

Classification of pulses in the LUX-ZEPLIN dark matter detector

An approach with Machine Learning Algorithms

Ronald Eduardo SOARES
Sunday Adeniyi ADEBAYO

INTRODUCTION

1.0 MAIN OBJECTIVE OF THIS STUDY

- To acquire practical understanding of some important concepts in "Statistical data analysis using Machine Learning algorithms" such as:
 - ❖ Features engineering and data classification;
 - ❖ Correlation and scatter matrices;
 - ❖ Supervised learning: decision trees & random forest;
 - ❖ Unsupervised learning: cluster analysis etc.

INTRODUCTION

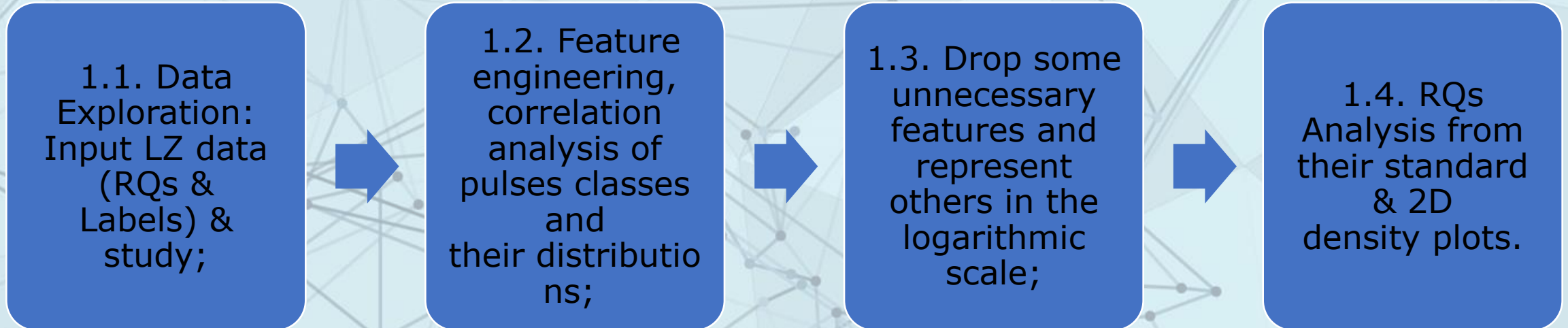
2.0 PROBLEMS DESCRIPTION

- To find the most efficient technique to classify different pulses measured in LZ dark matter detectors;
- To find the best sets of parameters (RQs), and selection criteria to obtain the most efficient classification of LZap pulse data from the LZ dark matter detector;
- To apply the understanding of Machine Learning Algorithms to solve a real-research data analysis problem.

2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP ONE:

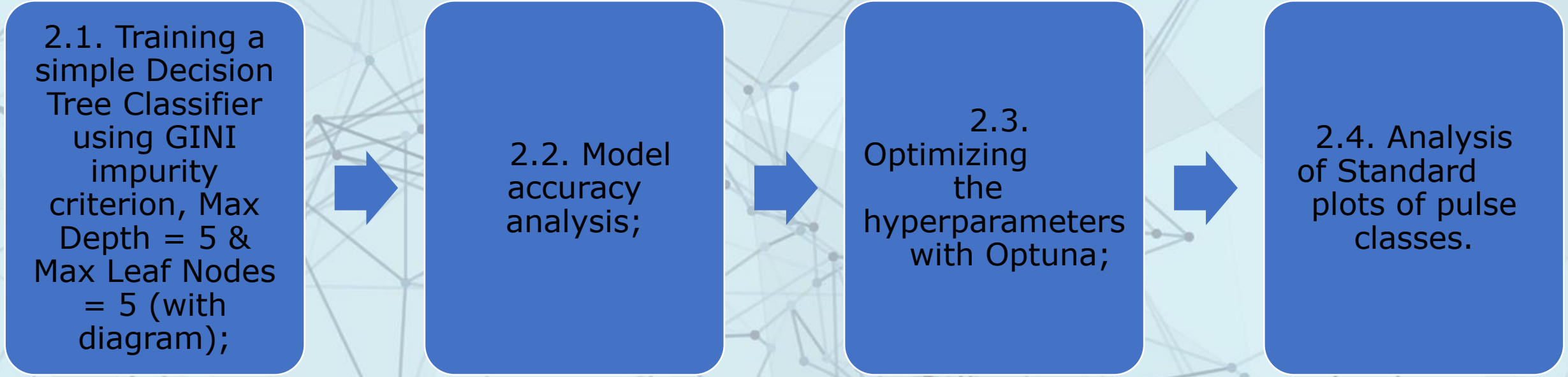


2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP TWO:

2.1. Training a simple Decision Tree Classifier using GINI impurity criterion, Max Depth = 5 & Max Leaf Nodes = 5 (with diagram);



```
graph LR; A[2.1. Training a simple Decision Tree Classifier using GINI impurity criterion, Max Depth = 5 & Max Leaf Nodes = 5 (with diagram);] --> B[2.2. Model accuracy analysis;]; B --> C[2.3. Optimizing the hyperparameters with Optuna;]; C --> D[2.4. Analysis of Standard plots of pulse classes.];
```

2.2. Model accuracy analysis;

2.3. Optimizing the hyperparameters with Optuna;

2.4. Analysis of Standard plots of pulse classes.

2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP THREE:

3.1. Preprocessing the data:
normalization process using
the standard scaler;

2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP FOUR:

4.1. Training a random forest classifier with 100 trees using GINI impurity criterion, Max Depth = 5 & Max Leaf Nodes = 5;



4.2. Model accuracy analysis;



4.3. Optimization of hyperparameters with Optuna;



4.4. Analysis of standard plots for the predicted pulse classes.

2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP FIVE:

5.1. Reducing the dimensionality of the data choosing some of the less correlated features;



5.2. Clustering the data with the Gaussian Mixture;



5.3. Checking the distribution of pulse classes for each cluster obtained;



5.4. Correlating one class with each cluster;

2.0 METHODS

SIX STEPS USED SOLVING PROBLEM IN THIS STUDY

STEP SIX:

6.1. Estimating the better hyperparameters to train a random forest with the labels that was extracted from the clusterization process;



