CERN summer studentship 2021

Prediction of RPC efficiency using machine learning techniques

Supervisor:

Dr. Andres Leonardo Cabrera Mora

Departamento de Física
Universidad de los Andes. Bogotá, Colombia.

Author:

Erfan Abbasgholinejad

Physics Department
Sharif University of Technology. Tehran, Iran.

Abstract

In this work, K-NN classification and regression models are used to estimate the efficiency for the Resistive Plate Chambers (RPC) of the Compact Muon Solenoid (CMS), one of the Large Hadron Collider (LHC) experiments located at CERN between the French and Swiss border. Measurable detector quantities are used as input parameters to estimate the efficiency. Ten datasets taken by the CMS experiment during 2018 are used. The first set (80% of the data) is used to perform a training for the K-NN model, and the second set (20%) is used to make an efficiency prediction using the training performed with the initial set. Finally, a comparison between the real value of the efficiency and the value of the prediction is performed. The results obtained are promising.

August 27, 2021

Contents

List of Tables

List of Figures

1 Introduction

1.1 CMS

The Compact Muon Solenoid (CMS) is one of the four main detectors built on the Large Hadron Collider (LHC) at CERN. The massive cylindrical detector has a diameter of 15 meters, a length of 28.7 meters, and weighs about 14000 tonnes. At the center of the detector, as shown in Figure 1, proton beams collide at a center of mass energy of 13 TeV. Each beam is formed by bunches of protons, containing approximately 10^{11} protons per bunch. These bunches cross simultaneously every 25 nanoseconds. In each of these crossings, many collisions occur simultaneously, an effect called pileup. During 2017 and 2018, the average number of pileup collisions was 32 [1]. CMS is designed to collect information on interaction outcomes. It consists of different subdetectors used to detect particles such as photons, hadrons, electrons and muons, as shown in Figure 2.

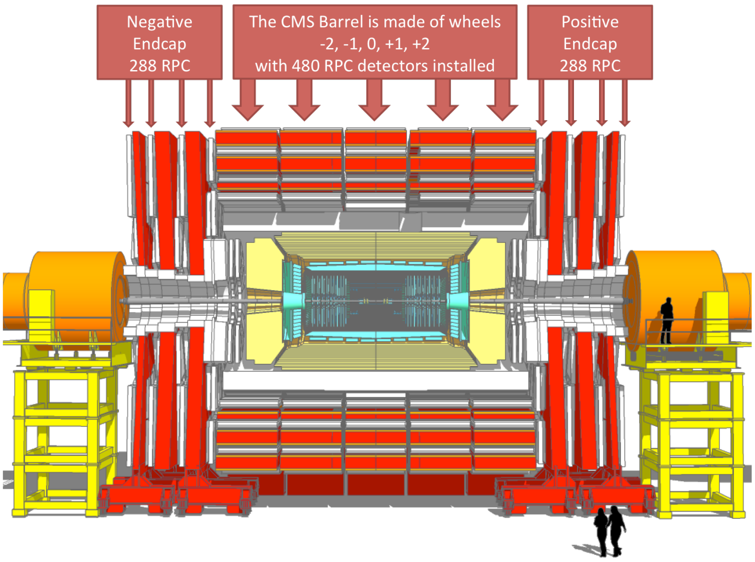


Figure 1: Distribution of the RPC detector in the CMS experiment. The detector is composed of five wheels (barrel region) at the center and two caps on the sides (endcap region). The central region (CMS barrel) contains 480 RPC detectors. The endcap region consists of 8 disks (4 disks on each side) containing 576 RPC detectors [2].

1.2 RPC

The Resistive Plate Chambers (RPC) are gaseous detectors [4] combining both a good spatial (10 mm) and time (1 ns) resolution [5]. Despite its fast time resolution, the electronics currently installed in the detector can only identify particles with a time resolution of 25 ns. Each RPC contains two gaps, each gap is formed by two parallel plates of high resistivity separated by 2 mm thick spacers in which the gas is enclosed. As a charged particle crosses

