Instituto Superior Técnico

Intelligent Systems

MEMec

Project

Group 7

João Moura, 83404
Hallvard Bjørgen, 105243

Professors: Susana Vieira, João Santos

21th October 2022

Contents

1 Introduction

There are multiple interesting possibilities to be explored within the dataset. This paper will explore 3:
- Is it possible to predict self-reported sleep quality based on physiological and subjective measurements of working memory tasks (N-back tests)?
- Which accuracy can we achieve for predicting what type of N-back test the participant is taking at a given time, based on the physiological and subjective measurements of the tests?
- Which is better at predicting these things; wrist-PPG or the fingertip-PPG?
- Does the accuracy of the predicting model increase when the half of participants with bad self-reported sleep quality are removed from the data?

To explore these questions, a neural network will be implemented. A fuzzy model (in matlab), recurring neural networks, long short-term memory and gated recurring units will also be explored for creating an optimal model.

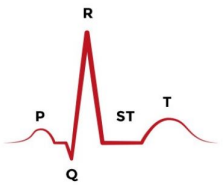
The possibility of correctly predicting which N-back test a person is taking at a given time, based on their physiological and subjective measurements of the tests is explored in great detail. Correctly predicting such outputs may be heavily influential in any workplace as it could further operational technology to encompass early detection of when a human is having greater mental workload than what the current task is supposed to lead to, possibly predicting underlying stressors meaning the worker needs a break or should be put to do less demanding tasks.

This could be, for example, used in synergy with dynamic scheduling of a pharmaceutical quality control (QC) lab. A study performed by great academics looks at the implementation of dynamic scheduling in a QC lab (Coito et al., 2022). Here, analysts time spent at each workstation is measured by wrist RFID bracelets. In such an industry, mistakes due to fatigue or underlying stressors may be disastrous if they go unnoticed. Furthermore, having the most effective usage of time is important for the profit of the operations. A possible benefit of being able to correctly predict mental workload from the MAUS dataset in this paper can lead to more accurately predicting the time spent at each workstation for each analyst, further optimizing the dynamic scheduling algorithm to increase utilization of resources (analysts).

2 The Dataset

The dataset used in this report is the MAUS dataset, retrieved through the IEEE web page for open access datasets (Beh et al., 2021). The dataset contains data collected by simple sensor measurements and subjective ratings of participants during tests of different difficulty, giving the participants varying levels of mental workload. The sensors used to measure their stress level was:

- ECG: Sample rate: 256Hz, 76800 data points per participant.
- GSR: Sample rate: 256Hz, 76800 data points per participant.
- PPG on fingertips: Sample rate: 256Hz, 76800 data points per participant.
- PPG on writs: Sample rate: 100Hz, 30750 data points per participant

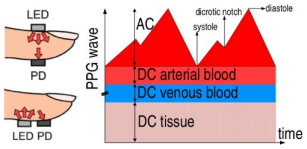


ECG

Electrocardiogram is used for measuring the heart. Sensors (electrodes) attached to the skin are used to detect the electrical signals produced by your heart each time it beats.

GSR

Galvanic Skin Response measures the change in electrical activity, which takes place due to the shift in sweat gland activity. An increase in sweat gland activity can take place due to both positive ("joyful") and negative ("scary") events. Therefore the GSR signal does not represent the type of emotions. The hands have a high density of sensitive sweat glands; thus GSR data are collected from the finger, wrist, or palm. A well-known application of GSR is a lie detection test. (Lazar et al., 2017)



PPG

Photoplethysmography is a simple and low-cost optical technique that can be used to detect blood volume changes in the microvascular bed of tissue. It is often used non-invasively to make measurements at the skin surface (Allen, 2007). A study (Rajala et al., 2018) compared PPG measurements from the finger and from the wrist, in measuring "pulse arrival time" (PAT). The result (in short) was that the PPG fingertip measurements are more accurate, but that wrist PPG may also be used for measuring PAT.

The hardware to record ECG, GSR and PPG (fingertip) was the "Procomp Infiniti", while a "PixArt PPG Watch" was used for the PPG wrist measurements.

