

# MACHINE LEARNING FOR IDENTIFICATION OF SOURCES OF IONIZING RADIATION DURING SPACE MISSIONS

Ricardo Vilalta<sup>1</sup>, Sujeet Kuchibotla<sup>1</sup>, Francisco Ocegueda-Hernandez<sup>1</sup>, Song Hoang<sup>1</sup>, and Larry Pinsky<sup>2</sup>

<sup>1</sup>*Dept. of Computer Science, University of Houston, 501 Philip G. Hoffman Hall, Houston TX 77204, USA*  
*Email: {vilalta,ksujeet,focegueda}@cs.uh.edu, smhoang@uh.edu*

<sup>2</sup>*Dept. of Physics, University of Houston, 4800 Calhoun Rd., Houston TX 77204, USA*  
*Email: pinsky@uh.edu*

## ABSTRACT

We describe the current state of a project implementing a machine learning tool for the automated identification and classification of tracks of ionizing radiation during space missions. The tracks of ionizing radiation are measured by the hybrid semiconductor Medipix2 pixel detector system. Our analysis makes use of data taken in beams of heavy ions at HIMAC (Heavy Ion Medical ACcelerator Facility) in Chiba, Japan. The classification problem consists of predicting the nature of the source of radiation after analysis of track structures on pixel images; the track structure is formed by an incident traversing charged particle. Our main challenge lies in extracting relevant features that can facilitate discriminating among different sources of radiation.

Key words: classification, pattern recognition, pixel detector, radiation dosimeter.

## 1. INTRODUCTION

While recent advances in hardware technology promise a major step forward in the development of active portable space radiation dosimeters, little effort has been devoted toward software tools for analysis and classification of sources of radiation. Coupling radiation dosimeter hardware with machine learning tools has the potential to greatly improve current applications on space projects, future Extra Vehicular Activity (EVA) suits and lunar habitats, predicting conditions that can damage electronic equipment, and on assessment of radiation environments during solar particle events.

Overall our goal is to provide a local alarm capability that can be set based on logical combinations of relatively sophisticated dose-related factors. For example, an astronaut performing a repair task on the International Space Station can receive a prompt warning of the nature and intensity of any incoming radiation field. The ability to monitor particle tracks directly with an electronic device

provides the capability to respond within milliseconds to significant increases in dose. Our focus, however, is not to measure any dosimetric values directly, but rather simply to determine the physical nature of the radiation field accurately enough to allow the subsequent calculation of any dosimetric endpoint; this would allow to specify a detailed function of the nature of the radiation field itself.

From a hardware perspective, an instance of a portable space radiation dosimeter is the MediPix2 chip device, developed through a collaboration based at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. The MediPix2 chip is a hybrid semiconductor pixel detector made of readout CMOS-based integrated circuit; such technology is able to survive and perform for extended periods in strong radiation fields, and has an enormous dynamic range in sensitivity (from minimum ionizing muons through very heavy ions). But despite its advanced sensing capability, the technology currently lacks a software tool that can accurately predict the type of source of radiation captured by the pixel detector device.

In this paper we describe a machine learning tool for the automated classification of sources of ionizing radiation as captured by the Medipix2 device. Our tool first uses a segmentation operator that identifies groups of connected pixels corresponding to a track left by a charged particle captured by the Medipix2 device. Each set of connected pixels corresponds to a track that can be characterized through a feature vector, casting the problem as one of classification. This paper is an extension of a previous paper where we used features capturing properties of track structures [1]. Different from such previous work, here we introduce a new set of features based on how the charge disseminates throughout the track. Our experimental results show improved performance when we add features that measure (indirectly) the diffusion of collected charge.

The paper is organized as follows. Section 2 describes the Medipix2 device, and our pattern recognition tool. Section 3 explains our approach to extract features from tracks structures. Section 4 shows experimental results. Lastly, section 5 provides a summary and conclusions.

