International Journal of Innovative Engineering Applications

Journal homepage: https://dergipark.org.tr/ijiea

ARTIFICIAL INTELLIGENCE BASED MACHINE LEARNING APPROACH IN HIGH ENERGY PHYSICS

Serpil Yalçın Kuzu[*1]

[1]*Firat University, Department of Physics Firat University 23119 Elazig - Turkey*

Abstract

Conference paper

In high energy physics experiments data quality plays a significant role for particle identification. Methods used in particle analysis are mainly based on high level knowledge and complex computation skills of human experts and require long time for data quality assurance. Artificial intelligence (AI) applications in various fields are getting important to improve the speed, accuracy and efficiency of human efforts. For this purpose, artificial intelligence-based machine learning approach can be used in particle physics analysis. Dielectrons (e^-e^+) are electromagnetic probes that provide information about dynamics of the medium formed in high energy collisions due to lack of final state interactions. A high purity sample of e^-e^+ pairs can be obtained by traditional cut-based methods resulting in low efficiency. In this contribution, application of machine learning approaches in dielectron analysis is discussed.

Keywords: *Dielectron, machine learning approach, random forest.*

1 Introduction

The purpose of particle collisions at ultra-relativistic energies is to understand the evolution of the universe by creating little 'Big Bang' under laboratory conditions. When two nuclei collide, the nucleons at the collision zone interact initially resulting production of high momentum particles. The non-interacting particles in the collision region begin to thermalize and form dense and hot quark and gluon soup, QGP phase. This fireball expands and cools down until the chemical freezeout temperature (T_c) at which particle species are fixed. At this stage, the particles are in the form of partonic and hadronic states [1, 2, 3]. The medium continues its expansion until the kinetic freezeout temperature (T_k) at which the particle yields are fixed [1, 2, 3]. In the end of these stages the particles are identified by detectors. Dynamics of the medium formed in the high energy particle collisions are shown in Figure 1 [4].

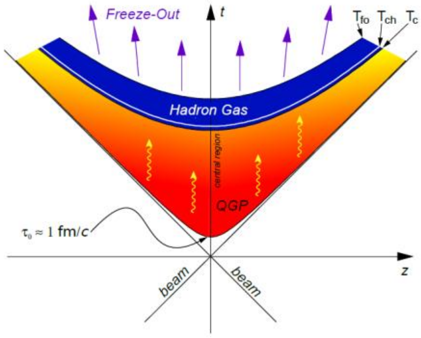

Figure 1. Dynamics of heavy ion collisions [4].

1.1 Dielectrons

The medium evolution in high energy collisions may be investigated by studying experimental probes which provide information about the characteristic features of each phase. Dielectrons are electron – positron pairs that are the unique tools to study different stages of the collisions. Since dielectrons are leptons they do not participate in strong interactions resulting lack of medium effect on their production. Therefore they can be used to probe the inner regions of collisions. In addition, their production any stage of the collision makes them a significant tool to investigate the whole dynamics of the system.

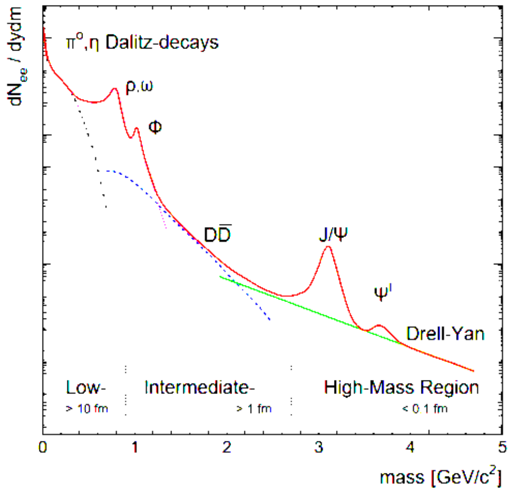

Figure 2. The dielectron spectrum in high energy collisions [5].

[*] Corresponding author.