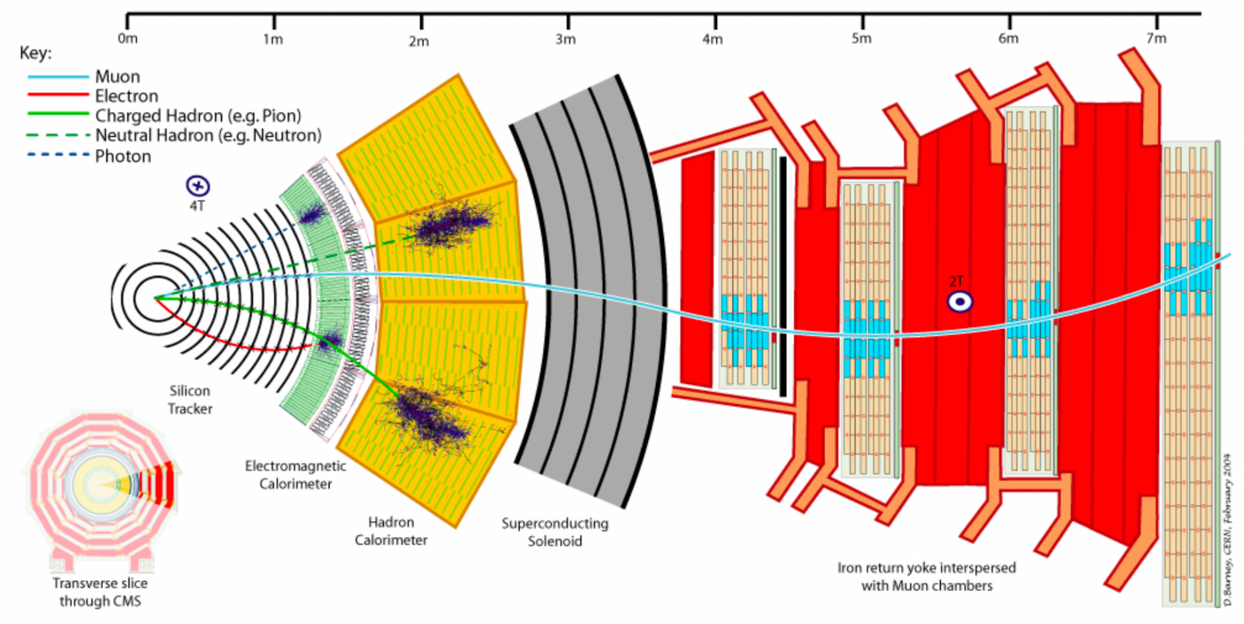
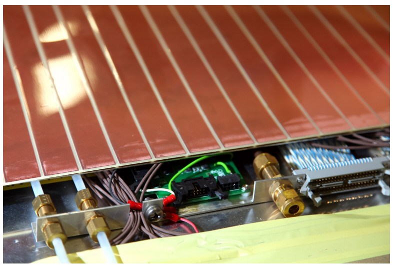
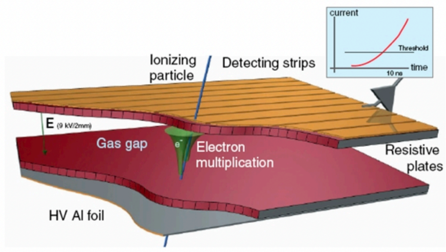


Figure 2: CMS experiment sectional view. The CMS experiment uses different subdetectors to identify different types of particles [3].

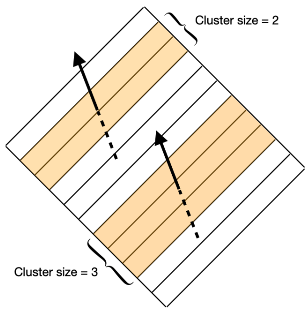
through the RPC, some free electrons are produced by the ionization of the gas between the plates. The movement of this electron avalanche induces a fast signal in the strips that represents the received signal due to ionization. The information provided by each strip is of binary type, where “1” represents signal due to ionization and “0” no signal. These ionization signals can trigger several consecutive strips as shown in Figure 3.



(a)



(b)



(c)

Figure 3: Readout and pictorial representation of an RPC detector. (a) Image of the RPC readout strips [6]. (b) Schematic diagram of the operation of an RPC. The crossing of a charged particle causes ionization of the gas, the subsequent avalanche induces a signal on the readout strips [7]. (c) Pictorial representation of the crossing of two charged particles producing a signal on adjacent strips (in this figure: number of clusters = 2, multiplicity = 5, and cluster sizes = {2, 3}).

