Gamma Ray Classification

Eesha H Sanjay

9/10/22

Abstract

The focus of this research project is to be able to classify high-energy gamma particles as signal or background. It is important to solve this issue, as it allows us to explore phenomenas that we are unable to see on Earth. High energy gamma particles are also a major part of particle physics. Research in particle physics has helped us in the world of medicine, in drug development, along with exploring more about matter that makes up our world.[1] I created a model that classifies a particle as signal or background based on given test features that I was able to get from the NASA MAGIC Gamma Telescope experiment.[2] This was accomplished by creating an ML model and following the process of training, predicting, and evaluating to test various models. I found the accuracy score of the model, created a confusion matrix, and calculated the f1 score for all the models. For the final model, I got an accuracy score of 88.44% and an f1 score of about 0.883, allowing me to finally find a model that classifies the high energy gamma particles with the highest accuracy.

1. Introduction

Particle Physics is about the forces and particles that make up matter. This study is extremely important for understanding more about our world. For example, scientists use this to understand more about protein structure or other biological processes or explore phenomenons in outer space.[3] A significant example of this is the study of high-energy Gamma particles. These particles are gamma-ray photons that have very high amounts of ultra energy when coming from outer space.[4]

This research paper is in this domain of physics, specifically using Artificial intelligence (AI) to classify high-energy Gamma particles as background or signal. I was able to create an AI model to classify these gamma particles with high accuracy. I was able to accomplish this by using a wide range of data, from the MAGIC Gamma Telescope Dataset, which consists of key features, to construct a model.[5] This experiment uses a detector to detect Cherenkov radiation leaks, which can then be reconstructed. Based on the energy of the gamma, many photons get collected, which by analyzing the features we can determine if it is caused by signal or background. This data was then analyzed by comparing the predictions obtained from the model to the real values.

2. Background

There has been some research conducted in creating models to classify high energy gamma particles, and in this section I will talk more about relevant research done in the field of

[1] https://science.osti.gov/hep/-/media/hep/pdf/files/pdfs/hep_benefits_v2.pdf
[2] https://www.kaggle.com/datasets/abhinand05/magic-gamma-telescope-dataset
[3] https://www.fnal.gov/pub/science/particle-physics/benefits/industry.html
[4] https://science.nasa.gov/ems/12_gammarays
[5] https://www.kaggle.com/datasets/abhinand05/magic-gamma-telescope-dataset

particle physics and AI. The use of Convolutional Neural Networks (CNNs), a neural network primarily designed to use image data, is very common for this type of classification. In one of the articles, it discussed how a deep learning approach that combines CNNs with a recurrent neural network (RNN) has been shown to improve background rejection.[6] In an experiment on this topic, they based their classification on the assumption that images with the largest size parameter are closest to the shower core.[7] Furthermore, in another research paper, I was able to find that in classical methods astrophysicists first create categories based on the physical characteristics of the particle, to isolate interesting events.[8] To do this they needed to first compute Hillas Parameters, a set of geometric features used in gamma ray astronomy, and apply cuts to the obtained results.[9] They can also use CNNs to provide a direct event classification method that uses all relevant information as is that is contained within the Cherenkov shower image and this speeds the processing of data done in classical methods by removing the need to bypass the Hillas Parameters. The Cherenkov shower image is created by the collection of the photons in patterns.[10] These images allow it to be easier to differentiate between signal and background.

This was important to understand the other methods that were used to solve this classification problem, and how they were able to produce pretty accurate results. I used artificial intelligence to see if I could create a better and more efficient model than the models previously discussed. I utilized RandomizedSearchCV to find the optimal hyperparameters to produce the best accuracy.

3. Dataset

For this project, I needed to use a dataset that contains 19020 particles to ensure that the model is built on a wide variety of data. My chosen dataset contains numerical data that can be used to classify high-energy particles in the atmosphere as either signal or background. It identifies some of the key features for each labeled particle, as well.

Key Features	Meaning
fLength	major axis of the ellipse
fWidth	minor axis of the ellipse

[6] https://www.semanticscholar.org/reader/061af43faca95f4fe43e8a323beb70f0695a5e70
[7] https://arxiv.org/pdf/1803.10698.pdf
[8] https://arxiv.org/pdf/2103.06054.pdf
[9]https://www.researchgate.net/publication/1788402_Analysis_methods_for_Atmospheric_Cerenkov_Telescopes
[10] https://www.kaggle.com/datasets/abhinand05/magic-gamma-telescope-dataset

fSize	total size
fConc	ratio of the sum of two highest pixels over the fSize
fConc1	ratio of the highest pixel over the fSize
fAsym	distance from the highest pixel to center
fM3Long	3rd root of the third moment along the major axis
fM3Trans	3rd root of the third moment along the minor axis
fAlpha	angle of the major axis with vector to origin

Table 1: This Table shows the key features that were used in my dataset, along with the meaning of each of the features.