6.2. Training a random forest;



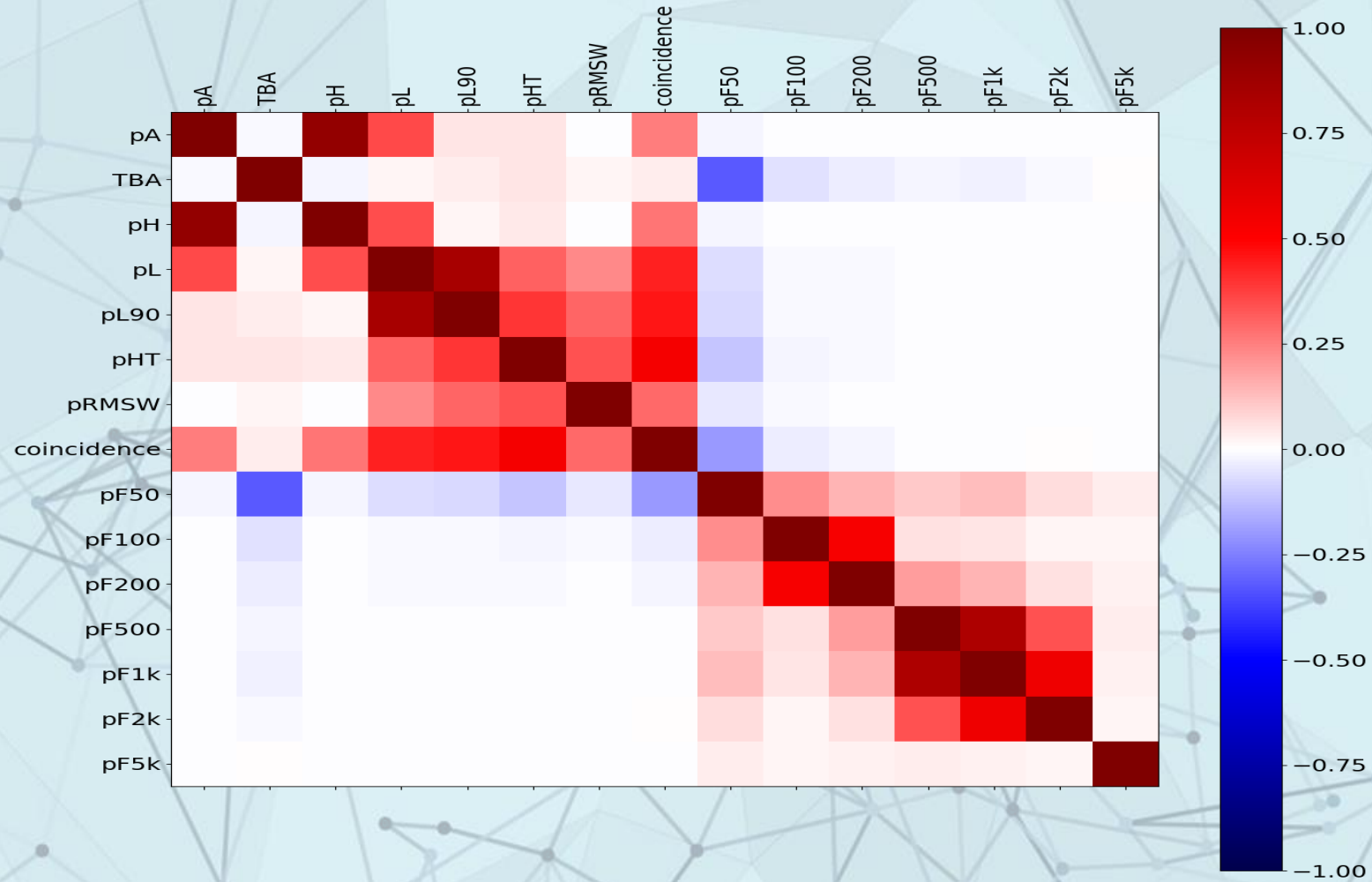
6.3. Calculating the accuracy of the model;



