**A Project Report on**

# CONVEX HULL TO CHARACTERISE IRRADIATION DAMAGE

Submitted to

## BHABHA ATOMIC RESEARCH CENTRE, VISAKHAPATNAM


Submitted by:

**SIVA RISHITA PAPPALA**

**YASWANTH AMBARUKHANA**

Under the guidance of:

**UTKARSH BHARADWAJ**

***Scientific Officer D,***

***Computational Analysis Division (CAD),***

***BARC, Visakhapatnam***

# ABSTRACT

The study of collision cascades database is helpful in radiation damage study of different materials. It is of immense practical use in predicting changes in properties of fission and fusion reactor materials. We present a novel method to characterise damage areas with volume, density and structure using computational geometric algorithms like convex hull and classify damage areas based on geometric properties. The methods are applied on a database of collision cascades in Fe and W at energies ranging from 10 keV to 200 KeV. The results show interesting relations among geometric properties of convex hull providing new insights and parameters that can be used in simulations at higher scales. They also manifest whether collision cascades contains sub-cascades or not with noticeable accuracy. We discussed each step, starting with application of the computational geometry algorithms on simulation output of collision cascades to reduction of physics problems to machine learning stages viz. feature extraction, dimensionality reduction and supervised classification.

We also accord structural visualisation of convex hull in different dimensions. Additionally, we performed qualitative analysis on collision cascades' characteristics (energy, substrate), convex hull features like area, volume, density, no. of indices, etc., using different statistical plots. We can infer interesting trends from statistical plots that can further be used to carry out classification. The supervised machine learning classification figures the existence of sub-cascades if present in a cascade.

We discuss the key points and computational efficiency of the algorithms along with various prominent results of the application. The open-source software implementation of the methods along with the supporting analysis and visualisations gives a supervised approach for data exploration and classification of collision cascades. The geometrical properties of convex hull like volume, density, etc., for different elements and energies can be used as input to higher scale models in a multi-scale radiation damage study.

# INDEX


| S. NO. | CONTENTS | PAGE NO. |
|---|---|---|
| 1 | Introduction | 1 |
| 2 | Motivation | 3 |
| 3 | Objectives | 4 |
| 4 | Software Requirements and Skills | 4 |
| 5 | Methods and Algorithms<br>2.1 Finding Convex Hull and its Features<br>2.2 Analysis with various types of Statistical Plots<br>2.3 Correlations<br>2.4 Logistic Regression - Binary Classification<br>2.5 Logistic Regression - Multi Classification | 5 |
| 6 | Results and Observations | 9 |
| 7 | Conclusion | 33 |
| 8 | References | 34 |

# 1. INTRODUCTION

Defects caused by primary knock-on atoms in materials can be simulated using molecular dynamics. Primary damage of materials due to neutron irradiation occurs via energetic cascades caused by energetic primary knock-on atoms (PKA), created by energetic neutrons as they pass through the material. These cascades result in creation of Frenkel Pairs (interstitial-vacancy pairs). The interstitials and vacancies diffuse to (I) nullify the damage when an interstitial recombines with a vacancy, (II) form interstitial clusters when two or more interstitials recombine, and (III) form vacancy clusters when several vacancies come together. The PKA is a lattice atom that gets energy by interacting with a high energy neutron. It is generally placed at the centre of the cubic box.

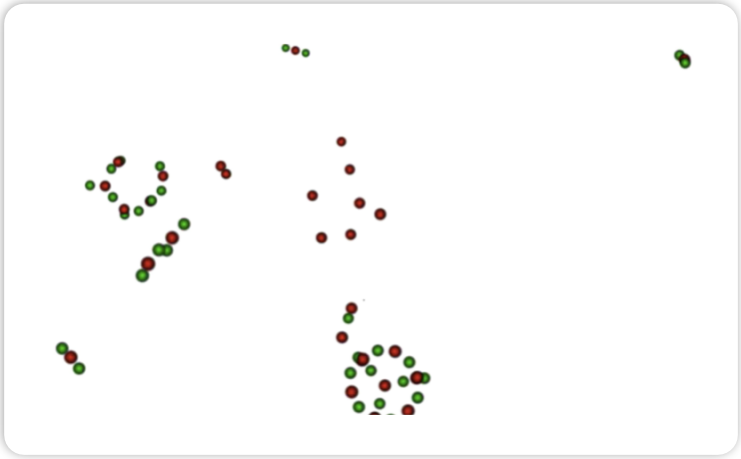


*Figure 1: A picture of the Frenkel pairs in W at PKA energy of 5 keV. Vacancies are in red and the interstitials are in green colour. It shows dumbbells, crowdions and ring like arrangement of interstitial clusters.*

The primary cascades are modelled by Molecular Dynamics (MD) Simulations. MD simulations of collisional cascades takes into consideration many body effects and therefore is accurate. It also provides insight into the atomistic details like in-cascade clustering of interstitials, their configuration and also their dynamic evolution up-to several nanoseconds.

However, it has inherent limitations in the number of atoms it can simulate (at most a hundred billion atoms, which corresponds to a metallic solid of size equal to a fraction of a micron, on advanced supercomputers), the time it can span (around 10 ns) and the huge computational costs involved.

A technical Meeting of the Code Centre Network on Molecular Dynamics Data of Collisional Cascades after Irradiation was organised by the International Atomic Energy Agency (IAEA), 16-17 November 2017, IAEA Headquarters, Vienna, Austria. The meeting brought together experts on the theoretical modelling of radiation damage in materials relevant to fusion reactor design in order to plan a database of collisional cascade molecular dynamics (MD) simulations. This database is to be hosted by the Atomic and Molecular Data Unit and will provide a central repository for the results of MD simulations of the evolution of a material's structure in the "ballistic phase" following impact by a high-energy particle.

The data which we explored and performed various analysis is obtained from primary knock-on atoms (PKA) collision cascades database using Molecular Dynamics (MD) Simulations which is picked up from data files for the IAEA Challenge on Materials for Fusion. The data provided in the IAEA Challenge are the positions of the atoms of either tungsten, W, or iron, Fe, after a collisional cascade molecular dynamics simulation run for 40 picoseconds. The initial state of the material is taken to be a perfect crystal, and the final state is provided as the (*x, y, z*) locations of the atoms after the energy of the impacting neutron has been absorbed. Four different impact energies are considered for each material and at least seven different simulations are run (for different impact directions) for each energy. The challenge was to come up with innovative ways to visualise, analyse and explore the provided data. As to address the challenge CSaransh is created. It is a software suite to post-process, explore and visualise Molecular Dynamics (MD) simulations of collision cascades. It is an elaborate software solution for studying MD results of radiation damage simulations, starting from identifying defects from xyz file to finding correlations, visualising sub-cascades to pattern matching clusters, etc.

In our project, analysis and methods were performed on a data, which contains defect coordinates obtained from results of MD simulations on Fe and W of PKA in different directions

at energies 10, 20, 50, 100, 150 & 200 KeV which yielded coordinates of all atoms. And after post processing to find the interstitials and vacancies, only the defect coordinates are left out which consisted of our data.

Cascades are formed as previously mentioned above due to primary knock-on atoms (PKA). 76 cascades containing only defect coordinates of Fe and W constituted the data. The defects in metals with body-centered cubic structure are produced in the form of single point defects (interstitials and vacancies) or clusters of such defects.

All cascades formed either consist a single cascade or multiple sub-cascades. Sub-cascades are disturbances formed in cascades. Classification of sub-cascades in different irradiated samples is the first step in the systematic study of properties of sub-cascade in collision cascades and their effects. The basic classification of sub-cascades in collision cascades can be put in the following machine learning template steps:

1. Feature extraction.
2. Correlational analysis on characteristic features of the convex hull and cascades' attributes.
3. Dimensionality reduction.
4. Classification with a supervised machine learning algorithm.

## 2. MOTIVATION

Classification of sub-cascades is insightful for further classification of classes of sub-cascades shapes which in-turn can be used to predict (predict the class of sub-cascade given few parameters) and visualise the distribution of disturbances in collision cascades simulation at higher scales, as the number of sub cascades affect the cluster size and shape distribution. Thus, the study can help in defining a high energy cascade as a combination of different lower energy sub-cascades.

# 3. OBJECTIVES

- Data exploration on collision cascades database
- Visualisation of collision cascades
- Data Analysis on features extracted from collision cascades' database
- Finding trends among features extracted
- Classification of sub-cascades using machine learning algorithms

# 4. SOFTWARE REQUIREMENTS AND SKILLS

| S. NO. | ITEM | ITEM TYPE | ITEM DESCRIPTION |
|---|---|---|---|
| 1 | Anaconda Distribution | Python/R Data Science Platform | The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. |
| 2 | Python | Programming Language | Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. |
| 3 | Plotly | Open Source Data analytics and Visualisation tool. | Main use of Plotly in our project is Plotly's Python graphing library which makes interactive, publication-quality graphs like scatter plots, box plots, histograms, heatmaps, subplots, etc. |

# 5. METHODS AND ALGORITHMS

## 2.1 Finding convex hull and its features:

**To read file and get data :-**

We load the json file containing the data (i.e., 76 cascades of Fe and W) into a dictionary using which we find convex hull features like area, volume, density, no. of indices, vertices, simplices, etc., Using ConvexHull library from scipy.spatial we enclose all the vacancies coordinates which are non-annihilated i.e., true defects to form a convex hull in-order to know and draw few important features of these points. Using these coordinates, we also plot convex hull either in 2D or 3D.

**2D plotting:-**

We can plot convex hull in 2D either using 2D coordinates of non-annihilated vacancies or 3D coordinates of non-annihilated vacancies. Convex hull is generated with either of the coordinates. The points are plotted using pyplot in matplotlib and the lines joining those points (the hull points) are plotted using simplices of the convex hull. We can visualise the convex hull either by using notebook or qt5 from matplotlib.

We have improvised the visualisation of convex hull by plotting it in 3-dimensional space using graph_objs of plotly.

**3D plotting:-**

For plotting in 3D we used the 3D coordinates of non-annihilated vacancies. The convex hull is generated using these points. The hull indices are points which contains the vertices of the convex hull. Non hull points are found by performing symmetric difference between the sets of all points and the hull indices/hull points. Using hull simplices we join lines among the hull points. With the help of plotly library we visualised our convex hull in either scatter 3D or mesh 3D. Scatter 3D just shows hull points, non-hull points and lines joining the hull points in three dimensional view whereas, mesh 3D shows the facets of the convex hull in three dimensional space.

## 2.2 Analysis with various types of statistical plots:

We analyse the most important features of the convex hull like area, volume, density and no. of indices which were previously generated using ConvexHull library with various statistical plots (histograms, box-plots, etc.).

Firstly, we plotted convex hull features mentioned above using histograms separately. These histograms are drawn with the help of pyplot from matplotlib. Next, the same features of plotted with histograms according to energies of cascades in the data i.e., a histogram plot is made for each feature according to unique energies present in the data. Similarly, histogram plots are drawn for all features according to all unique elements present in the data.

Secondly, histogram plots are made for all features according to unique elements which are in-turn again according to unique energies correspondingly. That means, a plot is drawn for each element which again contain a number of energies. This is done for better understanding of trends followed by each element according to energies. Instead of previous method which was a tedious approach while analysing, we came up with drawing histograms for all features according to elements and within the plot according to energies represented with different colours. In this manner one can easily identify how features according to elements display different trends .

