International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Enhancing Gamma Particle Prediction from MAGIC Telescope Data through Machine Learning Techniques

M.Anitha[1], K.Baby Ramya[2],SK.Moulali[3]
#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.
#2 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada.
#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada.

Abstract – The Cherenkov gamma telescope observes high energy gamma rays by taking advantage of the electromagnetic showers initiated by the gammas. The detector records and allows for the reconstruction of the shower parameters. The reconstruction of the parameter values was achieved using a Monte Carlo simulation algorithm called CORSIKA. The problem statement is to classify the gamma particles from the background/hadron. In this paper I have mentioned the designed and evaluated multiple machine learning based classification algorithms and measured the models' performances. The performance metrics and the outcomes have also been included. The models used are Decision Tree Classifier, Random Forest Classifier and Naïve Bayes Classifier and ensemble techniques Adaboost Classifier for Decision Tree and voting classifier including all the above methods.

Key Words – Machine Learning, Classification, Gamma Telescope,Feature Extraction, Gamma Particles.

I. INTRODUCTION

THE presented data were generated using Monte Carlo simulations to replicate the process of detecting high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope through imaging techniques. This type of telescope observes high-energy gamma rays by utilizing the radiation emitted by charged particles generated within the electromagnetic showers initiated by the gamma particles and developing in the Earth's atmosphere. The resulting Cherenkov radiation, which falls within the visible to ultraviolet range, permeates through the atmosphere and is captured by the detector. This recorded information enables the reconstruction of various shower parameters. Specifically, the data consists of pulses left by the Cherenkov photons upon interaction with the photomultiplier tubes arranged in a two-dimensional plane known as the camera. Depending on the energy of the primary gamma particle, a varying number of Cherenkov photons, ranging from a few hundred to several thousand, are collected and form discernible patterns known as the shower image. These patterns allow for statistical differentiation between those caused by primary gamma particles (signal) and the images of hadronic showers initiatedby cosmic rays in the upper atmosphere (background).

The data set was generated by a Monte Carlo program, Corsika, described in: D. Heck et al., CORSIKA, a Monte Carlo code to simulate extensive air showers, Forschungszentrum Karlsruhe FZKA 6019 (1998) [1].

II. LITERARY REVIEW

Previous research in the field of gamma particle separation from background noise has shown promising results. Dadzieand Kwakye [2] conducted a study comparing the effectiveness of two classification algorithms, namely the multiple-layer perceptron (MLP) and the self-organizing tree algorithm (SOTA). They found that using a hybrid approach.combining these techniques improved classification results while reducing training time. Unity embeddings were also explored to enhance classification accuracy.

Another group of researchers focused on the application of the Random Forest (RF) tree classification method for analysing data from a ground-based gamma telescope. They compared RF with other semi-empirical techniques and observed superior performance. The researchers discussed important considerations and challenges associated with RF, particularly its application in estimating continuous parameters using other variables.

SUMMARY RESULTS OF THE PERFORMANCE OF MODELS BASED ON ROC_AUC.

Data	LR	LDA	KNN	CART	NB	SVM	Mean
Raw	0.8394	0.8364	0.8264	0.7901	0.7558	0.6979	0.7910
Clean	0.7958	0.7975	0.7768	0.7502	0.7342	0.6672	0.7536
Norm	0.7976	0.7975	0.8441	0.7460	0.7342	0.8312	*0.7918
Stand	0.7967	0.7975	0.8410	0.7473	0.7342	0.8964	***0.8022**
PCA	0.7780	0.7789	0.7793	0.6711	0.7816	0.6661	0.7425
ICA	0.7757	0.7789	0.7892	0.6944	0.7813	0.7759	0.7659
UFS	0.7813	0.7816	0.7887	0.7085	0.7486	0.7389	0.7579
RFE	0.7863	0.7865	0.7653	0.7258	0.7451	0.8079	0.7695

Fig.1 Results from Emmanuel A. Dadzie, and Kelvin K. Kwakye's paper.