6.4. Producing the standard plots for each identified class.

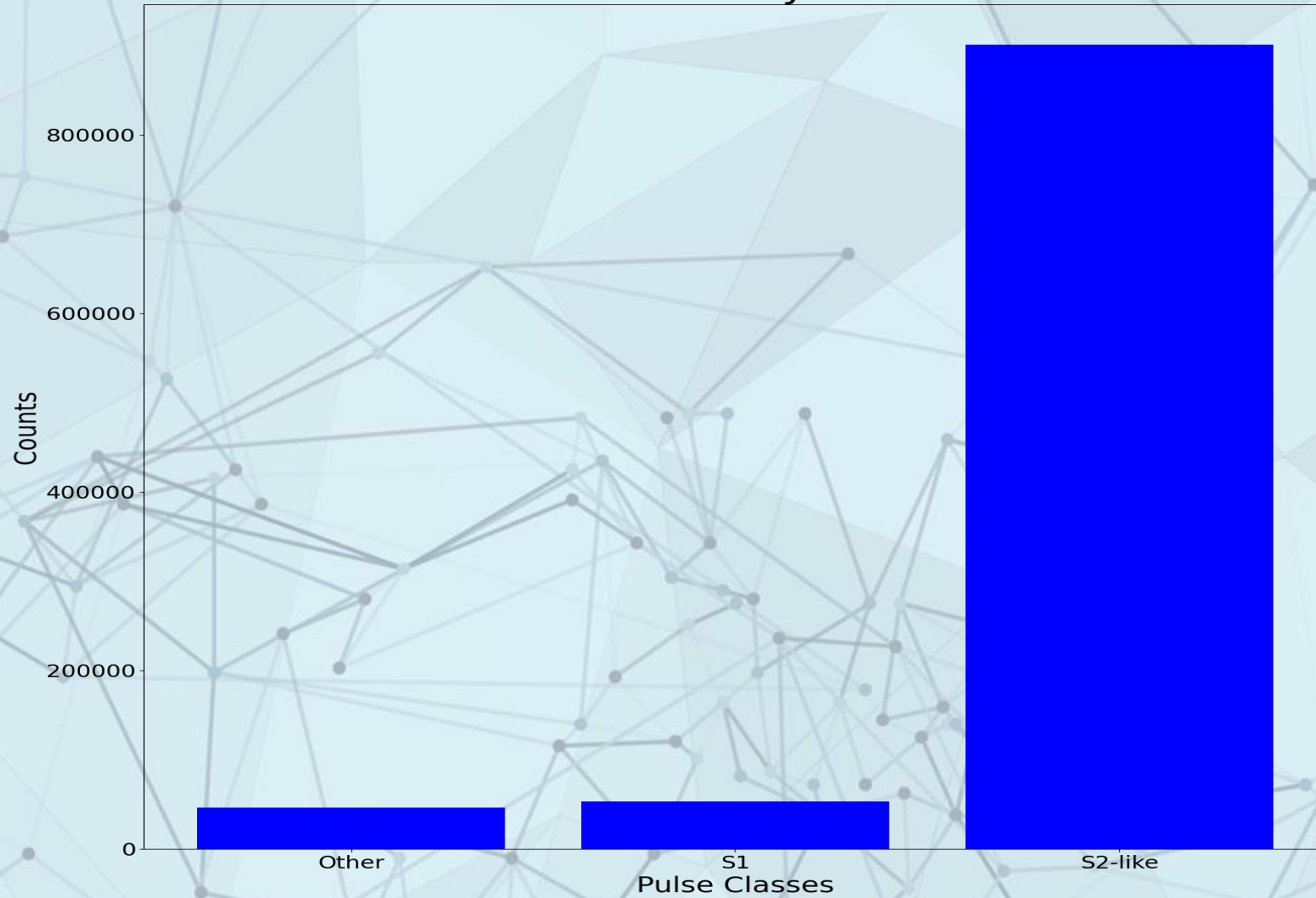
3.0 RESULTS

Correlation matrix for some of the features

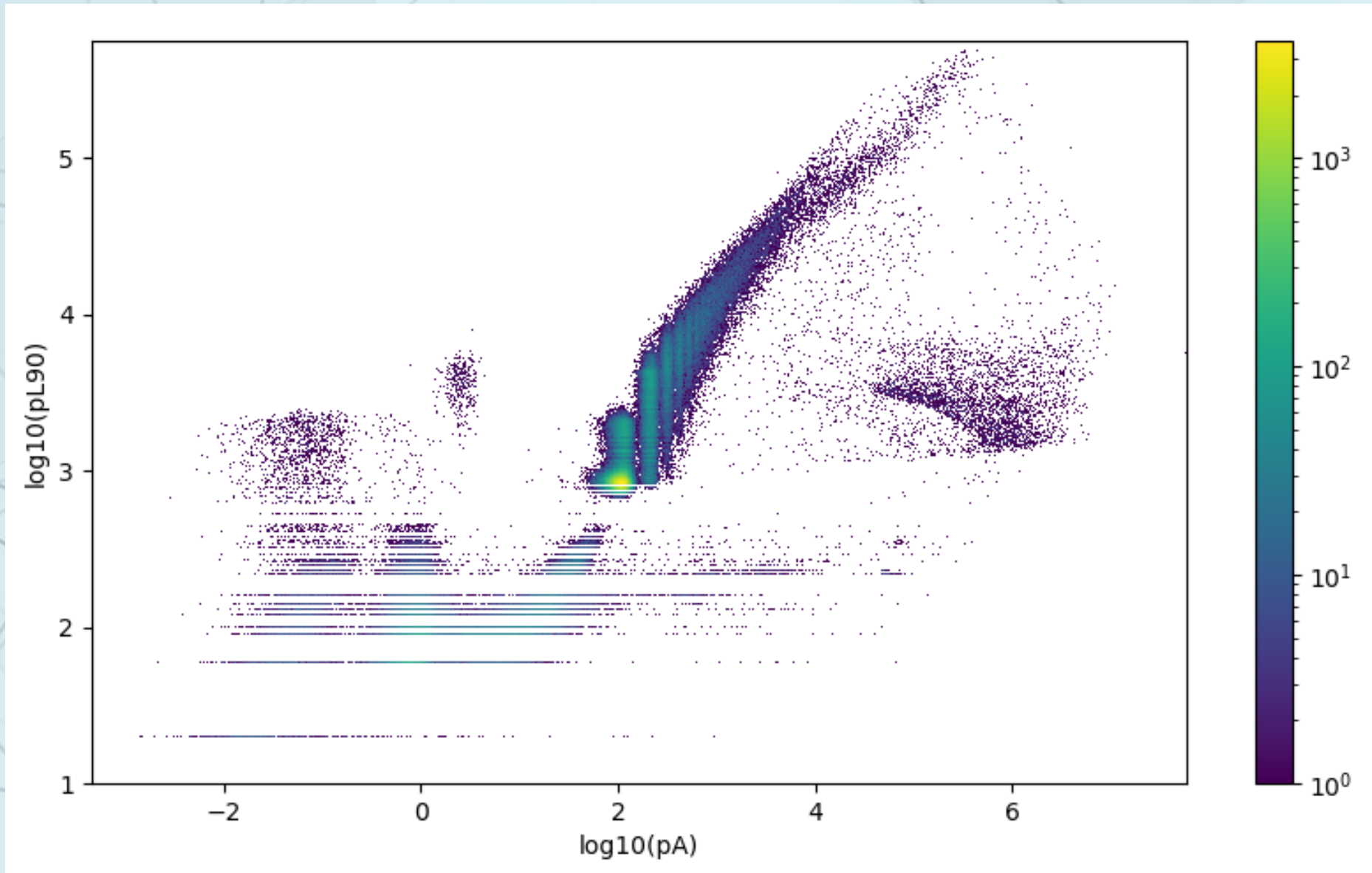


3.0 RESULTS

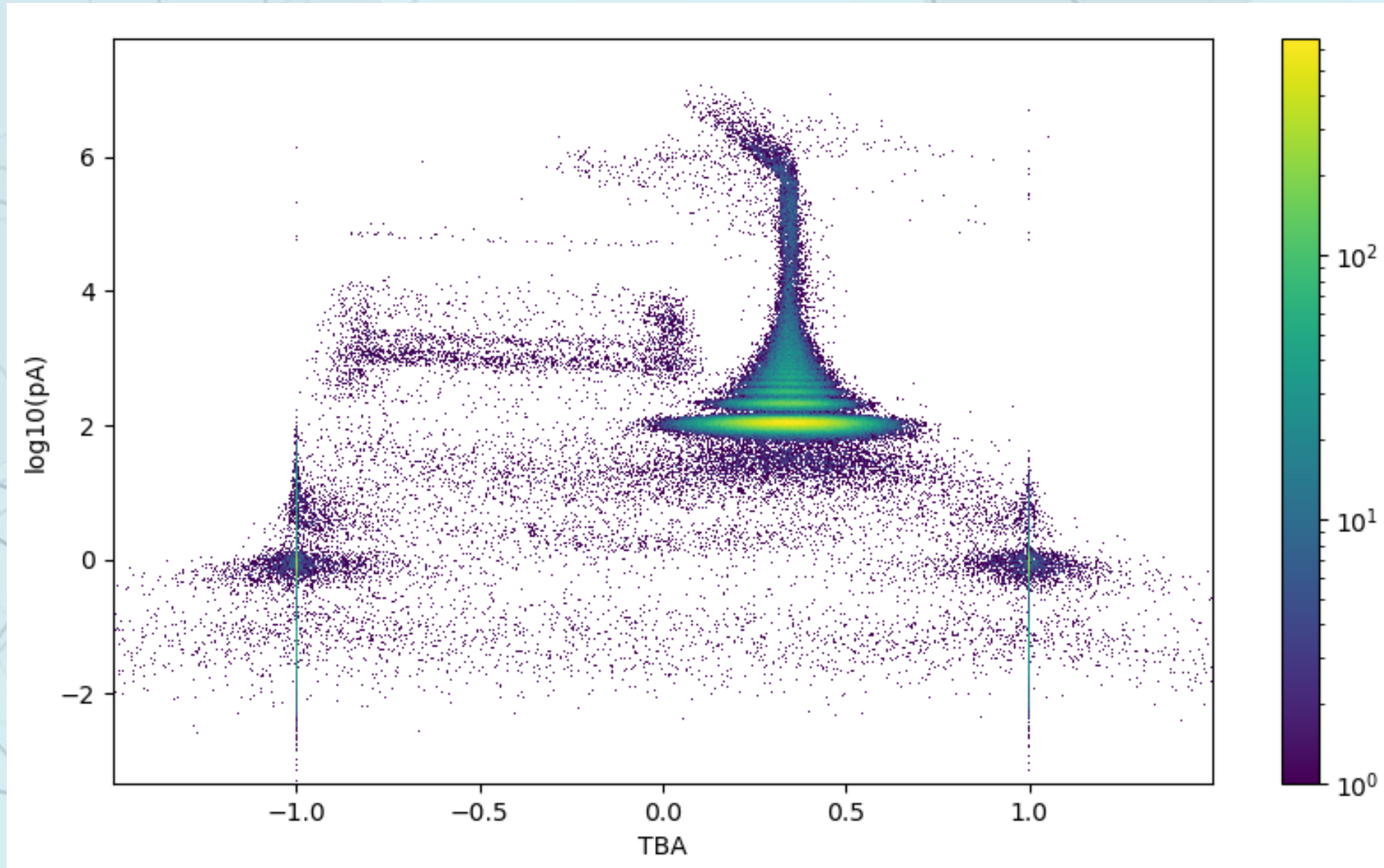
Pulse count by classes



3.0 RESULTS



3.0 RESULTS

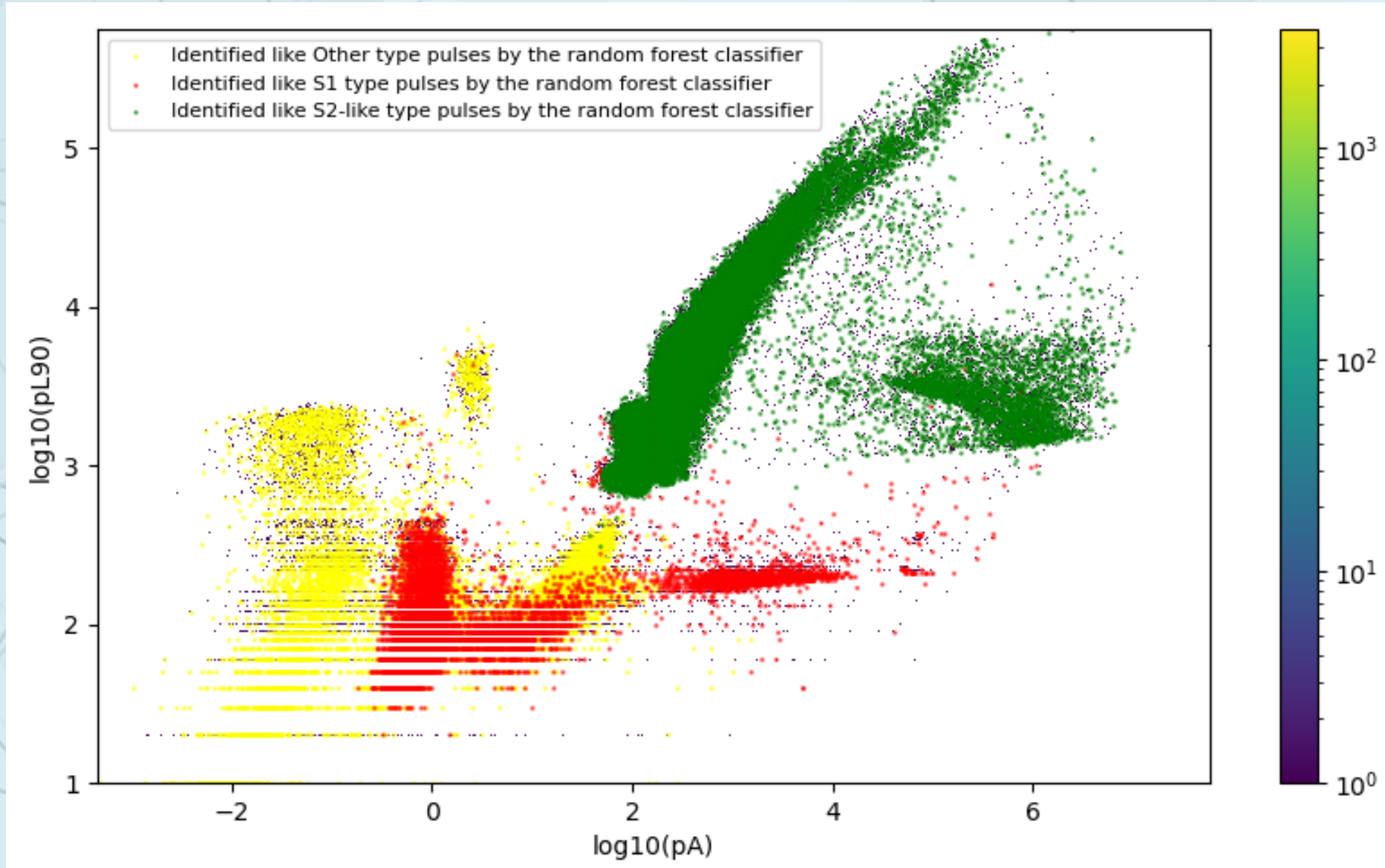