To understand more about how these features affect whether or not a particle is a signal or background, I needed to visualize my dataset.

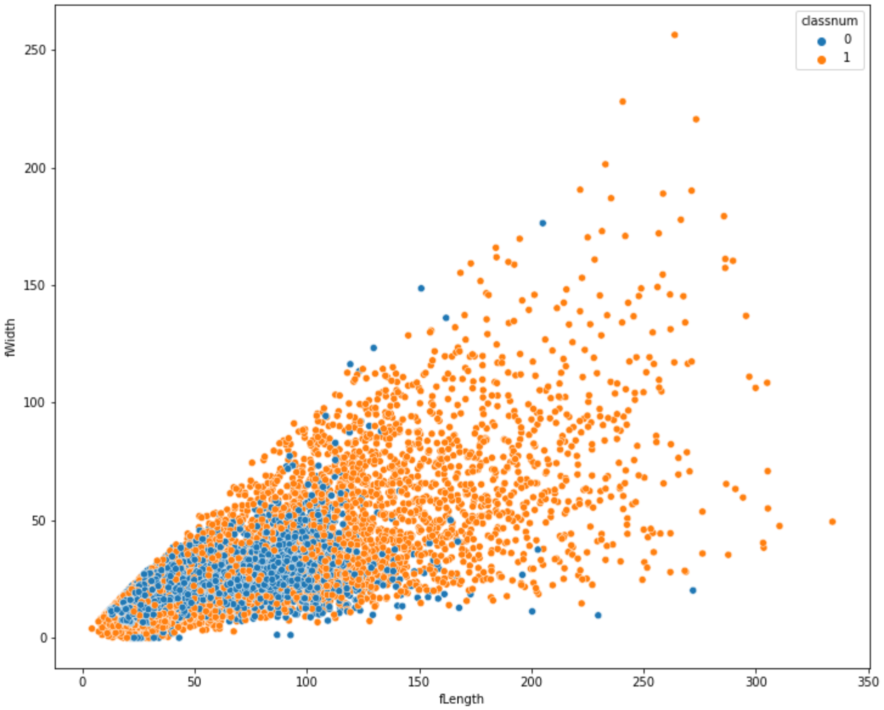

Figure 1: In this chart the blue particles represent the gamma/signal particles, and the orange represents the hadron/ background particles. This plot shows that there is not a clear distinction between the signal and background when you look at the relationship between the minor axis of the ellipse and the major axis. An ellipse is the shape that these particles are in.

I graphed many other pair plots as well, but in all of these pair plots, there was no clear distinction between what feature values meant that the class was signal vs. background. This was why I needed to use Logistic Regression to solve this problem.

I also noticed in my dataset that there was an unequal distribution of the number of signal data and background data.

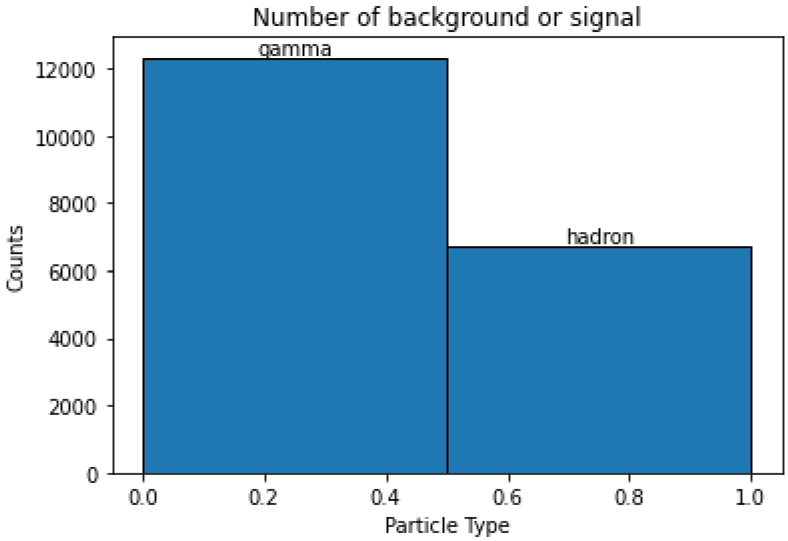

Figure 2: Shows the ratio of Signal to Background. As shown there are more signal values than background values, so it is necessary that there is the same amount of each to create an accurate model. This meant I needed to do data augmentation to equalize the distribution of signal and background.