The study developed multiple classification models using different machine learning algorithms and different data transformations. The results suggest that models created with the raw data have similar classification performance as the models created with the transformed data. However, the bestmodel for the detection of high-energy gamma particles is the support vector machine (SVM) algorithm on a standardized dataset. The results indicated similar performance levels for all the models across the different datasets (i.e., raw, clean, normalized, standardized, PCA, ICA, UFS, and RFE transformed data). The ANOVA test reveals that the performance levels across the different transformations were not significantly different. The pairwise comparison of the performance (i.e., AUC) between the raw data and all the other data forms were not significantly different. Hence, none of the transformations increase the performance significantly.However, the mean accuracy for normalized, standardized, UFS, and RFE transformations were higher than that of the raw (baseline) data. Similarly, in the case of AUC, only the normalized and standardized were higher than the value for the raw data. In both cases of the performance metrics, the standardization transformation produced the highest score.

In a research by T. Hassan and others [3], researchers have looked at the various possibilities of machine learning algorithms that can be applied to classify the galactic nucleus type of the ray. This research helped me understand how to

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

compare different classification techniques with a single attribute in place rather than the sole task of classifying gamma particles from the hadron. However, there is still an extensive scope of study in the same context to build sustaining and reproduceable machine learning models.

III. DATASET AND PREPROCESSING

Source: Bock,R. (2007). MAGIC Gamma Telescope. UCI Machine Learning Repository. https://doi.org/10.24432/C52C8B.

The data is generated basing on a Monte Carlo program CORSIKA in order to simulate the registration of high-energy gamma particles in a Cherenkov gamma telescope with imaging technique. The telescope observes high-energy gamma rays using the radiation emitted by charged particles that are produced inside the electromagnetic showers initiated by the gammas. There are a total of 10 attributes that are continuous and a binary attribute class which is our target variable to classify. The shape of the data is 19020 x 11.

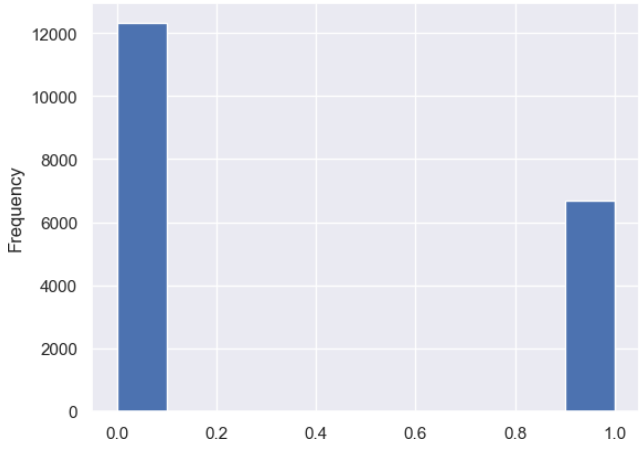


Fig. 2. The imbalance in the values of target variable 'class.'

The dataset is imbalanced with 12332 instances for gamma and 6688 instances for hadron, The preprocessing of the data includes the identification and removal of missing values, which existed the data as black cells, and the numerical value of 99999. The data void of missing values serves as the raw baseline data, which is later compared with the pre-processed and transformed data in the performance evaluation during the model development. Further, outliers are removed, and the null values are filled with mean of the column data. All the features had a significant impact on the class of the ray, so all the 11 attributes were used to make a prediction.

Attribute Name	Variable Type	Explanation
fLength	continuous	major axis of ellipse [mm]
fWidth	continuous	minor axis of ellipse [mm]
fSize	continuous	10-log of sum of content of all pixels [in #phot]
fConc	continuous	ratio of sum of two highest pixels over fSize [ratio]
fConc1	continuous	ratio of highest pixel over fSize [ratio]
fAsym	continuous	distance from highest pixel to center, projected onto major axis [mm]
fM3Long	continuous	3rd root of third moment along major axis [mm]
fM3Trans	continuous	3rd root of third moment along minor axis [mm]
fAlpha	continuous	angle of major axis with vector to origin [deg]
fDist	continuous	distance from origin to center of ellipse [mm]
class	g, h	gamma (signal), hadron (background)

TABLE I. Information regarding the attributes.

In order to best explain the relationship between the attributes I have chosen to go with pair plot for the whole dataset. Correlation plot is important for finding patterns and relationships between variables. It can also be used to make predictions and decisions based on data. Low correlation coefficients show that the two variables do not have a strong relationship with each other.