3.0 RESULTS

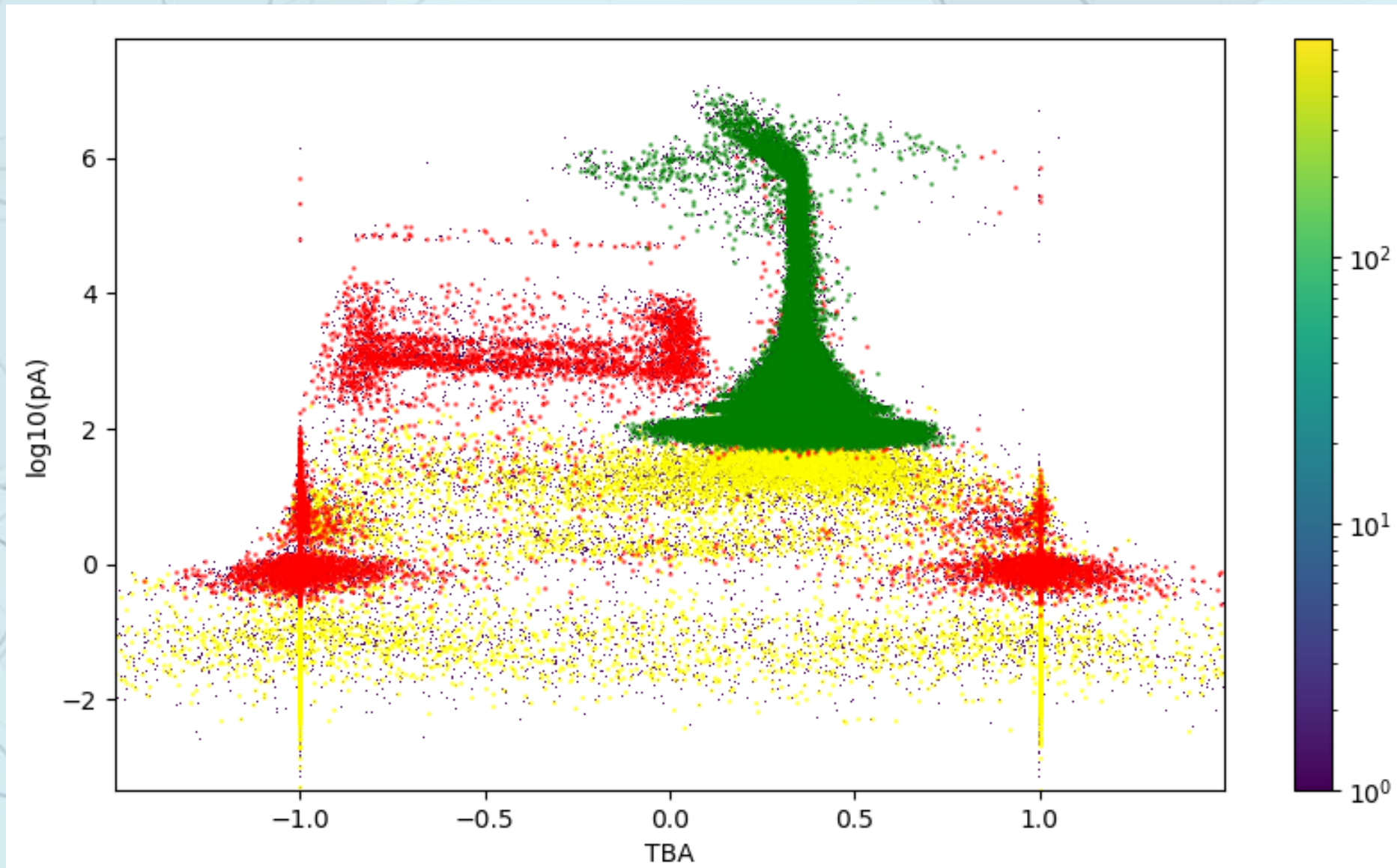
Simple Decision Tree			
	Max depth	Max leaf nodes	Accuracy
Suggested Hyperparameters	5	5	0.978
Optimized Hyperparameters	105	150	0.995

```
DecisionTreeClassifier  
DecisionTreeClassifier(class_weight='balanced', max_depth=105,  
max_leaf_nodes=150, random_state=0)
```