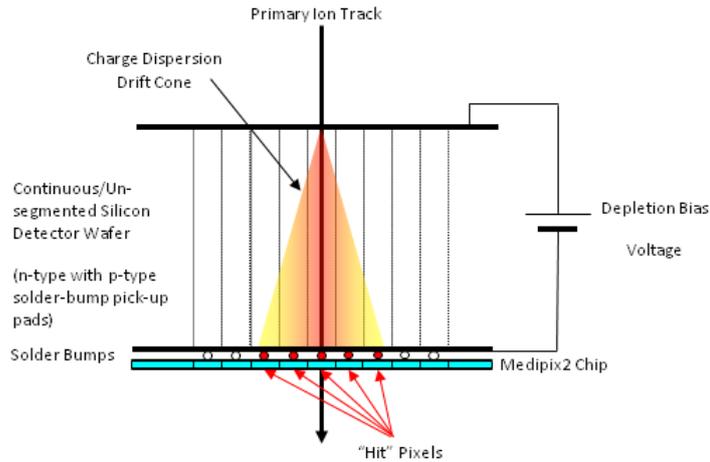


Figure 1. A schematic view of the formation of a pixel image of the track structure of an incident traversing charged particle. The drift cone of charge is dependent on the LET and the track structure of the particle, and the footprint of the image on the pixels is in general distributed in a conic section depending upon the angle of incidence of the particle.

## 2. BACKGROUND INFORMATION

We begin by providing a short description of the pixel electronics circuitry behind Medipix2, and of our machine learning tool, as described in previous work [1]. The interested reader can extract additional details from several sources [2, 3, 4].

### 2.1. The Medipix2 Device

The Medipix2 Collaboration is the second generation of a CERN-based consortium of university and institute-based physicists and electronic particle detector design groups that have evolved a pixel-based charged sensitive detector technology into devices that can be used for space applications, medical imaging, as well as a number of other commercial and research applications. The basis of the Medipix2 technology is the Medipix2 chip [5, 6]. It is a 256 x 256 pixel readout CMOS-based integrated circuit with a 55 $\mu$ m square pixel pitch in which the readout circuitry for each pixel is embedded totally within the footprint pixel area. This property allows multiple Medipix2 chips to be tiled effectively seamlessly together to create larger sensitive areas with no gaps. For use as a charged particle detector, the technology behind MediPix2 enables us to employ silicon (Si) detector layers up to several mm thick. When a penetrating heavy ion traverses the Si and produces a core of charge carriers surrounded by a halo of carriers associated with the track structure, the diffusion of the collected charge by the time it reaches the Medipix2 pixels is generally cone shaped with the charge from the most distant part of the track being disbursed the widest and that from nearest to the pixel contact remaining closest to the track itself. Models of this diffusion predict and measurements confirm that

effectively one is seeing an analog-amplified picture (albeit very non-linear) of the track structure cross-section.

A schematic view of the formation of a pixel image of the track structure of an incident traversing charged particle is shown in Fig. 1. The *drift cone* of charge is dependent on the Linear Energy Transfer (LET) and the track structure of the particle, and the *footprint* of the image on the pixels is in general distributed in a conic section depending upon the angle of incidence of the particle. Upward moving particles will yield very similar, but slightly modified footprints, depending upon their energy.

Fig. 2 shows an image of a frame from a normally incident beam taken at HIMAC, along with a magnification of one of the individual track images. The visible features in the track structures can be used to distinguish the charge and energy of the incident particles. Our goal is to automate the identification and classification of track images to predict the source of the ionizing radiation. In our experiments, we had five classes corresponding to tracks produced by beams of Si, Fe, O, N, and Ne. Identifying the right element source is the task of the machine learning software, which we describe next.

### 2.2. Automatic Identification of Sources of Radiation

We briefly explain the inner mechanism of our software tool [1]. The tool divides into two modules: a signal extraction module and a classification module.

**Signal Extraction Module.** The signal extraction module assumes as input an image generated by a TimePix-based detector after exposure to a source of ionizing radiation (see Fig. 2). The first step is to identify all tracks