E-mail address: skuzu@firat.edu.tr (S. Yalçın Kuzu)

Received 26 April 2021; Received in revised form 10 November 2021; Accepted 15 November 2021

2587-1943 | Doi: https://doi.org/10.46460/ijiea.929292

A schematic view of dielectron mass spectrum with corresponding sources in ultra-relativistic heavy-ion collisions is represented in Figure 2 [5]. Basic properties of the collision medium such as evolution of hadronic matter, phase transition and medium temperature can be determined by studying spectrum of dielectron pairs. In addition, different mass ranges of the distribution are sensitive for different stages and physical properties of the medium. In the spectrum, higher mass region gives information about early stage of the system evolution since these pairs are production of virtual photon produced by quark – anti quark interactions called Drell-Yan process [6]. Quarkonium decays such as J/ψ and Υ are other sources of the pairs in this region providing information about hard scattering process. Dielectron pairs between 1 – 3 GeV/c^2 are produced during the thermalization stage by D^+D^- meson decays. Since D^+D^- may decay semileptonically, at the intermediate mass region the pairs have continues distribution. Low mass dielectron pairs are produced due to ρ, ω, φ resonances and Dalitz decays. In order to investigate initial state of the medium, higher mass region of the pair spectrum between 2 – 5 GeV/c^2 can be studied.

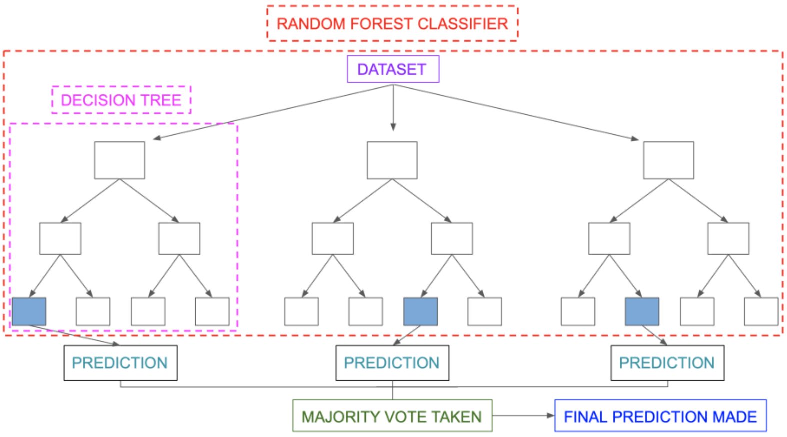

Figure 3. Schema of random forest classifier algorithm [14].

A high purity sample of electron – positron pairs are required to identify the dielectron spectrum. Due to various large sources of background, it is challenging to extract the pair signals. Rejection of those background components needs sophisticated analysis techniques that bring high purity samples with low signal efficiency resulting in high systematic uncertainties. For this reason, implementation of artificial intelligence (AI) based machine learning (ML) tools for pair identification is the necessity to improve dielectron spectrum with high efficiency.

1.2 Machine Learning Approach in Particle Analysis

In this study machine learning approach based on Random Tree method [7] was developed to enhance dielectron pair identification in particle and high energy experiments. For this purpose Random Forest Classifier [7], one of the supervised learning algorithm, was used.

1.2.1 Random Forest Classifier

Random Forest Classifier is a bagging classifier including Decision Trees. In the model, there is an ensemble of trees producing decisions according to a set of sub-decisions. In each decision tree there are nodes and leaves representing features and decisions respectively. The nodes are generated by choosing the best features from the subset of features applied to train current tree. The quality of the node split for each feature can be evaluated by estimation of entropy gain or Gini index, probability of wrong classification for a given property. [8] The model has two steps: the creation and prediction [9]. In the classifier firstly a set of trees are generated from subset of randomly chosen training sample. After this process the votes from different decision trees are collected to give final decision of the test sample [10, 11, 12, 13]. Schema of the classifier algorithm is represented in Figure 3 [14]. In ML approach depending on the model hyper-parameters, a set of parameters initiated at the beginning of the learning process, can be adjusted. In Random Forest Classifier hyper-parameters such as number of decision trees inside the forest, maximum depth of a tree and minimal impurity of a node can be tuned [8].

There are several advantages of using Random Forest Classifier. First of all the classifier can be used for both classification and regression. In ML studies, overfitting that is loss of correct classification ability of the model for the samples out of training set is one of the main problems [15]. Since it is a forest of decision trees the model is resistant for overfitting. In addition, the measurement of the relative importance of each feature on the prediction makes model interpretable. Lastly, compared to a Decision Tree method Random Forest Classifier is more precise due to having forest of decision trees. The cost of having forest is the long process time which is the main disadvantage of the model.