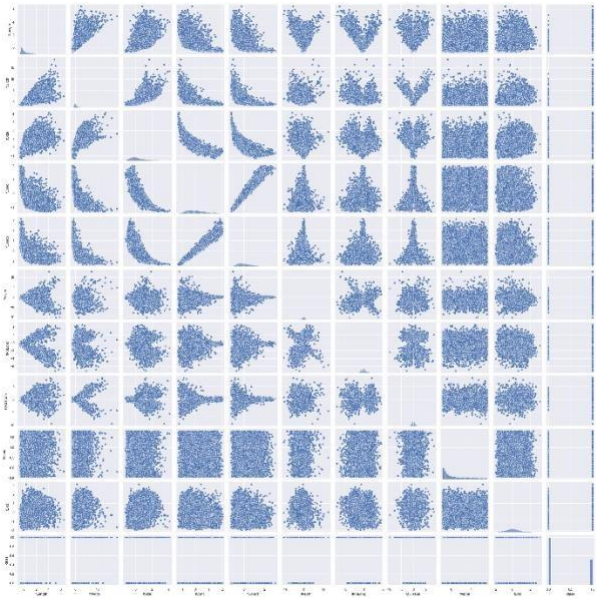


Fig. 3. Pair plot for displaying relationship between the attributes.

Basing on the above pair plot and below correlation plot, we can establish a relationship between the attributes and with respect to our target variable class, therefore we can establish

a relationship and then choose which attributes to choose to build a model.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362
IJARST

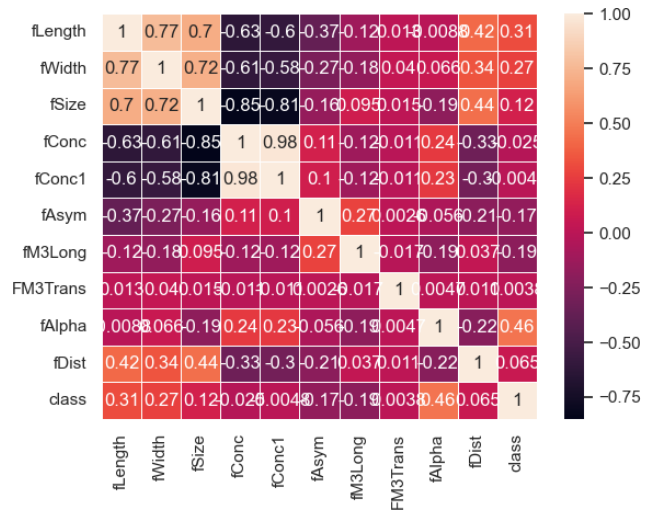


Fig. 4. Correlation Plot to find patterns and relationships between the dependent variable and the independent variables.

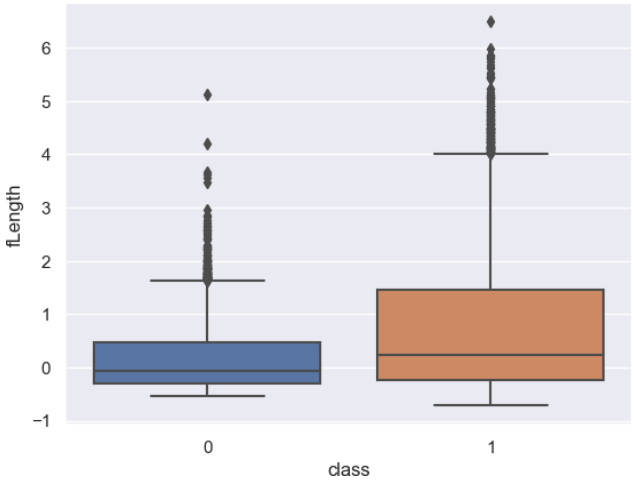


Fig. 5. Boxplot between the variables fLength and Class in order to show the distributions of numeric values.

IV. EXPERIMENTAL SETUP

On cleaning the dataset by filling null values by mean of that specific column and dropping the redundant values from the dataset, as there were not many missing values present in the dataset it could be dealt in a rather easy manner by using the best statistical measure mean to fill the columns in. Imputing the null values with zero would leave us with a relatively lesser number of values of data to work with. Then the data has been run through robust scaling by using *'RobustScaler()'* from *'scikit-learn'* [4], which scaled the numerical input variables that contain outliers. The target variable 'class' has been converted into binary equivalents '0''s and '1's for the values 'h' and 'g' respectively by using the *'LabelEncoder()'* from *'scikit-learn'*.