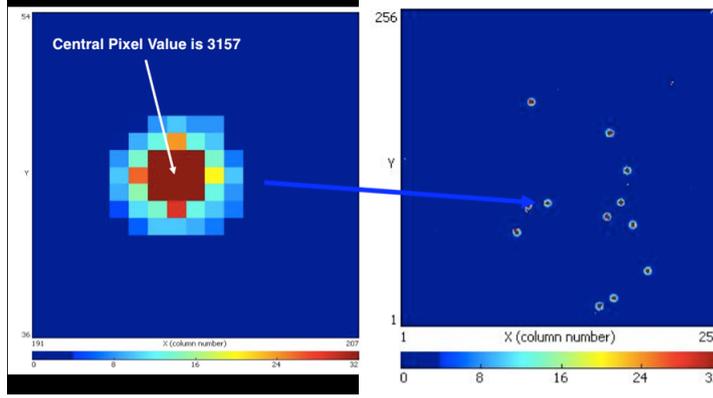


Figure 2. A single frame (right) from a normal incidence beam taken at HIMAC (11B beam with a 1 ms shutter time). The frame on the left is a magnification of one of the track images.

in the image using a segmentation operator. Our algorithm identifies clusters of pixels assuming an elliptical form that vary in size and degree of elongation based on multiple factors, including the type of radiation source and the angle of incidence. Our algorithm extracts clusters using Thresholding [7] and Connected Component Analysis [8]; each cluster becomes a track amenable to analysis. We employ Matlab as the image analysis tool.

Generating relevant features from each track is challenging because different sources of radiation produce very similar tracks. Subtle features carry high discriminatory power such as the rate of change of the intensity cone from inner ellipses to outer ellipses. Our set of feature descriptors is as follows:

- *Area*, defined as the total number of pixels within the track.
- *Volume*, computed as the sum of the energy of each pixel within the track.
- *Eccentricity* of the bounding ellipse of the track.
- *Major and Minor Axis* of the bounding ellipse of the track.
- *Mean energy intensity* of four bands within the bounding ellipse.

All features excluding the last are easily derived once an elliptical boundary is assigned to each track. For the last feature, the idea is to capture the flow of energy (i.e., charge) within the track; our algorithm creates four bands for each ellipse, as shown in Figure 3, and computes the average energy within each band (band 4 is the average energy for the entire track, whereas band 1 corresponds to the energy of the innermost band only). We employed a technique called distance transform [9]

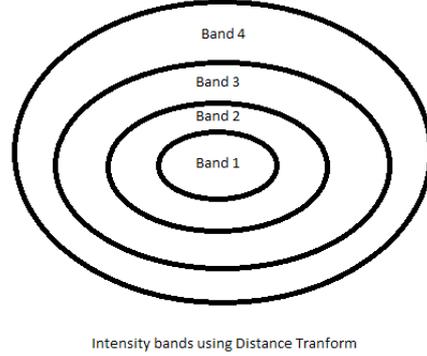


Figure 3. Intensity bands; mean intensity of band 2 includes that of band 1; mean intensity of band 4 implies mean intensity of the entire cluster.

to obtain the bands. Distance transform involves computing the perimeter pixels of the track and then labeling each pixel within the track with the distance to the nearest perimeter pixel; we used Euclidean distance as our distance metric. The result is a new pixel representation within a track where higher values correspond to pixels farther away from the perimeter or border (i.e., closer to the center of the track) and low values correspond to pixels closer to the border. Such distances can then be used to generate different concentric bands.

**Classification Module.** The classification module aims at the automatic identification of sources of ionizing gas from tracks. We assume an input training set  $T = \{(\mathbf{x}, y)\}$ , where each element  $\mathbf{x} = (a_1, a_2, \dots, a_9)$  in the set is a vector of nine features (area, volume, eccentricity, major axis, minor axis, mean energy within band  $\{1,2,3,4\}$ ). Each vector  $\mathbf{x} \in \mathcal{X}$  is attached a label  $y \in \mathcal{Y}$ . Our classes are the elements acting as sources of ioniz-

ing radiation (Si, Fe, O, N, and Ne). The outcome of the classifier is a function  $f$  mapping a track (characterized by a feature vector) to one element class,  $f : \mathcal{X} \rightarrow \mathcal{Y}$ . Function  $f$  can then be used to predict the class of new unseen tracks.

We invoked several classifiers to train on dataset  $T$ , including non parametric techniques (decision trees, support vector machines, and neural networks) and one parametric technique (Bayesian classifier). All of these techniques are able to delineate complex decision boundaries over the feature space, while avoiding models overly complex (i.e., while avoiding over-fitting).

