# User documentation – ClusterProcessor

# About

Cluster Processor is a set of tools for processing output of the TimePix3 detector. It was developed during the of created by TimePix3 is a hybrid pixel detector which is used in Large Hadron Collider for capturing the trajectories of the sets of elementary particles, also known as clusters. However, the amount of data generated in this proces is too large for manual processing, which creates the needs for automated tools which would help with processing and extracting information from the clusters. The extracted data cant hen be used for classification of the clusters.The applications in the ClusterProcessor are the folowing:

* **ClassifierForClusters** is a console application which runs a given trained classifier on a selected dataset.
* **ClusterFilter** serves forselecting clusters based on their attributes for further processing.
* **ClusterViewer** provides a graphic user interface which visualizes clusters. It allows for interactive cluster visualization (one by one).
* **Description Generator** is a tool for generating the attribute file for the collection of clusters. This output file can be consequently used for training the classifiers.
* **ClassifierUI** provides a graphical interface with two main functionalities. Firstly, users can train a new classifier or train an existing classifier. And secondly, user can merge single-level classifiers into a cascade classifier.
* **ClassifierTrainer** is a console application with the similar capabilities as ClassifierUI. It trains a classifier based on passed command line arguments.
* **ClusterExperiment** serves as a tool which performs classification experiments on a dataset of clusters. It trains multiple classifiers and evaluates their results.

In the following sections we will show how to install the set of tools and how to use each its individual applications.

# Installation

In order to run all of the primary applications in the ClusterProcessor solution, user has two options.

1. Download the archived build (for Windows x64) of the solution accessible on the link: <https://github.com/TomSpeedy/ClusterProcessor>. This method is suitable when you plan just to use the existing implementation.
2. Clone the solution from https://github.com/TomSpeedy/ClusterProcessor … This method is for those who woud like to modifiy any of the tools of ClusterProcessor and requires SDK for .Net version ….:

* For compiling some (or all?) of the tools, you will need the .Net Framework version 4.7. For further information about *.Net Framework,* visit <https://docs.microsoft.com/en-us/dotnet/framework/get-started/system-requirements>
* The library ClusterCalculator is downloaded (cloned). This library is a part of the solution Cluster Processor available at https://github.com/TomSpeedy/ClusterProcessor

### Data formats

The tools work with clusters that are traces of one particle with possible traces of its decay detected by TimePix3 as shown in Figure 3.1. Raw data from TimePix3 detector contains information on hits, energy, time of arrival and time over the threshold for each pixel. These pieces of information are grouped together by the Clusterer application []. Hence the base format of input is **MM cluster format.** This file format consists of 3 separate files – ini, cl, and px file.

|  |  |  |
| --- | --- | --- |
| **File** | **Format** | **Example** |
| ini | [Measurement (or any string ending with a newline char)] PxFile=[Relative path from the parent direcotry of .ini to px file]  ClFile=[Relative path from the parent direcotry of .ini to cl file] | Measurement 123 PxFile= Clusters\_px.txt ClFile= Clusters\_cl.txt |
| cl | [First ToA (decimal)] [Pixel Hit Count ( integer 0-232)] [LineOfStart in px file ( integer 0-232)] [Byte of start in px file (integer 0-232)] | 12345.647 100 5 30 |
| px | [x coordinate of the pixel ( integer 0-255)] [y coordinate of the pixel (integer 0-255)] [ToA (decimal)] [Energy (decimal)] | 123 128 15540 14.235 |

Another kind of data format we use in our applications is **JSON data format.** This format is used for these two purposes:

* capturing information about calculated attributes of a cluster, and
* storing multiple parameters for the classifier training.

The last data format we use in our applications are **csf** and **csf\_support**. While these are not standardized data formats, these contain information about the trained classifiers, so we can reconstruct them completely from these files.

