6

Charge Classification with Convolutional Neural Networks

Contents

The correct categorisation of particle interactions in the detector is a major problem faced by any particle physics experiment. This typically starts with identifying low level features of the interactions, which can then be used to gradually build up a picture of the full interaction. In DUNE, the classification of neutrino interactions requires identifying the lepton content of the final state, and in ProtoDUNE–SP, cross section analyses rely on accurately identifying the particles in the interaction. Therefore, it is important to be able to distinguish muons and pions from electrons in LArTPCs, or more generally, tracks from showers.

In order to build up a complete picture of an event, it is useful to begin

by identifying the small features of the interaction, which can then be used to gradually build an understanding of the full event. In ProtoDUNE–SP, the smallest reconstructed features are the hits, which correspond to small charge depositions collected on individual wires. Classifying these hits provides useful information for future analyses, and can potentially be used to aid decision making during event reconstruction.

This chapter will describe an approach to hit classification in ProtoDUNE–SP using machine learning techniques. Section 6.1 will detail an approach to hit classification in LArTPCs based on identifying the source of energy depositions with a convolutional neural network. The performance of this approach will be analysed with ProtoDUNE–SP simulation and data in sections 6.2 and 6.3 respectively. Finally, Section 6.4 will briefly mention some of the current applications of the network in ProtoDUNE–SP analyses.

6.1 Hit Classification with Convolutional Neural Networks

Effective track shower separation forms the basis of many reconstruction challenges in DUNE and ProtoDUNE–SP; it is used to define pure calibration samples, such as minimum ionising muons and π^0 decays, and it is an important part of neutrino event reconstruction. Each event sample leaves a unique signature in the detector, but the first step in reconstructing these samples is the same, reconstructing tracks and showers, which can be combined to build the final state.

In a LArTPC, tracks and showers are built from collections of hits, these hits have to be clustered and identified as track or shower objects. In this section, we will describe a method for identifying the source of hits in the ProtoDUNE–SP LArTPC. The classification of each hit is stored as part of the reconstructed output in LArSoft, and can be used by subsequent reconstruction and analysis algorithms.

In addition to track and shower objects, Michel electrons are a useful calibration sample in LArTPCs. Michel electrons are electrons produced when a muon decays at rest, which have an energy spectrum in the range of 1–50 MeV. As discussed in

Chapter 4, the critical energy for electrons in liquid argon is at around 30 MeV. Therefore, Michel electrons have a unique signature in LArTPCs, and they were included as a unique category in the classification algorithm.

6.1.1 Data Preparation

A CNN was designed for the hit classification, the network was trained to predict,

$$[p_t, p_s, p_e] \text{ and } [p_m]$$

where p_t, p_s, p_e, and p_m, are the probabilities for track, shower, empty, and Michel electron classifications respectively. The empty category is included to ensure that the network doesn't learn to assign track–like or shower–like classifications to empty or noisy regions of the data. In addition, because the Michel electron category has an overlap with the track and shower categories, the Michel electron probability is decoupled from the other probabilities. The track, shower, and empty (TSE) probabilities are constrained to sum to one, such that every hit is classified into one of these categories.

An input image was produced for every reconstructed hit, with the hit being classified at the centre of the image. These images were produced from the wire readout data, after the noise removal and 2D deconvolution steps described in Chapter 3. Each input image is 48×48 pixels, with each pixel being filled with the ADC value from the wire readout data. The width of the image corresponds to one wire per pixel, and the height is the time coordinate. The time data is downsampled, using an average over time samples, such that the spatial dimensions of the image are the same in both directions. Therefore, each image represents around 24×24 cm^2 of wire data. Examples of an input image from each none empty class are shown in Figure 6.1, which also demonstrates the relationship between the images and the detector readout.