Then, I converted the categorical features to their numerical counterparts. I also was able to remove unnecessary information from the data, and remove the class and unnamed columns/ features of the dataset because this information was unnecessary for classification purposes because they only contained information that already existed in a different column. I split this dataset into training and testing sets, and the percentage of testing data was 33%.

4. Methodology / Models

To solve this research problem of accurate classification, there were many different models that I created. It was clear when I graphed the dataset with the pairplots that there was no visible correlation with the specific features for classification. It was also significant to understand that the data was not distributed equally, so we needed to make sure that there was the same ratio of signal to background.

As part of Data reprocessing for this data set, I decided to convert the categorical features to their numerical counterparts. This meant that I made the signal equal to 0 (g=0) and the background equal to 1 (h=1). Then, I could perform feature extraction. I had made the category numerical counterparts a new feature, so I was able to remove the class and unnamed columns/ features of the dataset because this information was unnecessary for classification purposes. Finally, I split this dataset into training and testing sets.

After preparing the data, I was able to compare the accuracy of all the different models I created to figure out which ones classify the best. For each of these models I trained these models, then allowed them to predict/ classify correctly, and then evaluated them. This helped me stay consistent while assessing all of the different models. I used the Logistic Regression model, Random Forest Regressor, Random Forest Classifier model, SMOTE, BorderlineSMOTE, ADASYN, and RandomizedSearchCV.

The Logistic Regression model was created after splitting my data into training and testing data. This was made as a baseline model to see how much I can improve from its accuracy. I was able to fit a logistic model x and y training data and then make predictions using my model with the x testing data.

Random Forest Regressor "fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting".[11] Random Forest Classifier uses a very similar approach, using decision tree classifiers. These two models were necessary because based on the predictions from each decision tree used in the model, the best solution was given.

SMOTE is a Synthetic Minority Oversampling Technique, and it helps with overcoming an imbalance in class.[12] For example, as seen in the dataset section, there is an imbalance in class size, so SMOTE allows this to be overcome. It also overcomes the overfitting problem by using random oversampling. ADASYN generates synthetic instances for samples. BorderlineSMOTE and ADASYN use SMOTE techniques.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[12]https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques

RandomizedSearchCV first creates a base model to tune and then runs through multiple different combinations, using a random grid, to search for the best hyperparameters. Then I was able to evaluate the model using these new hyperparameters. I chose this as the final model.

5. Results and Discussion

To understand which models I developed worked best, I needed to evaluate all of them and calculate their accuracy.

For the Logistic Regression model, the accuracy score was about 78.86%. This score determines the percentage of accurate predictions made by the model.[13] For the Random Forest Regressor and Random Forest Classifier models, I was able to follow the same steps of training, predicting, and evaluating, but to evaluate these models I calculated the accuracy score, along with using a confusion matrix and the f1 score. A confusion matrix allows you to visually see true positives, false positives, true negatives, and false negatives, allowing you to compare all these values to get an holistic view of the performance of the model. The F1 score determines the precision and recall of the model.[14] For the Random Forest Regressor Model, I was able to get an accuracy score of 87.21%, and an f1 score of about 0.805, and for the Random Forest classifier I was able to get an accuracy score of 87.49% and an f1 score of about 0.808. These two models resulted in much higher accuracy than the Logistic Regression model, but I still tested more models to see if I can find a model with even higher accuracy.

Then. I was able to test three models from imblearn.over_sampling: SMOTE, BorderlineSMOTE, and ADASYN. Imblearn.over_sampling helps with data augmentation, allowing classification to occur between unequally distributed classes.[15] Out of these models SMOTE had the highest accuracy score of 88.26% and an f1 score of about 0.880.

Finally, I used RandomizedSearchCV. This allowed me to understand which were the best hyperparameters to use that would optimize this model. I was able to determine that the random best hyperparameter values, which can be used to define a network structure and changing these values have an effect on the output.

'n_estimators'	200
'min_samples_split'	2

[13] https://www.obviously.ai/post/machine-learning-model-performance#:~:text=So%2C%20What%20Exactly%20Does%20Good,not%20only%20ideal%2C%20it's%20realistic.

[14] https://www.myaccountingcourse.com/accounting-dictionary/f1-score

[15] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

'min_samples_leaf'	1
'max_features'	sqrt
'max_depth'	None
'bootstrap'	False

Table 2: This table displays all the optimal hyperparameter values for this model.