In addition, the participants filled out Pittsburgh Sleep Quality Index (PSQI) questionnaires as a way to measure their sleep quality, and during the test, NASA Task Load Index measurement were subjectively recorded for each of them.

The stimuli of the participants was to have them take N-back tests. Here, the participant has to remember the N last numbers of a number series being flashed quickly before them. As an example, for the 2-back tests with this number sequence: 6383616 each being flashed one by one before the participant; he/she has to press the space bar when being presented with the second "3" and for the last "6", as the numbers 2 before them are the same. The participants went through 0-Back, 2-Back and 3-Back tests, in an ordering like so: 0-2-3-2-3-0. What's the 0-back test? "In the 0-back condition, the target was any letter that matched a pre-specified letter (i.e., "c"). Thus, this condition required sustained attention but no working memory demand." p.712 Miller et al., 2009.

Before the tests there was a five minute resting session (where they filled out the PSQI), plus 5 minute trials + 2 minute rest between each trial. The complete testing lasted for 47 minutes, and sensor measurements were collected for the whole duration.

WRITE ABOUT TLX VALUES!!!

The database contains 22 healthy participants (2 females) from the university's graduate students. The average age was 23 years with a standard deviation of 1.7.

To sum up, the MAUS data includes: Physiological measurements ECG, GSR, and PPG, and subjective ratings PSQI and NASA-TLX. These are all linked with the different N-back tasks/resting periods the participants were undergoing.

3 Methodology

3.1 Preprocessing the raw data

As previously stated, the dataset contains data sampled with 100Hz and 256Hz, this results in a total amount of points equal to 5 728 800. The idea here is that for each individual submitted to the trials, thousands of sensor data entries will be produced, according to their corresponding sample time, however, the NASA Task Load Index (TLX) values are measured for each trial, so a total amount of 6 sets of evaluations for each person. This *TLX* data is obviously relevant for model building when predicting which trial a person is performing, so it must be included as features. In this case, 6 different

features, for *"Mental Demand"*, *"Physical Demand"*, *"Temporal Demand"*, *"Performance"*, *"Effort"* and *"Frustration"*, which would then be repeated throughout the dataset, for each correspondent trial. This results in a lot of repeated data for these features, which will obviously create bias in the model and diminish the importance of the sensorial data.

The way that this problem is resolved in this project is by subsampling the original sensor samples, by averaging out 1 minute of data into one single sample, which will result in two things: firstly, the ammount of repeated *TLX* data will be much smaller and secondly, the sensor noise will be filtered quite a bit, resulting in less samples whilst preserving some of the original data's information. After applying this subsampling of the data points, the preprocessed dataset has size (1320,16) with 10 features as well as 6 columns resulting from one hot encoding the 6 classes (trials) considered as output.

The preprocessing is done using a function and saves a finished, preprocessed, *.csv* file to make the testing more efficient (as one does not have to preprocess each time one runs the code).

3.2 Feature Selection

Before jumping into the model building methods and their parametrization, we must first select the most relevant features through some feature selection. Through an initial evaluation of the data, one would assume that all features are beneficial for the accuracy and performance of the models that will be built, however this requires confirmation. The sensor data is highly nonlinear so correlation methods are not advisable, one should approach this in a different manner, for example by removing specific features and checking the effect on the produced model's performance. To do this, we will use a Fuzzy Takagi-Sugeno model using the *fmid* toolbox in MATLAB (Abonyi, 2022), with 6 clusters, product-space MFS for the antecedent and global LS for the consequent. For each considered scenario, 5 models will be built and their accuracies averaged out and shown. The scenarios and results are shown in table 1.

Feature Selection	
Dropped data-variables	Model average accuracy
None	**0.95192**
All sensor variables	0.88282
All *TLX* variables	0.23028
inf_ecg_csv	0.94598
inf_gsr_csv	0.94544
inf_ppg_csv	0.95152
pixart_csv	0.93586
Mental Demand	0.91366
Physical Demand	0.92980
Temporal Demand	0.90806
Performance	0.95150
Effort	0.91162
Frustration	0.93990

Table 1: Accuracy feature selection

It must be noted that since this is a multi classification problem, the accuracy is measured by a precise match of the predicted output for all 6 trials with the real result, i.e., the output vectors (size 6) must match.