# Cluster Viewer

***Setup and input***

**Requirements:**

* Classifier for Clusters project, which is part of the repository at <https://github.com/TomSpeedy/ClusterProcessor>

Third-party software for three-dimensional plotting Chart Director for .Net library accessible at [https://www.advsofteng.com/download.html](https://www.advsofteng.com/download.html%20)

**Application Dependencies:**

Table 3.1 Dependencies of the projects in the solution

Cluster Viewer

Cluster Calculator

Classifier

Trainer

ClassifierForClusters

Descr Generator

Cluster Filter

ClassifierUI

Cluster Experiment

The viewer enables to display clusters built out of raw data from TimePix3 by the program Clusterer. However, apart from MM data format, viewer also supports JSON data format, but only in the case, where clusters in this format have the following attributes: ClFile, PxFile and ClIndex.

## Loading cluster file

To view clusters, we run the viewer and either enter the path to our .ini or .json file or click the Browse button, and select the cluster file.

If the message “File was loaded successfully”appears, it means that our collection of clusters is now ready for viewing. Otherwise, a problem occurred during the loading of the file. The error is further specified in a message. Some of the common causes are:

* .ini file does not exist or is inaccessible
* .cl and .px files referenced by the .ini file do not exist or are inaccessible
* .cl, .px or .ini file is not in the correct format

## Browsing clusters

To display a particular cluster, …The image of the specified cluster displays the energy of each pixel logarithmically mapped to the color spectrum starting from white (for pixels with energy below 2eV) through yellow (for pixels with energy greater than 2eV and less than 15eV) to orange and red (pixels with energy up to 500eV). If the same pixel is hit twice in the cluster, the pixel with more energy is displayed – see Figure 3.1.



Pixel Energy around 3keV

Pixel Energy around 20keV

Energy more than 200keV

Figure 3.1 Displayed cluster

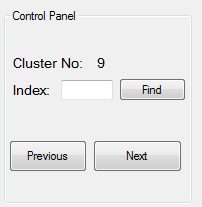


Figure 3.2 Panel for navigating through the collection

After the cluster collection is loaded, we can navigate through the collection using the Previousand Nextbuttons. To find the -th cluster (clusters are numbered from 1) in the collection, we can select the index of the cluster in the Control Panel box and click the Findbutton.

### Histograms

Apart from collection browsing, users can also view histograms. There are two histograms available in the viewer:

* Collection Histogram displays the histogram of the whole collection for the pixel count attribute.



Figure 3.3 Example of a cluster collection histogram

* Pixel Histogram represents the histogram of the pixels of the currently viewed cluster for the energy.

Figure 3.4 Example of a pixel histogram

Both histograms are calculated and displayed after clicking Show Collection Histogram (or Show Pixel Histogram).

### 3D visualization

Figure 3.5 Cluster trajectory visualization

Based on the time of arrival and callibration data, it is possible to reconstruct the trajectory of particles in the cluster. To get a better idea of how the trajectory of the cluster looked like, we can visualize the cluster in 3D. The visualization is based on the ToA of each pixel.

* **Viewing:** To view the three-dimensional image of the cluster user can click the View 3D button.
* **Rotation**: After the image is displayed, the user can rotate the image around the -axis and -axis by clicking the Up, Down, Left, and Right buttons.

### Cluster Attributes

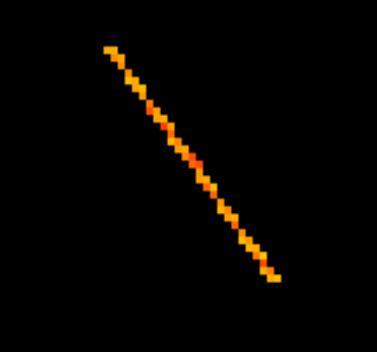
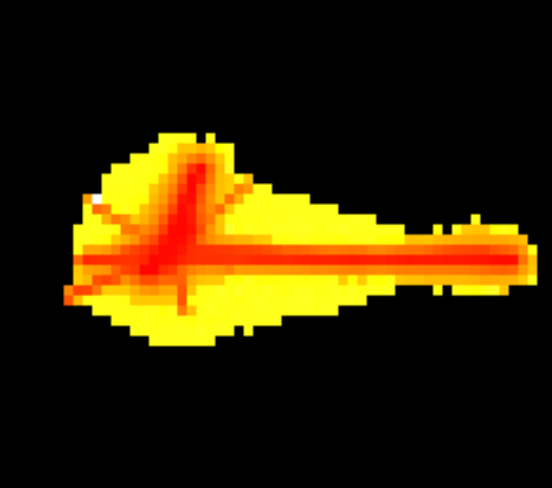
By clicking the button Show Attributes located in the Cluster Details section, various cluster properties are displayed. All these properties must be calculated in advance and stored in the cluster file (extension .cl). These attributes range from straightforward like Total Energy and Maximum Energy to the more sophisticated ones like Branch Count and Relative Halo Size.