2 Experimental Setup

2.1 Data Set

For ML based classifier development, proton proton collision at the center-of-mass energy 7 TeV with the integrated luminosity 41.47 pb-1 data set collected by the CMS experiment in 2010 was used to study electron pairs from 2 – 5 GeV/c^2 [16 - 20]. In the analysis 10015 pairs were analyzed. 67% of these pairs were e^-e^+ pairs, called signal, and 33% of these pairs were background including e^-e^- and e^+e^+ pairs. The spectrum of signal and background pairs is represented in Figure 4.

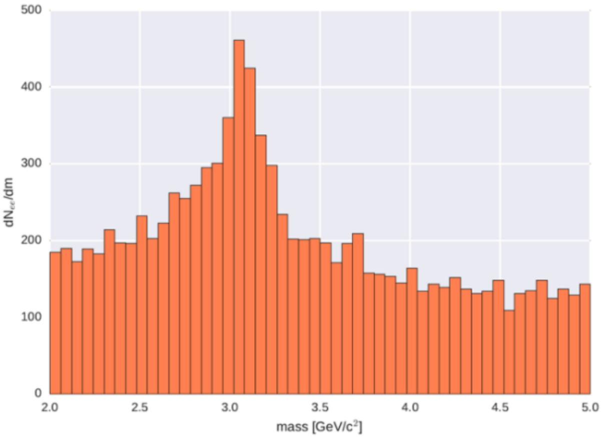

Figure 4. Invariant mass spectrum of all pairs including signal (e^-e^+), and background (e^-e^- and e^+e^+).

For identification of dielectron pairs python [21] implementation of the Random Forest classifier by scikit-learn [22] package was used.

2.2 Application of the Classifier

ML application in particle physics is challenging as compared to its implementation in other research areas due to the nature of quantum mechanics resulting in interference between the particle signal and background [23]. Different ML approaches have been studied to advance the physics results such as higgs boson estimation [24, 25], and beam dynamics analyzation at the CERN Large Hadron Collider (LHC) [26]. It is revealed that with random forest classification particle recognition [8] and data quality investigation [27] can be accomplished. Therefore the random forest classification model was selected to analyze the production of dielectrons.

In the end of high energy collisions, detectors identify the particles with the help of global features (GF) such as charge (q), pseudorapidity (η) and transverse momentum (p_T) that are directly determined by the detectors. By using GF of the particles, characteristic features (CF) such as momentum (p), invariant mass (M) of pairs and opening angle between pair partners (θ) can be calculated for the pair identification [28]. Global and characteristic features are listed in Table 1.

Table 1. Global and characteristic features used in the analysis.

Global Features (GF)	Characteristic Features (CF)
q_1,q_2 (charge)	M (invariant mass of pairs)
η_1,η_2 (pseudorapidity)	p (momentum of pairs)
φ_1, φ_2 (azimuthal angle)	θ (opening angle)
p_{z1}, p_{z2} (z component of momentum)	
p_{T1}, p_{T2} (transverse momentum)	

Since Random Forest Classifier provides information about relative importance of each feature on the prediction, in this study GF and GF+CF were implemented to model separately to understand the impact of features on classification. By using the classifier with and without CF it can be understood if detector responses are good enough for reconstructing pairs, if the highest importance feature matches with the ones used in traditional cut-based methods and if the pairs are derived with the highest efficiency. In both scenarios, hyper-parameters of the classifier were tuned to have the best prediction. In the experiment 60% of data was selected for training and 40% of data was selected for test.

Thanks to the model, feature importance on classification of dielectron pairs in GF and GF+CF implemented models were studied and shown in Figure 5 and 6, respectively. Comparison of feature importance represented in Figure 5 illustrated that GF implemented classifier highly used charges of particles to make prediction of the pairs which is also used in traditional cut – based pair identification method. GF+CF implemented model used characteristic features dominantly to predict dielectrons as shown in Figure 6.

