# Cognitive Load Estimation using Machine Learning


Aditi Khandelwal
*Electrical Engineering*
*IIT Delhi*
ee1180434@iitd.ac.in

Akshat Jain
*Deptt of Management Studies*
*IIT Delhi*
bsy207632@iitd.ac.in

Shauryasikt Jena
*Electrical Engineering*
*IIT Delhi*
ee1180500@iitd.ac.in



***Abstract*—Human-computer interaction applications like autonomous vehicles, medical devices etc can use the estimated cognitive load and greatly benefit from it. However the data and modelling process for such a task is complex. In this paper we propose a neural network model which uses four physiological measurements (Heart rate, skin conductance, R-R intervals and skin temperature) which are available at low cost and at large scales. The dataset we use is CogLoad2020. We benchmark the baseline model and also propose our technique. Our method outperforms the existing methods by 5.2%. The code will be available at: https://github.com/aditi184/CognitiveLoadEstimation.**



***Index Terms*—Cognitive Load Estimation, Machine Learning, Neural Network, Physiological Measurements**


## I. Introduction

Our cognitive resources, such as our attention, working memory and decision-making abilities are frequently called as Cog. Load. (Sweller, 2011). Human-computer interaction (HCI) applications that automatically adapt to a user's cognitive state might benefit from forecasting and quantifying cognitive load, thanks to new multimedia technologies, cognition-aware computing, and human centred systems (HCI). Furthermore, the capacity to accurately determine where a user's tension, annoyance and boredom may be avoided would open up new paths and opportunity, which intelligent user cantered systems an improved end user knowledge. Intelligent tutoring technologies could have been vital in aiding learning and teaching. Medical professionals may be able to benefit from the ability to detect cognitive overload in order to devise supportive methods and systems. More than three decades of research have focused on the measurement and prediction of cognitive load, which is considered a critical indicator of human cognition and performance. Self-reports, on the other hand, have a major drawback because of the subjective character of user behaviour, and due to which it difficult to identify ground truth labels. This type of questionnaire has been shown to put the user under additional strain and is not appropriate for use in online environments and a one's cognitive load should be measured while they are working (Abdelrah-man et al., 2017). It has become increasingly important to estimate cognitive load either by the technology or the procedural map i.e. using physiological data or image data. Brain imaging procedures are very precise and doable to predict users cognitive load but level of ubiquity and economical viability is still a big question to answered and subject of investigation. As a result, these methods cannot be put to use outside of the lab just yet, despite their promise.

## II. Related Works

A number of studies have lately studied the link between smooth pursuits and cognitive stress, and showed that factors obtained from smooth pursuits have a high predictive power for cognitive load.

In past decades, many approaches in relation to cognitive load have been investigated (Kramer, 1990), from conventional methods to (HartStaveland,1988) advanced Neural network (Fridman et al., 2018).

## III. Data and Preprocessing

In this section we first provide details about the generation of the dataset and then preprocessing pipeline of the data.

### A. Dataset

The approach we present in this paper was developed and tested using the CogLoad data set [1] given by UbiComp 2020 for global challenge CogLoad [2]. The data set includes four distinct physiological measurements taken by a Microsoft Band 2 wristband from 23 people while they completed six psychological activities on a computer with increasing levels of toughness. The measurements are also taken when the participants were resting. The trials were carried out in a calm, room with a constant temperature, with only one participant at a time. The dominant hand of twenty two participants is the right, whereas one member is left-handed. The wrist-band device was worn by all participants on their left hand. The average age of the participants was 29.51 with standard deviation being 10.10. There were two sections to the experimental situation. The first phase of the study focused on determining the individuals' cognitive abilities. The participants completed two N-back tasks [3], i.e., 2-back and 3-back tasks, with a 3-min rest in between them to test their cognitive capacity. The participants were given six key tasks to complete in Part 2. Three variations of a randomly chosen major cognitive-load task were shown to subject for each task. The complexity of the variations varied (easy, medium, and difficult). The cognitive load condition corresponds to 50% of the samples for each participant, while the resting state corresponds to 50%. The response is represented by a binary value in the data set, i.e., a '1' represents a low cognitive load, i.e., resting, and a '0' represents a high cognitive load, i.e., performing.