The data is split into training and testing datasets randomly, in the current setup the splitting has been achieved by the *train_test_split* function from the *scikit-learn* package in python. In this research, from the total dataset 70% of the data is used for training and the rest of 30% has been used for testing which is generally an ideal proportion of data to use.

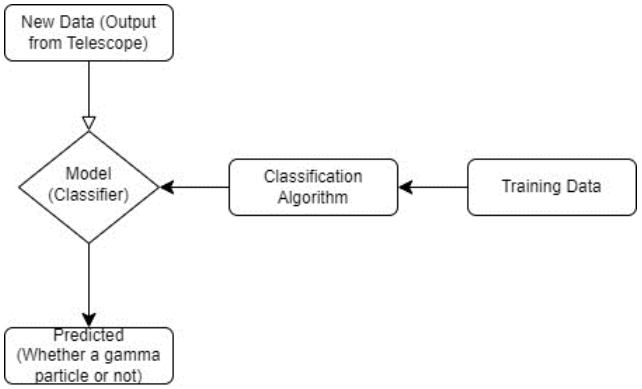


Fig. 6. Machine Learning Workflow for the currentexperiment.

The code used is attached in Appendix A and GitHub link has been presented.

V. METHODS USED

The classification techniques used in this experiment are,

I. Decision Tree Classifier
II. Adaboost Classifier
III. Random Forest Classifier
IV. Naïve Bayes Classifier
V. Voting Classifier

I. Decision Tree

The first algorithm chosen is Decision Tree [5]. It is one of the widely used classification models in the machine learning realm that is an easy to visualise classification model. A decision tree works just as the name suggests in the structure of a tree with leaves and roots. The idea being to break the whole dataset into smaller subsets based on homogeneity of the data (examples).

- Decision Nodes – For example, the attribute has branches according to the attributes affecting the class which in this case are fm3trans which is having one of the highest impacts.
- Leaf Nodes - Whether the particle belongs to class 'g' or 'h.'

By using Decision Tree on our dataset, the results were as follows,

```
Classification Report of Decision Tree Induction:

Accuracy of DT:  84.6477392218717 %

              precision    recall  f1-score   support

           0       0.85      0.93      0.89      3731
           1       0.84      0.69      0.76      1975

    accuracy                           0.85      5706
   macro avg       0.85      0.81      0.82      5706
weighted avg       0.85      0.85      0.84      5706
```