Till now, we have analysed important features of the convex hull by using histograms alone. Histograms can tell us about the underlying distribution of data and shows the variability in data. They are normally used to represent moderate to large amounts of data. They give an insight about the nature of the data i.e., the minimum and maximum frequencies, the shape of distribution. We can also adjust the look of histograms by modifying the range attribute and bin sizes.

We have also made box-plots to characterise damage areas (non-annihilated vacancies) with important features (area, volume, density, no. of indices) using computational geometric convex hull algorithm. Box plots can give us more detailed information of the distribution by giving the mean, median, minimum and maximum values, interquartile ranges and outliers.

These are primarily used when we compare several distributions among each other. They are good for moderate to large amounts of data.

## 2.3 Correlations:

Correlation is often used as a preliminary technique to discover relationships between two variables. Once correlation is known it can be used to make predictions. When we know a score on one measure, we can make a more accurate prediction of another measure that is highly related to it. The stronger the relationship between/among variables, the more accurate is the prediction.

We found correlations among important features of convex hull like area, volume, density, no. of indices and energy, sub-cascades count, n defects with the help of corr() function in pandas library. We also visualised it with correlation scatter matrix graph from pandas.plotting. Additionally, we have customised the correlation scatter matrix plot according to energies (each in different colour) available in the data. Further, we have made scatter matrix plot for the data according to elements which are in-turn according to energies (each in different colour). By this we get better visualisation and understanding of how the correlation features display different trends which vary according to elements and in-turn also according to energies of the corresponding elements.

## 2.4 Logistic regression - Binary classification:

Logistic regression falls under the category of supervised learning; it measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic/sigmoid function. In spite of the name ‘logistic regression’, this is not used for regression problem where the task is to predict the real-valued output. The dependent variable should be dichotomous in nature (e.g., presence vs. absent). It is a probabilistic classification problem which is used to predict a binary outcome (1/0, -1/1, True/ False) given a set of independent variables.

We have chosen logistic regression as, we had created labelled data (csv) from a given data file (json) and we need to predict whether a cascade contains a multiple sub-cascade or not. We performed Logistic Regression with the help of linear model in sklearn library before which we checked for null values, allocated independent variable (important features of convex hull, energy and n-defects) & dependent variable (multiple_sub-cascades which is 0 if cascade contains single cascade else 1), had split the data into train data & test data using train_test_split from sklearn.model_selection library by giving the desired test size and performed standardisation using Standard Scalar from sklearn.processing library. After which we generated classification report accuracy and training accuracy (to know if our model overfits, under-fits or best-fits).

To reduce the number of features used as independent variables we have used PCA. Principal Component Analysis (PCA) is a linear transformation technique and is considered as Un-supervised machine learning method. By giving The N-components value we can transform the variables (independent variable features - correlated attributes) to given number of new set of variables which are known as the principal components (uncorrelated attributes) while possibly keeping as much as variance in the data.

## 2.5 Logistic Regression - Multi classification:

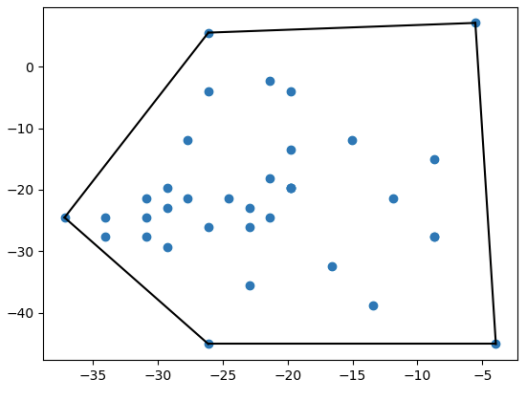
Binary classification is quite useful if we just want to know the presence of multiple sub-cascades in a cascade. But if we want a fair prediction of number of sub-cascades a cascade contain then, we need to perform logistic regression with multi classification which is just a further extension of logistic regression with binary classification. Multinomial logistic regression/Logistic regression with multi classification is a classification method that generalises logistic regression to multi-class problems, i.e., with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a dependent variable, given a set of independent variables.

Multivariate logistic regression is performed similarly as that of logistic regression with binary classification with the only difference being the possible outcome. Here the possible

outcome which is the dependent variable classifies the quantity/number of sub-cascades existing in a cascade (i.e., the output can be 0,1, 2,… sub-cascades in a cascade). We also perform PCA in the same way as we did in logistic regression with binary classification.

For much better accuracy we remove data of cascades with sub-cascade count 0 or 1 and predict just multiple sub-cascades present in a cascade (i.e.,-the output can be 2, 3,… sub-cascades in a cascade).

# 6. RESULTS AND OBSERVATIONS



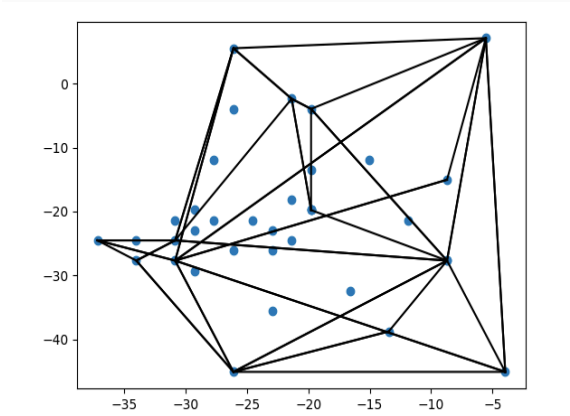




*Fig. 2(a)* *Fig. 2(b)*

*Figure 2: 2D plotting of Convex Hull with (a) 2D-coordinates and (b) 3D-coordinates.The coordinates/ points are from a cascade with substrate W at 50 KeV.*

***Note**: The data used here (or anywhere, if nothing else is mentioned) consists of non-annihilated vacancy defects from 76 cascades of Fe, W at energies 10, 20, 50, 100, 150, 200 KeV.*

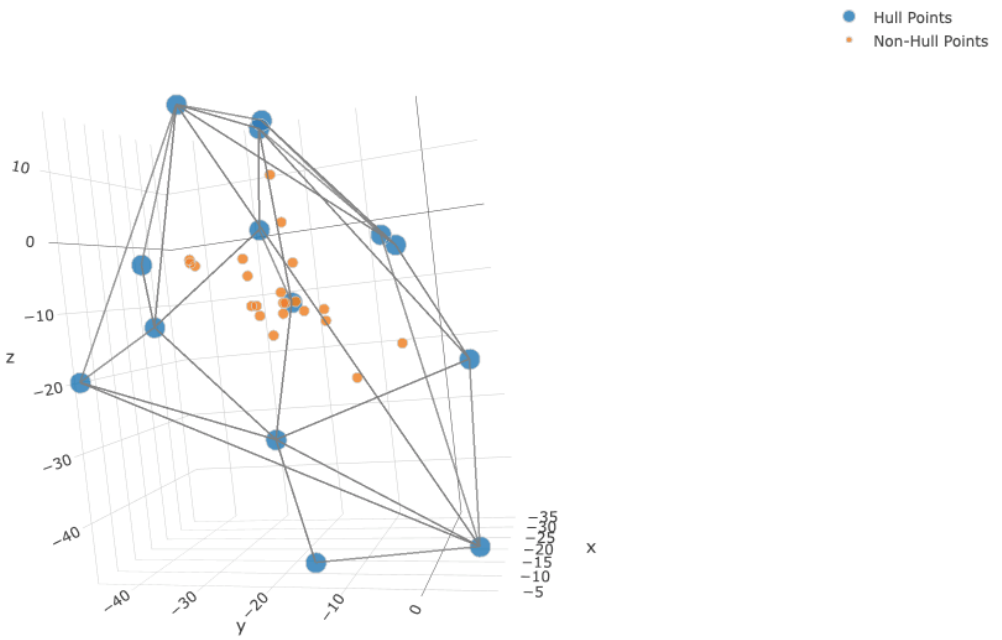



*Figure 3: 3D scatter plot of the convex hull using coordinates from a cascade with substrate W at 50 KeV.*

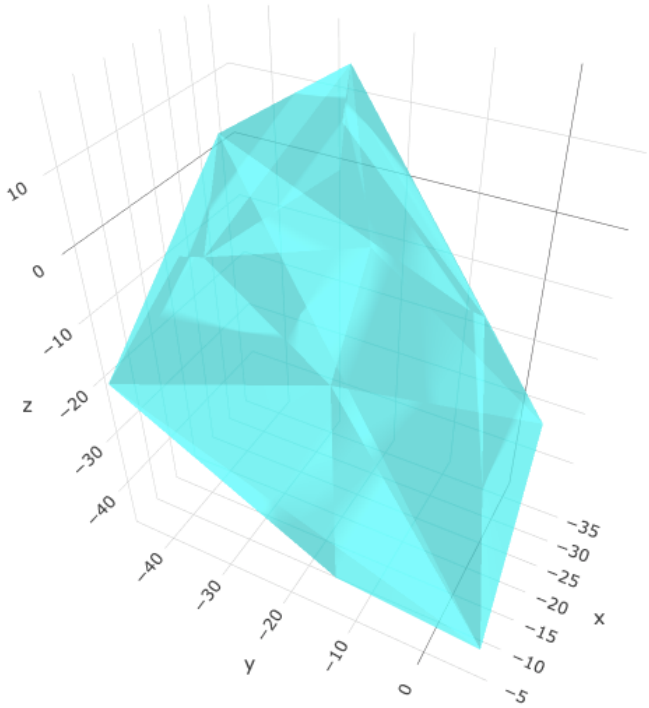



*Figure 4: 3D mesh plot of the convex hull using coordinates from a cascade with substrate W at 50 KeV.*