2 Data

2.1 RPC detector parameters in DQM

The Data Quality Monitoring (DQM) Software [8], being a central tool in the CMS experiment, contains recorded data of previous collisions at CMS. For each dataset [9] in DQM, there is a number that represents the total number of lumisections in which data were taken, since each lumisection is equivalent to 23.3 seconds [10], that number is a kind of representation of the total duration of data collection. In this project, we use the data of 10 datasets. For selecting the datasets, we considered the lumisections to be as high as possible (≥ 1700) since higher values of lumisections mean more data to analyze.
To deal with the data at DQM, initially, we need to know the RPC detector configuration at CMS as shown in Figure 1. Each wheel contains 96 RPC detectors distributed around the central axis. The RPC chambers located in the barrel are divided into 2 or 3 eta partitions along the z-axis, so that the 480 chambers in the barrel result in 1020 eta partitions called "rolls". These rolls take the names "forward" and "backward". In the case of chambers with three rolls, the name "middle" is used for the additional roll (see Figure 4).
The RPC workspace in DQM provides different detector parameters for each RPC chamber, the parameters and their description are listed in Table 1. An illustration of these parameters is shown in Figure 3c. One of the tools used in the past to analyze datasets [11] was modified to extract the data for a given wheel in tabular format into a CSV output file.

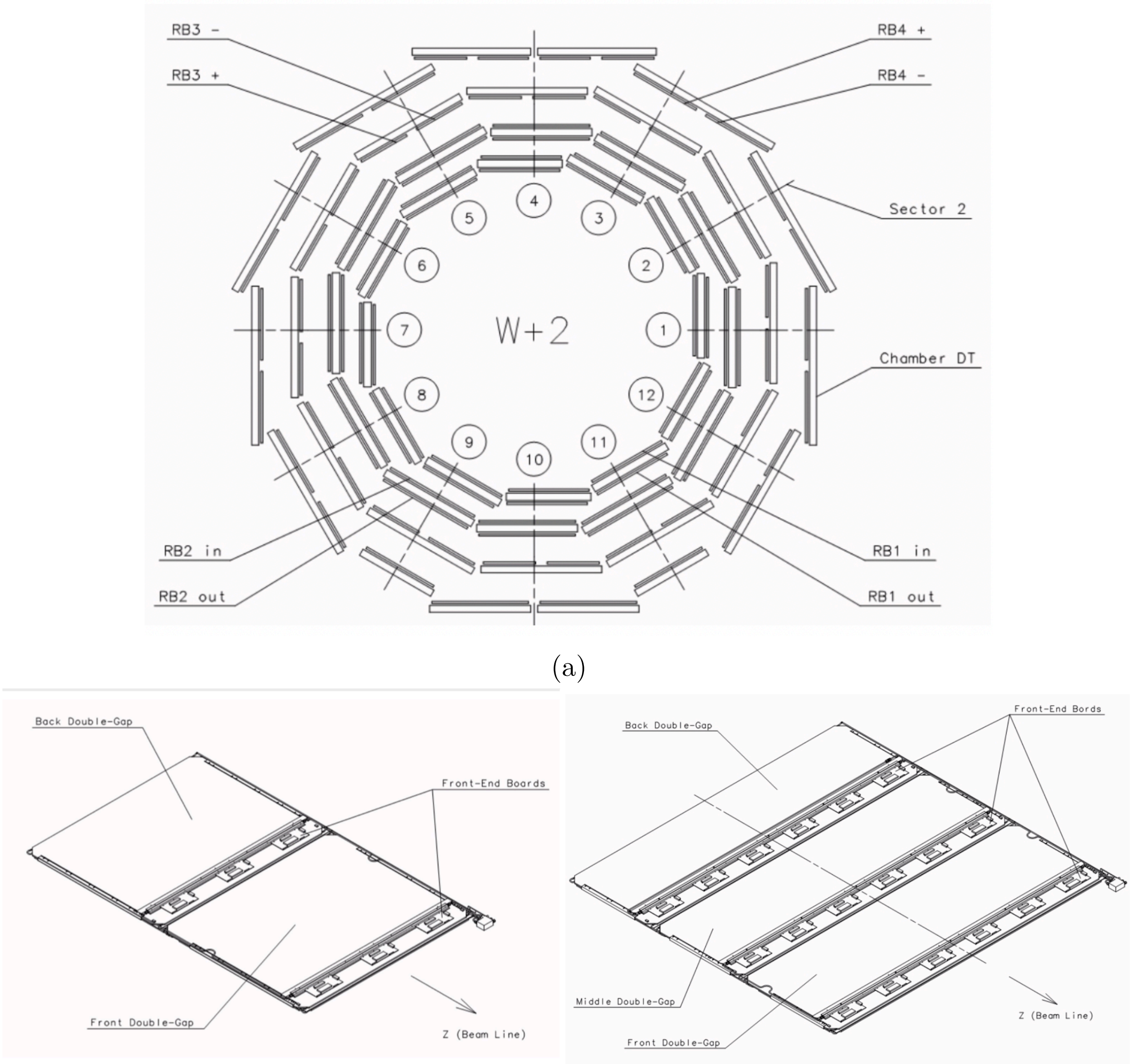
Feature	Description
Bunch crossing distribution	Depends on arriving time of muons
Cluster Size	Number of strips fired when a charged particle crosses the detector
Multiplicity	Activated strips in a window of 25 ns
Number of clusters	Activated strip clusters in a window of 25 ns
Occupancy	Total activated strips during the run