Fig. 7. Classification Report for Decision Tree

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

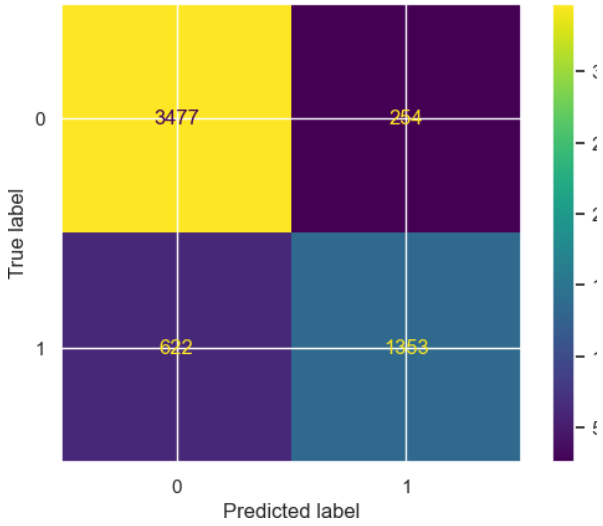


Fig. 8. Confusion Matrix for the output from Decision Tree

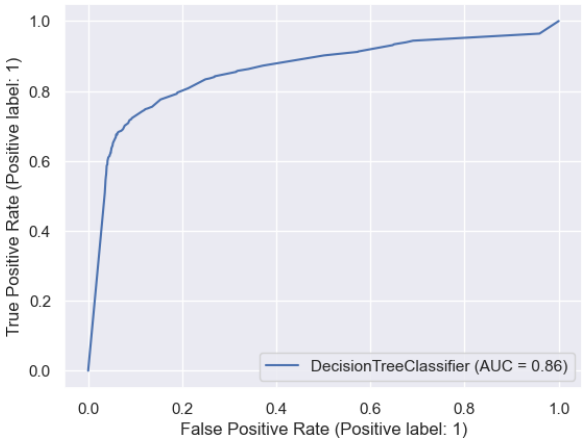


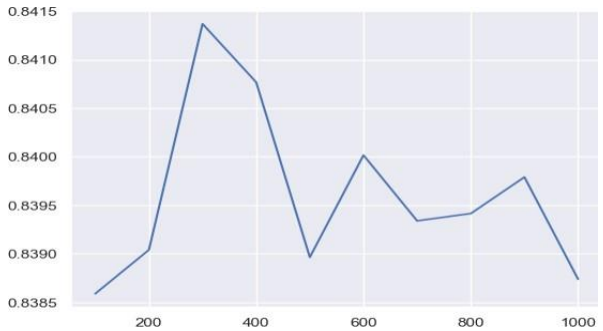
Fig. 9. ROC Curve for Decision Tree

II. Adaboost Classifier

Adaboost [7], is a boosting technique that is used over weak learners to improve the accuracy score. Adaboost is used as an ensemble technique. In our case we were able to achieve almost the same kind of accuracy as in Decision Tree.

	precision	recall	f1-score	suppor
0	0.86	0.91	0.88	373
1	0.81	0.73	0.76	197
accuracy			0.84	576
macro avg	0.83	0.82	0.82	576
weighted avg	0.84	0.84	0.84	576

Fig. 10. Classification Report for Adaboost



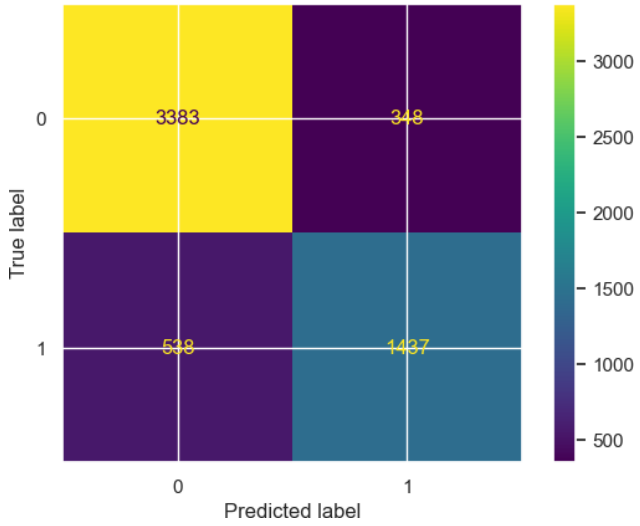


Fig. 11. Adaboost Classifier with n=10 estimators

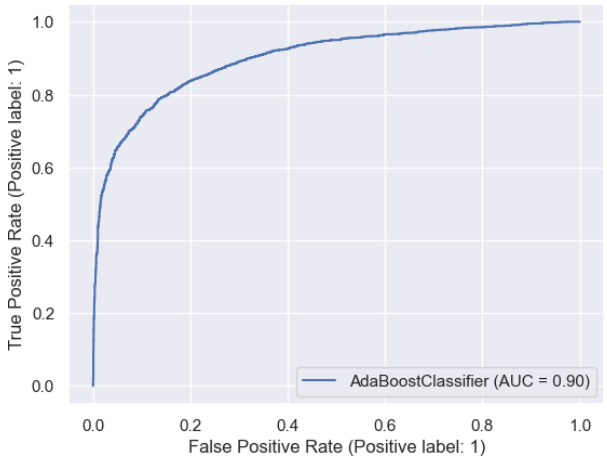
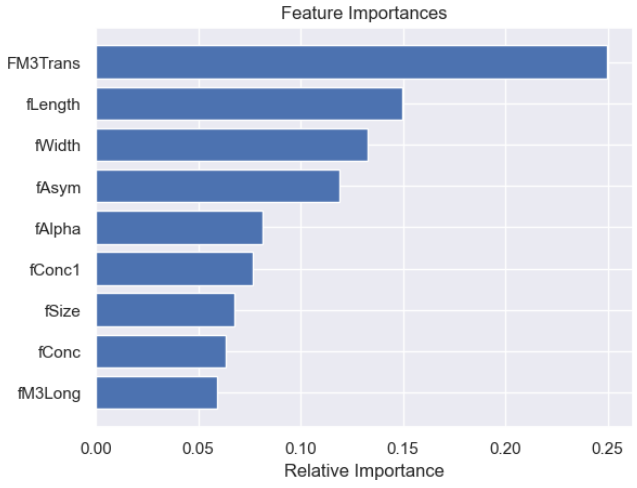


Fig. 12. Confusion Matrix for Adaboost

Fig. 13. ROC Curve for Adaboost

III. Random Forest Classifier

Random forest[6] operates as an ensemble technique where it has a collection of decision trees that result in a better accuracy than a regular Decision Tree. Random Forest utilizes bagging technique that enables it to train on a random sampling of the original dataset and takes the majority selection from all the trees. The important measure in random forest is to calculate the feature importance.