### 3. EXTRACTING NEW TRACK FEATURES

Currently, we are investigating how to derive new features from track structures to improve on our problem representation. Different element beams may produce very similar tracks; looking at the shape of tracks alone seems insufficient to discriminate among classes. Our attention is now focused on the charge dispersion. The diffusion of collected charge produced by a heavy ion penetrating the particle detector can be used as a relevant signature to identify the right source of radiation. We mentioned before how the diffusion of the collected charge is generally cone shaped (Fig. 1), with the charge from the most distant part of the track being disbursed the widest. Our goal is to capture properties that can differentiate among different forms of charge dispersion.

The spread of charge in a track can be captured by simply counting the number of pixels within the track having a certain range of charge, i.e., we use histograms to capture energy levels within a track. These features are easier to compute than the distance transform based energy band feature (Sect. 2.2). Our approach simply divides the amount of charge captured at each pixel within a track in four non-uniform bins, ranging between zero and the maximum charge in the track. The non-uniformity of bins guarantees each bin is likely to be populated. Bins are created as follows. The first bin corresponds to the frequency of pixels within the track with energy or charge values in the top 15%. The second bin captures the frequency of pixels with charge values in the top 15% – 50%. The third and fourth bins capture the frequency of pixels with charge values in the top 50% – 70% and 70% – 100%, respectively.

Fig. 4 shows sample histograms for representative clusters of all five classes. The frequency of pixels on each bin stands for a new feature. One can see important information captured by the histogram representation. For example, the 1st bin (top 15%) and 4th bin (70% – 100%) for iron (Fe) tend to be much larger than any other elements. Although there are more complex forms of capturing the spread of charge on each cluster, in this paper we assess the usefulness of a simple histogram to characterize track structures.

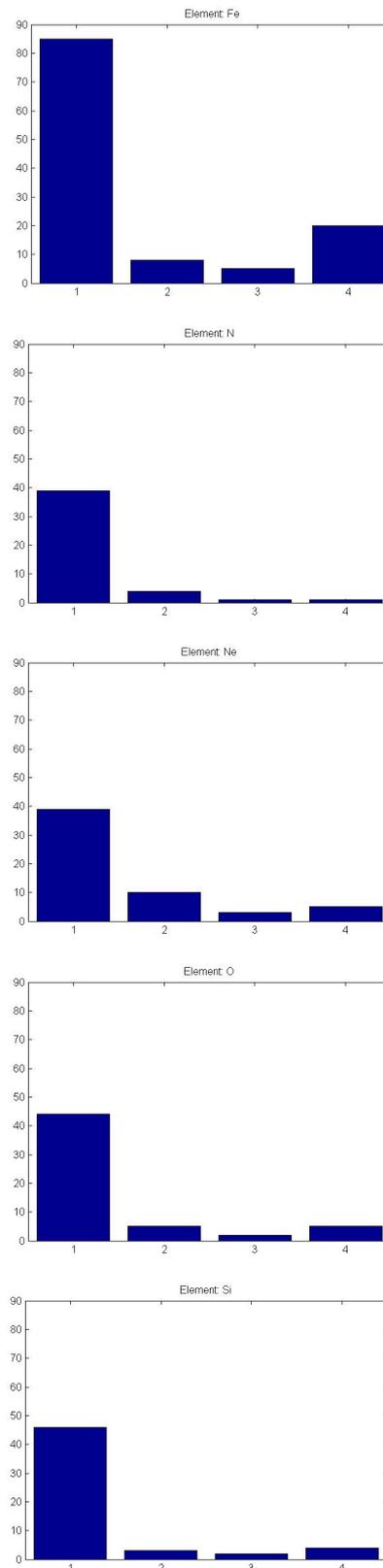


Figure 4. Histograms derived from sample clusters (one for each class).

Table 1. A comparison of average predictive accuracy with beams coming at different incident angles.