Figure 3.6 Example of a cluster with high energy (red), non-trivial halo (yellow) and branch count

Figure 3.7 Example of a simple cluster with lower total energy. It contains only a single branch without a significant halo effect.

### Skeletonization

Skeletonization is often referred to as a thinning process, “transforms an input binary image into a skeleton by reducing the original image which contains different thicknesses to a thin representation (a set of curves and lines).“ Skeletonization can be used to remove the halo effect of the clusters while preserving the shape of the original cluster. During the skeletonization process, as the image is thinning, pixel energy is split among its neighbors. This means the total energy of the cluster and its skeletonized version is the same. To view the skeleton of the original cluster, we click Skeletonize.

Figure 3.8 Cluster after skeletonization

Figure 3.9 Cluster before skeletonization

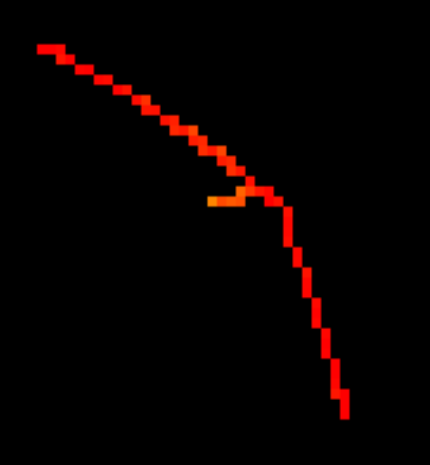


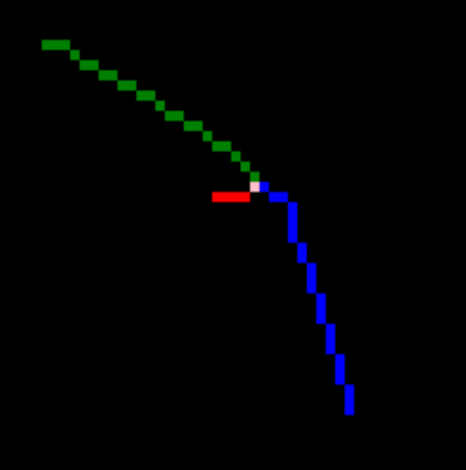
Figure 3.10 Example of a cluster before and after the skeletonization

Skeletonization

### Branch analysis

After finding the skeleton of the cluster, we can try to detect particle trajectories in the cluster. To do so, we can click the Show Branches button. The center point of the cluster is represented by the white dot, while the separate branches are denoted by distinct colors – blue, red, and green. Each branch can have its sub-branches – the starting point of the sub-branch is contained in its parent branch. The sub-branches are highlighted by a lighter shade of their parent branch color.

Branch 1



Branch 2

Branch 3

Center

Figure 3.11 Branches of a cluster

### Classification

Using all the features, we are able to calculate for the cluster, we can classify the cluster into various categories. This classification is done via machine learning using neural networks. The default classes of particles that are implemented in the default bestClassifier.csf classifier are:

* Lead,
* Iron,
* Helium,
* muon,
* electron,
* pion,
* low energy electron,
* and protons.

In the viewer application, a user can load a classifier by clicking Load Classifier and choosing the right trained classifier in .csf file. When the classifier is successfully loaded, it can be used for classification of the currently viewe cluster.