Table 1: Detector parameters for each RPC chamber available in the RPC workspace in DQM.

2.2 Efficiency estimation

The efficiency is defined to be the ratio of the number of expected to detected muon hits. This number is obtained by muon trajectories extrapolation using other sub-detectors data, such as DT[1] in the Barrel and CSC[2] in the Endcap. The extrapolation of a muon trajectory allows us to expect a muon hit at a specific time or not [12]. We obtained the efficiency using the segment extrapolation method [13], in this method the results are obtained per roll. Since the RPC workspace of DQM provides detector parameters per chamber, the average for each chamber is obtained taking into account the efficiency values for each roll

[1] Drift Tubes
[2] Cathode Strip Chambers



(a)

(b) (c)

Figure 4: RPC detector distribution in the barrel. (a) Geometrical distribution of the RPC detector in the barrel region. (b) An example of an RPC detector with two rolls: “forward” and “backward”. (c) An example of an RPC detector with three rolls: “forward”, “backward” and “middle”.

of that chamber.
For each chamber, a list is generated including all the parameters from Table 1, the average efficiency value obtained in the previous step and the name of each chamber. This name provides information about the spatial location of the chamber within the CMS detector. In this project is used only data coming from the barrel, i.e. 480 chambers in total for each dataset, as 10 datasets are used, 4800 records in total. The objective is to achieve an efficiency prediction using the detector parameters and compare the obtained results with the efficiency provided by the segment extrapolation method.

3 Exploratory Data Analysis

3.1 1D Histograms

Over 8% of all the records in the dataset contain zero efficiency. There are two types of such records; first, records that have all the detector parameters equal to zero, and secondly, records that have non-zero detector parameters (for that case all the entries belong to the chamber located in Wheel +2, Sector 7, station 2). Subsequently, all these zero efficiency cases are removed from the records.
Figure 5 shows the histograms for all the detector parameters and the efficiency. It can be observed that there is more than one peak in the histograms of occupancy, average number of clusters and efficiency. To find the origin of these peaks, the efficiency is plotted for different data sets in Figure 6. It can be concluded that there are two kinds of efficiency distribution, one with lower average efficiency and wider variance and the other with higher average and tighter dispersion. A proper machine learning model distinguishes between these two types of distribution.

3.2 2D Histograms

Figure 7 shows two different 2-dimensional histograms with color bars showing the efficiency range. From figure 7a, we can conclude that the average number of clusters is a proper parameter to divide records having different efficiencies (there are two separate regions colored light green and dark blue). Moreover, in figure 7b, there is a separated region including low-efficiency records(red and orange points).

3.3 Correlations

Figure 8 shows linear correlations between different detector parameters and the efficiency. Average multiplicity and average cluster size are correlated. This is to be expected since both variables represent somewhat the same detector parameter, the number of fired strips. Average cluster size and average multiplicy are the most correlated parameters to efficiency. However, the coefficient of linear correlation is 0.47, which suggests that the relationship between the two quantities does not behave linearly.