**********PLOTS FOR IMP. HULL FEATURES SEPARATELY**********

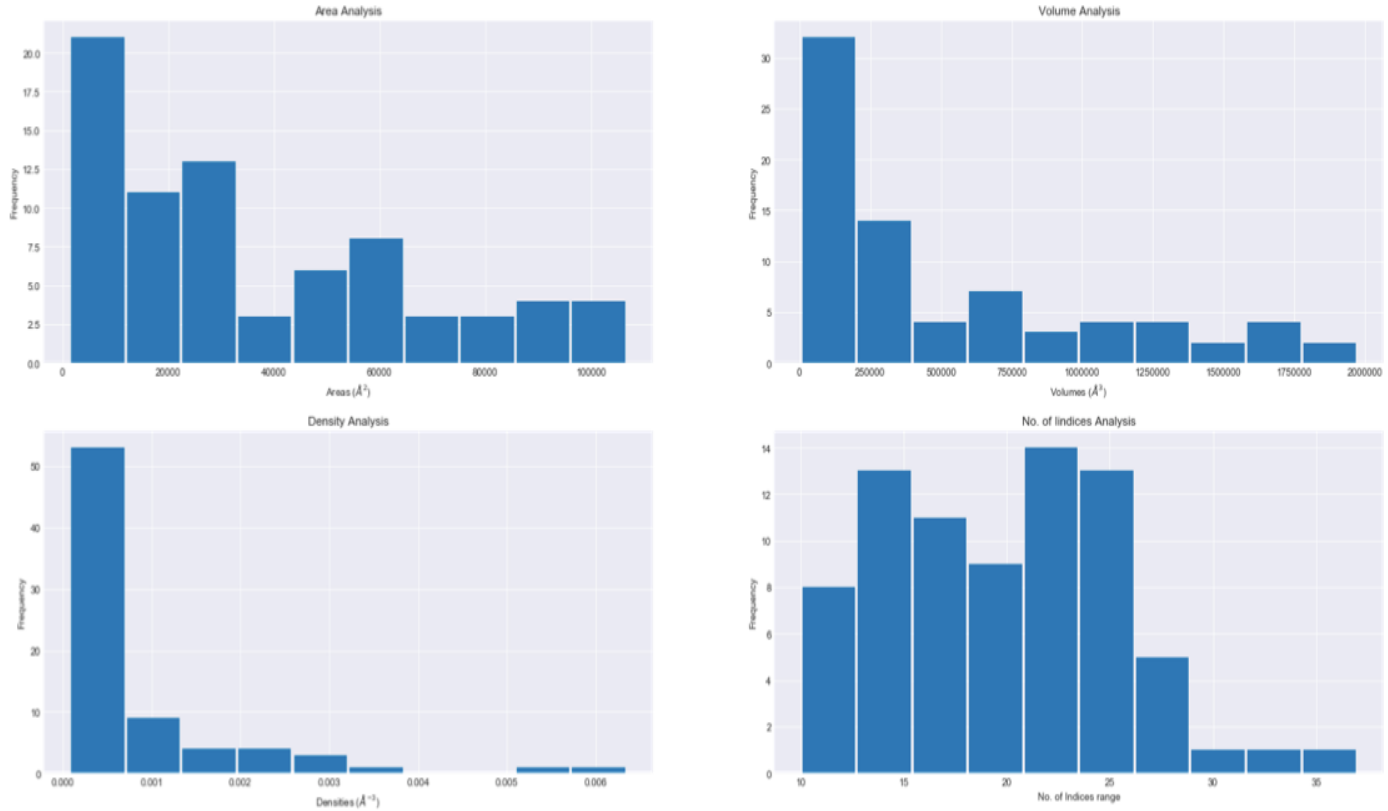
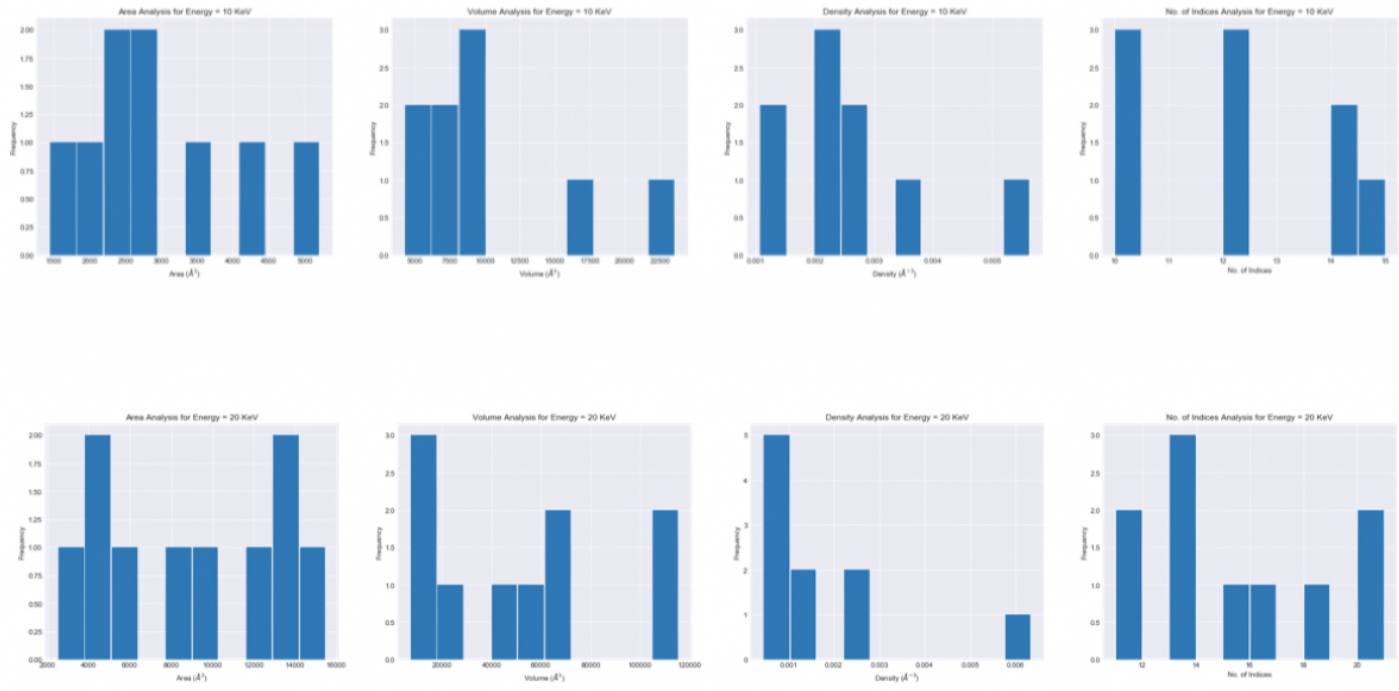



*Figure 5: Histogram sub-plots for convex hull features like area, volume, density and no. of indices separately.*

**********PLOTS FOR THE WHOLE DATA ACCORDING TO ENERGIES (KeV)**********

AREA, VOLUME & NO_OF_INDICES PLOTS ACCORDING TO ENERGIES (KeV) : [10, 20, 50, 100, 150, 200]



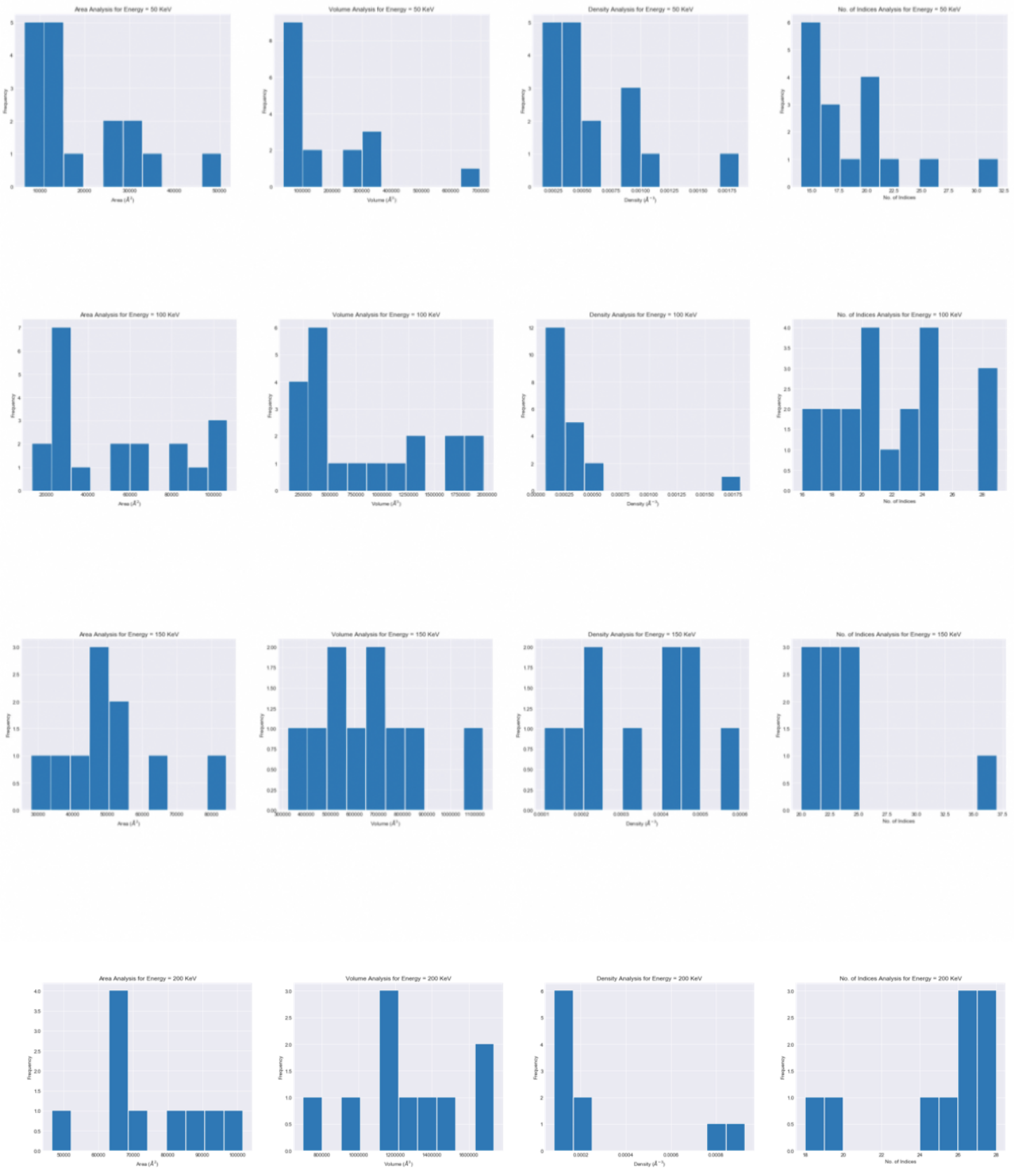



*Figure 6: Histogram plots for convex hull features like area, volume, density and no. of indices according to different energies available in the data.*

**********PLOTS FOR THE WHOLE DATA ACCORDING TO ELEMENTS**********

Fe PLOTS

W PLOTS

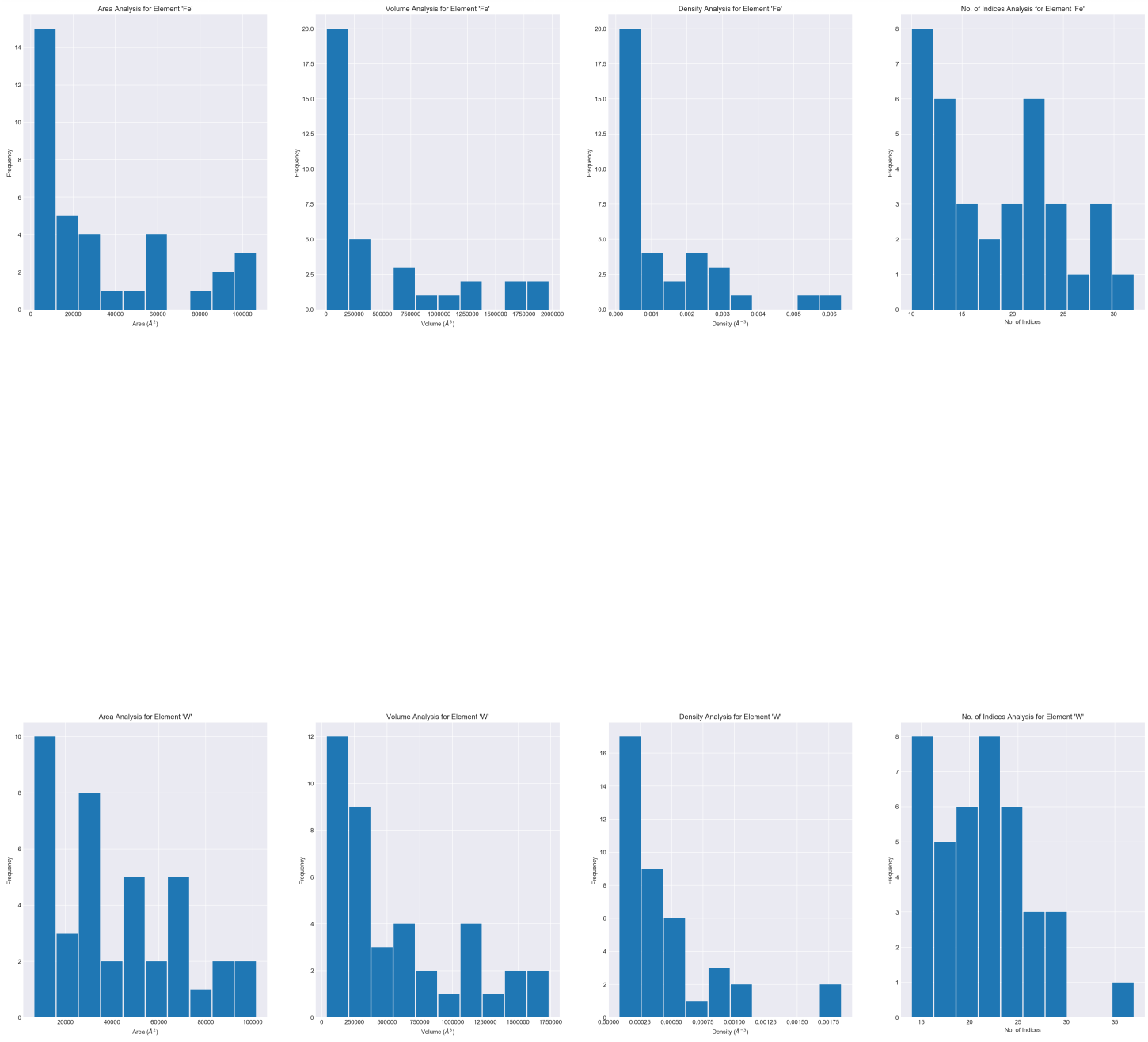



*Figure 7: Histogram plots for convex hull features like area, volume, density and no. of indices according to different substrates present in the data.*