The true classification for each sample was obtained from the simulation, by associating the measured ionisation energy depositions to the corresponding simulated particle. If the reconstructed hit is not associated to a true particle, then

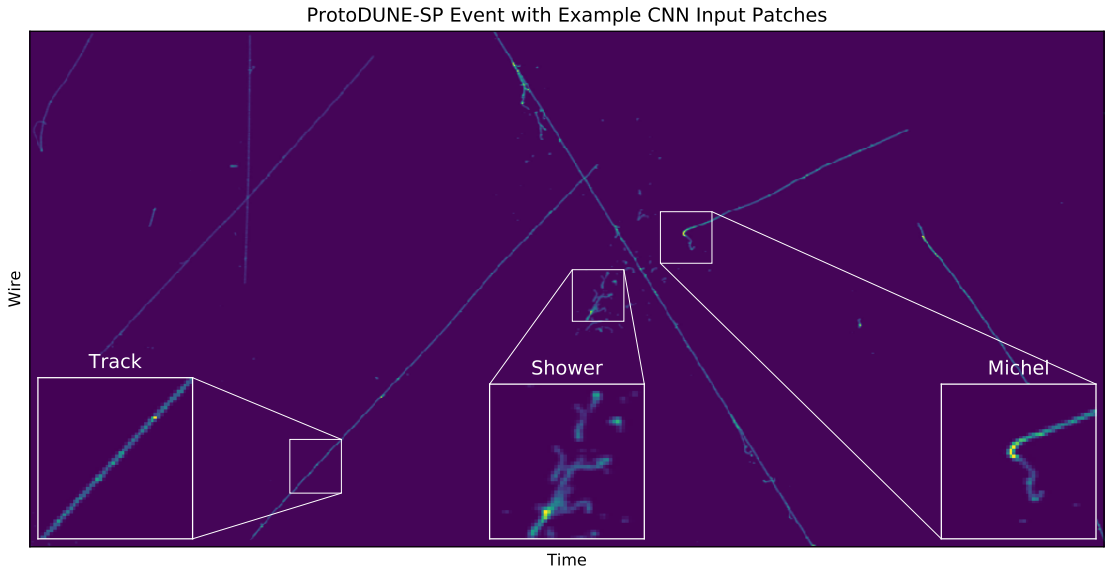


Figure 6.1: Example CNN input images for each class.

Ionisation Source	Track, shower, empty	Michel
Muons	[1,0,0]	[0]
Hadrons	[1,0,0]	[0]
Michel Electrons	[0,1,0]	[1]
Other Electrons	[0,1,0]	[0]
Noise	[0,0,1]	[0]

Table 6.1: Truth labels for different particles.

this hit is due to noise and the true classification is empty. Additional images are produced in empty regions of the detector, in the vicinity of charged particles, to increase the training sample for the empty category. The truth vectors for different true particle sources are detailed in Table 6.1. These values were chosen based on the typical interaction type in ProtoDUNE–SP for these particle species, which is based on the electromagnetic energy loss in liquid argon for energies in the GeV range.

The training data for the CNN was built using simulations of the ProtoDUNE–SP detector in the LArSoft framework; the simulations used were under beam operating conditions, and therefore included simulations of cosmic–rays, and test beam particles in the range of 1–7 GeV. Around 29 million input images were produced in total for the training. This sample was split into three datasets, the training,

Dataset	Shower	Track	Empty	Michel	Total
Training	13,493,982	9,727,604	2,517,882	731,456	26,470,925
Validation	734,673	562,038	141,388	42,727	1,480,826
Test	764,659	518,805	139,987	39,674	1,463,125
Total	14,993,314	10,808,447	2,799,257	813,857	29,414,876

Table 6.2: Number of input images with each truth label.

validation, and test sets. The training set is used to train the CNN, the validation set is used to monitor the performance of the CNN during training, and the test set is used as an initial verification of the performance of the network after training. Details of the number of patches of each type in these three datasets are detailed in Table 6.2.

6.1.2 Network Architecture

The network architecture for this CNN was designed to provide the best possible performance given constraints on running time and memory usage during network evaluation. This CNN is run on CPUs as part of the low level reconstruction chain for ProtoDUNE–SP, and it's run time is required to be on the order of 10 seconds per event. In addition, the CNN should not increase the maximum memory usage during reconstruction beyond around 4 GB. There is currently ongoing work looking into using GPUs during the network evaluation, which would allow more complex architectures to be used, and therefore potentially increase performance.

The network architecture used is shown in Figure 6.2. The images are first processed by convolutional layer with 48 5×5 filters, this layer extracts feature maps from the data. The responses from the convolutional layer are passed through the Leaky ReLU activation function, which is discussed in Chapter 5, before being processed by the dense layers. The feature maps are processed by a pair of dense layers, with 128 and 32 nodes respectively. These layers also use the Leaky ReLU activation function. After the second dense layer the network is split into two branches in order to make it's prediction. The first branch returns the prediction for the TSE categories, and the second branch for the Michel electron category.