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

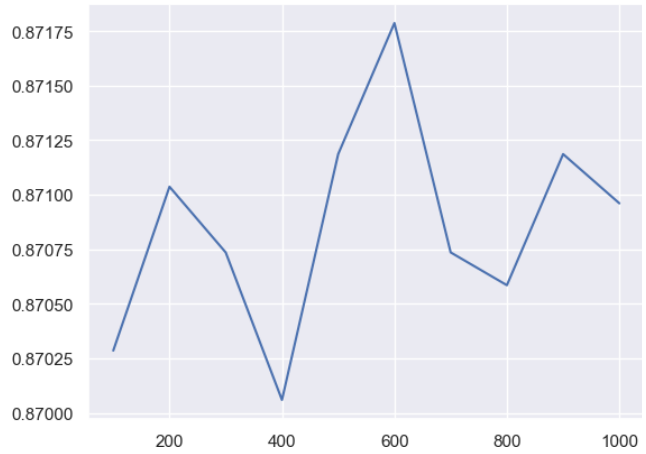


Fig. 14. Feature Importance for each feature obtainedby Random Forest Classification.

	precision	recall	f1-score	support
0	0.88	0.93	0.91	3731
1	0.86	0.76	0.81	1975
accuracy			0.87	5706
macro avg	0.87	0.85	0.86	5706
weighted avg	0.87	0.87	0.87	5706

Fig. 15. Random Forest Classifier with n=10 estimators

Fig. 16. Classification Report for Random Forest Classifier.

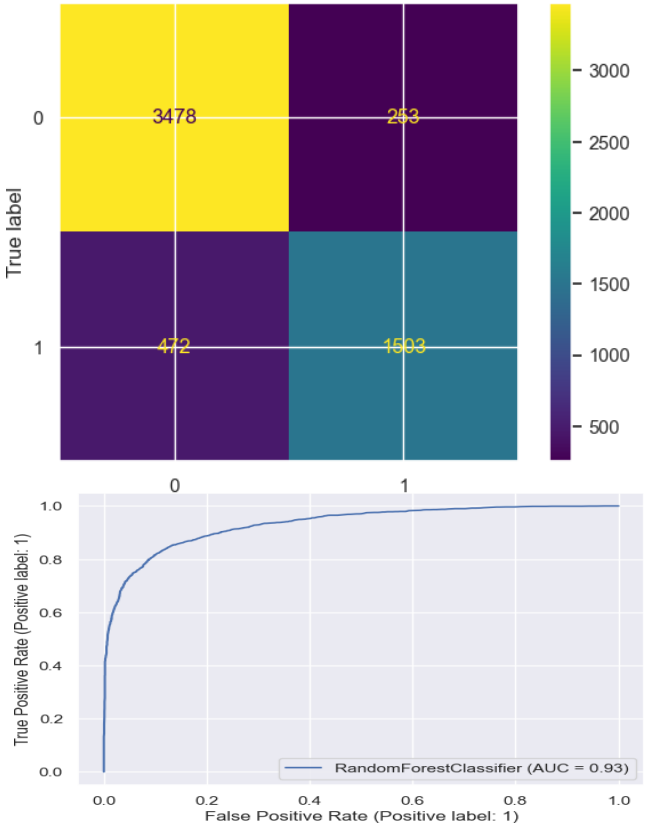


Fig. 17. Confusion Matrix for Random ForestClassification

Fig. 18. ROC-AUC Curve for Random Forest Classifier

Naïve Bayes Classifier

Naïve bayes [8], classification is a supervised machine learning technique that is ideal for solving anysort of multi-class prediction problems. It performs on an assumption of the independence of features due to which even with a lesser amount of data, the model performs well. The simple calculation is,

- P(A|B) = P(B|A) * P(A) / P(B)