Learning Technique	Previous Predictive Accuracy			New Predictive Accuracy		
	All angles	0 degrees	60 degrees	All angles	0 degrees	60 degrees
Decision Trees	80.44 (0.50)	79.41 (0.44)	81.10 (1.47)	82.09 (0.23)*	80.92 (0.24)*	83.06 (0.23)*
Support Vector Machines	71.13 (0.72)	71.24 (1.33)	76.38 (0.78)	74.10 (0.33)*	74.80 (0.33)*	81.72 (0.32)*
Neural Networks	77.20 (1.21)	79.17 (1.19)	81.16 (1.52)	81.53 (0.22)*	80.92 (0.22)*	84.38 (0.20)*
Bayesian Classifier	65.34 (0.79)	67.56 (0.96)	72.58 (1.10)	64.23 (0.28)	70.43 (0.28)*	76.08 (0.25)*

#### 4. EXPERIMENTS

Our experimental analysis uses images obtained from beams of heavy ions at HIMAC. We work with five classes corresponding to tracks of Si, Fe, O, N, and Ne. Each class is represented by 4000 clusters, for a final sample size of 20,000. We compute predictive accuracy (i.e., the fraction of all correct predictions) using 10-fold cross validation. Results are shown in Table 1. Numbers enclosed in parentheses represent standard deviations. The 1st column shows the classification algorithms used for our experiments; the rest of the columns show results with beams of radiation at various angles: all angles, 0°, and 60°. Columns 2-4 show accuracy results using the concentric-bands information as features (see Fig. 3), discarding the histogram bins as features. Columns 5-7 show equivalent results when the histogram bins are part of the analysis, replacing the use of concentric-bands. An asterisk to the right of a result means there is a significant difference with the corresponding column matching the beam angle. For most cases there is a significant increase in predictive accuracy with the new set of histogram-bin features. The difference goes from 1.51% to 5.32% with an average value of 2.95%. We take these results as evidence of the relevance behind the histogram-bin feature information.

Regarding the incidence angles, predictive accuracy reaches a maximum of around 82% when all angles are used for analysis. As observed in previous work, varying the incidence angle does produce different results, but differences with the top performance algorithm (decision trees) are about the same. Overall an angle of 60° produces the best setting to discriminate among classes.

As a second experiment, we investigate the idea of transforming the original classification problem to a simpler one, by identifying the most conflictive classes, merging them into a single class. This entails a loss of information when a prediction points to the new class, but can be useful when other classes are predicted accurately. We have observed that three of the five element classes, namely Si, Ne, and O, produce very similar tracks. Our second experiment groups these elements into a single class, and proceeds as before under the new class representation. Table 2 shows our results, where no distinction exists among incidence angles, and the new histogram-bin features are incorporated. The second column shows

Table 2. New predictive accuracy with beams coming at all angles and classes merged.

Learning Technique	New Predictive Accuracy, all angles	
	4 classes	3 classes
Decision Trees	95.63 (0.14)	99.3 (0.06)
Support Vector Machines	92.26 (0.32)	99.19(0.27)
Neural Networks	94.87 (0.14)	99.60(0.04)
Bayesian Classifier	82.50 (0.25)	95.97(0.14)

Table 3. Features ranked based on information gain.

Features	Information Gain
Volume	1.33
Histogram Bin 4	1.24
Major Axis	0.92
Area	0.87
Minor Axis	0.87
Histogram Bin 1	0.67
Histogram Bin 3	0.65
Eccent.	0.45
Histogram Bin 2	0.19

results when Ne and O are merged into a single class (reducing the total number of classes to four). The third column shows results when Si, Ne, and O are all merged into a single class (reducing the total number of classes to three). In both cases we observe a significant improvement in performance, with the highest gain corresponding to the case where all three conflictive classes are merged together.

We are also interested to determine the relevance of our features by measuring the amount of mutual information between each feature and the target class, using a metric known as information gain [10]. Table 3 shows results with our set of features (ordered from top to bottom). The high rank of histogram-bin four shows how the outer part of every track is the most important region to differentiate among sources of ionizing radiation. This information is relevant to understand efficient approaches to represent class signatures in radiation environments, and complements previous studies along this research direction [11, 12, 1].

## 5. SUMMARY AND CONCLUSIONS

This paper describes a machine learning tool to identify sources of ionizing radiation by extracting useful information from pixel images produced by the MediPix2 device. A direct application is to have a local alarm capability that can be set based on logical combinations of relatively sophisticated dose-related factors (e.g., during a mission on the International Space Station).