# Filter

The ClusterFilter application serves as a tool for picking the clusters that fit the given criteria. The user interface of the Cluster Filter application is very similar to the Cluster Viewer.

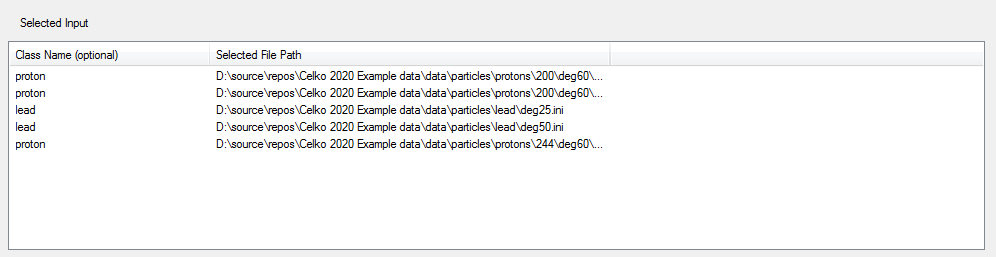
1. As the first step, we select the input .ini file. This can be done either by typing out the path to the file or clicking the Browse button.
2. Then, we choose the name for the output .ini file.
3. After that, we can select the properties to filter by and set the lower and upper bounds for the property. If no bound (upper or lower) is specified, the filter automatically sets the bound to the maximum (or minimum) possible value.
4. To start the processing, we click the Process button. During the filtering process, a new .cl file is created. The newly created file contains only those clusters, which fit the filtering criteria. However, no .px file is created because the output reuses the original pixel file to speed up the calculation and prevent unnecessary copying.
5. After the filtering is done, the filter displays the message “Filtering completed successfully”. If any other message is shown, it means there was an error. The message should provide more information about the problem. In most cases, it is either an incorrect data format or file inaccessibility.

# Description Generator

The Cluster Description Generator is a tool for calculating attributes for the whole collection of clusters at once. This can be useful, e.g., when creating training data for the machine learning algorithms. The collection of clusters can be composed of clusters from several source files. The application combines clusters from several sources and produces a collection of clusters (.ini, .cl and .px file) containing the selected clusters together with computed attributes. The datailed instructions how to use the tool follow:

1. We start by clicking the Browse and Add… button, where we select one or more .ini files and add them to the Selected Input collection.
2. To add more input files, we can repeat the process until all desired files are shown in the Selected Input box as shown in Figure 5.1.
3. To remove elements in the collection, we click the Remove Selected button.
4. For every single file we select, we can edit its Class Name column by triple-clicking – there can be multiple files containing the same class. This way, we set the name of the particle present in the input file, which can then be used for the supervised machine learning algorithms.

“Proton“ class partitions



“Lead“ class partitions

Figure 5.1 Selected partitions in Description Generator

1. Then we can choose the output file name in the Select output text field and also tick the attributes that will be calculated. To include the cluster class name as an attribute, remember to tick the Class attribute (which will not be actually calculated, it will just be copied from the provided class name).
2. After that, we can choose either the **even** distribution of each class in the output or the distribution **proportional** to the particular class size. Even distribution of the class mean that each class has the same chance to be processed as the next cluster. In constrast, proportional distribution means the chance of picking clusters from a class with smaller datasets is smaller while classes with bigger datasets are more likely to be picked.
3. For each file with given class (further referenced to as a class partition), we can choose whether we want to process those in **a serial order** (provided by the order on input – next partition is processed after the previous was already fully processed) or in **parallel order** (after a single cluster from current partition is processed, a cluster from the next partition is set to be processed next).
4. The user can also select the **ending condition** of the process. When this condition is satisfied, the program finishes the calculation – for a large number of clusters, the process may take several minutes to complete. There are three types of ending condition:
   1. Any partition is fully processed (**First partition**)
   2. Any class is fully processed (**First class**)
   3. All classes are fully processed (**Last class**)

By default, no cluster on the input will appear in the output more than once. When generating the data for imbalanced classes (there are huge differences in class sizes), this could lead to problems with machine learning the classification of the less frequent classes. To compensate for that, we can choose the Align class. By selecting the align class, any other class partition that is processed will not be discarded but will be processed repeatedly until the specified Align class is fully processed.