**********PLOTS FOR THE WHOLE DATA ACCORDING TO ELEMENTS -> ENERGIES (KeV) {EACH IN DIFFERENT COLOR}**********

AREA, VOLUME, DENSITY & NO_OF_INDICES PLOTS ACCORDING TO ELEMENTS -> ENERGIES (KeV) {EACH IN DIFFERENT COLOR}

['Fe', 'W'] -> [[10, 20, 50, 100], [50, 100, 150, 200]]

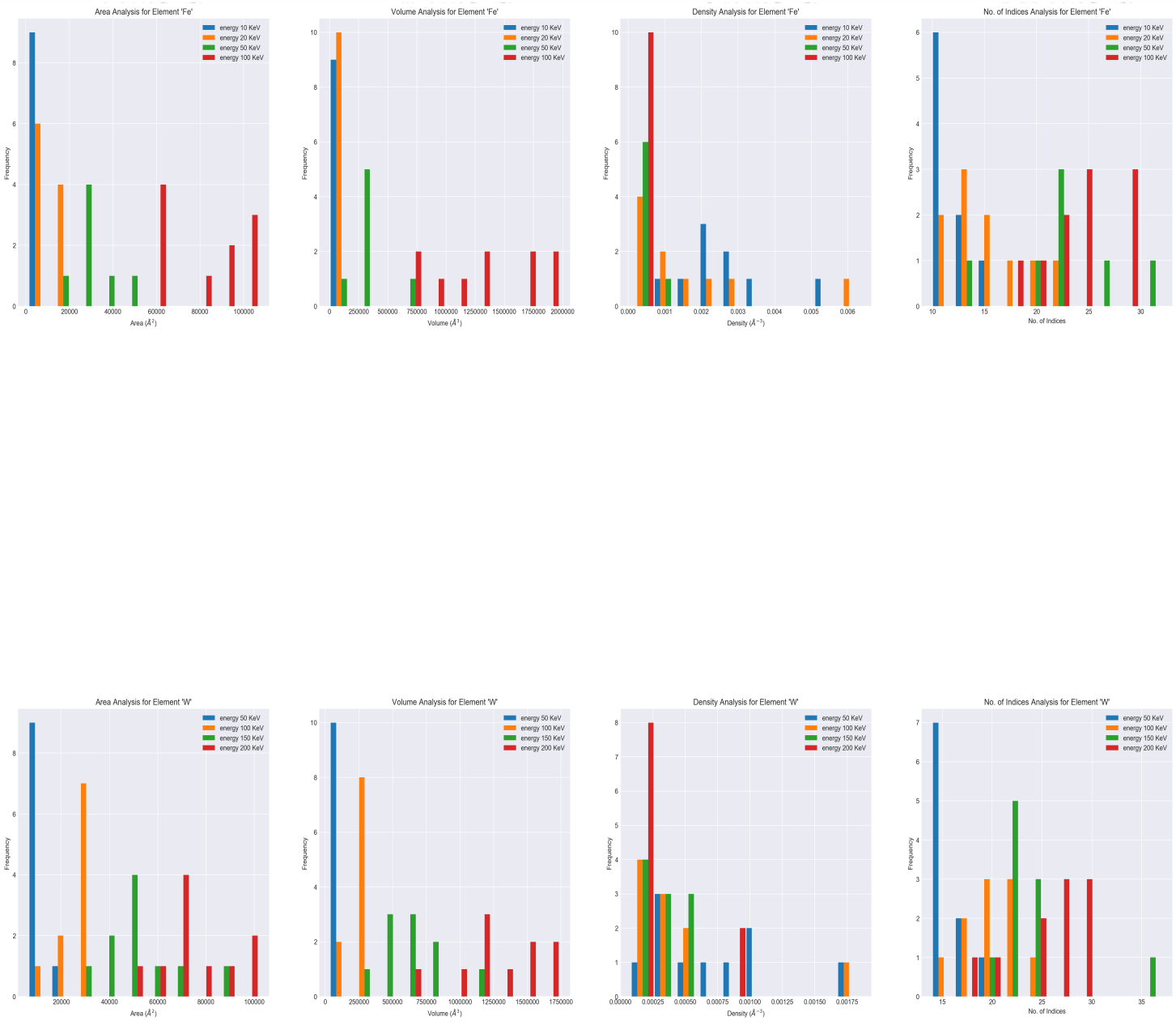



*Figure 8: Histogram plot for convex hull features like area, volume, density and no. of indices according to substrates present in the data which are again according to their respective energies available in the data, each represented by a different colour.*

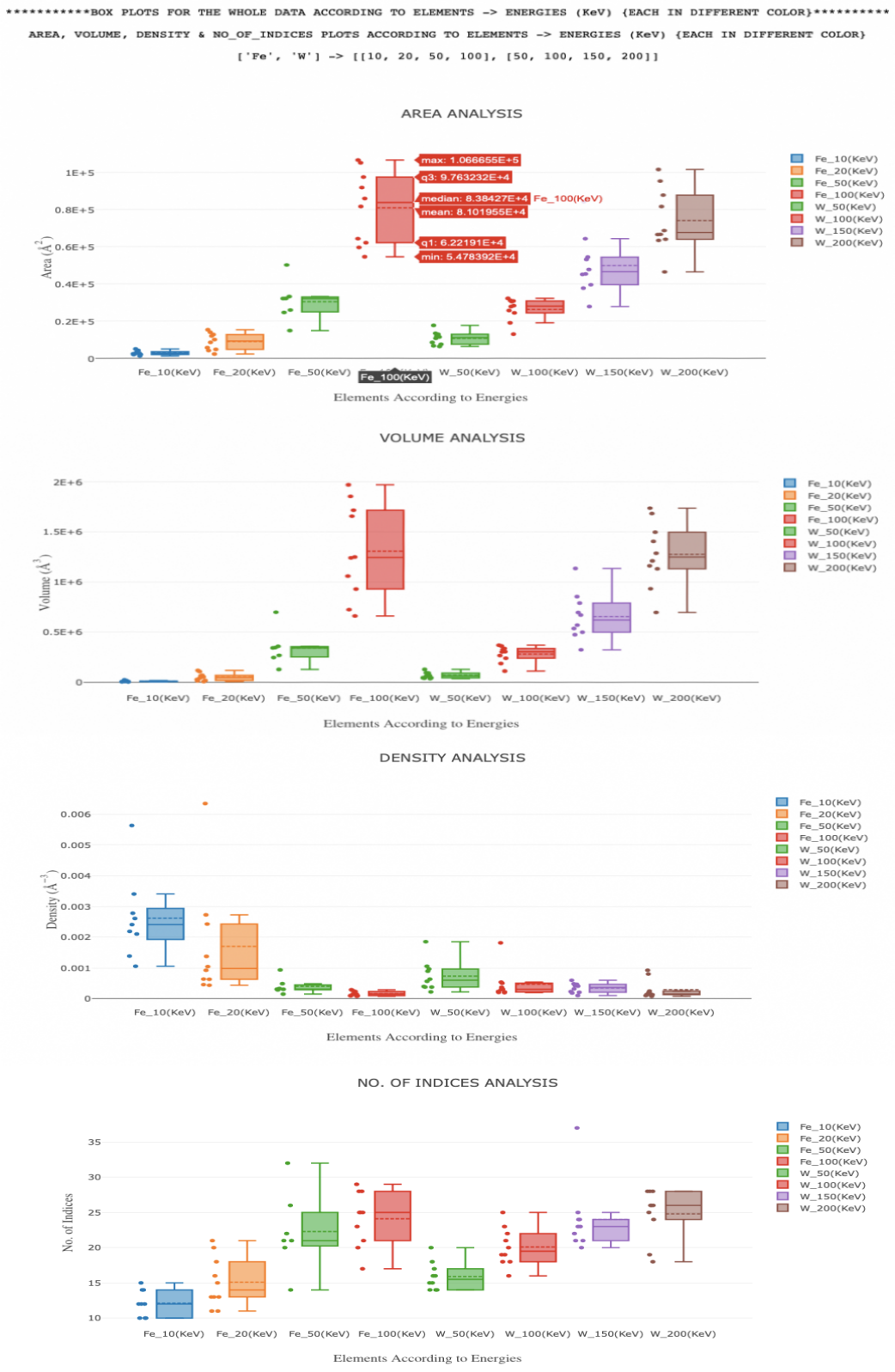
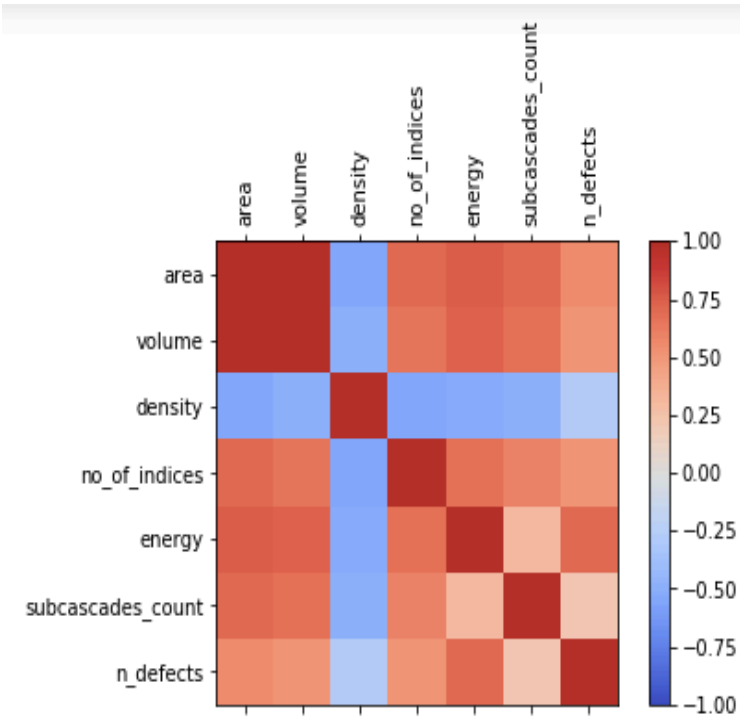