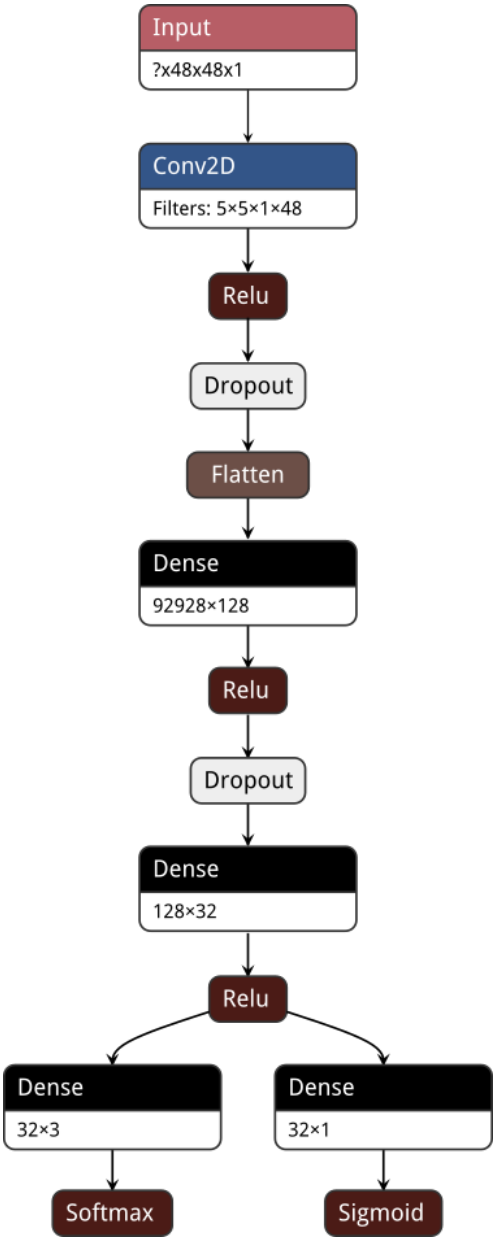


Figure 6.2: Network architecture used for hit classification.

The first branch uses a Softmax activation function, which ensures that the scores for the TSE categories sum to one. The second branch uses a Sigmoid activation function, which ensures that the Michel electron scores are bounded between zero and one. The choice of activation functions in the output layers allows for a pseudo–probabilistic interpretation of the scores from the CNN.

Two dropout layers are used in the network, these layers are used as part of the regularisation of the network. A dropout probability of 0.5 is used in both of

the dropout layers. The dropout algorithm is discussed in Chapter 5.

The loss of the network was the weighted sum of the losses for the two output branches,

$$L = 0.1 \cdot L_{TSE} + L_M,$$

where the Michel electron loss is given higher precedence due to the smaller training dataset available for the Michel electron output. The loss function for the TSE branch is the categorical cross entropy loss[103], and for the Michel electron branch it is the mean squared error[104],

$$L_{TSE} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=0}^{2}(t_j)_i \log(p_j)_i,$$
$$L_M = \frac{1}{N}\sum_{j=1}^{N}(t_j - p_j)^2$$

where t_j and p_j are the truth and the prediction for the j^{th} sample in the training batch, and i sums over all outputs in the TSE branch.

6.1.3 Training and Validation

The TensorFlow[105] library and the Keras[106] application programming interface (API) were used to design and train the CNN. The training was completed on a dedicated ProtoDUNE–SP server at CERN, with an NVIDIA GTX 1080 GPU. Training was monitored using the TensorBoard visualisation toolkit[107], which is part of the TensorFlow library.

The CNN was trained using the mini–batch stochastic gradient descent (SGD) algorithm, including both the momentum and decay algorithms[83]. The momentum algorithm reduces the oscillations of the weights during learning, while the decay of the learning rate allows for rapid learning during early stages of SGD, and increased precision as the model converges.