3.0 RESULTS



3.0 RESULTS



3.0 RESULTS

➤ After normalising the data using the standard scaler:

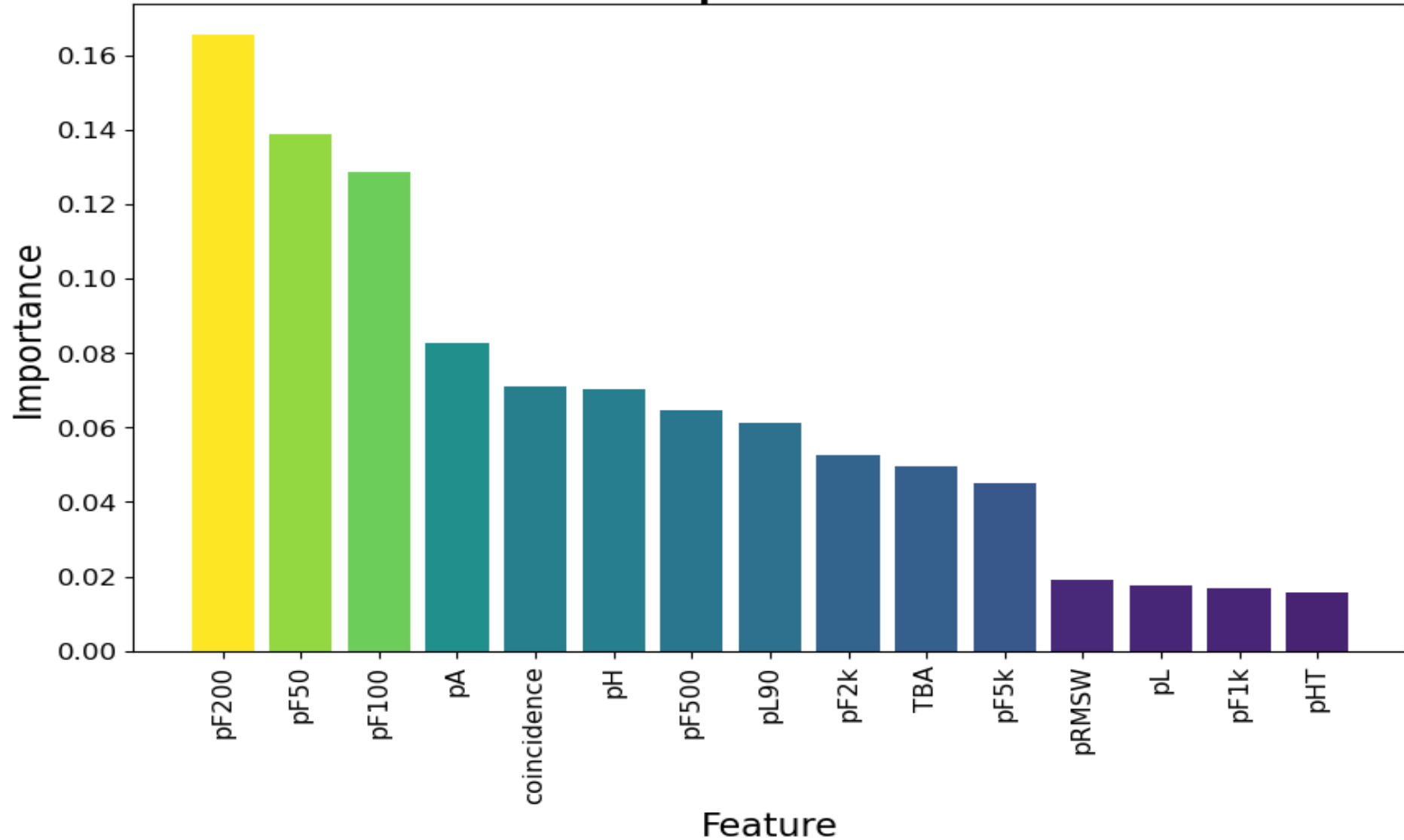
Random Forest Classifier				
	Number of Components	Max depth	Max leaf nodes	Accuracy
Suggested Hyperparameters	100	5	5	0.984
Optimized Hyperparameters	39	23	100	0.995

3.0 RESULTS

```
----- Confusion matrix -----  
-----y_test (<class 'pandas.core.series.Series'> - 200000)  
-----y_pred (<class 'pandas.core.series.Series'> - 200000)  
Predicted Class    0      1      2  
Class Label  
0                 8691   500     0  
1                 479  10104   32  
2                  0    13  180181
```