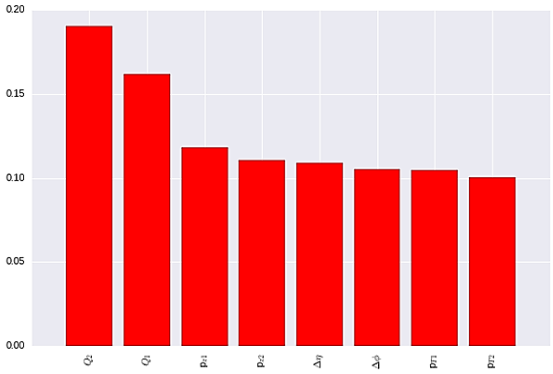

Figure 5. Feature importance of GF implemented Random Forest Classifier study.

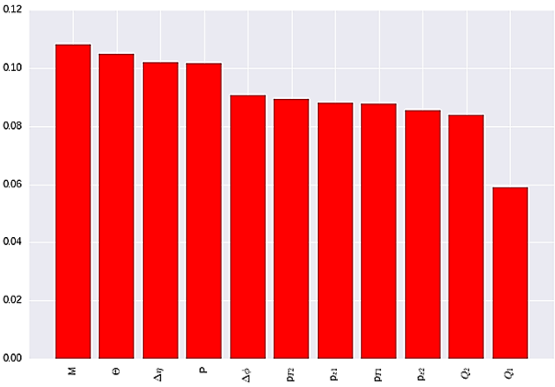

Figure 6. Feature importance of GF + CF implemented Random Forest Classifier study.

3 Results

In ML approaches precision, sensitivity and F-1 scores are widely used metrics to evaluate the success of implemented models. They are defined in Eq. (1), Eq. (2) and Eq. (3), respectively:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{F} - 1\ \text{Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{3}$$

In the equations TP, FP and FN represent number of correctly classified dielectron pairs, misclassified dielectron pairs and misclassified background pairs. As it is understood from Eq. (3) F-1 Score is the harmonic mean of precision and sensitivity. The average precision, sensitivity and F-1 Score of GF and GF+CF implemented models are listed in Table 2. As it is demonstrated in the table compared to GF+CF implemented model, GF implemented classifier showed almost 20% more precise and sensitive results.

Table 2. The average precision, sensitivity and F-1 Score of GF and GF+CF implemented classifiers

Features	Average Precision	Average Sensitivity	Average F-1 Score
GF+CF	0.78	0.72	0.65
GF	0.93	0.93	0.92

Another popular metric for ML applied models is Area Under Receiver Operating Characteristic Curve (ROC-AUC) [15] that is shown in Table 3 for GF and GF+CF implemented classifiers separately. It is concluded that by using global features Random Forest Classifier find e^+e^- 12.43% better than CF applied model which is also used in traditional pair identification method in high energy experiments.

Table 3. Comparison of ROC-AUC for GF and GF+CF implemented classifiers.

Features	ROC-AUC
GF+CF	0.874
GF	0.998

4 Conclusion

In this study, Random Forest Classifier model is applied for e^-e^+ pair identification produced in high energy collisions to understand early stage of universe. The classifier is selected due to having forest of decision trees preventing overfitting problem in ML models. To understand effect of features on prediction, GF and GF+CF were implemented separately in the model. It is shown that global features implemented Random Forest Classifier determined e^+e^- 12.43% better than CF applied model. The results showed that features directly from detectors are good enough to be used in ML based pair identification without further human effort. In addition, comparison of two different set of features implemented model showed that selection of features has an important role on predictions. The results also proved that without hard and time consuming background analysis the pairs can be identified with high efficiency. Application of machine learning techniques is promising and may enhance the quality of particle experiment results.

Acknowledgements

This work is supported by TÜBİTAK-1001 119F302 project.

This study titled "Artificial Intelligence Based Machine Learning Approach in High Energy Physics " was presented at PCFM 2021.

The authors declare no conflict of interest and ethics committee approval.

References

[1] Markert, C. (2005). What do we learn from resonance production in heavy ion collisions?. *Journal of Physics G: Nuclear and Particle Physics,* 31(4), S169.

[2] Torrieri, G., & Rafelski, J. (2001). Strange hadron resonances as a signature of freeze-out dynamics. *Physics Letters B,* 509(3-4), 239-245.