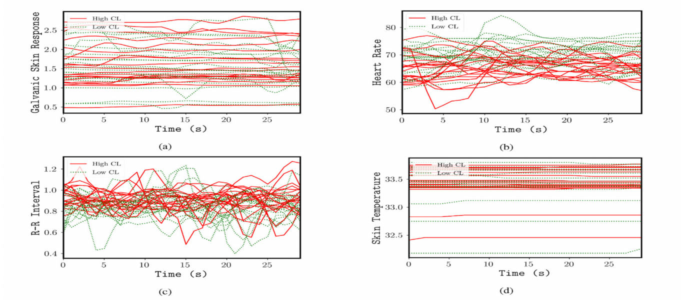


Fig. 1. Visualizing data for 1 sample. Figures represent galvanic skin responses (a), heart rates (b), R-R Intervals (c), and skin temperature (d) respectively [4]

Galvanic skin response (GSR), heart rate (HR), R-R intervals, and skin temperature are among the physiological measurements (ST). All of these readings are taken at a 1Hz sampling rate. Time windows of 30 seconds are used to create the dataset. The dataset can be found in the link provided[1].

### *B. Preprocessing the data*

The relatively small CogLoad's training data provides a severe difficulty, as time series preprocessing strategies are unable to develop discriminatory features for robust predictions. As a result, we opted to employ simple aggregate statistics as feature functions for after conducting some research on feature creation methodologies. For each of the four physiological signals (i.e., GSR, HR, RR, and ST), these include the maximum, minimum, median, mean, standard deviation, total, and skew values from the 30-second series. Finally, for feature importance analysis, we used the CancelOut mechanism [5] and the Shapley additive explanations (SHAP) framework [6]. The standard deviation of physiological data, as expected, is consistently one of the most discriminative properties.

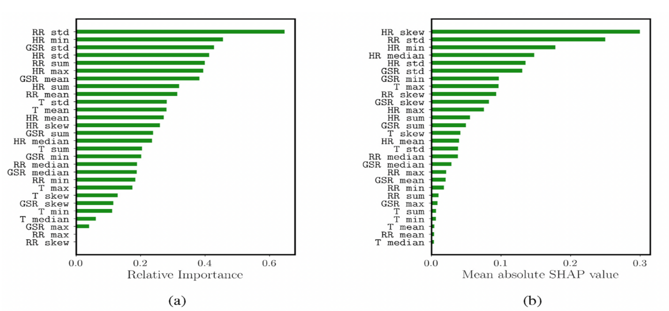


Fig. 2. Visualizing data for 1 sample. Figures represent galvanic skin responses (a), heart rates (b), R-R Intervals (c), and skin temperature (d) respectively [4]

The preprocessing pipeline has been adapted from the mentioned link[2].

[1]Dataset - https://www.ubittention.org/2020/data/Cognitive-load%20challenge%20description.pdf

[2]https://github.com/unnir/CogLoad_UbiComp2020/blob/main/pipeline.ipynb

## IV. Model

### *A. Baseline*

The ensemble learning technique to accurate and robust cognitive load detection is presented in this part, which is based on the CogLoad 2020 dataset utilized by the competition's winning entry. We treat this approach as our baseline and try to improve upon this approach. Baseline approach has the following pipeline mentioned in the Fig 3.

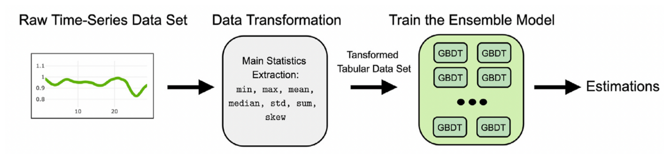


Fig. 3. Baseline model pipeline for feature engineering and ensemble learning [4]

The baseline proposes an ensemble model which employs 8 GBDT models. Based on the LightGBM implementation, Baseline creates each of the eight models [7]. Different hyper-parameters are used to train the base learners. Random search and Bayesian optimization algorithms are employed to pick hyper-parameters [8]. Because adding more models to the ensemble did not increase predictive performance, the final meta-model is made up of 8 individual GBDT models, with the final prediction being the average of all of them. The baseline model has an accuracy of 62.5%[3].

### *B. Our Method*

We try different approaches like Gaussian Naïve Bayes classifier, KNN classifier, support vector machines and neural network. We empirically found that the neural network outperforms any other method. This can be explained with the fact that neural network is also known as universal function approximator as it can fit the underlying non-linear target function. We have single hidden layer with 2 neurons having ReLU activation function. The following table I shows the results of various models that we tried and tested. As can be seen from Table I, neural network outperforms all other methods and the baseline model by 5.2%.

| Model | Accuracy | ROC-AUC | F1-Score | Recall | Precision |
|---|---|---|---|---|---|
| NN | 67.71% | 65.81% | 68% | 68% | 68% |
| GNB | 63.77% | 72.9% | 61% | 64% | 69% |
| SVM | 64.56% | 64.5% | 65% | 65% | 65% |
| KNN | 58.26% | 63.33% | 58% | 58% | 58% |