3.0 RESULTS

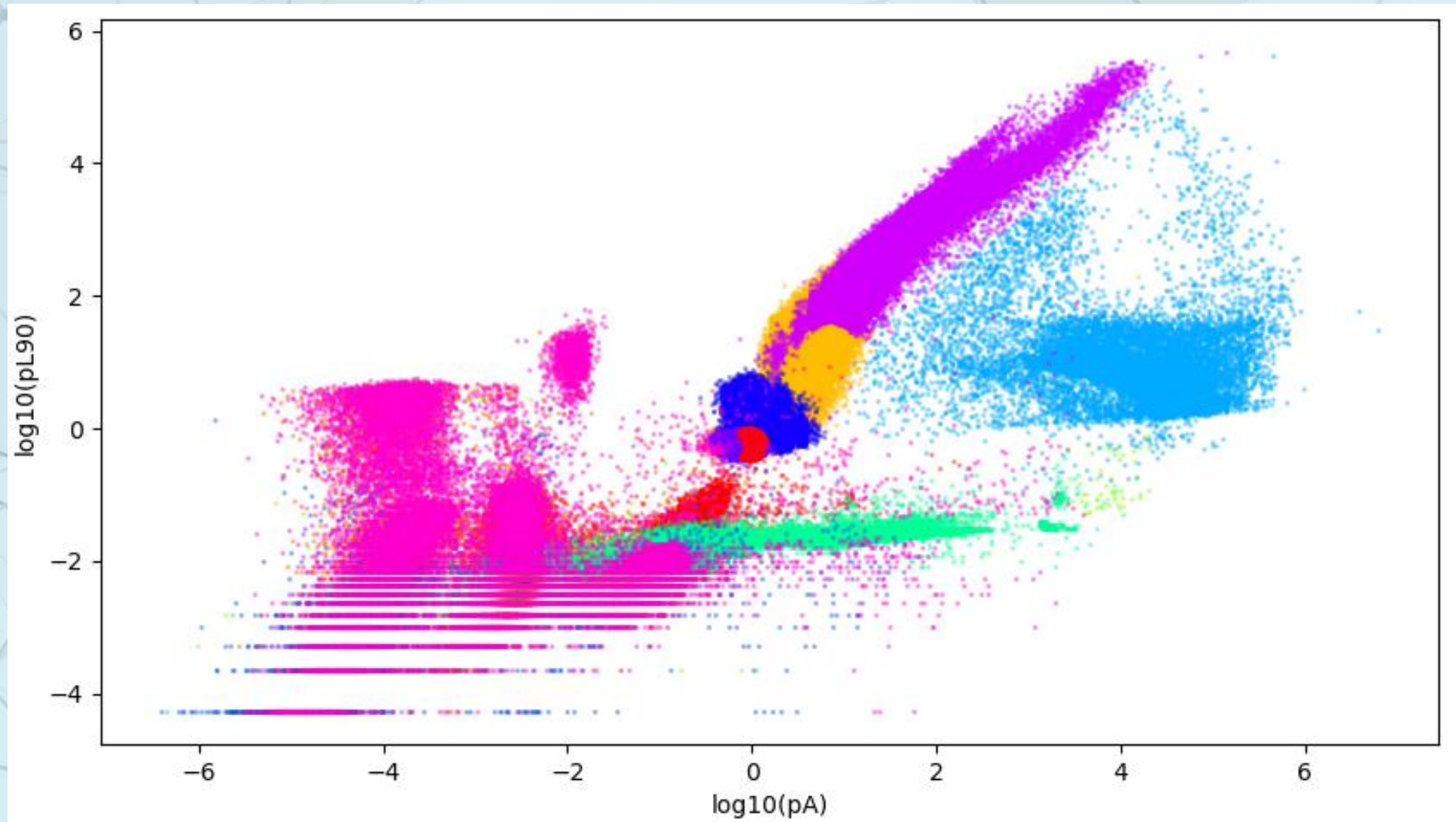
Features Importance Scores



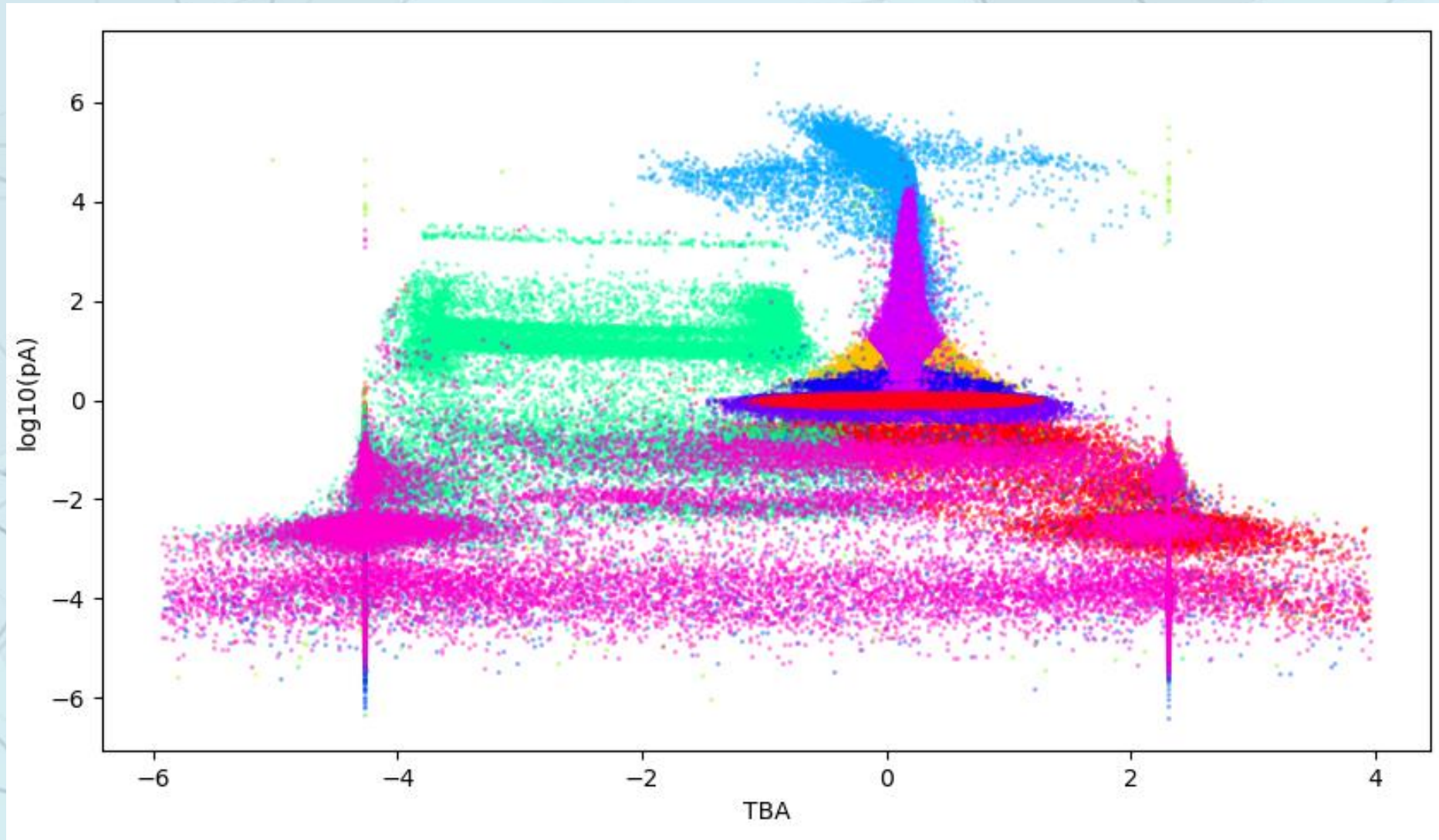
3.0 RESULTS

- Unsupervised learning process of data:
- First step: select some features with less correlation;
- Search for an ideal number of clusters;
- Study the correlation between the clusters and the label obtained from HADES;
- Analyse what cluster is correlated with each class;
- Optimize hyperparameters with optuna to train a Random Forest with the labels extracted from the clustering process;
- Train the forest;
- Analyse the accuracy. If isn't good:
 - ❑ Change the features or change the number of components in the clustering process.