# Classifier

The ClassifierForClusters is a console application that provides an interface for the classification process of the selected clusters. The prerequisite for the classifier is to have Accord.net NuGet packages installed.

Syntax:

**ClassifierForClusters.exe [trained\_classifier.csf] [file\_to\_classify.json] [options]**

|  |  |
| --- | --- |
| Command line options | |
| **--simple** or **--multi** | Specification of the classifier type (single-level or multi-level), default is single |
| **--distr** | Print only key-value pairs of the class name and its frequency. |
| **--classes** | Split the clusters into separate JSON files according to the predicted class and also print key-value pairs of the frequencies to the console. In order to be able to view the results in the ClusterViewer , cluster must contain attributes ClFile, PxFile and ClIndex. |
| **--specials** | Split the clusters into separate JSON files (same as --classes), print frequencies, while also create a file for non-trivial unclassified clusters. |

Table 6.1 Options of ClassifierForClusters application

# ClassifierUI

Classifier UI is a tool that enables training classifier models based on the input parameters. User needs to set:

* **Training Json file** = training data in JSON format
* **Classifier config file** = parameters of the classifier in JSON format
* **Trained model** = If we want to continue in training with an existing model of single-level classifier, we fill in this field with its configuration .csf file.
* **Minimal Accuracy** = When this level of accuracy (or better) is reached on the validation set (the data which the classifier did not see during training), the classifier is stored into a file. It is a value between 0 and 1.
* **Maximal repetition count** = The maximal number of times we will try to train our classifier to reach the target minimal accuracy.
* **Seed** = Integer value, based on the value of the seed, data is split randomly into training and validation set (where 90% if the data is used for training and the remaining 10% for validation).

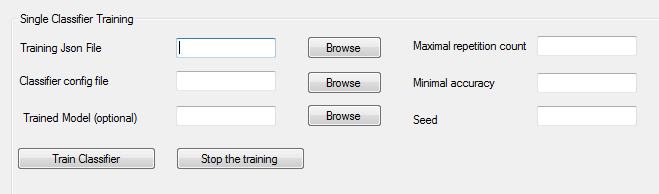


Figure 7.1 Classifier training dialog

The UI also provides an option to merge up to 4 simple trained classifiers into a multi-level classifier. User can choose the root classifier that will be applied first. The first classifier classifies the input clusters into several classes. One of these classes can be selected as „split class“. All clusters classified into the split class by the first classifier are fed into the classifier Trained Model Lv1 for further classification. Among the output classes of Trained Model Lv1 classifier, another split class can be selected and so on. That means, for combining classifiers, we need to fill in split classes because the bottom level classifier cannot have a split class. If you use only two classifiers (root and Lvl 1), make sure all the other fields for classifiers are empty.

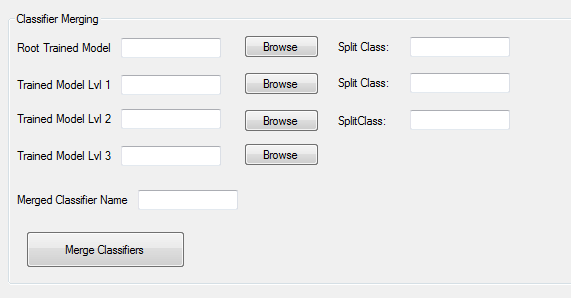


Figure 7.2 Classifier merging dialog

In the example below, we can see a multi-level classifier that uses three simple classifiers and two split classes.

Figure 7.3 Example of a multi-level classifier

# ClassifierTrainer

ClassifierTrainer is a console application that can train a classifier. This application can be used instead of ClassifierUI. Both tools can train the same classifier and differ only in their interfaces.

