ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

THESIS SUPERVISOR'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Automatic event recognition for Higgs boson detection**
Author's name:	**Bc. Jakub Malý**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Cybernetics
Thesis supervisors:	prof. Dr. Ing. Jan Kybic (supervisor), doc. Dr. André Sopczak (co-supervisor)
Supervisor's department:	Department of Cybernetics (JK)

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — challenging

How demanding was the assigned project?

While the main focus of this project was in machine learning, it also required the student to familiarize himself with the concepts of high-energy particle physics, the Large Hadron Collider in CERN, and the data formats and some software used in CERN. Another difficulty came from the necessity of communicating in English with one of the supervisors. The machine learning itself is challenging because Higgs boson events are exceptionally rare with respect to other events. Last but not least, the anti-coronavirus measures reduced the chances of frequent in-person meetings and thus reduced the possibility of interaction between the student and the supervisors.

Fulfilment of assignment — fulfilled with minor objections

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The student succeeded in familiarizing himself with the new domain, learned how to obtain, curate and process the data, and performed an extensive experimental evaluation of several classification algorithms, including some parameter tuning, On the other hand, he did not use any of the more advanced methods (e.g. deep learning) and the obtained accuracy is not very high.

Activity and independence when creating final thesis — A - excellent.

Assess whether the student had a positive approach, whether the time limits were met, whether the conception was regularly consulted and whether the student was well prepared for the consultations. Assess the student's ability to work independently.

The student was very active and motivated and interacted with both supervisors regularly. He was able to work independently and did not hesitate to ask questions. The student was generally willing to accept our suggestions for changes and future work, except at the very end, where we started to run out of time.

Technical level — B - very good.

Is the thesis technically sound? How well did the student employ expertise in his/her field of study? Does the student explain clearly what he/she has done?

The student demonstrated his ability to learn the basics of a completely new domain (particle physics) and to work with an international team. He mastered all the necessary online tools for low-level processing and management of the large data sets, as well as for the machine learning itself. He successfully applied several standard machine learning classification algorithms. The thesis will certainly serve as a starting point for future work. On the other hand, the student did not manage to employ any advanced methods (e.g. deep learning) nor any non-trivial features. Both the data and the classifiers are treated as black boxes - there is little evidence of the student getting a deeper insight and applying a well justified strategy for improving the performance. I also have doubts whether the relative frequencies of the classes in the simulated data were properly taken into account.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

THESIS SUPERVISOR'S REPORT

Formal level and language level, scope of thesis — C - good.

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The presentation is the weakest part of the theses. It must be rather hard to understand for people unfamiliar with the work, since many concepts are not clearly defined or not defined before being used. The terminology is sometimes confusing. The text is written more as a diary, with later text sometimes superseding the previous one. The thesis is not well structured - there is a lot of near-repetitions and boilerplate text. There are a lot of images in the appendices but without almost any comments or analysis. Previous work, general theory, and the student's own contribution are intermixed. The results are scattered throughout the text and not being properly discussed. What I am missing most is a joint comparison of the final results of the different methods with the state the art.

Selection of sources, citation correctness — B - very good.

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

The sources are chosen correctly, although only a minority of them concerns machine learning. However, the formatting leaves to be desired - missing bibliographical data, lowercase letters instead of uppercase, etc.

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

I appreciate that the student created a practically usable software pipeline, which implements the discussed techniques. He showed that machine learning techniques can be used to find events of interest in the collider data and can be competitive with classical handcrafted rule-based classifiers. While I would have preferred a more groundbreaking results, I feel that the work done more than satisfies the requirements for a master thesis. And I believe that given a little more time, the presentation could also be improved to be more clear, concise, and focused, to do justice to do work which has been done.

The review of my co-supervisor, Andre Sopczak, is attached to this document.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading.

In spite of the reservations expressed above, I am happy with the results. The student was motivated, has worked diligently and has fulfilled the set goals.

The grade that I award for the thesis is A - excellent.

Date: 05/28/20

Signature:

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

THESIS SUPERVISOR'S REPORT

Review by the co-supervisor Andre Sopczak:

The goal of the master thesis by Bc. Jakub Maly is the separation of signal Higgs boson events produced in proton-proton collisions at the LHC at CERN from background events which resemble the signal. Higgs boson research remains at the forefront of particle physics. In 2012 the Higgs boson was discovered, and in the subsequent years several hundred physicists are working on the determination of the Higgs boson properties, for example to determine the production and decay mode, as well as their relative rates. For the study of the Higgs boson is it crucial to obtain samples of detected Higgs boson events with little as possible contamination of non-Higgs bosons, called background. The challenge of this thesis project is to separate pre-selected Higgs boson events which were produced in association with two top quarks (ttH) from events where the Higgs boson is replaced by a W boson or a Z boson, called ttW and ttZ, respectively.

Machine learning techniques are well suited as the separation can be performed by using features of the events which on only differ slightly for the signal and background reactions. In order to optimize the separation several machine learning algorithms were applied. Jakub Maly worked very systematically on each algorithm and tuned the algorithms for best performance.