3.0 RESULTS



3.0 RESULTS



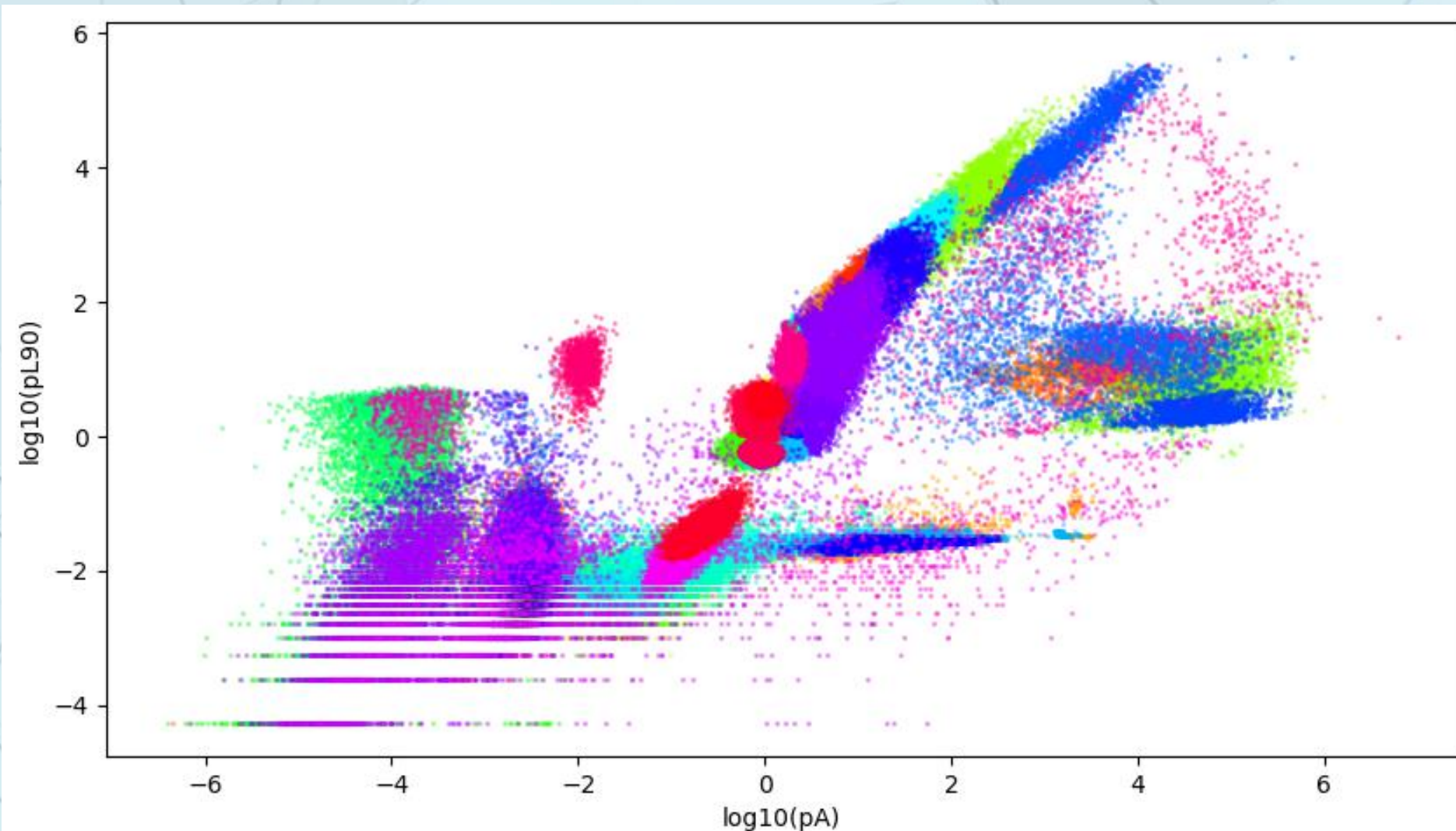
3.0 RESULTS

Cluster Label	Total Number Of Pulses	Identified Like Other Type Pulses (%)	Identified Like S1 Type Pulses (%)	Identified Like S2-like Type Pulses (%)	Assigned Class
0	29228	48%	52%	0%	1
1	18494	31%	69%	0%	1
2	161318	0%	0%	100%	2
3	7	100%	0%	0%	0
4	528	67%	31%	2%	0
5	1	100%	0%	0%	0
6	6	100%	0%	0%	0
...					
12	220902	0%	0%	100%	2
13	43404	0%	0%	100%	2
14	25277	68%	32%	0%	0
15	4	100%	0%	0%	0
16	369537	0%	0%	100%	2