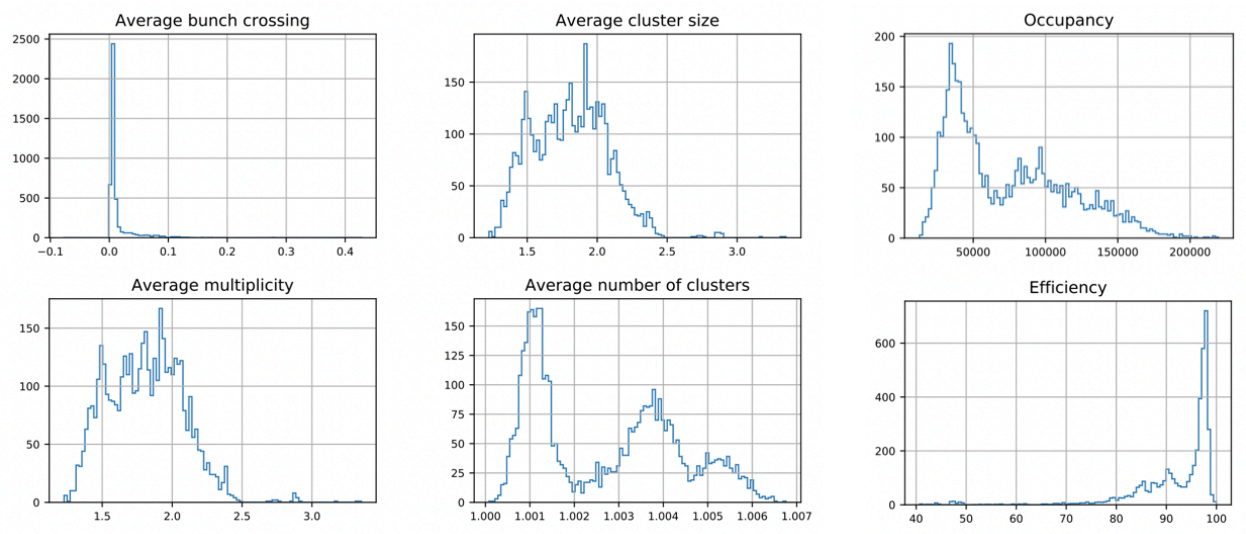


Figure 5: Histograms for different detector parameters of the entire dataset. All the zero efficiency cases were removed from the histograms. The total number of entries for each histogram is 4407. (a) average bunch crossing distribution, (b) average cluster size, (c) occupancy, (d) average multiplicity, (c) average number of clusters and (e) efficiency (averaging for each roll within a chamber)

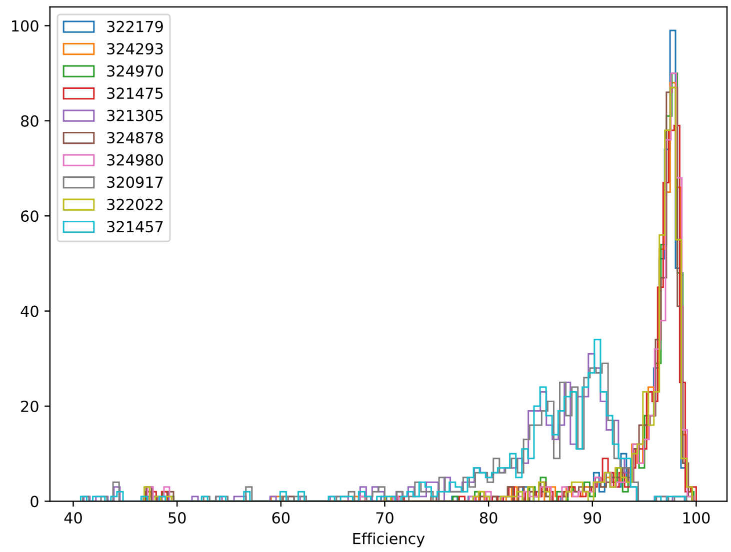
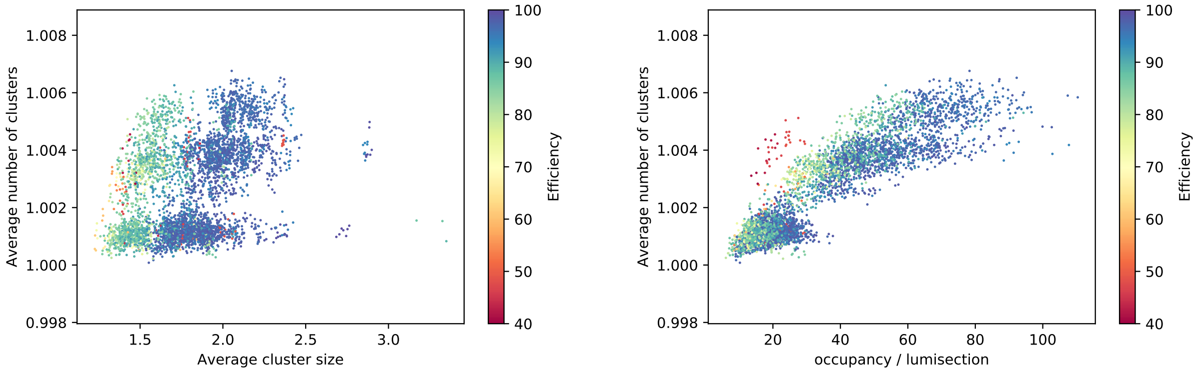


Figure 6: Efficiency histogram for different datasets.