As we can see, the accuracy is highest for the case where no variables are dropped, which suggests all variables provide relevant information, albeit in different senses. Also, we can immediately notice that removing *TLX* data completely kills the model performance as this data proves to be extremely important in prediction. Still, sensor data proves to be relevant as excluding it lowers performance quite a bit. This might be explained by the fact that *TLX* data is enough to predict most of the "easier" cases, but is not enough for some more specific cases in which sensor data is extremely

important as it provides additional information serving as a "tiebreaker" of sorts, aiding the model in correctly predicting these cases, resulting in a quite considerable improvement in accuracy from sensor data, even though it does not perform well by itself.

Another aspect that must be noted here is that removing one of the sensor variables does not worsen the performance by a lot. This suggests that each sensor variable is not very relevant for prediction individually, but rather the combination of them is.

The takeaway in this brief feature selection process is that discarding any features will most likely produce a worse outcome than keeping them in, and thus the data utilized in the models that will be built will include all of them.

3.3 The Neural Networks

There are a multidue of available neural networks to choose from (see figure 1 from (van Veen and Leijnen, 2019)). Some good options are recurring neural network (RNN), long short term memory (LSTM), and gated recurring units (GRU) (that are different variations of each other).

From further discussions, however, two other seemingly promising neural networks that are not on Veen and Leijnen's list (van Veen and Leijnen, 2019) were selected; the single-link chaining and the multi-link chaining neural networks (Zaamout and Zhang, 2012). These neural networks and their implementation details will be discussed in further detail in chapters 5,6. The results will be compared to those of a typical neural network (see chapter **??**) and the results of a fuzzy model (see chapter 7).

4 Deep neural network

From assignment 2, a neural network was developed to predict income level of random U.S. residents. That is a different problem from what is the focus of this paper and therefore the parameter and implementation might not be perfect for this case as well.

Therefore, the parameters needed some changing for it to be a usable benchmark for the other networks.

The parameters to achieve an accuracy of **85.61%** were 300 epochs, a batch size of 30 and a learning rate of 0.005. The layers are shown in table 2. This result was obtained after running several times, and calculating the accuracy from $(TP + TN)/N$.

Layer	**Neurons**
ReLU	40
ReLU	20
ReLU	20
Sigmoid	6

Table 2: Parameters for neural network

5 Single-link chaining neural network

Instead of just "wasting" the outputs of our model for each target prediction, this neural network will do a binary classification of each output based on previous output. The neural networks are chained together, using the outputs of the previous trial to more accurately prediction of the next.

This is what's referred to as a single-link chaining (SLC) neural network (Zaamout and Zhang, 2012). In this article, an experiment was conducted on 16 datasets to establish whether single-link chaining or multi-link chaining (MLC) is the better method for classification. For their experiment, MLC outperformed SLC and a typical neural network in 8/16 experiments, while the SLC outperformed MLC and the typical network in 6/16. The typical neural network was better in 2 of the experiments. According to the article, this aligns with previous research. Therefore, there is much hope that chaining the networks together may lead to increased predictions in total.