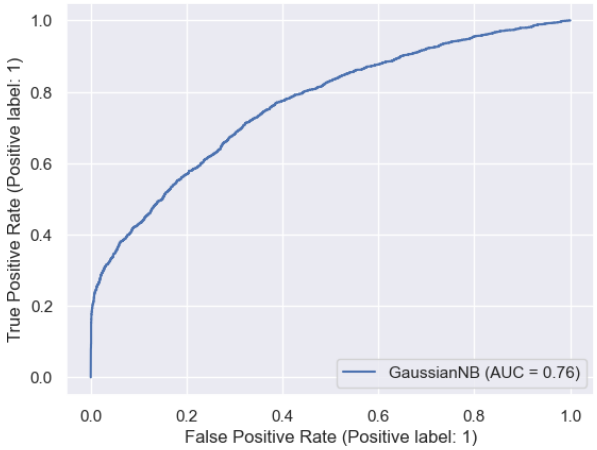


Fig. 19. ROC for Naïve Bayes Classifier

IV. Voting Classifier

Voting classifier in an estimator in machine learning realm that trains various base models and then predicts basing on the aggregate findings of each base model. There are two criteria for voting,

- Hard Voting: Voting is based on the predicted output class.
- Soft Voting: Voting is based on the predicted probability of the output class.

Accuracy: 84.7704%

❖ Evaluation Metrics

On finishing up any experiment and obtaining the results, we need to have a measure or a metric in order to measure the performance. For the same we have some commonly used evaluation metrics,

- **Accuracy:** The measurement of the closeness ofthe predicted value to the original value.
- **Precision:** The measure of closeness in measurements of the same item to each other.
- **Recall:** The model's ability to correctly predictpositives out of actual positives.

- **F1 score:** A combination of both the precisionand recall scores of a model.
- **Support:** The number of actual occurrences ofthe class in the dataset.
- **ROC-AUC:** The ROC (Receiver operating characteristic) curve is a graph that shows the performance of a classification model with two parameters *'True Positive Rate'* & *'False Positive Rate'* where True Positive Rate is just another name for recall. AUC (Area Under the ROC Curve) is the entire tw0-dimensional area underneath the ROC curve.
- Confusion Matrix: A confusion matrix represents the prediction summary in matrix form, as visualisation is the easiest representation it is better to visualise the results.

VI. EXPERIMENTAL RESULTS

Model	Score
Decision Tree	0.846477
Adaboost Classifier	0.870961
Random Forest Classifier	0.872941
Naive Bayes Classifier	0.741325
Voting Classifier (Ensembled accuracy)	0.847704

TABLE II. Accuracy Scores of all the applied Classification Techniques.

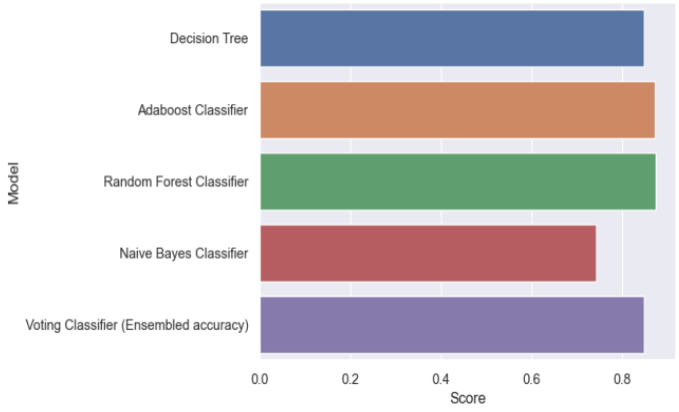


Fig. 20. Models in the order of their performances.

As per the TABLE I, we can see the accuracy scores that each of the models have achieved with the MAGIC telescope dataset. From which we can understand that Random Forest Classifier has achieved the highest accuracy of 87.29%, Adaboost classifier has achieved the next highest accuracy of 87.0961. Then in the next order there are Voting Classifier (Ensembled) Decision Tree and Naïve Bayes classifier with accuracies of 84.77%, 84.64% and 74.13% respectively.

These results indicate that all four techniques performed reasonably well in classifying the dataset. However, the Random Forest technique demonstrated the highest overall performance, outperforming the other techniques in terms of accuracy, precision, recall, and F1 score. It should be noted that the specific performance of these techniques may vary depending on the dataset and the problem at hand.