During training the learning metrics where monitored with TensorBoard. The losses for each branch and the total loss were monitored for the training and validation datasets. The validation loss was calculated once per training epoch,

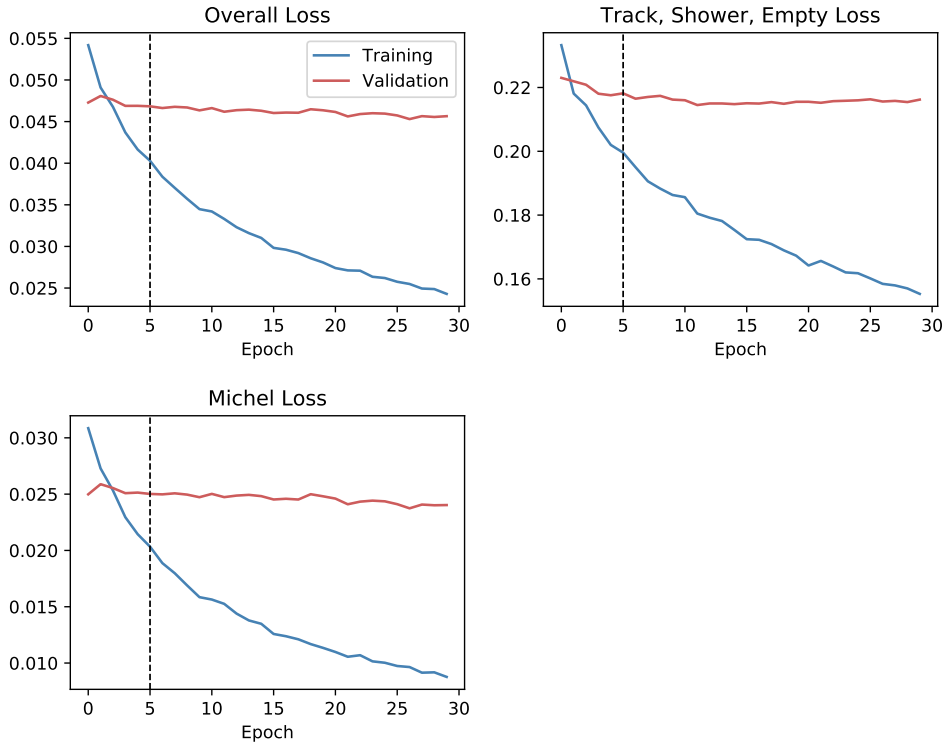


Figure 6.3: Evolution of the validation and training set losses during the training process.

which is one iteration through the full training dataset, and the training loss is averaged at the end of each epoch. The training and validation losses as a function of training epoch are shown in Figure 6.3. The weights of the network were saved at the end of each epoch. The final weights were chosen based on the early stopping algorithm discussed in Chapter 5, focussing on the TSE branch of the network, this is highlighted in Figure 6.3.

During training, the validation loss remains stable over a number of epochs, which suggests that the dropout algorithm was successful in preventing over–fitting. After training, the networks performance was verified against the test set, which found that the test set losses were compatible with the validation set loss. The final test set losses were,

$$L = 0.033,$$

$$L_{TSE} = 0.155,$$

$$L_M = 0.017.$$

6.2 Performance on ProtoDUNE–SP Simulation

The performance of the hit tagging algorithm was evaluated with reconstructed events from ProtoDUNE–SP simulation, the dataset used for this performance analysis was distinct from the training, validation, and test sets.

The distributions of the shower score from the TSE classifier for true shower hits and all other hits is given in Figure 6.4. There is a strong separation seen between the distributions for the shower and track hits, showing that the network has strong discriminating power. In practice, the empty score of the TSE classifier was found to be on the order of 10^{-9} or smaller for all hits tested. As such,

$$\text{TrackScore} \approx 1 - \text{ShowerScore},$$

which means that the results of the analysis of the shower score are valid for the analysis of the track score. Therefore, we will only discuss the shower score from now on.

The shower classification threshold for subsequent algorithms should be tuned on a case by case basis, however, for this study a simple optimisation strategy is presented in order to quantify the basic network performance. This is based on the F_1 metric, a specific case of the F_β metric[108], which places equal importance on precision and recall. The F_1 metric is given by,

$$\frac{1}{F_1} = \frac{1}{\text{precision}} + \frac{1}{\text{recall}},$$

where we define the precision as the fraction of correctly classified shower hits in the sample of all selected shower hits, and the recall as the fraction of all true shower hits, which were selected as shower hits. The F_1 score was calculated across a range of selection thresholds, this is shown in Figure 6.4. The score peaks at a threshold of 0.72 where,

$$F_1 = \text{precision} = \text{recall} = 0.863.$$

The overall performance of the TSE classifier can also be evaluated with a receiver operating characteristic (ROC) curve[109]. The ROC curve shows the true positive rate vs the false positive rate for the classifier, as the selection threshold is