(a) (b)

Figure 7: Two-dimensional histograms colored with efficiency values in the range 40-100. (a) Average number of clusters as a function of the average cluster size. (b) Average number of clusters as a function of the occupancy divided by the total number of lumisections for each dataset.

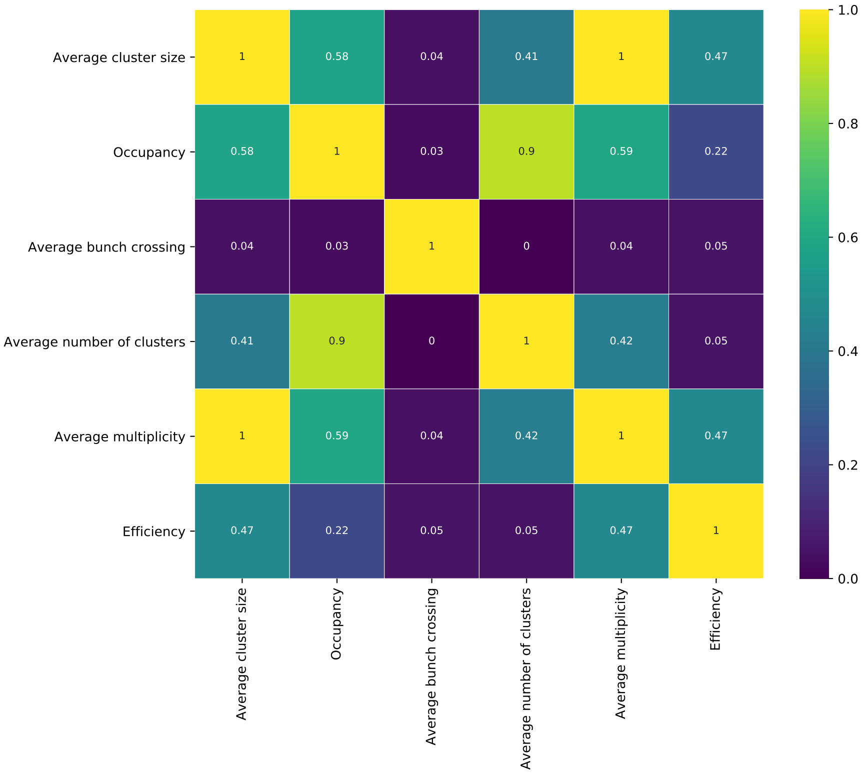


Figure 8: Linear correlations between different detector parameters and efficiency. The color scale represents different correlation values for each case.

4 Machine Learning Model

The K-NN algorithm [14], which considers the surroundings of a data point, was used to obtain a prediction of the efficiency. This algorithm is used to solve both classification and regression problems. The output of a k-NN classification is a class membership. An object is classified to the most common class among its k (a positive integer) nearest neighbours. For the special case k = 1, it is the class of its nearest neighbor. The output of a k-NN regression is an averaged number. It is the average of the values of k nearest neighbours.
The k-NN algorithm was used using a classification model with two classes (low efficiency and high efficiency), and then the k-NN algorithm for regression problems was used using the efficiency values for each chamber instead of the classification of lower and high efficiency used before.

4.1 Classification

The data were divided into two blocks of equal size with labels 0 (low efficiency) and 1 (high efficiency) representing efficiencies in the range $(40.718, 95.95]$ and $(95.95, 100]$ respectively. An accuracy of 92% was achieved on the test samples (20% of the total data) recorded in Table 2 by applying GridSearch algorithm [15] on the model using training samples (80% of the total data) and finding the best hyperparameter for our K-NN classification model as k = 1. Figure 9 indicates that more data can lead to higher accuracy and a better model as the validation score is ascending.

	precision	**recall**	**f1-score**	**support**
0	0.93	0.91	0.92	451
1	0.91	0.93	0.92	431
accuracy			0.92	882

Table 2: K-NN classification report

4.2 Regression

It is obtained a mean squared error (MSE) of 30 and an r2-score of 0.53 by training and tuning the K-NN regression model using the efficiency values and the training samples, these values are calculated using Equations 1 and 2 respectively:

$$MSE = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} (Y_{pred} - Y_{real})^2 \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{samples}} (Y_{pred} - Y_{real})^2}{\sum_{i=1}^{n_{samples}} (Y_{real} - \overline{Y})^2} \quad , \quad \overline{Y} = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} Y_{real} \tag{2}$$

Y_{real} and Y_{pred} represent the true and model predicted values for efficiency. Figure 10 shows a comparison of these two values using a test sample. Overall, the ratio between the distributions is compatible with 1, which shows that the model used is useful for predicting efficiency.