TABLE I
COMPARISON OF METHODS TRIED. NUMBERS REPORTED ARE ON THE VALIDATION SET, NN: NEURAL NETWORK, GNB: GAUSSIAN NAÏVE BAYES, KNN: K NEAREST NEIGHBOUR CLASSIFIER, SVM: SUPPORT VECTOR MACHINE,
BASELINE ACCURACY: 62.5% AND ROC-AUC:68.4%

[3]https://github.com/unnir/CogLoad_UbiComp2020

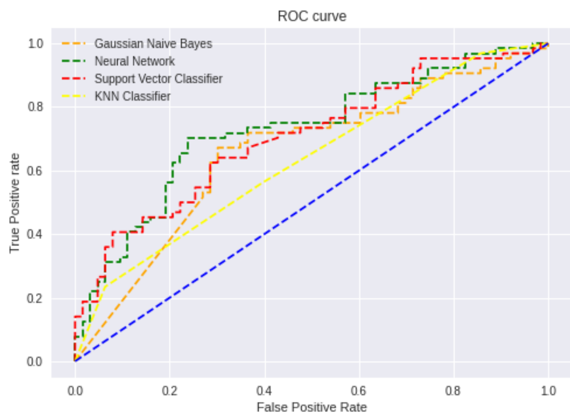


Fig. 4. ROC Curves for methods implemented

## V. Experimental Details

We split the dataset 80:20 ratio for training and validation. Our neural network architecture has 1 hidden layer of size 2 with ReLU as the activation function. We set the initial learning rate as 0.001 and use Adam optimizer for training and run the MLP for a maximum of 300 epochs. The experiments were performed on Google Colab notebook with training and inference on CPU, no GPU was used since the data set is small.

**Evaluation Metrics**: We utilized the widely used metrics accuracy, ROC-AUC, F1-score, recall and precision.

## VI. Conclusion

In this study, we proposed a neural network model which outperforms the baseline by 5.2%. It is also a better model as it requires less space and memory in comparison to the ensemble where we have to learn eight different weak learners which increases the training time and memory to store those parameters. Some future directions to this study include collection and analysis of more data and modeling the data using deeper networks which can lead to a significant boost in the accuracy of the task.

## Acknowledgment

We would like to extend our gratitude to Prof. Tapan Gandhi for designing and making course ELL890 intriguing. We would also like to thank him for giving us an opportunity to work on this project. Thanks to TA Sapna Ma'am also for sharing papers related to the datasets for the given task at hand.

## References

[1] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, V. Pejović Datasets for cognitive load inference using wearable sensors and psychological traits Applied Sciences, 10 (2020), p. 3843
[2] N. van Berkel, A. Exler, M. Gjoreski, T. Kolenik, T. Okoshi, V. Pejovic, A. Visuri, A. Voit, Ubittention 5th international workshop on smart & ambient notification and attention management (2020)
[3] F. Schmiedek, M. Lövdén, U. Lindenberger A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context Frontiers in Psychology, 5 (2014), p. 1475
[4] Borisov, Vadim Kasneci, Enkelejda & Kasneci, Gjergji. (2021). Robust cognitive load detection from wrist-band sensors. Computers in Human Behavior Reports. 4. 100116. 10.1016/j.chbr.2021.100116.
[5] V. Borisov, J. Haug, G. Kasneci Cancelout: A layer for feature selection in deep neural networks International conference on artificial neural networks, Springer (2019), pp. 72-83
[6] S.M. Lundberg, S.I. Lee A unified approach to interpreting model predictions I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in neural information processing systems, Vol. 30, Curran Associates, Inc. (2017), pp. 4765-4774
[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu Lightgbm: A highly efficient gradient boosting decision tree Advances in neural information processing systems (2017), pp. 3146-3154
[8] J. Bergstra, D. Yamins, D.D. Cox Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms Proceedings of the 12th Python in science conference, Citeseer (2013), p. 20
[9] Kramer, A. E. (1990).Physiological metrics of mental workload: A review of recent progress.
[10] Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. Multiple-task Performance, 279–328.
[11] Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results ofempirical and theoretical research. InAdvances in psychology(Vol. 52, pp. 139–183).
[12] Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Putyour thinking cap on: Detecting cognitive load using eeg during learning. InProceedings of the seventh international learning analytics&knowledge conference(pp.80–89).
[13] Abdelrahman, Y., Velloso, E., Dingler, T., Schmidt, A., & Vetere, F. (2017). Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1,1–20