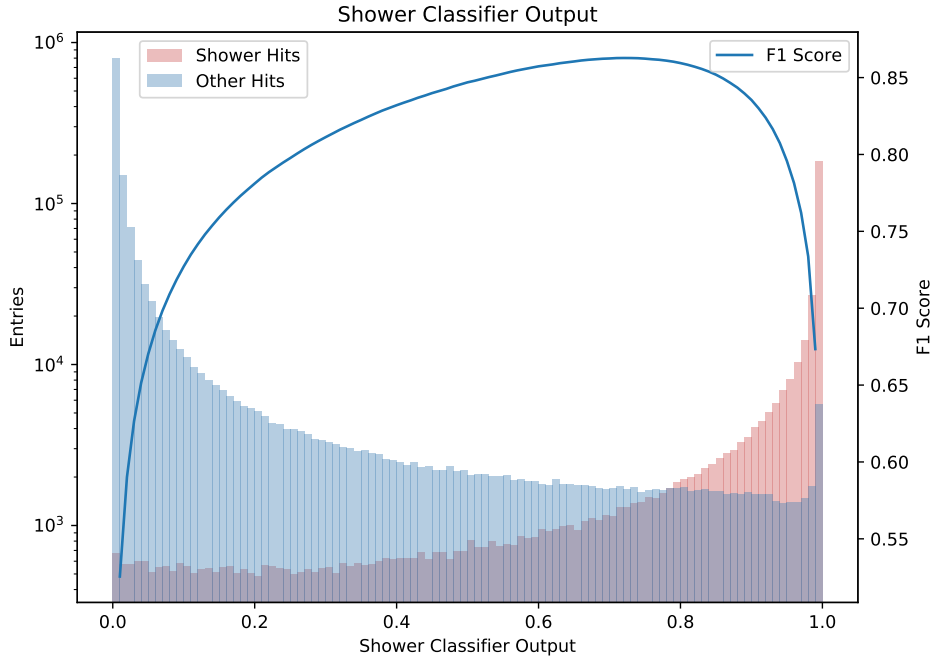


Figure 6.4: Shower score output distributions for the TSE classifier. Threshold optimisation was done using the F1 score metric, which is also plotted.

varied. Figure 6.5 shows two ROC curves for the TSE classifier, one is evaluated in simulation including the space charge effect (SCE), and the other excludes the SCE. Both curves demonstrate that the network is capable of achieving high true positive rates, while maintaining low false positive rates. In addition, their is a very close agreement between the two curves, which suggests that the TSE classifier is robust to changes in the SCE model.

The performance of the Michel electron classifier was analysed with the same methods as the shower classifier. The Michel electron score distribution for true Michel electron hits and all other hits is shown in Figure 6.6. In this case the large discrepancy in sample size between Michel electron hits and other hits, leads to a low F1 score of around 0.2 when considering the performance on a hit–by–hit basis. However, in Chapter 7, we will see that despite the low performance of the classifier for individual hits, a pure sample of Michel electron events can be selected by searching for clusters of hits with high Michel electron scores.