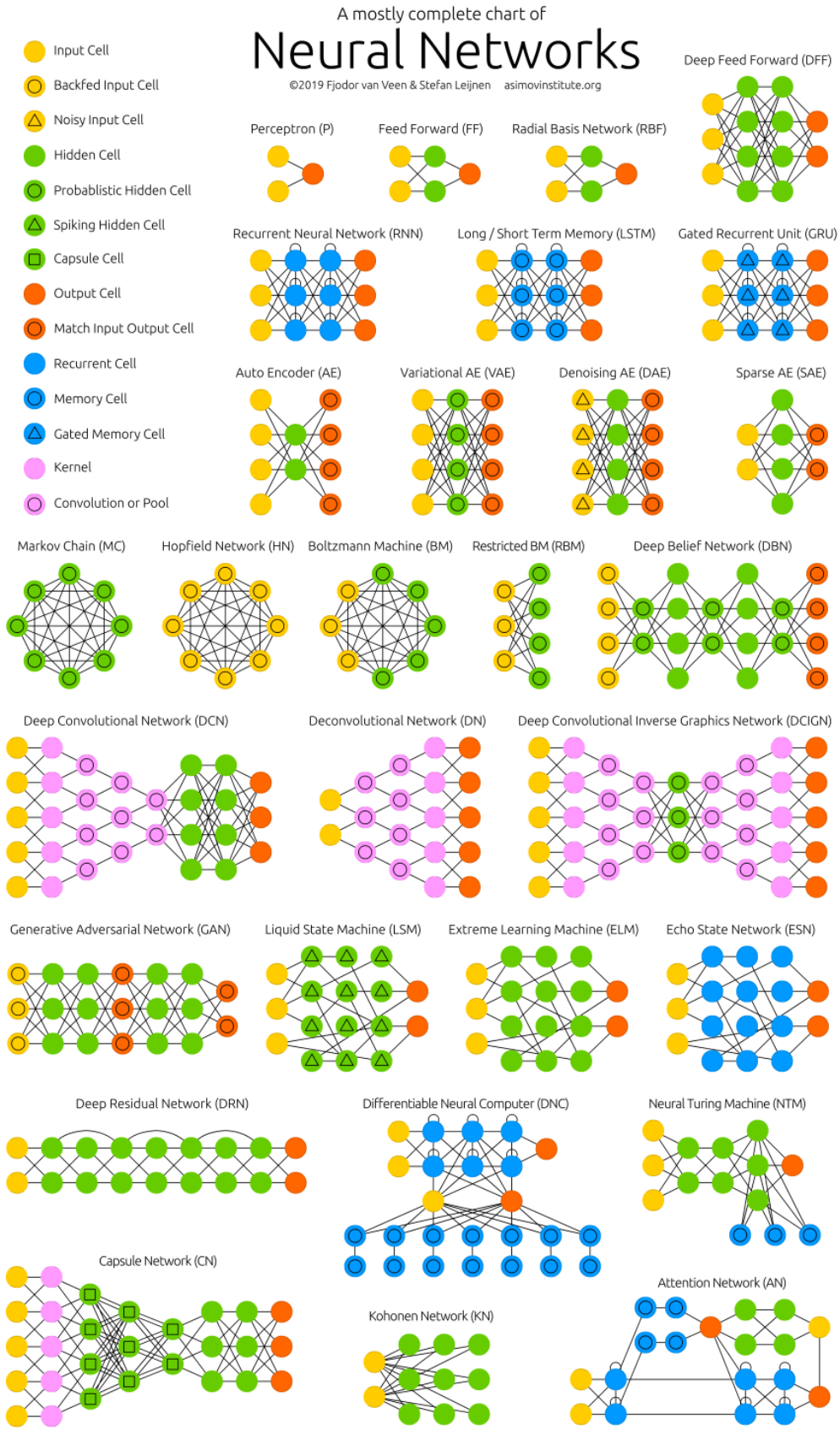


Figure 1: Neural Networks Zoo (van Veen and Leijnen, 2019)