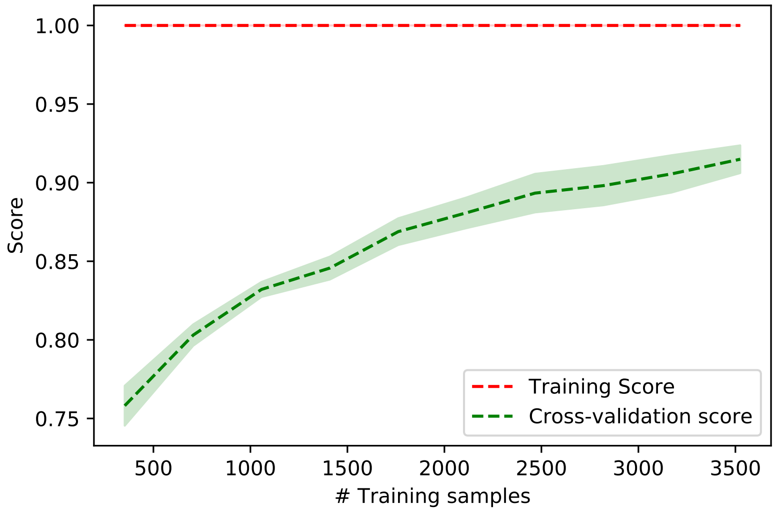
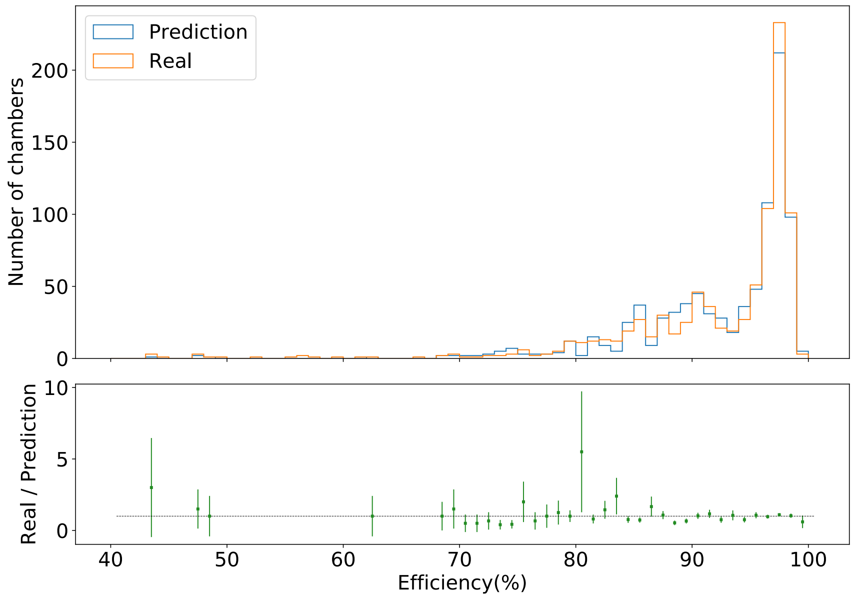