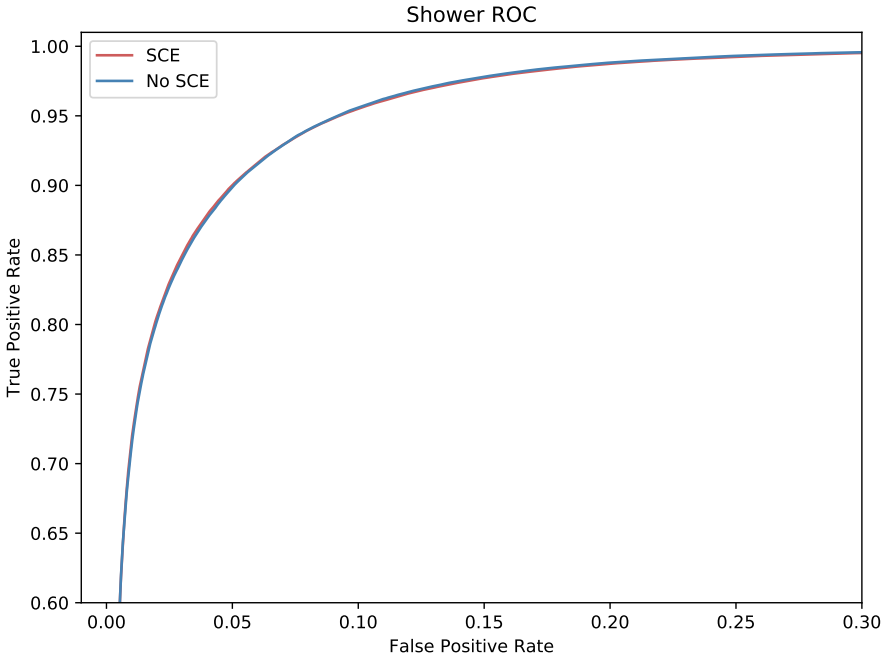


Figure 6.5: ROC curves for the shower score from the TSE classifier.

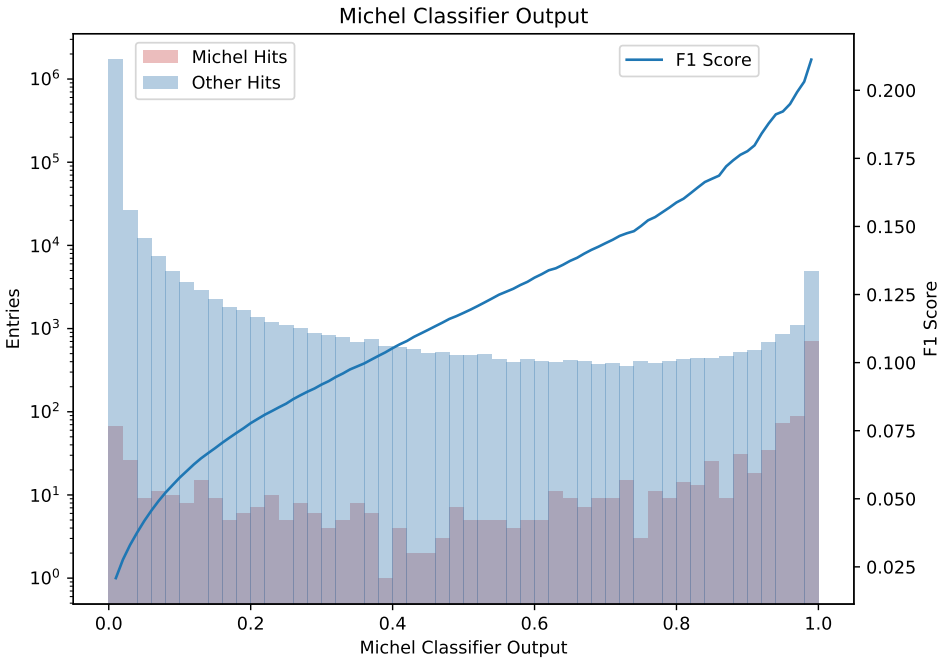


Figure 6.6: Michel electron score distributions for the Michel electron classifier. The F1 score metric was used to assist threshold selection, and is also plotted. The final threshold was modified when combined with a clustering algorithm, see the analysis in Chapter 7

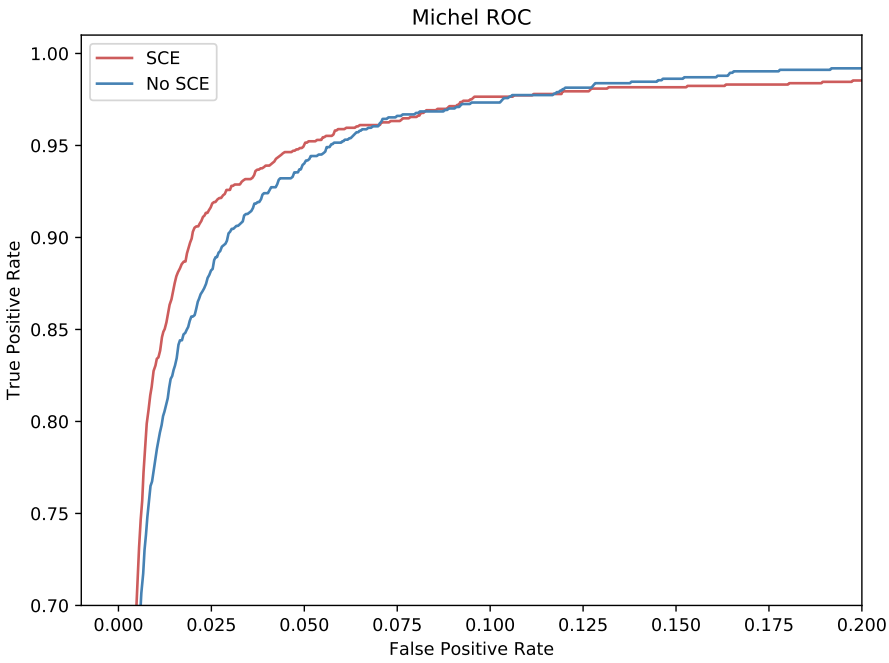


Figure 6.7: ROC curves for the Michel electron classifier.