*Figure 9: Box plots for convex hull features like area, volume, density and no. of indices according to substrates and their respective energies available in the data, with each energy being represented by different colour.*



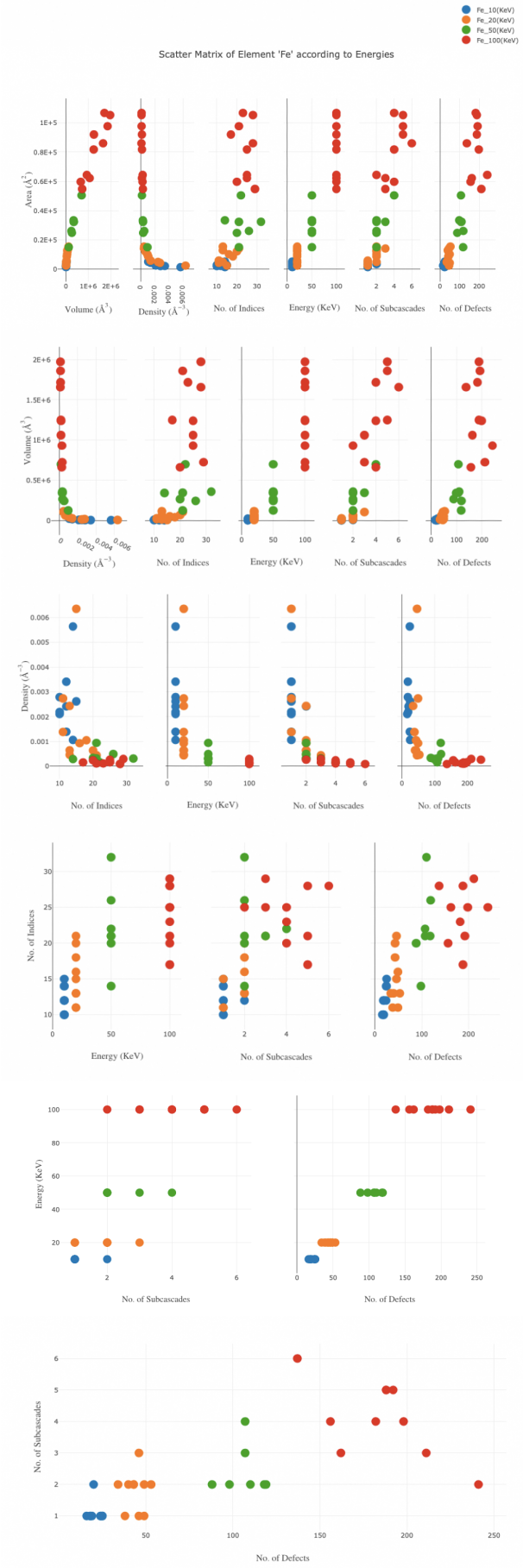

| | area | volume | density | no_of_indices | energy | subcascades_count | n_defects |
|---|---|---|---|---|---|---|---|
| **area** | 1.000000 | 0.977643 | -0.546015 | 0.703123 | 0.752263 | 0.704934 | 0.547059 |
| **volume** | 0.977643 | 1.000000 | -0.477247 | 0.656050 | 0.726845 | 0.670735 | 0.508359 |
| **density** | -0.546015 | -0.477247 | 1.000000 | -0.535701 | -0.507754 | -0.482769 | -0.274268 |
| **no_of_indices** | 0.703123 | 0.656050 | -0.535701 | 1.000000 | 0.679026 | 0.593027 | 0.522234 |
| **energy** | 0.752263 | 0.726845 | -0.507754 | 0.679026 | 1.000000 | 0.305458 | 0.706108 |
| **subcascades_count** | 0.704934 | 0.670735 | -0.482769 | 0.593027 | 0.305458 | 1.000000 | 0.215770 |
| **n_defects** | 0.547059 | 0.508359 | -0.274268 | 0.522234 | 0.706108 | 0.215770 | 1.000000 |

*Figure 10: Correlation matrix and table for data consisting of important convex hull features like area, volume, density & no. of indices and cascades' attributes like energy, sub-cascades count & no. of defects.*


*Figure 11: Scatter matrix plot of the data according to energies available, each represented by different colour.*

**********SCATTER MATRIX PLOT OF THE DATA ACCORDING ELEMENTS --> ENERGIES (KeV) {EACH IN DIFFERENT COLOR}**********

['Fe', 'W'] --> [[10, 20, 50, 100], [50, 100, 150, 200]]



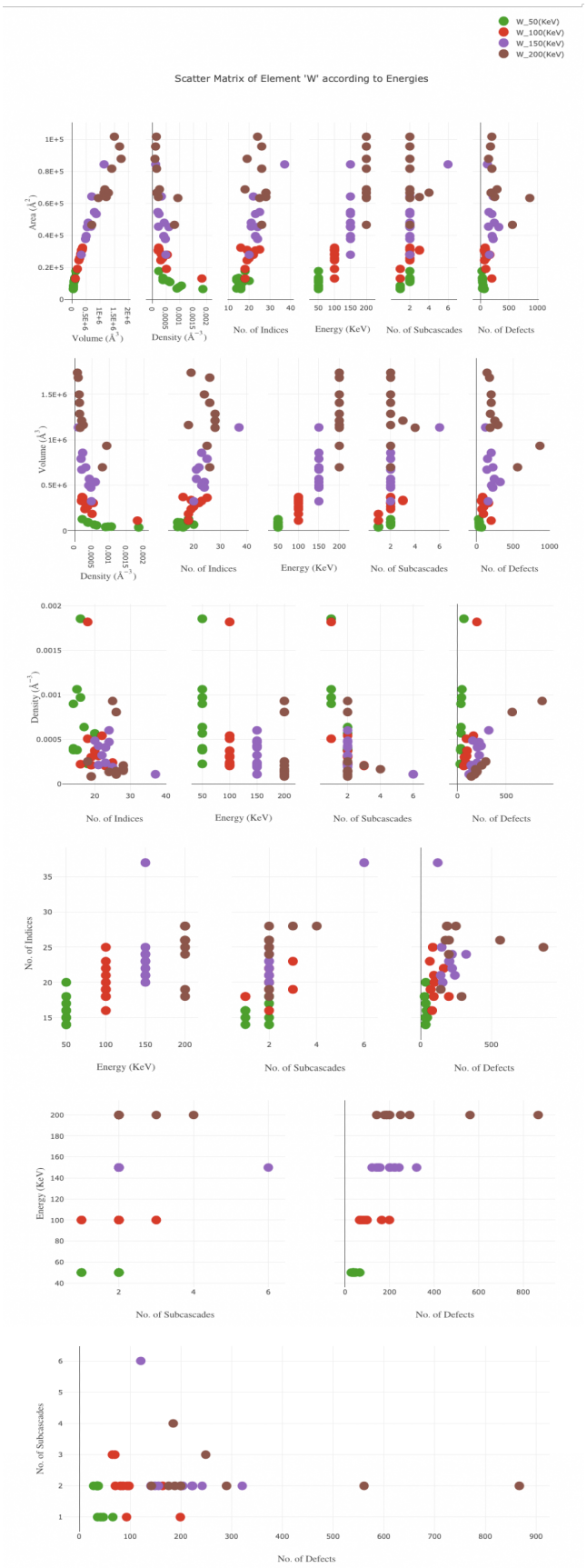
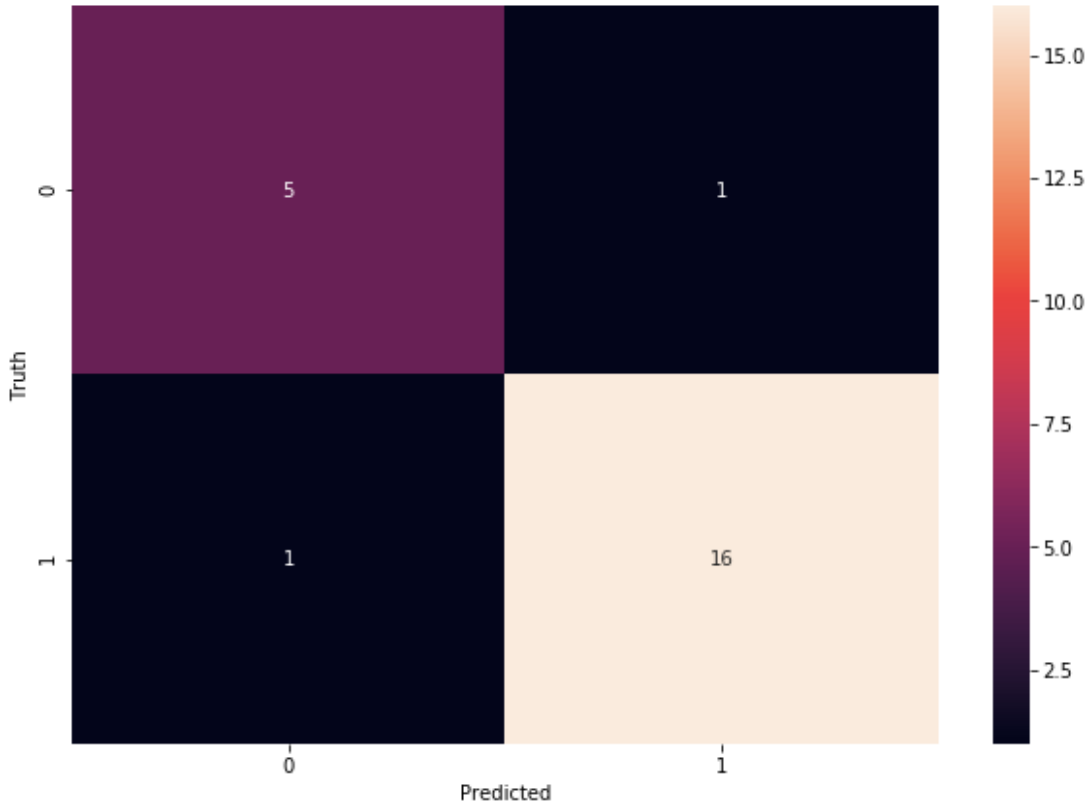



*Figure 12: Scatter matrix plot of the data according to substrates present which in-turn according to energies available respectively, each being represented by different colour.*

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.83 | 0.83 | 6 |
| 1 | 0.94 | 0.94 | 0.94 | 17 |
| micro avg | 0.91 | 0.91 | 0.91 | 23 |
| macro avg | 0.89 | 0.89 | 0.89 | 23 |
| weighted avg | 0.91 | 0.91 | 0.91 | 23 |

***CONFUSION MATRIX***




***ACCURACY***

91.30434782608695

***TRAINING ACCURACY***

92.45283018867924

*Figure 13: Logistic Regression - Binary Classification on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) with test-size = 0.30.*

Logistic regression with binary classification predicts whether a cascade contains multiple sub-cascades (i.e., > 1 sub-cascades) or not (i.e., single cascade). Our model's accuracy is 91.30% when the test-size is 30% (which is usual as, we need to train the model with as much of the data as possible i.e., 70% of the whole data). The training accuracy of our model is 92.45% which portraits that our model was neither over-fit nor under-fit but, is best-fit.

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as as the ratio of true positives to the sum of true and false positives. Said another way, "for all instances classified positive, what percent was correct?"

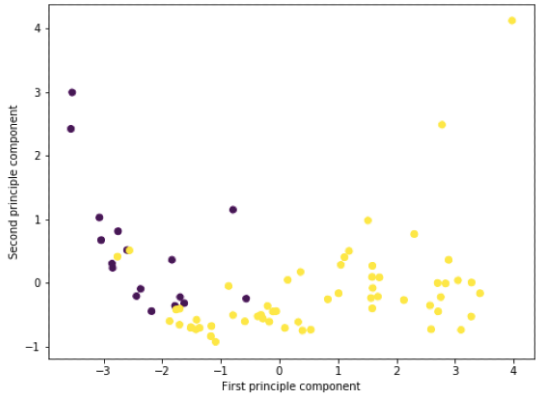
Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. Said another way, "for all instances that were actually positive, what percent was classified correctly?"

f1 score

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

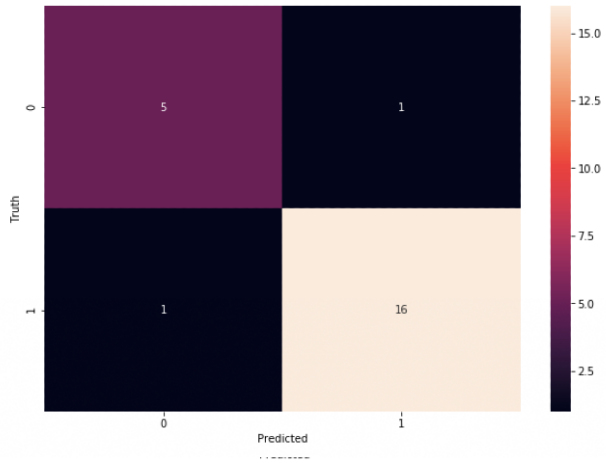
```
Shape of Independent variable
(76, 6)

Explained Variance Ratio
[0.68238084 0.12602696]

Components of PCA
[[ 0.45998138  0.44366199 -0.32413717  0.41074027  0.43983782  0.35263492]
 [-0.07450007 -0.04346835  0.71217395 -0.09716709  0.21064695  0.6569284 ]]

Shape of Scaled Independent variable
(76, 6)

Shape of PCA Independent variable
(76, 2)
```




***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.83 | 0.83 | 6 |
| 1 | 0.94 | 0.94 | 0.94 | 17 |
| micro avg | 0.91 | 0.91 | 0.91 | 23 |
| macro avg | 0.89 | 0.89 | 0.89 | 23 |
| weighted avg | 0.91 | 0.91 | 0.91 | 23 |

***CONFUSION MATRIX***




***ACCURACY***

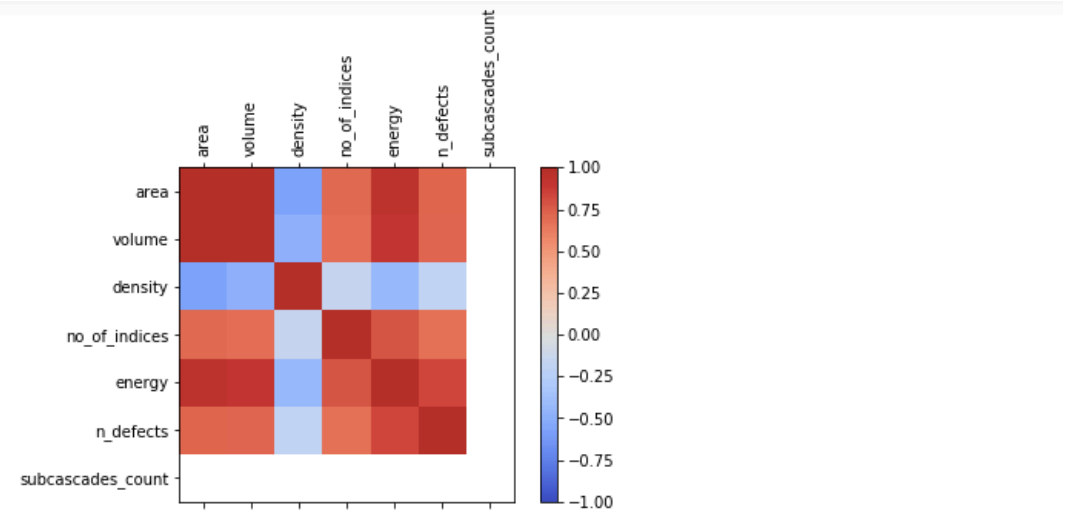
91.30434782608695

***TRAINING ACCURACY***

92.45283018867924

*Figure 14: Logistic Regression - Binary Classification with Principal Component Analysis (PCA) on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) with test-size = 0.30.*

Our model displays accuracy of 91.30% which is same as that of Logistic Regression - Binary Classification without Principal Component Analysis (PCA) as, the no. of features used are very less (i.e., just 6) for dimensionality reduction. Usually, works best when there are more no. of features.




| | area | volume | density | no_of_indices | energy | n_defects | subcascades_count |
|---|---|---|---|---|---|---|---|
| **area** | 1.000000 | 0.980982 | -0.576775 | 0.707838 | 0.934325 | 0.719133 | NaN |
| **volume** | 0.980982 | 1.000000 | -0.474420 | 0.690322 | 0.913125 | 0.713844 | NaN |
| **density** | -0.576775 | -0.474420 | 1.000000 | -0.135837 | -0.431661 | -0.184576 | NaN |
| **no_of_indices** | 0.707838 | 0.690322 | -0.135837 | 1.000000 | 0.785424 | 0.671533 | NaN |
| **energy** | 0.934325 | 0.913125 | -0.431661 | 0.785424 | 1.000000 | 0.843242 | NaN |
| **n_defects** | 0.719133 | 0.713844 | -0.184576 | 0.671533 | 0.843242 | 1.000000 | NaN |
| **subcascades_count** | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

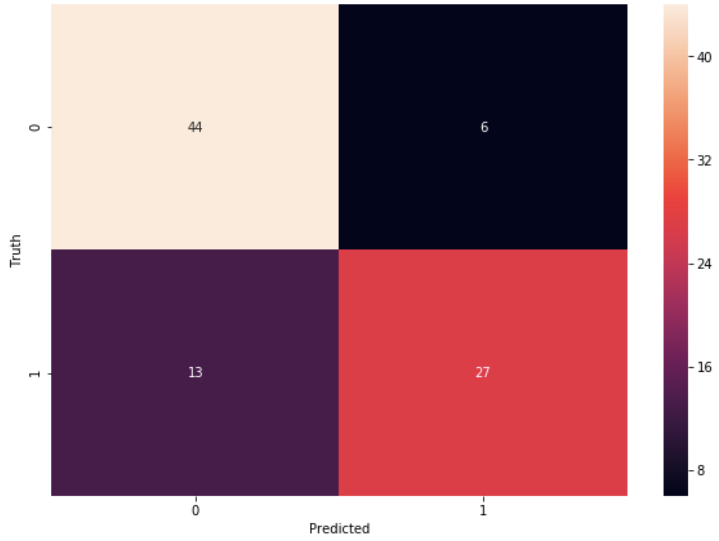
*Figure 15: Correlation matrix and table for the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) which contains cascades with only single cascade.*

The data consisting of cascades with multiple sub-cascades are removed (i.e. only cascades with sub-cascade 1 are left. Hence, the 'subcascades_count' column shows NaN value *("Not a Number")* means 0/0 as, random variable 'subcascades_count' displays zero variance). The table is just for a better understanding of correlations among the characteristic features when a cascade contains single sub-cascade.

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.88 | 0.82 | 50 |
| 1 | 0.82 | 0.68 | 0.74 | 40 |
| micro avg | 0.79 | 0.79 | 0.79 | 90 |
| macro avg | 0.80 | 0.78 | 0.78 | 90 |
| weighted avg | 0.79 | 0.79 | 0.79 | 90 |

***CONFUSION MATRIX***




***ACCURACY***

78.88888888888889

***TRAINING ACCURACY***

90.0

***BEST ACCURACY***

87.4074074074074

***TRAINING ACCURACY OF THE BEST ACCURACY***

86.66666666666667

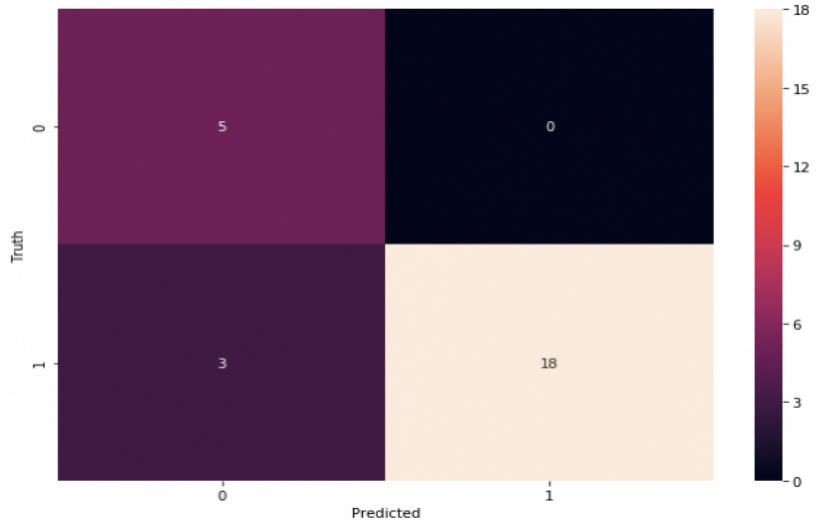
*Figure 16: Logistic Regression - Binary Classification on the data (consisting of non-annihilated vacancy defects from 300 cascades of Fe at 5, 10, 20 KeV) with test-size = 0.30 & 0.90.*

Our model's accuracy and training accuracy is 78.89% and 90% respectively, when the test-size is 30%. Surely, the model overfits but this can be turned to best-fit by increasing the test-size as the no. of samples are more (i.e., 300 cascades). At test-size 0.90 it yields the best accuracy and corresponding training accuracy to be 87.41% & 86.67% respectively, displaying best-fit. With PCA, the accuracy reduced to 72.22% and corresponding training accuracy is 86.68% at test-size 0.30.

**CLASSIFICATION REPORT**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.62 | 1.00 | 0.77 | 5 |
| 1 | 1.00 | 0.86 | 0.92 | 21 |
| micro avg | 0.88 | 0.88 | 0.88 | 26 |
| macro avg | 0.81 | 0.93 | 0.85 | 26 |
| weighted avg | 0.93 | 0.88 | 0.89 | 26 |

**CONFUSION MATRIX**




**ACCURACY**

88.46153846153845

**TRAINING ACCURACY**

87.03703703703704

*Figure 17: Logistic Regression - Binary Classification, training on data-1 (consisting of non-annihilated vacancy defects from 300 cascades of Fe at 5, 10, 20 KeV) and testing on data-2 (consisting of non-annihilated vacancy defects from 36 cascades of Fe at 10, 20, 50, 100 KeV) with default test-size for data-1 to be 0.10 and test-size for data-2 = 0.70.*

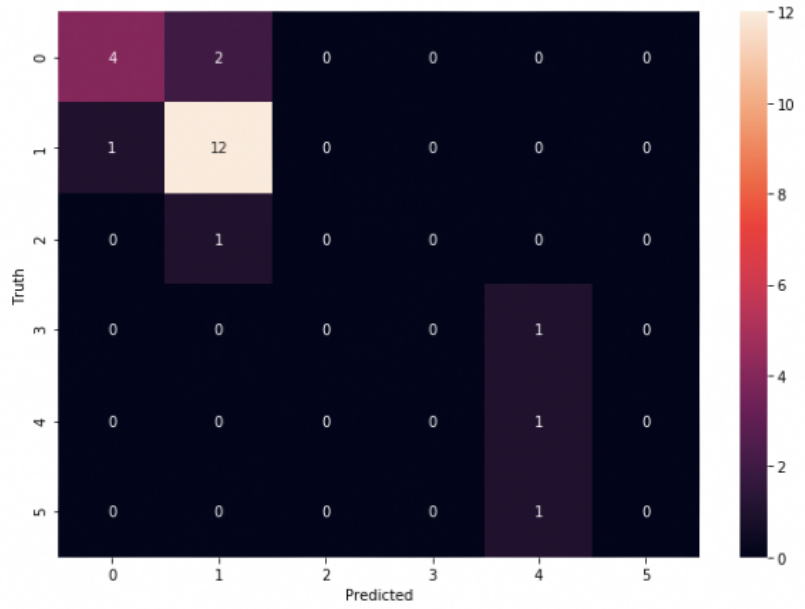
Data-2 is derived from the data (i.e., original data or data used still now ) consisting of 76 cascades of Fe [36 cascades] & W [40 cascades] at 10, 20, 50, 100, 150, 200 KeV ([Fe at 10, 20, 50, 100 KeV] & [W at 50, 100 150, 200 KeV]). The default test-size for data-1 is taken to be 0.10 which implies that the train size of data-1 is 90% (We will be training the model with most of the given data-1). The test-size for data-2 is taken as 0.70 which means that the model will be tested on most of the data-2 (i.e., there will be very good no. of samples to predict).

Our model's accuracy is 88.46 when the test-size is 70%. The training accuracy is 87.04 which portraits that our model was neither over-fit nor under-fit but, is best-fit. With PCA, the accuracy and training accuracy is 65.38% and 81.48% respectively, at test-size 0.70.

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.80 | 0.67 | 0.73 | 6 |
| 2 | 0.80 | 0.92 | 0.86 | 13 |
| 3 | 0.00 | 0.00 | 0.00 | 1 |
| 4 | 0.00 | 0.00 | 0.00 | 1 |
| 5 | 0.33 | 1.00 | 0.50 | 1 |
| 6 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.74 | 0.74 | 0.74 | 23 |
| macro avg | 0.32 | 0.43 | 0.35 | 23 |
| weighted avg | 0.68 | 0.74 | 0.70 | 23 |

***CONFUSION MATRIX***




***ACCURACY***

73.91304347826086

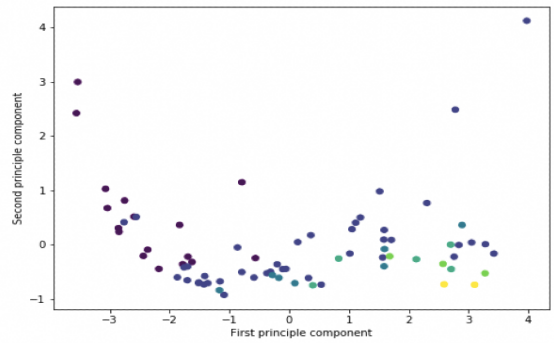
***TRAINING ACCURACY***

69.81132075471697

*Figure 18: Logistic Regression - Multi Classification on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) with test-size = 0.30.*

Logistic regression with multiple classification predicts the number of sub-cascades a cascade contains(i.e., 0, 1, 2,… sub-cascade/s). Our model's accuracy is 73.91 when the test-size is 30%. The training accuracy of our model is 69.81 which shows that our model is under-fit. But this is due to less no. of cascades (i.e., only 76 cascades) present in the data. Hence, there is very less data for training the model.

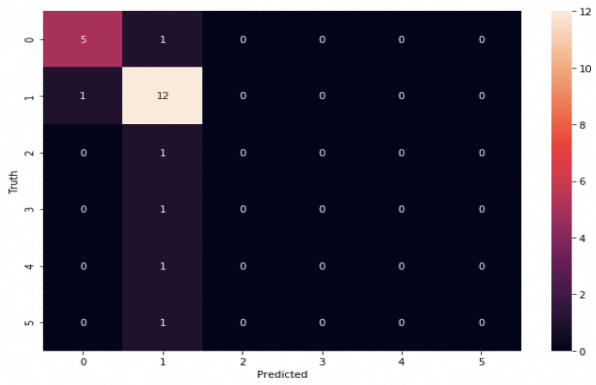
```
Shape of Independent variable
(76, 6)

Explained Variance Ratio
[0.68238084 0.12602696]

Components of PCA
[[ 0.45998138  0.44366199 -0.32413717  0.41074027  0.43983782  0.35263492]
 [-0.07450007 -0.04346835  0.71217395 -0.09716709  0.21064695  0.6569284 ]]

Shape of Scaled Independent variable
(76, 6)

Shape of PCA Independent variable
(76, 2)
```




***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.83 | 0.83 | 0.83 | 6 |
| 2 | 0.71 | 0.92 | 0.80 | 13 |
| 3 | 0.00 | 0.00 | 0.00 | 1 |
| 4 | 0.00 | 0.00 | 0.00 | 1 |
| 5 | 0.00 | 0.00 | 0.00 | 1 |
| 6 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.74 | 0.74 | 0.74 | 23 |
| macro avg | 0.26 | 0.29 | 0.27 | 23 |
| weighted avg | 0.62 | 0.74 | 0.67 | 23 |

***CONFUSION MATRIX***




***ACCURACY***

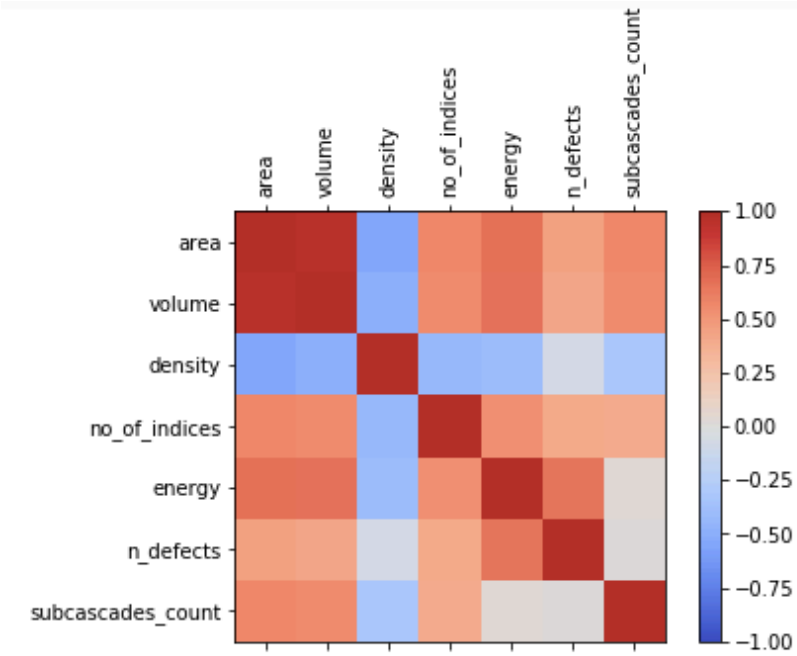
```
73.91304347826086
```

***TRAINING ACCURACY***

```
67.9245283018868
```

*Figure 19: Logistic Regression - Multi Classification with Principal Component Analysis (PCA) on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) with test-size = 0.30.*

The model's accuracy is 73.91 when the test-size is 30% which is same as the previous model (i.e., MLR without PCA). This is due to very less (i.e., just 6) no. of features used for dimensionality reduction. But the training accuracy of our model is 67.92 which is a bit less and better than (as it should close to accuracy) than that of the previous model.




| | area | volume | density | no_of_indices | energy | n_defects | subcascades_count |
|---|---|---|---|---|---|---|---|
| **area** | 1.000000 | 0.975621 | -0.540672 | 0.575295 | 0.673061 | 0.453780 | 0.576091 |
| **volume** | 0.975621 | 1.000000 | -0.497519 | 0.547718 | 0.661323 | 0.419290 | 0.558304 |
| **density** | -0.540672 | -0.497519 | 1.000000 | -0.439419 | -0.405401 | -0.058125 | -0.327996 |
| **no_of_indices** | 0.575295 | 0.547718 | -0.439419 | 1.000000 | 0.541972 | 0.408883 | 0.399993 |
| **energy** | 0.673061 | 0.661323 | -0.405401 | 0.541972 | 1.000000 | 0.649375 | 0.041466 |
| **n_defects** | 0.453780 | 0.419290 | -0.058125 | 0.408883 | 0.649375 | 1.000000 | 0.020970 |
| **subcascades_count** | 0.576091 | 0.558304 | -0.327996 | 0.399993 | 0.041466 | 0.020970 | 1.000000 |

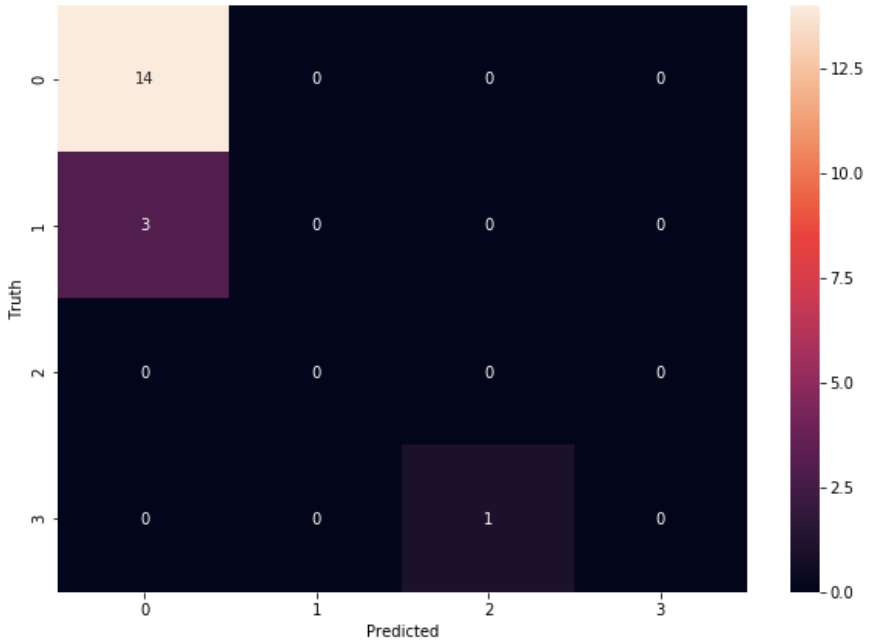
*Figure 20: Correlation matrix and table for the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV) which contains cascades with only multiple sub-cascades (i.e., only cascades with sub-cascades > 1).*

The data consisting of cascades with single (or even 0) sub-cascades are deleted. The table is just for a better understanding of correlations among the characteristic features when a cascade contains multiple sub-cascades.