Syntax:

**ClassifierTrainer.exe [training\_data.json] [network\_parameters\_config.json] [options followed by their value (space separated)]**

Example

ClassifierTrainer.exe lead\_training data.json lead\_network\_config\_parameters.json --acc 0.8 --seed 42 --maxrep 5

|  |  |
| --- | --- |
| Command line options | |
| **--trained** | (Optional) Indicates existing .csf file should be trained again instead of creating a new classifier. A path to the trained model should follow this option. |
| **--acc** | Sets the target minimal accuracy of the classifier on the validation set to save the result. Must be followed by the accuracy value between 0 and 1. |
| **--seed** | Sets the seed, according to this seed, the data is split into the training and validation sets. It is followed by an integer. |
| **--maxrep** | Sets the maximal number of passes through the dataset. The training ends when either minimal accuracy on the validation set is achieved or the number of passes the through training set reached maxrep. |

Table 8.1 Command line parameters of the ClusterTrainer

# ClassifierExperiment

The ClassifierExperiment is a console application for performing experiments with the classifiers. It accepts only a single parameter – a path to the directory where the training and testing data, together with trained model directories are located. (Note: The data and models are not in the same folder, it is their parent directories which are in the specified folder).

Syntax:

**ClassifierExperiment.exe [path to the directory that contains the train\_data, test\_data and trained\_models directories]**

To verify that the classifier bestClassifier.csf was successfully loaded, it is tested on the test dataset and the result is displayed in the console. To check the performance of each single-level classifier we trained the following types of classifiers:

1. – classifies the particles intto lead particles, and the rest. This classifier was trained on roughly 80000 clusters.
2. – separates fragments, helium, and iron from each other and the rest. The dataset for this classifier consists of 450000 clusters.
3. – separates low energy electrons from protons and from the rest. The total number of clusters we used during the training is more than one million.
4. – splits clusters into 4 classes – electrons, muons, pions and the artificial class elPi0. It is trained on the dataset of approximately 500000 clusters.
5. – splits clusters into all categories directly (including the artificial category elPi0) and is trained on roughly 100000 clusters.

During the training proces user can see the training (mean squared) error after the given number of iterations. The experiments executed are the following:

1. **Evaluation of the best trained model on the test dataset** – this is displayed in a form of confusion matrix of classified particles. We decided to use this as a “sanity check“, to quickly find out whether the correct data directory was downloaded and if it has the expected structure (It is more of a quick check, rather that rigorous check of the whole dataset).
2. **Comparison of the accuracies of the classifier models with different learning parameters, by altering their default values diplayed in** Table 9.1**:**

|  |  |
| --- | --- |
| **validAttributes** | TotalEnergy, AverageEnergy, MaxEnergy, PixelCount, Convexity, Width, CrosspointCount, VertexCount, RelativeHaloSize, BranchCount, StdOfEnergy, StdOfArrival and RelLowEnergyPixels |
| **layerSizes** | [13, 13] |
| **learningAlgorithm** | backProp |
| **epochSize** | 8 |
| **learningRate** | 0.6 |
| **momentum** | 0.5 |
| **activationFunction** | sigmoid |

Table 9.1 Default model parameters

1. ***Lower learning rate and momentum*** 
   * learning rate: 0.1, and
   * momentum: 0.1.
2. ***Medium learning rate and momentum***
   * learning rate: 0.5, and
   * momentum: 0.5.
3. ***Higher learning rate and momentum***
   * learning rate: 1, and
   * momentum: 1.
4. ***Small number of hidden layers***
   * A single hidden layer with one neuron.
5. ***Medium number of hidden layers***
   * Two hidden layers with 13 neurons each.
6. ***Higher number of hidden layers***
   * Three hidden layers with 13 neurons each.
7. **Comparison of the single-level classifier accuracy with the multi-level classifier.** We trained 10 multi-level and 10 single-level classifiers (we verified the models were properly trained by observing the training error which was not decreasing) and compared their accuracies.
8. **Calculation of the k-fold cross-validation for each type of the single-level classifiers used in the bestClassifier.csf.** To check the performance of each single-level classifier we trained the following types of classifiers:

Then, we calculated their accuracy on a test fold (we performed a 6-fold validation)

Note: the execution of the tests could take a significant amount of time (around an hour for the first two tests each, the last test took less than 10 minutes) because of the large training datasets. The experiments were run on the desktop with the following parameters:

CPU: Intel i5-66000K, 3.5GHz,

GPU: NVIDIA GeForce GTX 1060

RAM: 8GB,

SSD: 256GB,

Address space: 64-bit,

OS: Windows 7.

