REVIEWER´S ASSESSMENT OF FINAL WORK

I. IDENTIFICATION DATA

Title:	**Application of Machine Learning Methods for the Identification of Proton Decay in Liquid Argon Detector**
Author´s name:	**Anna Guľa Gartman**
Type of assignment:	Bachelor Project
Faculty:	Faculty of Nuclear Sciences and Physical Engineering (FNSPE)
Department:	Department of Mathematics
Reviewer:	Ing. Kristina Jarůšková
Reviewer´s affiliation:	FNSPE Czech Technical University in Prague; CERN

II. ASSESSMENT OF CRITERIA

Work assignment **demanding**

Assess how demanding the work topic is.

The work assignment follows a standard structure – familiarize oneself with the high energy physics experiment DUNE and with the task of the proton decay classification, prepare the provided data and choose suitable ML methods to perform classification. However, I classify the assignment as rather difficult because the individual steps are technically demanding, and it mostly involves knowledge and skills not covered in the bachelor's courses.

Fulfilling the assignment **fulfilled**

Consider whether the work submitted meets the assignment. If necessary, give your comments on items of the assignment not fully answered, or judge whether the scope of the assignment has been broadened. If student failed to fully treat the assignment, try to assess the importance, impact and/or the reasons for the failings.

The student fulfilled the assignment without reservations.

Chosen approach to solution **appropriate with reservations**

Assess whether student applied a correct approach or method of solution.

The use of deep learning for event reconstruction became a standard approach in high energy physics already a few years ago. ResNet-type of architecture was already successfully applied to a classification task based on a different set of simulations from the same experiment, so the ResNet model can be considered a reasonable choice for the proton decay classification as well.

I do have reservations about the use of the EfficientNet B2 model with weights pre-trained on the ImageNet. The ImageNet data and the proton decay data are of a very different nature. For this reason, I do not consider the pre-trained EfficientNet B2 model to be a suitable choice for the demonstration of transfer learning.

Professional standard **average**

Assess the professional standard of the work, application of course knowledge, references, and data from practice.

The theoretical chapters of the thesis are in general of a good quality, especially the introduction of the DUNE experiment. Chapter 3 could be improved with a better mathematical description of the operations performed by the network layers. The given mathematical formulas should be more precise (e.g. stating the possible values of the parameter beta in the F-measure formula 3.12). In chapter 3, I would expect a description of the binary classification task based on the score returned by a ML model, as well as a description of the influence the choice of the classification threshold has on various classification metrics.

I appreciate that the student managed to process in this thesis actual simulations from the DUNE experiment, even though the original format of the data was unsuitable for use with the standard deep learning frameworks.

The last two chapters (methodology and results) could be organized in a better manner. Similar information is sometimes repeated, and the results of individual experiments are scattered in the text. Overall, it could use more clarity. I appreciate that chapter 6 shortly discusses results of other deep learning applications of a similar type that were performed on simulations from the same experiment.

Unfortunately, I must reproach the quality of most images included in the thesis. Several images have low resolution and most of the graphs contain text in a font size that is too small.

Level of formality and of the language used **average**

Assess the use of scientific formalism, the typography and language of the work.

I am pleased with the level of English demonstrated in the thesis. There was only a negligible number of typos and grammar mistakes, except for a few mistakes in the punctuation (e.g. p. 16). A few typographical mistakes occur in the thesis (missing non-breaking spaces, a few abbreviations used before being explained).

Choice of references, citation correctness **average**

Assess student´s effort in finding and using study sources for completing their work. Give characteristics of the references chosen. Assess whether student made use of all the relevant sources. Verify whether all items used are properly distinguished from the results obtained by student and their deliberations, whether there are no violations of citation ethics, and whether the bibliography presented is complete and complies with the citation usage and standards.

The choice of references is adequate for a bachelor thesis, the sources are varied enough. However, there are a few instances of missing references in the text (e.g. some of the claims on p. 2, gated fusion network on p. 29).

Further comments and assessment

Give your opinion on the quality of the main results obtained in the work, e.g. the theoretical results, or the applicability of the engineering or programming solutions obtained, publication outputs, experimental skills, and the like.

The student managed to train two ensemble models for binary classification. The late fusion model seems to perform well, the ROC and precision-recall curves look very promising, as well as the AUC values. I have reservations about the processing of the output scores of the model, as the decision threshold for the final classification was not optimized to maximize any specific metric. The threshold was set to 0.5, which is not an optimal choice for the final models (see fig 6.2 and 6.4). The calculated values of the F1-score are then lower than they could be.
I also disagree with the choice of recall as the guiding metric for the hyperparameter optimization. This metric reaches it maximum value when everything is classified as signal (which can be clearly seen in the precision-recall curves).

On the other hand, I do acknowledge that the thesis required obtaining a lot of new technical skills (python, PyTorch, Optuna and Weights&Biases for hyperparameter optimization) and the student used advanced architectures and methods (ResNet, ensemble model, transfer learning).

III. OVERALL ASSESSMENT, QUESTIONS TO BE ASKED DURING THE WORK DEFENCE, SUGGESTED GRADE

Summarize those aspects of the work that were significantly influential for your overall assessment. Suggest questions to be answered by student during the defence of the work before the examination board.
From the formal point of view, I negatively evaluate the quality of the images and graphs, and several missing references. Concerning the content, a definition of the classification task in terms of probability scores.

In the practical part, I would prefer to see the classification threshold being optimized to maximize a chosen classification metric. I do not support the choice of recall as the evaluation and optimization metric and the use of the pre-trained EfficientNet B2 for reasons stated above.

However, the assignment itself was ambitious and technically demanding (data preparation, deep learning framework, frameworks for hyperparameter optimization). I am pleased that the student was able to process data from the DUNE experiment in this thesis and I also appreciate the level of English.

Taking all the positive and negative aspects into account, I suggest evaluating this thesis with the grade C, if the student is able to answer the following questions.

1. Could you provide more details on how you split the data to training, validation, and test sets? In figure 5.6, we can observe a significant difference between the training loss and validation loss. This may point to differences between the training and validation data. Were the data shuffled before the splitting?

2. The individual ResNet models perform better on the training data than the validation data (figure 5.6). Contrary to that, the late fusion model (built on top the ResNet models) performs better on the validation data. Have you inspected this discrepancy and tried to explain it?

3. Can you justify the choice of the recall as the guiding metric for the hyperparameter optimization?

4. Can you elaborate on your choice of the pre-trained EfficientNet B2 model? Have you tried training the EfficientNet B2 on the proton decay data from random initialization (given that one epoch takes 14 minutes)?

5. Given the size of the models, the amount of training data is rather small. Do you plan (if you have the possibility) to train the models on larger amounts of data?

Suggested grade: **C - good.**

Date: 24/01/2024

Signature: