[16] X. Luo, X. Li, Z. Wang, and J. Liang, "Discriminant autoencoder for feature extraction in fault diagnosis," *Chemometrics and Intelligent Laboratory Systems*, vol. 192, p. 103814, Sep. 2019.

[17] M. R. Islam, S. Barua, M. U. Ahmed, S. Begum, P. Aricò, G. Borghini, and G. Di Flumeri, "A novel mutual information based feature set for drivers' mental workload evaluation using machine learning," *Brain Sciences*, vol. 10, no. 8, p. 551, Aug. 2020.

[18] G. Di Flumeri, G. Borghini, P. Aricò, N. Sciaraffa, P. Lanzi, S. Pozzi, V. Vignali, C. Lantieri, A. Bichicchi, A. Simone, and F. Babiloni, "EEG-based mental workload neurometric to evaluate the impact of different traffic and road conditions in real driving settings," *Frontiers in Human Neuroscience*, vol. 12, p. 509, Dec. 2018.

[19] S. Barua, M. U. Ahmed, C. Ahlstrom, S. Begum, and P. Funk, "Automated EEG artifact handling with application in driver monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1350–1361, Sep. 2018.

[20] R. Elul, "Gaussian behavior of the electroencephalogram: changes during performance of mental task," *Science*, vol. 164, no. 3877, pp. 328–331, Apr. 1969.

[21] A. W. Corcoran, P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky, "Toward a reliable, automated method of individual alpha frequency (IAF) quantification," *Psychophysiology*, vol. 55, no. 7, p. e13064, 2018.

[22] N. Tintarev, S. Rostami, and B. Smyth, "Knowing the unknown: visualising consumption blind-spots in recommender systems," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18. New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1396–1399.