# Demonstration

In order to guide the user through the whole ClusterProcessor solution, we decided to create a brief demonstration. User can go step by step and gain an intuition on how to use the application.

## Preparing the solution and the data

* + Download the binary files or clone the repository at <https://github.com/TomSpeedy/ClusterProcessor>. For further information see Chapter 2.
  + Download the sample dataset from <https://github.com/TomSpeedy/ClusterProcessorData>.
  + Observe the downloaded data file – it consists of 4 folders:
    1. **train\_data** are used for training of the neural network classifiers in JSON file format. These contain both the examples of clusters used for training as well as training parameters of the networks.
    2. **test\_data** are data in JSON file format used for testing of the classifiers trained on a different dataset (eg. **train\_data**).
    3. **demo\_data\_MM** represents the clusters in MM format, which are split into multiple directories based on their class (eg. proton, electron…).
    4. **trained\_models** contain the classifier models which were already trained and can be used for classification or (in case of single-level classifiers) for further training.

## Filtering

# When we have the data and applicaitons prepared, we can navigate to the folder with ClusterFilter Binary (/bin/Release). Its structure should match the one displayed in Figure 10.1.

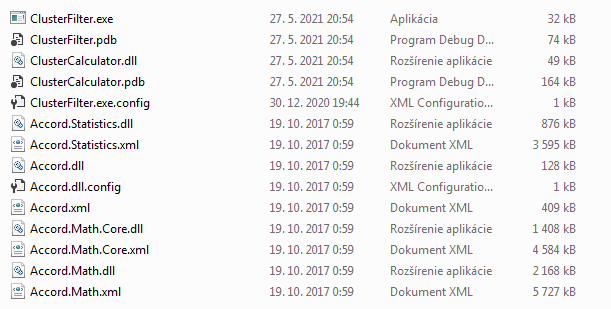
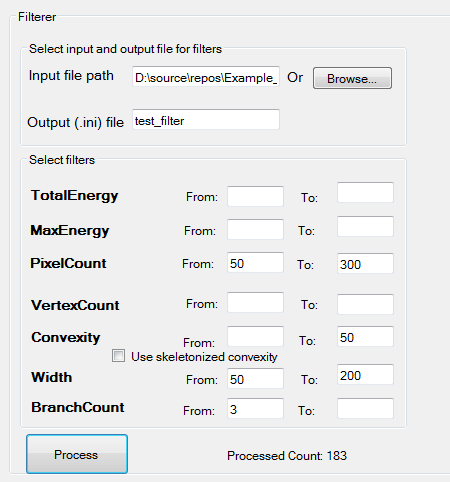
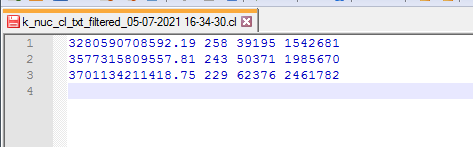


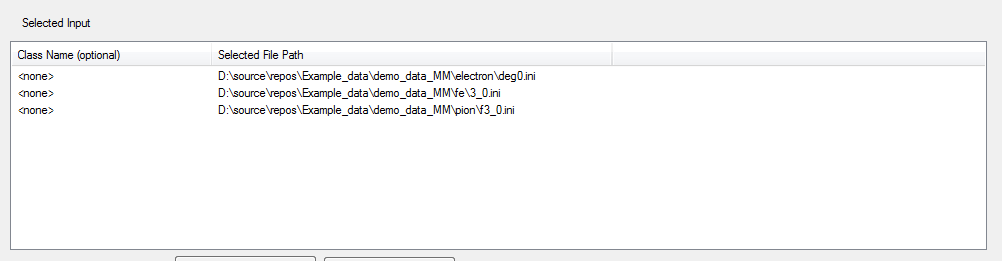
Figure 10.1 Filterer application folder with executable binary