VII.DISCUSSION & CONSCLUSION

The present study addresses a highly specific problem within a particular domain, limiting its applicability to a niche user base. The challenge lies in accurately classifying gamma particles from hadrons, which poses a significant difficulty. Moreover, the dataset used in this research contains attributes such as 'FM3Trans' and 'fAsym' that exhibit negative values, requiring appropriate handling techniques. Additionally, the target variable necessitates transformation, and the entire dataset must undergo scaling before any analysis can be performed. This study aims to enhance the classification performance of algorithms for detecting high-energy gamma particles through the implementation of various techniques. The proposed methods are designed to optimize the accuracy of the classification models, with the evaluation metrics primarily focusing on the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) [9] and overall accuracy. Initially, individual techniques were applied to the dataset, followed by an ensemble approach that incorporated multiple classifiers, namely Decision Tree, Adaboost Classifier, Random Forest Classifier, and Naïve Bayes Classifier. Among these methods, the Random Forest Classifier demonstrated the highest accuracy. Consequently, based on the experimental results, it can be concluded that Random Forest is the most suitable technique for effectively classifying high-energy gamma particles in the presence of background or hadronic interference. The findings of this research contribute to the advancement of gamma particle classification and provide valuable insights for future studies in this domain. Further investigations may involve exploring alternative algorithms or refining the existing techniques to achieve even better classification performance. Ultimately, the application of accurate classification models is crucial for accurately identifying high-energy gamma particles, which has implications for various fields, including particle physics and radiation detection.

REFERENCES

[1] Heck, D., Schatz, G., Knapp, J., Thouw, T., & Capdevielle, J. N. (1998). CORSIKA: a Monte Carlo code to simulate extensive air showers (No. FZKA-6019).
[2] Emmanuel Dadzie, Kelvin Kwakye. Developing a Machine Learning Algorithm-Based Classification Models for the Detection of High-Energy Gamma Particles. 2021. ffhal-03425661.
[3] T. Hassan and others, Gamma-ray active galactic nucleus type through machine-learning algorithms, Monthly Notices of the Royal Astronomical Society, Volume 428, Issue 1, 1 January 2013, Pages 220–225, https://doi.org/10.1093/mnras/sts022.
[4] https://scikit-learn.org/stable/
[5] https://scikit-learn.org/stable/modules/tree.html

[6] Polamuri, S. (2017) How the Random Forest Algorithm Works inMachine Learning [online] available from [11 December 2017].
[7] https://www.analyticsvidhya.com/blog/2021/09/adaboost- algorithm-a-complete-guide-for-beginners/#:~:text=AdaBoost%2C%20also%20called%20Adapt i ve%20Boosting,are%20also%20called%20Decision%20Stumps.
[8] https://towardsdatascience.com/a-mathematical-explanation-of- naive-bayes-in-5-minutes44adebcdb5f8#:~:text=The%20Naive%20Bayes%20Classifier%20is,y%20given%20input%20features%20X.
[9] https://developers.google.com/machine-learning/crash- course/classification/roc-and-auc
[10] J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, C. Baixeras, J.A. Barrio, H. Bartko, D. Bastieri, J. Becker, W. Bednarek, K. Berger, C. Bigongiari, A. Biland, R.K. Bock, P. Bordas, V. Bosch-Ramon, T. Bretz, I. Britvitch, M. Camara, E. Carmona, A. Chilingarian, S. Ciprini, J.A. Coarasa, S. Commichau, J.L. Contreras, J. Cortina, M.T. Costado, V. Curtef, V. Danielyan, F. Dazzi, A. De Angelis https://doi.org/10.1016/j.nima.2007.11.068.
[11] Lyon, Robert (2015). Classification results for: A Study on Classification in Imbalanced and Partially Labelled Data Streams. figshare. Dataset. https://doi.org/10.6084/m9.figshare.1534548.v1

Author Profiles

Ms.M.Anitha Working as Assistant Professor & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with B.Tech, MCA,M. Tech in Computer Science .She has 14 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Ms.K.Baby Ramya completed her Master of Computer Applications. Currently working as an Assistant Professor in the department of MCA at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Networks, Machine Learning.

Mr.SK.Moulali is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. He has Completed Degree in B.Sc.(computers) from P.B Siddhartha College Of Arts & Science, Mogalrajapuram, Vijayawada, NTR District. His area of interest are DBMS and Machine Learning with Python.