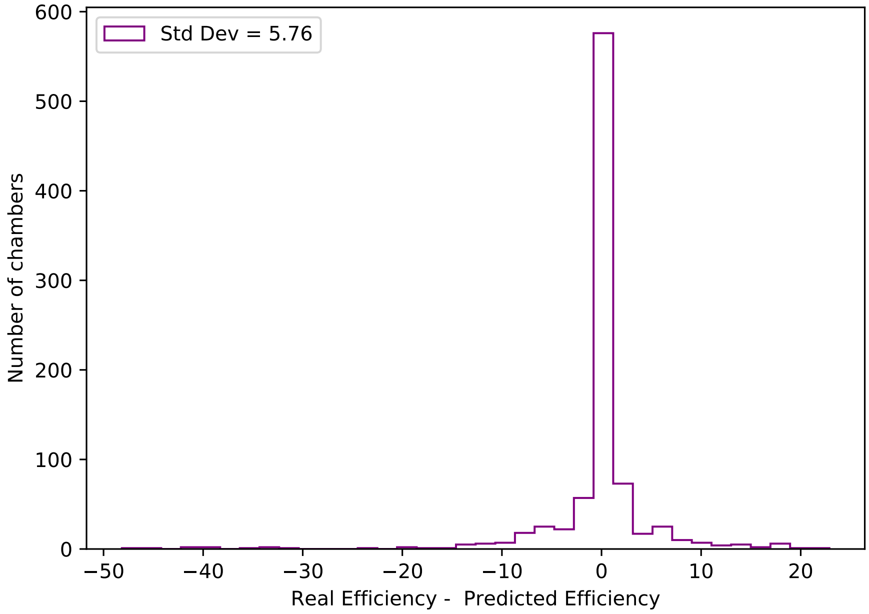
Figure 9: Learning curve for the K-NN classification algorithm using 5-fold cross validation with $k = 1$.

5 Conclusion and Future Remarks

The K-NN regression model was used to predict the efficiency of the RPC detector using as input values for the estimation some measurable detector parameters such as cluster size, bunch crossing distribution, occupancy, multiplicity and the number of clusters. Initially, the model received 3525 records as training values (80% of the total data), the next 20% (882 records) was used as a test sample to evaluate the model and an MSE value of 30 and an r2-score of 0.53 were obtained. The K-NN regression model achieved a reasonable prediction of efficiency as shown in Figure 10 despite the relatively small amount of data used for training. To improve the results it is advisable to use a larger data set and compare it with other models such as convolutional neural network (CNN) [16].



(a)



(b)

Figure 10: Histograms comparing actual and predicted efficiency. (a) Distributions of real and predicted efficiency obtained by the K-NN regression model on the test samples. (b) Distribution of the difference between real and predicted efficiency. An entry in the histogram is the difference between real and predicted efficiency for a single chamber.

References

[1] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic, J. Erö, A. E. D. Valle, M. Flechl, R. Frühwirth, and et al., "Pileup mitigation at cms in 13 tev data," *Journal of Instrumentation*, vol. 15, p. P09018–P09018, Sep 2020.

[2] A. L. Cabrera-Mora, "Cms resistive plate chambers performance at √s = 13 tev," *in Proceedings, 2016 IEEE Nuclear Science Symposium and Medical Imaging Conference: NSS/MIC 2016: Strasbourg, France*, p. 8069690, 2016.

[3] *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS, Geneva: CERN, 2006.

[4] R. Santonico and R. Cardarelli, "Development of resistive plate counters," *Nuclear Instruments and Methods in Physics Research*, vol. 187, no. 2, pp. 377–380, 1981.

[5] S. An, Y. Jo, J.-S. Kim, M. Kim, D. Hatzifotiadou, C. Williams, A. Zichichi, and R. Zuyeuski, "A 20ps timing device—a multigap resistive plate chamber with 24 gas gaps," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 594, pp. 39–43, 08 2008.

[6] M. Hoch, "RPC chamber with a real pannel of readout strips." CMS Collection., Nov 2011.

[7] P. V. et al., "A resistive plate chamber muon detector for the cms experiment at lhc..," *5th Conference on Nuclear and Particle Physics Cairo, Egypt*, p. 511, November 2005.

[8] V. Azzolini, B. van Besien, D. Bugelskis, T. Hreus, K. Maeshima, J. Fernandez Menendez, A. Norkus, J. F. Patrick, M. Rovere, and M. A. Schneider, "The Data Quality Monitoring software for the CMS experiment at the LHC: past, present and future," *EPJ Web Conf.*, vol. 214, p. 02003, 2019.

[9] C. Collaboration, "Dataset." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookGlossary#GlossD`.

[10] C. Collaboration, "Lumisection." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookGlossary#GlossL`.

[11] J. J. Goh, "Automatic e-log tool for rpc data managers." `https://gitlab.cern.ch/CMSRPCDPG/DataManagerTools`.

[12] "Performance of the CMS Muon Detectors in 2016 collision runs." `https://cds.cern.ch/record/2202964`, Jul 2016.

[13] S. Chatrchyan *et al.*, "The Performance of the CMS Muon Detector in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV at the LHC," *JINST*, vol. 8, p. P11002, 2013.

[14] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[15] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for nas," 2019.

[16] P. Kim, *Convolutional Neural Network*, pp. 121–147. Berkeley, CA: Apress, 2017.