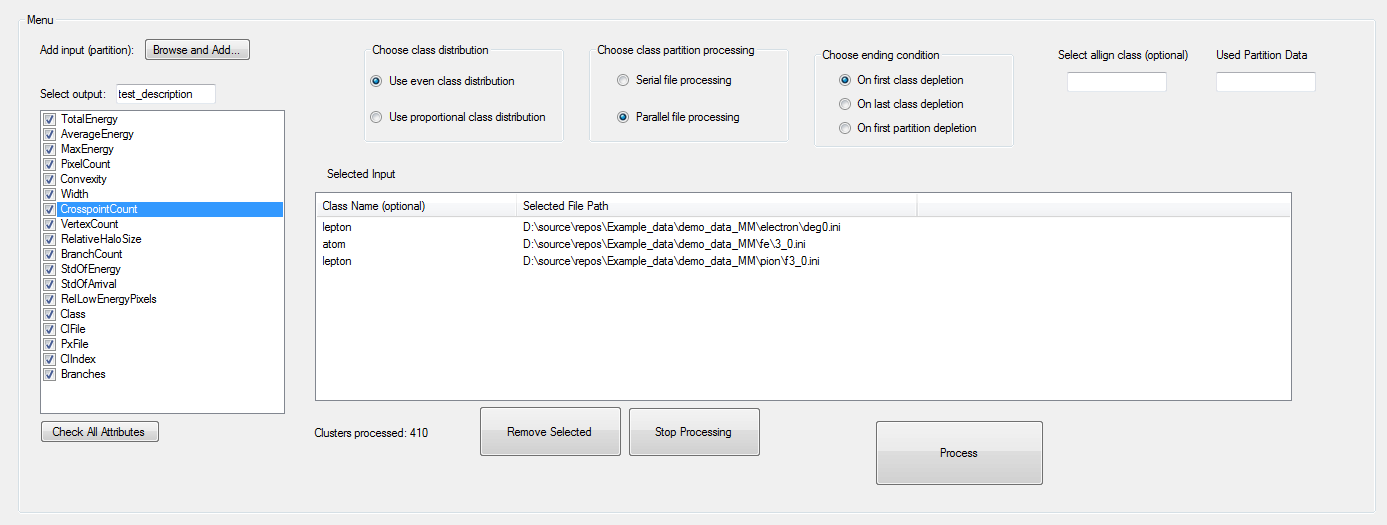
* By double clicking **ClusterFilter.exe** the application is run and the form is opened
* Then, we click Browse and in the dialog we select the file **demo\_data\_MM/frag/k\_nuc.ini**
* We also choose the name for our output file as **test\_filter**
* Now, we choose the filtering parameters as displayed in []
* We click Process button and we are immediately informed that the filtering is done.
* The output file is in the same directory as the input file and contains three clusters as displayed in []

## Description Generating

* We navigate to the folder with description generator executable binary (**/bin/Release**).
* By double clicking on **DescriptionGenerator.exe** we open the form.
* We click Browse and add button and select **demo\_data/electron/deg0.ini**
* We click Browse and add button and select **demo\_data/fe/3\_0.ini**
* We click Browse and add button and select **demo\_data/pion/f3\_0.ini**
* The selected partitions are displayed in the form as shown in []



* We can now choose class name for each of the three partitions – electron and pion are leptons while fe is an atom so we can edit the class name based on that by tripple clicking on the class name (which is <none> by default).
* After editing the partitions we select the name for the input as **test\_description**
* Then, we can decide to calculate all possible attributes by clicking Check All Attributes. The form should look similar to the one displayed in []
* After clicking Process the program starts processing the input files and creates the output file. It should only take a couple of seconds until we are shown the message box informing us that the processing is over.



* The ouput file contains the descriptions starting with the clusters as shown in []

(Note: the ClFile and PxFile values may be different based on the relative path of the output file to the original files in MM format)