Then, I evaluated the model using these hyperparameter values, and I found the accuracy score, created a confusion matrix, and calculated the f1 score for this model. The RandomizedSearchCV we used gave us the highest f1 score of about 0.883 and the highest accuracy score of 88.44%. This accuracy is pretty high, so it was reasonable for me to keep this as my final model.

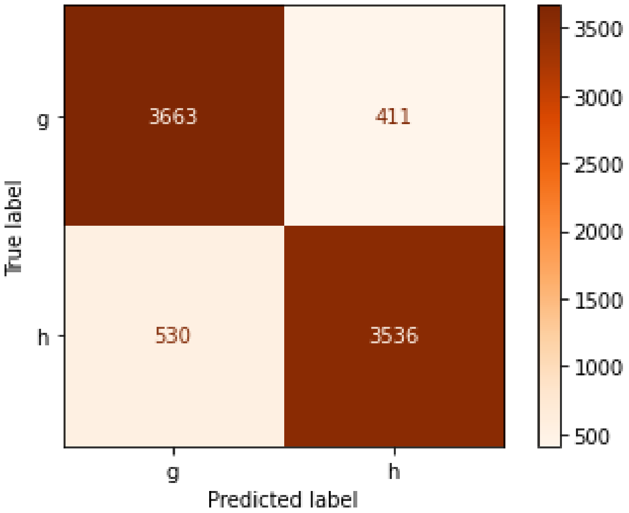

Figure 3: This figure shows the confusion matrix of my final model.

In the scope of my research I chose six different models to compare their accuracies. It is possible that there are other ML models that can even more accurately classify these gamma particles.

6. Conclusions

From these results, we can observe that the RandomizedSearchCV was able to give me the most optimal hyperparameters, which resulted in a model that classifies the particles with 88.44% accuracy. To get to this model it required a long process of testing multiple different models along with understanding more about current research that was conducted on this topic. Constructing a model to classify high energy gamma particles has allowed me to learn more about this topic and various machine learning methods. In the future, I hope to continue conducting more research and solving more issues in this field of study.

Acknowledgments

I would like to acknowledge my AI mentor, Victoria Lloyd, for guiding me through this entire project. I would also like to thank my family for giving me constant support.

References

Abhinand. “Magic Gamma Telescope Dataset.” Kaggle, 8 Nov. 2019, https://www.kaggle.com/datasets/abhinand05/magic-gamma-telescope-dataset.

A. Brill, et al. “Investigating a Deep Learning Method to Analyze Images from Multiple Gamma-Ray Telescopes.” *Semantic Scholar*, 2020, https://www.semanticscholar.org/reader/061af43faca95f4fe43e8a323beb70f0695a5e70.

Barkved, Kirsten. “How to Know If Your Machine Learning Model Has Good Performance? .” Data Science without Code- Obviously Ai, 2022, https://www.obviously.ai/post/machine-learning-model-performance

“Gamma Rays.” NASA, NASA, https://science.nasa.gov/ems/12_gammarays.

I. Shilon, et al. “Application of Deep Learning Methods to Analysis of Imaging Atmospheric Cherenkov Telescopes data” - ArXiv.org, https://arxiv.org/pdf/1803.10698.pdf.

Naurois, Mathieu de. “Analysis Methods for Atmospheric Cerenkov Telescopes (PDF).” Research Gate, 2006, https://www.researchgate.net/publication/1788402_Analysis_methods_for_Atmospheric_Cerenkov_Telescopes.

Office of Science. Particle Physics: Benefits to Society. https://science.osti.gov/hep/-/media/hep/pdf/files/pdfs/hep_benefits_v2.pdf.

“Science.” Fermilab, 2014, https://www.fnal.gov/pub/science/particle-physics/benefits/industry.html.

S. Khatchadourian, et al. “Efficient Level 2 Trigger System Based on Artificial Neural Networks (PDF) .” Research Gate , 2008, https://www.researchgate.net/publication/236032740_Efficient_Level_2_Trigger_System_Based_on_Artificial_Neural_Networks.

“Smote.” SMOTE - Version 0.9.1, https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

Spencer, Samuel, et al. “Deep Learning with Photosensor Timing Information as a Background Rejection Method for the Cherenkov Telescope Array.” ArXiv.org, 10 Mar. 2021, https://arxiv.org/abs/2103.06054.

“What Is an F1 Score? - Definition: Meaning: Example.” My Accounting Course, 19 June 2018, https://www.myaccountingcourse.com/accounting-dictionary/f1-score.