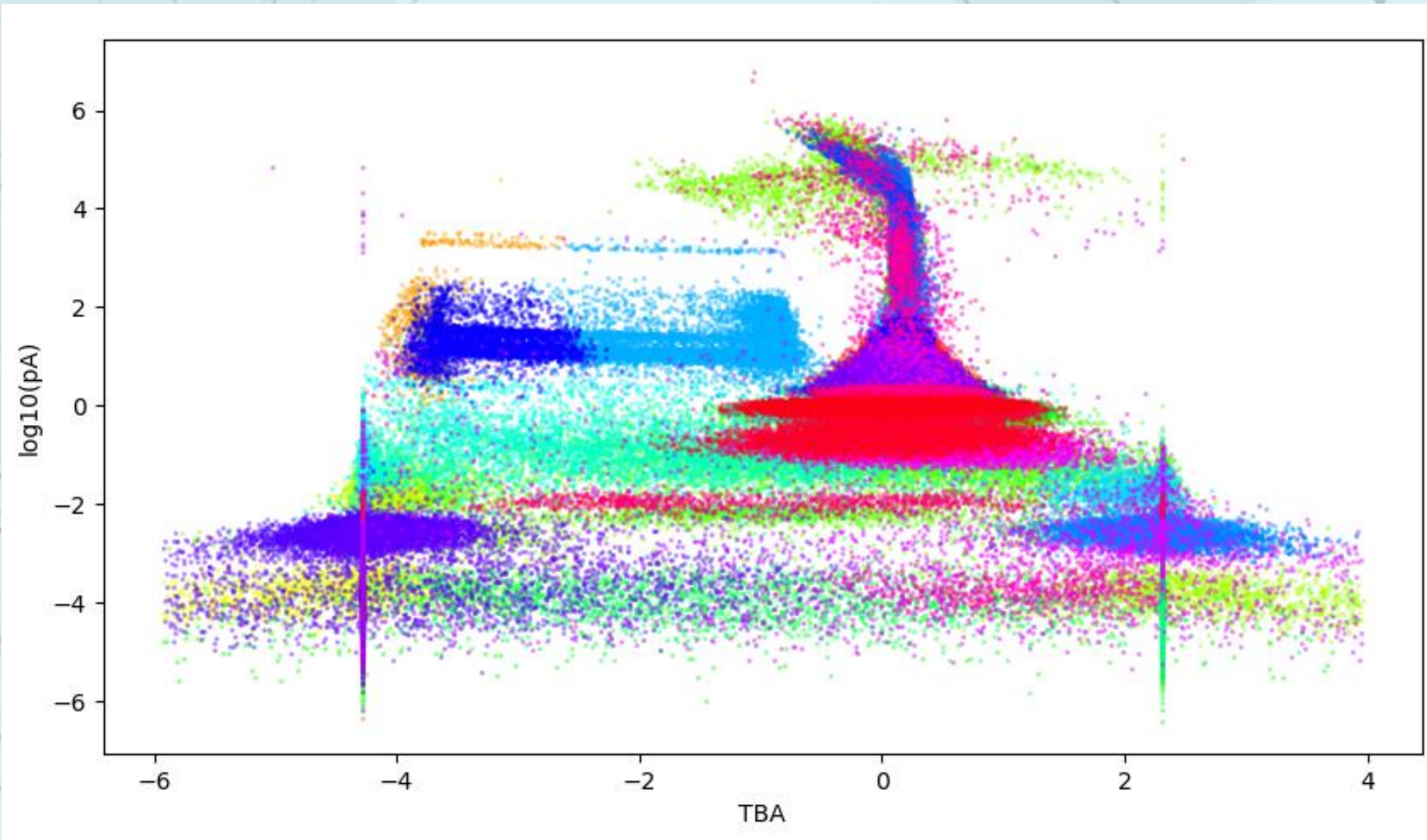
3.0 RESULTS

Features	Number of Components in the GMM (Unsupervised learning)	Accuracy of the trained random forests (Supervised learning)
pA, pF200, TBA, pL90, pF5k	17	0.974
pA, pF200, TBA, pL90, pH	17	0.985
pA, pF100, TBA, pL90, pH	17	0.986
	15	0.987
	20	0.990
	67	0.993

3.0 RESULTS



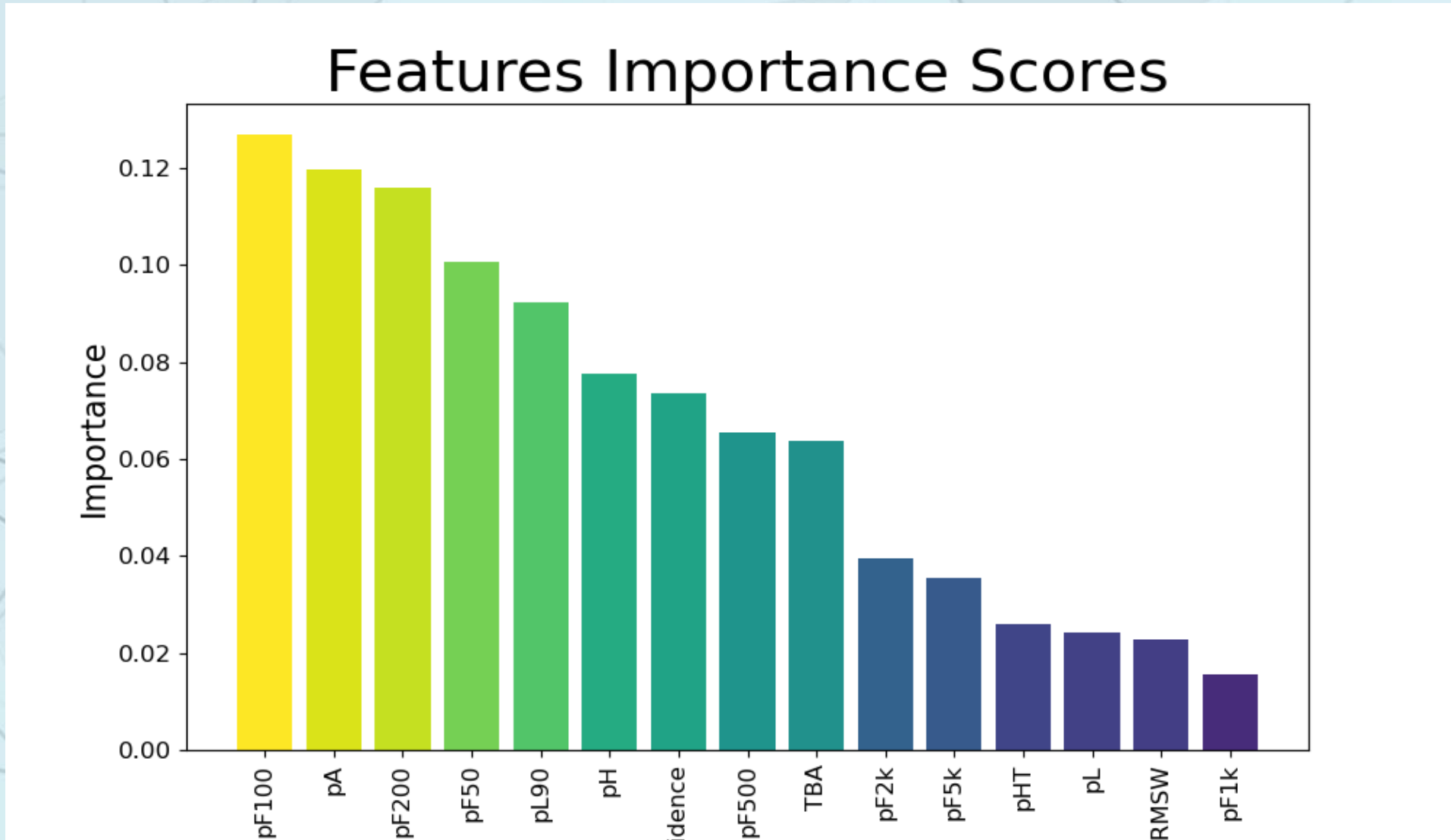
3.0 RESULTS



3.0 RESULTS

```
----- Confusion matrix -----  
-----y_test (<class 'pandas.core.series.Series'> - 1000000)  
-----y_pred (<class 'pandas.core.series.Series'> - 1000000)  
Predicted Class      0      1      2  
Class Label  
0          42358   3776      0  
1          3066   49894   138  
2           3      7  900758
```

3.0 RESULTS



4.0 Conclusion

- The three main developed models successfully classify this dataset;
- However, for general purposes, the one that can provide more information about the analyzed dataset is the training of the Random Forest using labels extracted from the unsupervised learning process;
- Choosing another clustering method might have been as efficient or even more effective than GMM, even with a smaller number of components.

References

- BRAZ, Paulo Alexandre Brinca da Costa. Sensitivity to the $0\nu\beta\beta$ decay of ^{136}Xe and development of Machine Learning tools for pulse classification for the LUX-ZEPLIN experiment. 2021. Tese de Doutoramento em Física, Astrofísica - Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Universidade de Coimbra.



**THANK YOU FOR
YOUR ATTENTION**