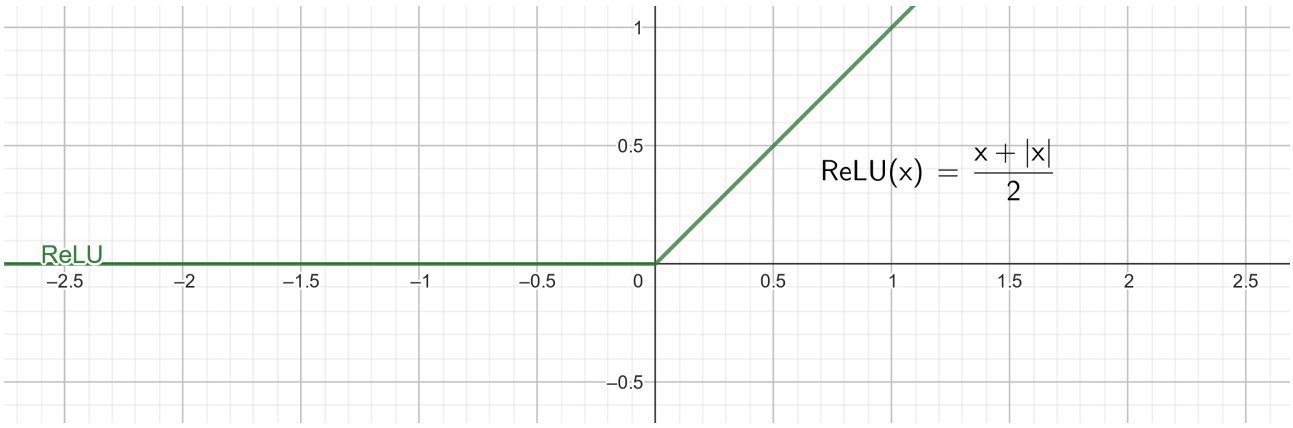


Figure 2: ReLU activation function. Illustration made in Geogebra.

Also, since we know from our deep neural network which trial has the highest accuracy in being predicted correctly, we can (hopefully) increase accuracy by placing this most accurate model first, and then place the trials in succession from most accurate to least. That this will increase the predictions is a hypothesis we have yet to test for...

5.1 Results and discussion

The result of running each in normal succession with the following parameters: Rectified Linear Unit (ReLU) activation function for the hidden layers (same amount of nodes as in the deep neural network) and Sigmoid activation function for the output layers, 200 epochs, 50 batch size, 10^{-3} learning rate is shown in figure 4.

SLC and MLC with binary outputs, Sigmoid should be the better output layer activation function according to previous literature.

A trap in the evaluation is to measure each of the 6 outputs by themselves. Since we always will have one trial that is correct and five that are not correct, it will, when predicting only one correct output, always get four outputs that are evaluated to being correct. This is not preferable, as the accuracy will not reflect the true accuracy of the model. Therefore, either evaluating the model by saying all predictions need to be correct at once for it to be evaluated as a correct prediction, or by using precision (looking at true positives) and recall (looking at true negatives) as evaluation method for our models.

Accuracy is bad measurement no matter what implementation we do.. right? So we should always look at precision or recall instead.

raw data has one output

The ReLU activation function (see figure) is effective for the hidden layers due to ...

The Sigmoid activation function (also known as "the logistic activation function") is used on the output layers to keep all values between 0 and 1, and to not change the outputs too much. This function moves all negative values to a positive range.

In figure 5 the training and validation accuracies, and figure 6 shows the plots for the losses....

The accuracies seem to fluctuate so much between the

6 Multi-link chaining neural network

7 Fuzzy model

Following the Neural Network models and their results, we now build a Fuzzy model for the dataset. These models consist mainly in multiple sets of "if then" rules that associate a given numerical input to a fuzzy set with a certain membership degree between 0 and 1. These fuzzy sets can be described and interpreted with words, which in contrast to Neural Networks, provides some interpretability of the obtained model. Typically, for a higher ammount of features, the number of rules will increase a

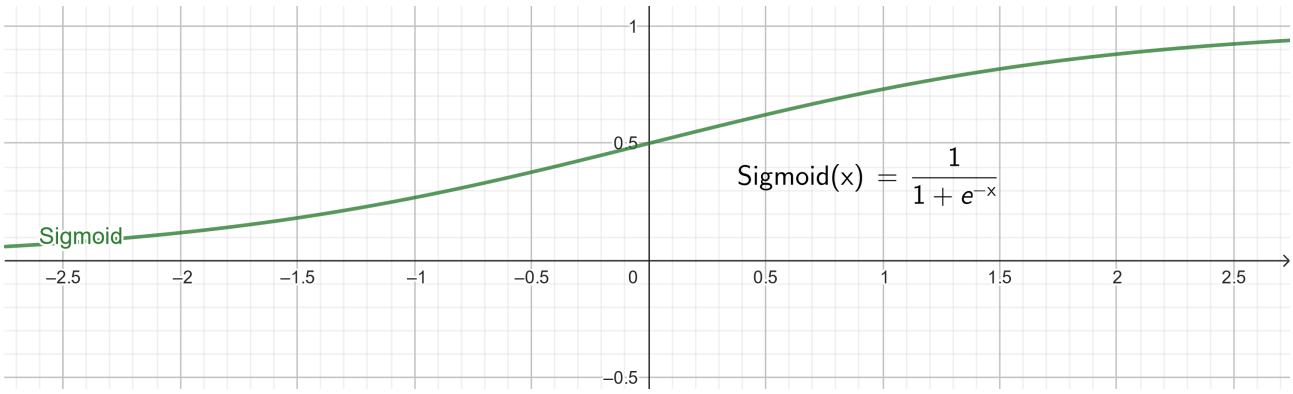


Figure 3: Sigmoid activation function. Illustration made in Geogebra.