As this project combined particle physics research and cybernetics, Jakub had to familiarize himself with basic terminology of particle physics. He demonstrated in the discussions during his thesis work and within his thesis that he understood very well how the definitions of for example efficiency, purity and significance relate. Furthermore, important are the correlation matrices and the ordering of the features regarding their performance.

Jakub has been very quick in responding to requests and proved on several occasions that he is capable of conducting independent research. He has been fast in understanding new concepts and follow up on specifics tasks. For example, the conversion of the particle physics data into a format accessible for machine learning algorithms, the choice of the machine learning algorithms.

A strong point in his research is his transparency and his effort to provide enough details that his result can be checked. This has been in particular been important when he converted ML results in efficiencies used in particle physics.

Jakub has always been punctual for discussion appointments and he was well prepared.

In the discussion it became also obvious that Jakub can express well the scientific work, and he asked the right questions.

Towards the end of his thesis project, he had the opportunity to present his research in a regular meeting at CERN by video. In this meeting experts discuss in particular the ttH and ttW analysis. Jakub prepared very well a 30min presentation, he gave a good rehearsal, and his actual presentation was well received. He showed that he had a good understanding and contribute to the advancement on a high level. A fruitful discussion with the experts followed his presentation. Jakub is a good communicator in English.

Overall, Jakub performed very well during his project. The task has been challenging scientifically as the separation of signal and background relies on small differences of their features. His systematic approach, and willingness to learn the basics terminology in particle physics contributed to the good result of his thesis, and the acceptance of his results by particle physics experts. A plus is that his research resulted also in questions which should be followed up in the future.

REVIEW OF MASTER THESIS

Name of the student: Jakub Malý
Thesis title: Automatic event recognition for Higgs boson detection
Name of the reviewer: Boris Flach
Institution: Czech Technical University in Prague, Faculty of Electrical Engineering

1. Results of the work and thesis structure

The master thesis presented by Jakub Malý aims at testing and analysing different machine learning approaches for classifying particle collision events obtained in the ATLAS detector of the CERN large hadron collider. In particular, it focuses on detecting events that may produce Higgs bosons. The prime motivation of the thesis is to analyse standard machine learning approaches and their suitability for predicting these rare events with high precision.

After a short introduction, the author gives an overview of the experimental setup and some relevant background of elementary particle physics. The next two chapters describe the structure of the available data and recall the basics of statistical pattern recognition. Chapter 6 describes the considered machine learning approaches, applies them on the data and validates their results. The spectrum of analysed approaches ranges from simple methods like k-neighbors classifiers over AdaBoost to more complex approaches like random forests and neural networks. The simulated data used for training and validation of the learning approaches were obtained from CERN. The last two chapters describe changes and adaptations that where necessary for applying the methods on real data obtained by the author from CERN in the last period of his work.

The thesis concludes with a summary of the authors findings and a personal comment on the impact of the pandemic restrictions on his work.

2. Comments

The thesis is written in a linguistically competent way. Its overall structure is appropriate. It reveals, however, weaknesses on a finer level that make it difficult to read. The main reason is that author has scattered concept details across several chapters. For instance, I would have expected to find all important facts about the data and their structure as well as all details about the chosen pre-processing in Chapter 4. (Data). Instead, relevant pieces are scattered in chapters 4-6 with partially wrong reference links.

The choice of the loss function/validation criterion is an important, application dependent design option for predictors and machine learning approaches. The author has chosen to follow CERN recommendations and proposes to use a significance score. Unfortunately, the thesis part explaining this loss is not well written and details remain unclear as a consequence. Again, I would have expected to find all related facts in Chapter 6. (Classification) and not scattered over several chapters.

Almost all machine learning approaches considered by the author are discriminative methods, i.e. methods that either learn a predictor by empirical risk minimisation or learn predictive posterior class probabilities in some model class. It remains unclear for me, why these approaches require class weights in case of unbalanced training sets, as long as the prior class probabilities do not change.

The list of references is appropriate. However, some references have missing bibliographical details.

3. Defense questions

Q1: Give a concise explanation of the significance score loss proposed in your thesis.

Q2: A predictor has been trained for classifying patterns by predicting the posterior class probabilities $p(y \mid x)$. Let us assume the 0/1 loss. Consider the situation that the class frequencies in the training set differ from the true prior class probabilities (at inference time). Explain how to use the predictor without re-training it.

Q3: Give possible options for losses that can be used in situations where we want detect rare events of some class on the background of possibly several other classes.

4. Conclusions

The thesis reflects a substantial amount of work performed by the author. Despite the weaknesses mentioned above, it fulfills the criteria of a graduation thesis. Moreover, I can imagine that working with data from a large research entity (as CERN), may require extra efforts due to possibly rigid regulations and procedures. Therefore I recommend to accept the thesis for the defense and grade it with 'C' (good).

Dresden, 15.06.2020

Dr.rer.nat.habil. Boris Flach