[3] Bleicher, M., & Aichelin, J. (2002). Strange resonance production: Probing chemical and thermal freeze-out in relativistic heavy ion collisions. *Physics Letters B, 530*(1-4), 81-87.

[4] Tawfik, A., & Shalaby, A. G. (2015). Balance function in high-energy collisions. *Advances in High Energy Physics,* 2015.

[5] Rapp, R., & Wambach, J. (2002). Chiral symmetry restoration and dileptons in relativistic heavy-ion collisions. *Advances in Nuclear Physics,* 1-205.

[6] Drell, S. D., & Yan, T. M. (1970). Massive lepton-pair production in hadron-hadron collisions at high energies. *Physical Review Letters, 25(5), 316*.

[7] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence,* 20(8), 832-844.

[8] Trzcinski, T., Graczykowski, L. K. & Glinka, M. (2019). Using Random Forest Classifier for particle identification in the ALICE Experiment. *Proceedings of Information Technology, Systems Research and Computational Physics,* pp. 3–17.

[9] Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R news,* 2(3), 18-22.

[10] Breiman, L. (2001). Random forests. *Machine learning,* 45(1), 5-32.

[11] Azhari, M., Alaoui, A., Achraoui, Z., Ettaki, B. & Zerouaoui, J. (2019). Adaptation of the Random Forest Method, *Proceedings of the 4th International Conference on Smart City Applications,* 1141–1146.

[12] Azhari, M., Alaoui, A., Abarda, A., Ettaki, B., & Zerouaoui, J. (2019). Using ensemble methods to solve the problem of pulsar search. *In International Conference on Big Data and Networks Technologies* (pp. 183-189). Springer, Cham.

[13] Azhari, M., Alaoui, A., Abarda, A., Ettaki, B., & Zerouaoui, J. (2019). A comparison of random forest methods for solving the problem of pulsar search. *In The Proceedings of the Third International Conference on Smart City Applications* (pp. 796-807). Springer, Cham.

[14] Mbaabu, O. (2020). Introduction to Random Forest in Machine Learning, Retrieved February 19, 2021, from https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

[15] Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.".

[16] McCauley, T. (2014). CMS releases open data for Machine Learning, Retrieved October 15, 2014, from https://cms.cern/news/cms-releases-open-data-machine-learning

[17] Krintiras G. (2021). CMS Luminosity - Public Results, Retrieved January 10, 2021, from https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults#2010_proton_proton_7_TeV_collisi

[18] CMS Collaboration, CMS luminosity information, for 2010 CMS open data, Retrieved November 22, 2017, from http://opendata.cern.ch/record/1050

[19] McCauley, T. (2014). J/psi to two electrons from 2010, CERN Open Data Portal.

[20] McCauley, T. (2014). Events with two electrons from 2010, CERN Open Data Portal.

[21] Python Software Foundation, The Python Language Reference, Retrieved February 5, 2021 from https://docs.python.org/3/reference/index.html.

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research,* 12, 2825-2830.

[23] Schwartz, M. D. (2021). Modern Machine Learning and Particle Physics. *arXiv preprint arXiv:2103.12226*.

[24] Chen, T., & He, T. (2015). Higgs boson discovery with boosted trees. *In NIPS 2014 workshop on high-energy physics and machine learning* (pp. 69-80). PMLR.

[25] Bourilkov, D., Acosta, D., Bortignon, P., Brinkerhoff, A., Carnes, A., Gleyzer, S., & Regnery, B. (2019). Machine Learning Techniques in the CMS Search for Higgs Decays to Dimuons. *In EPJ Web of Conferences (Vol. 214, p. 06002)*. EDP Sciences.

[26] Arpaia, P., Azzopardi, G., Blanc, F., Bregliozzi, G., Buffat, X., Coyle, L., ... & Wenninger, J. (2021). Machine learning for beam dynamics studies at the CERN Large Hadron Collider. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment,* 985, 164652.

[27] Trzcinski T. & Deja K. (2017) Assigning Quality Labels in the High-energy Physics Experiment Alice Using Machine Learning Algorithms, *Proceedings of NICA days,* 647-655.

[28] Nourbakhsh S. (2010). Studio degli eventi J/Ψ in due elettroni con i primi dati di CMS. (Doctoral dissertation, Sapienza University).