Predicted \ Actual	Positive	Negative	Total
Positive	109	2	111
Negative	11	10	21
Total	120	12	132

Accuracy: 0.90152

(a) Trial 1

Predicted \ Actual	Positive	Negative	Total
Positive	90	16	106
Negative	4	22	26
Total	94	38	132

Accuracy: 0.84848

(b) Trial 2

Predicted \ Actual	Positive	Negative	Total
Positive	116	0	116
Negative	7	9	16
Total	123	9	132

Accuracy: 0.94697

(c) Trial 3

Predicted \ Actual	Positive	Negative	Total
Positive	98	5	103
Negative	12	17	29
Total	110	22	132

Accuracy: 0.87121

(d) Trial 4

Predicted \ Actual	Positive	Negative	Total
Positive	108	3	111
Negative	9	12	21
Total	117	15	132

Accuracy: 0.90909

(e) Trial 5

Predicted \ Actual	Positive	Negative	Total
Positive	101	12	113
Negative	0	19	19
Total	101	31	132

Accuracy: 0.90909

(f) Trial 6

Figure 4: Confusion matrices and accuracies. Parameters and discussion are found in 5.1

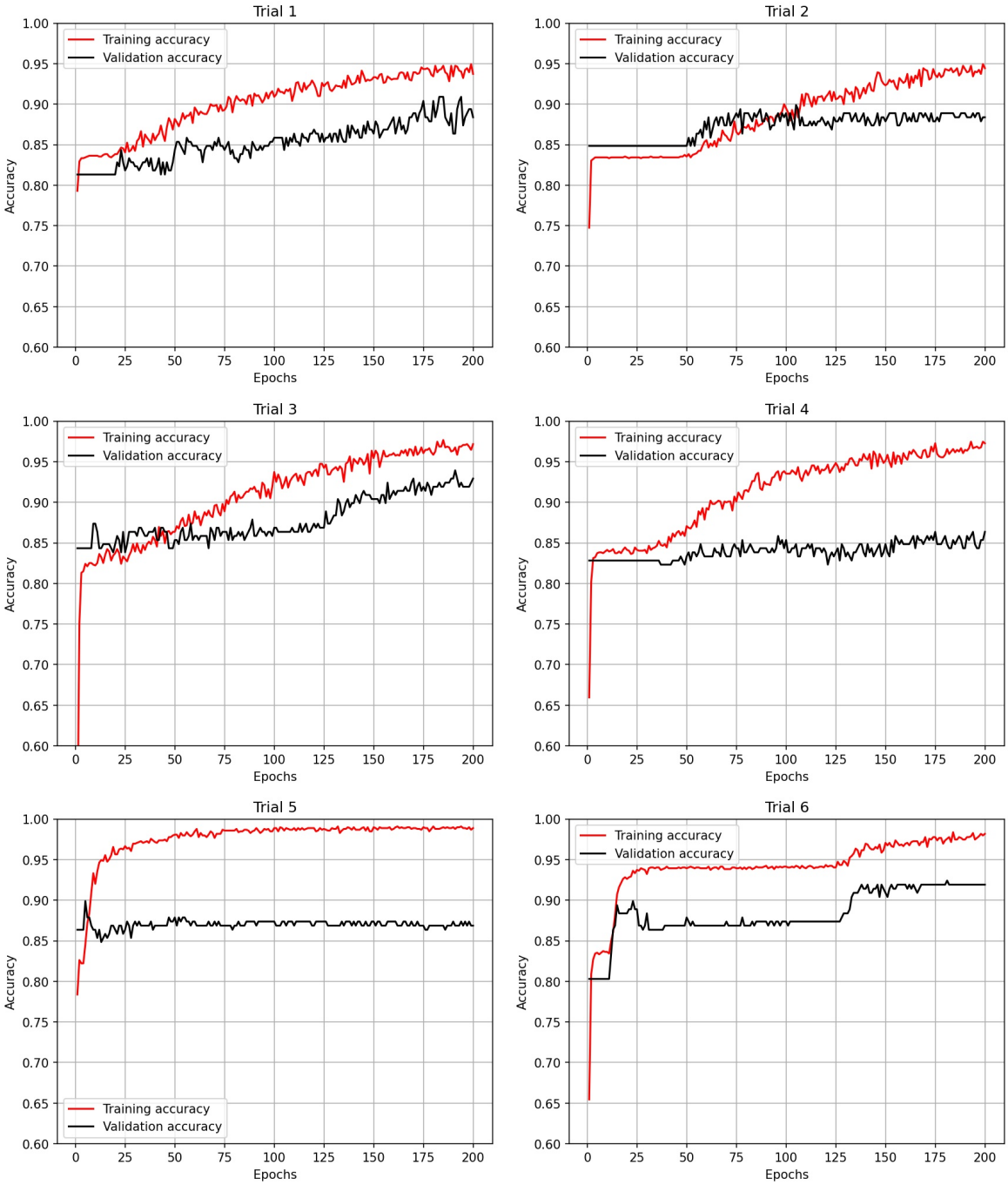


Figure 5: Training and validation accuracy plots

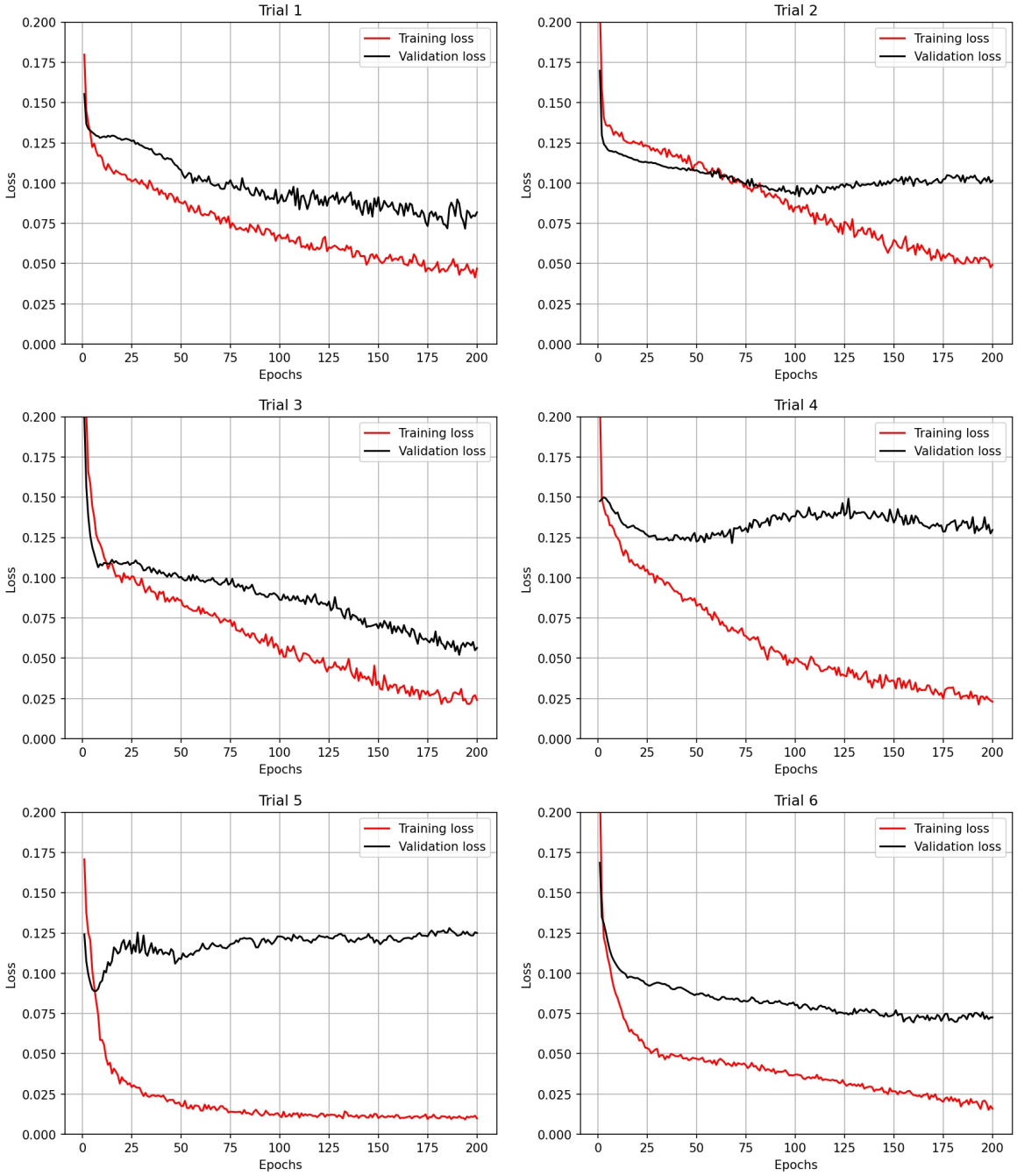


Figure 6: Training and Validation loss plots

lot which results in many fuzzy sets. With 10 features in the dataset, interpreting the rules and fuzzy sets is possible, however, without extensive knowledge of the medical and psychological aspects of the data and its origin, it would be very hard to do so, not only because of lack of expert knowledge, but also due to the fact it is a 10-dimensional problem.