We describe several features extracted from tracks on pixel images, with a focus on the diffusion of collected charge produced by a heavy ion penetrating the particle detector. In each case the diffusion of the collected charge is generally cone shaped (Fig. 1), with the charge from the most distant part of the track being dispersed the widest. Different beams of radiation vary on the degree of charge diffusion. Our work suggests a new set of features that capture different types of charge dispersion by looking at the frequency of pixels within a specific range of charge values. Empirical results show how histogram-bin information carries relevant information to differentiate among sources of ionizing radiation.

Our experiments show accuracy values around 82% (using decision trees) when classifying tracks produced by five possible elements. We investigate the performance improvement obtained when grouping together some of the classes, albeit with some loss of information. Merging two conflicting classes yields accuracy values around 95%. Merging three conflicting classes yields accuracy values around 99%.

As future work we plan to look for additional features to enhance our track representation, such as long range  $\delta$ -ray artifacts surrounding the main footprint images. We also plan to look for alternative representations of charge dispersion that provide a more granular description of the change of charge intensity along different track directions.

## ACKNOWLEDGMENTS

This work was supported by National Science Foundation (grant IIS-0812372).

## REFERENCES

[1] Vilalta, R., Kuchibhotla, S., Valerio, R. & Pinsky, L. (2010). Development of Pattern Recognition Software for Tracks of Ionizing Radiation In Medipix2-Based (TimePix) Pixel Detector Devices. *18th International Conference on Computing in High Energy and Nuclear Physics (CHEP-10)*, Taipei, Taiwan. Journal of Physics: Conference Series.

[2] Fiederle, M., Greiffenberg, D., Idarraga, J. et al. (2008). Energy Calibration Measurements of

MediPix2. *Nuclear Instruments and Methods in Physics Research A*, Vol. 591, pp. 75-79.

[3] Kroupa, M., Jakubek, J. & Krejci, F. (2008). Charge Collection Characterization with Semiconductor Pixel Detector Timepix. *Proc. of the IEEE Nuclear Science Symposium Conference Record*, pp. 259-262.

[4] Jakubek, J., Cejnarova, A., Holy, T., Pospisil, S., Uher, J. & Vykydal, Z. (2008). Pixel Detectors for Imaging with Heavy Charged Particles. *Nuclear Instruments and Methods in Physics Research A*, Vol. 591, No. 1, pp. 155-158.

[5] Llopart, X., Ballabriga, R., Campbell, M., Tlustos, L. & Wong, W. (2007). Timepix, a 65k Programmable Pixel Readout Chip for Arrival Time, Energy and/or Photon Counting Measurements. *Nuclear Instruments and Methods in Physics Research A*, Vol. 581, Issues 1-2, pages 485-594.

[6] Llopart, X., Campbell, M., Dingploi, R., Sansogundo, D., & Pernigotti, E. (2002). Medipix2: A 64-k Pixel Readout Chip with 55- $\mu$ m Square Elements Working in Single Photon Counting Mode. *IEEE Transactions on Nuclear Science*, Vol. 49, Issue No. 5, pp. 2279-2283.

[7] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62-66.

[8] Haralick, Robert M., Shapiro & Linda G. (1992). *Computer and Robot Vision*, Addison-Wesley, Vol. 1, pp. 28-48.

[9] Breu, H., Gil, J., Kirkpatrick, D. & Werman, M. (1995). Linear Time Euclidean Distance Transform Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 5, pp. 529-533.

[10] Rokach, L. & Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing.

[11] Holy, T., Heijne, E., Jakubek, J., Pospisil, S., Uher, J. & Vykydal, Z. (2008). Pattern Recognition of Tracks Induced by Individual Quanta of Ionizing Radiation in Medipix2 Silicon Detector. *Nuclear Instruments and Methods in Physics Research A*, Vol. 591, No. 1, pp. 287-290.

[12] Bouchami, J., Gutierrez, A., Holy, T., Houdayer, A., Jakubek, J., Lebel, C., Leroy, C., Macana, J., Martin, J., Pospisil, S., Prak, S., Sabella, P. & Teyssier, C. (2010). Measurement of Pattern Recognition Efficiency of Tracks Generated by Ionizing Radiation in a Medipix2 Device. *Nuclear Instruments and Methods in Physics Research A*.