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 2 | 0.82 | 1.00 | 0.90 | 14 |
| 3 | 0.00 | 0.00 | 0.00 | 3 |
| 4 | 0.00 | 0.00 | 0.00 | 0 |
| 6 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.78 | 0.78 | 0.78 | 18 |
| macro avg | 0.21 | 0.25 | 0.23 | 18 |
| weighted avg | 0.64 | 0.78 | 0.70 | 18 |

***CONFUSION MATRIX***




***ACCURACY***

77.77777777777779

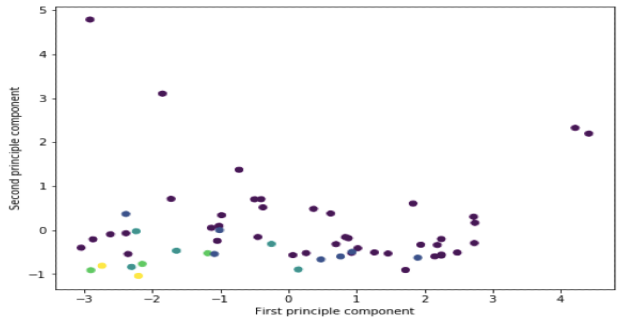
***TRAINING ACCURACY***

80.48780487804879

*Figure 21: Logistic Regression - Multi Classification on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV, containing only multiple sub-cascades i.e. > 1 sub-cascades) with test-size = 0.30.*

The model's accuracy and training accuracy is 77.78 and 80.49 respectively when the test-size is 30%. The model is best-fit and also better than previously discussed model which predicts no. of sub-cascades (i.e., whether 0, 1, 2, … sub-cascades present in a cascade).

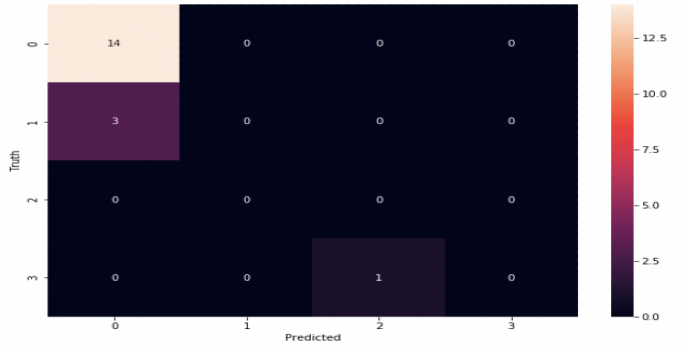
```
Shape of Independent variable
(59, 6)

Explained Variance Ratio
[0.61623414 0.16766667]

Components of PCA
[[-0.48020897 -0.46910644  0.31875188 -0.38698355 -0.43869714 -0.32484502]
 [-0.1297196  -0.12814874  0.64099355 -0.03627692  0.26005549  0.69780594]]

Shape of Scaled Independent variable
(59, 6)

Shape of PCA Independent variable
(59, 2)
```




***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 2 | 0.82 | 1.00 | 0.90 | 14 |
| 3 | 0.00 | 0.00 | 0.00 | 3 |
| 4 | 0.00 | 0.00 | 0.00 | 0 |
| 6 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.78 | 0.78 | 0.78 | 18 |
| macro avg | 0.21 | 0.25 | 0.23 | 18 |
| weighted avg | 0.64 | 0.78 | 0.70 | 18 |

***CONFUSION MATRIX***




***ACCURACY***

77.77777777777779

***TRAINING ACCURACY***

70.73170731707317

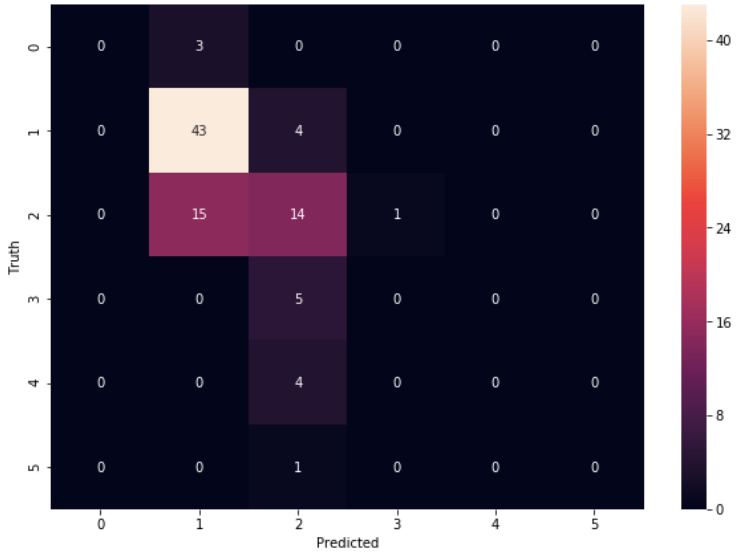
*Figure 22: Logistic Regression - Multi Classification with Principal Component Analysis (PCA) on the data (consisting of non-annihilated vacancy defects from 76 cascades of Fe, W at 10, 20, 50, 100, 150, 200 KeV, containing only multiple sub-cascades i.e. > 1 sub-cascades) with test-size = 0.30.*

The model's accuracy is 77.78 when the test-size is 30% which is same as the previous model (i.e., MLR without PCA). This is due to very less (i.e., just 6) no. of features used for dimensionality reduction. But the training accuracy of our model is 70.73 which less than accuracy and hence shows that the model somewhat falls under the category of under-fit. But this is due to less no. of cascades (i.e., only 76 cascades) present in the data. Hence, there is very less data for training the model.

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 3 |
| 1 | 0.70 | 0.91 | 0.80 | 47 |
| 2 | 0.50 | 0.47 | 0.48 | 30 |
| 3 | 0.00 | 0.00 | 0.00 | 5 |
| 4 | 0.00 | 0.00 | 0.00 | 4 |
| 5 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.63 | 0.63 | 0.63 | 90 |
| macro avg | 0.20 | 0.23 | 0.21 | 90 |
| weighted avg | 0.53 | 0.63 | 0.58 | 90 |

***CONFUSION MATRIX***




***ACCURACY***

63.33333333333333

***TRAINING ACCURACY***

80.47619047619048

***BEST ACCURACY***

75.43859649122807

***TRAINING ACCURACY OF THE BEST ACCURACY***

77.77777777777779

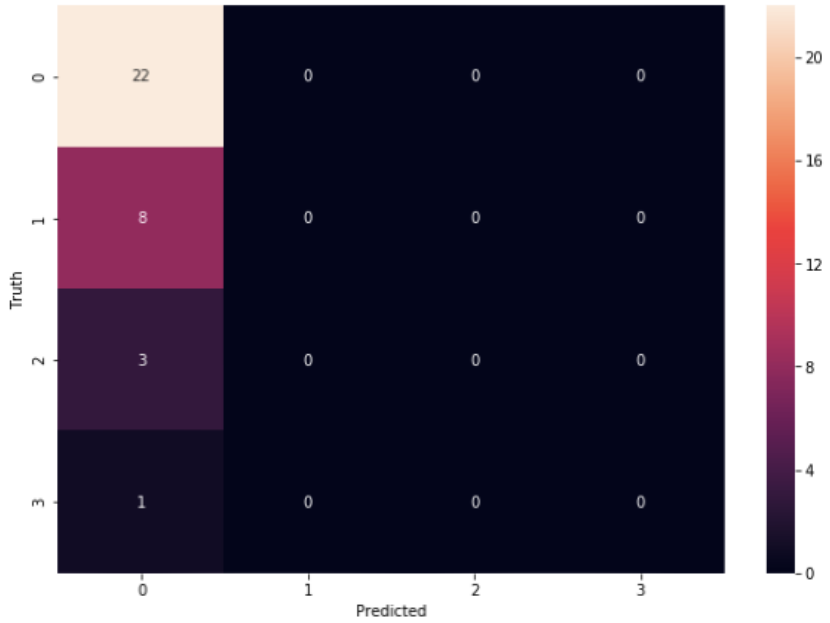
*Figure 23: Logistic Regression - Multi Classification on the data (consisting of non-annihilated vacancy defects from 300 cascades of Fe at 5, 10, 20 KeV) with test-size = 0.30 and 0.76.*

The model's accuracy and training accuracy is 63.33 and 80.48 respectively, when the test-size is 30%. The training accuracy of our model shows that our model is over-fit. But the overfitting can be decreased by increasing the test size as the no. of samples available in the data are more (i.e., by testing on more data the model performs its best). By doing so, the model fits to its best. With PCA, the accuracy and training accuracy at test-size 0.30 is 57.78% and 78.09%.

***CLASSIFICATION REPORT***

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 2 | 0.65 | 1.00 | 0.79 | 22 |
| 3 | 0.00 | 0.00 | 0.00 | 8 |
| 4 | 0.00 | 0.00 | 0.00 | 3 |
| 5 | 0.00 | 0.00 | 0.00 | 1 |
| micro avg | 0.65 | 0.65 | 0.65 | 34 |
| macro avg | 0.16 | 0.25 | 0.20 | 34 |
| weighted avg | 0.42 | 0.65 | 0.51 | 34 |

***CONFUSION MATRIX***




***ACCURACY***

64.70588235294117

***TRAINING ACCURACY***

81.0126582278481

***BEST ACCURACY***

72.09302325581395

***TRAINING ACCURACY OF THE BEST ACCURACY***

85.18518518518519

*Figure 24: Logistic Regression - Multi Classification on the data (consisting of non-annihilated vacancy defects from 300 cascades of Fe at 5, 10, 20 KeV, containing only multiple sub-cascades i.e. > 1 sub-cascades) with test-size = 0.30 and 0.76.*

The accuracy and training accuracy of the model is 64.70% and 81.01% respectively, when the test-size is 0.30. The model clearly over-fits but is better than previously discussed

model which predicts no. of sub-cascades (i.e., whether 0, 1, 2, … sub-cascades present in a cascade). But the overfitting can be controlled to a limit by increasing the test-size as there are more no. of samples (i.e., cascades) available in the data. The model displays best accuracy of 72.09% at test-size 0.76 with 85.18% as its corresponding training accuracy which is also better than previous test-size.

With PCA at test-size 0.30 the model gives an accuracy about 64.70% which is same as without PCA and training accuracy 79.75% which is better than that of without PCA. At test-size 0.80 the model's accuracy is 72.53% and training accuracy is 86.36%.

# 8. CONCLUSION

We have described a method to efficiently process and analyse the structures of cascades from MD simulations of collision cascades. We have discussed efficient geometric algorithms starting from representation of damaged areas in collision cascades, structural visualisation of convex hull to characterise damage areas with area, volume, density & structure and sub-cascades detection, to their classification. The results can be used to study and classify shape and provide structure based information of sub-cascades in collision cascades to higher scale models of radiation damage.

We have applied our methods and discussed elaborate results for collision cascades in Fe and W for a wide range of PKA energies. The correlations were helpful in visualising the relationships among the geometric features of collision cascades. The supervised classification algorithms predicts whether the collision cascades contain multiple sub-cascades or not. If yes it also estimates the number of sub-cascades with decent accuracy, provided few geometric features of collision cascades and geometric properties of the convex hull of the collision cascades. The classification can be used to quickly study the number of sub-cascades formed in new elements and energy ranges. A next step in a multi-scale model can be to further classify the sub-cascades into classes by shape and study them.

# 9. REFERENCES


1) Statistical study of defects caused by primary knock-on atoms in fcc Cu and bcc W using molecular dynamics by M. Warrier, U. Bhardwaj, H. Hemani, R. Schneider, A.Mutzke, M.C. Valsakumar,Journal of Nuclear Materials, Elsevier, Dec 2015.

2) The 5th Biennial Technical Meeting of the International Atomic and Molecular Code Centre Network on Molecular Dynamics Data of Collisional Cascades after Irradiation. [https://www-amdis.iaea.org/CCN/Meetings5/]

3) As a part of IAEA Challenge on Materials for Fusion (i.e. to study, explore and visualise collision cascades), CSaransh was used. [https://github.com/haptork/csaransh]