The Fuzzy model that will be built for this project is a first-order Takagi-Sugeno model, using the *fmid* MATLAB toolbox and Fuzzy c-means clustering that will assign each input to various clusters through membership degrees. For the feature selection process previously shown and described, 6 clusters (one for each class) were used together with product-space MFS for the antecedent and global LS for the consequent. We will now see if these parameters are indeed ideal, and we will also play with the fuzziness, m, parameter and observe its effect on accuracy and variance.

7.1 Parametrization

The parametrization process will consist in producing multiple models with various combinations of parameters in an attempt of optimizing them to produce the best model possible. We will first play with the antecedent and consequent model parameters, that is, the operators and methods used for them, next with the number of clusters and lastly, the fuzziness parameter.

7.1.1 Antecedent and Consequent methods

For the computation of the antecedents, the provided toolbox allows for two choices: product-space MFS and projected MFS.

7.2 Best model and discussion of results

Abonyi, 2022

8 Results and Discussion

9 Bibliography

References

Abonyi, J. (2022). *Constrained fuzzy model identification - files for fuzzy modeling and identification toolbox*. https://doi.org/10.21227/q4td-yd35

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, *28(3)*. https://doi.org/https://doi.org/10.1088/0967-3334/28/3/R01

Beh, W.-K., Wu, Y.-H., & Wu, A.-Y. ((2021). *Maus: A dataset for mental workload assessment on n-back task using wearable sensor*. https://doi.org/10.21227/q4td-yd35

Coito, T., Firme, B., Martins, M. S. E., Costigliola, A., Lucas, R., Figueiredo, J., M.Vieira, S., & Sousa, J. M. C. (2022). Integration of industrial iot architectures for dynamic scheduling. *Computers & Industrial Engineering*, *171*. https://doi.org/https://doi.org/10.1016/j.cie.2022.108387

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Chapter 13 - measuring the human. In J. Lazar, J. H. Feng, & H. Hochheiser (Eds.), *Research methods in human computer interaction (second edition)* (Second Edition, pp. 369–409). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-805390-4.00013-3

Miller, K., Price, C., Okun, M., Montijo, H., & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, *24*(7), 711–717.

Rajala, S., Lindholm, H., & Taipalus, T. (2018). Comparison of photoplethysmogram measured from wrist and finger and the effect of measurement location on pulse arrival time. *Physiological measurement*, *39*(7), 075010.

van Veen, F., & Leijnen, S. (2019). *The neural network zoo*. https://www.asimovinstitute.org/

Zaamout, K., & Zhang, J. Z. (2012). Improving neural networks classification through chaining. *International Conference on Artificial Neural Networks*, 288–295.