The ROC curve is independent of the relevant size of each true sample, and therefore, it provides a more instructive evaluation of the performance of the Michel electron classifier than the score distribution and F1 score. The ROC curves for the Michel electron classifier are shown in Figure 6.7, where both SCE and no SCE samples are shown. The jitter in the lines is due to the smaller sample size in this case. In both cases the Michel electron classifier is able to achieve a true positive rate of over 90%, while maintaining a false positive rate of less than 2.5%. There is a larger difference between the SCE and no SCE curves for the Michel electron sample, however, the difference is still no bigger than 4%.

6.2.1 Comparison with Pandora

Pandora is the primary reconstruction framework used in ProtoDUNE–SP, it was discussed in Chapter 3. One of the goals of the CNN is to provide supplementary information to Pandora, to assist analysers in defining pure event samples. This is possible because Pandora and the CNN have slightly different goals, Pandora aims to

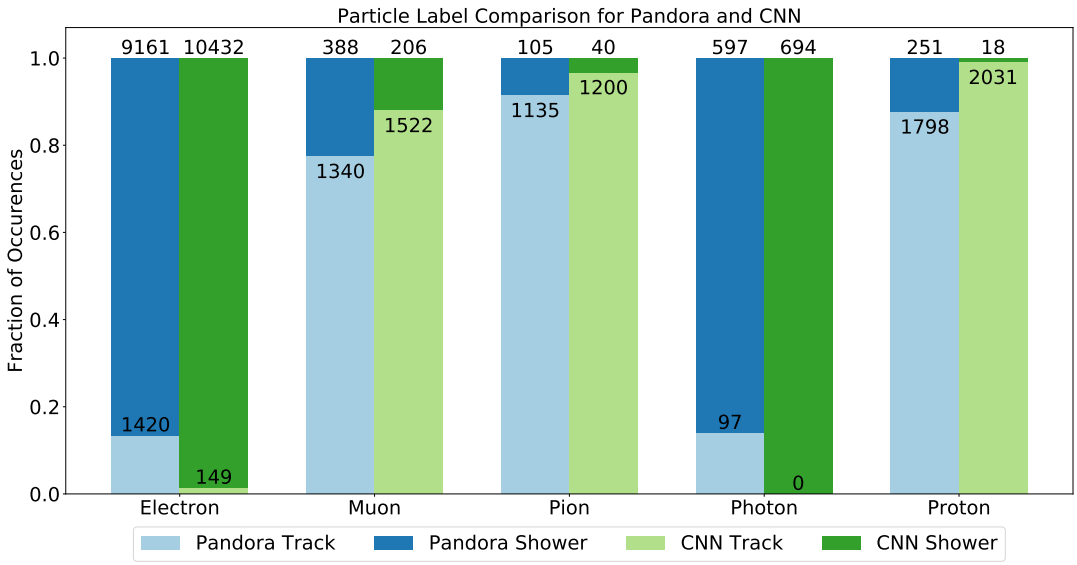


Figure 6.8: Comparison of track and shower classifications by Pandora and the CNN.

cluster hits in the most appropriate way based on their spatial distribution, while the CNN aims to classify the hits based on the particle that caused the energy deposition.

The comparison of the particle classification to Pandora was done with a beam particle sample in ProtoDUNE–SP simulation, including the primary particle and any daughters. The true particle type was obtained from the simulation, and the track–shower categorisation was compared between Pandora and the CNN. For Pandora, the type of reconstructed cluster was taken as the classification. For the CNN the average shower score for the hits in the particle was calculated and compared to a threshold of 0.72, based on the previous F1 score calculation. Any particles with an average shower score above the threshold were classified as showers, and any with score below the threshold were classified as tracks.

The particle classification by particle species based on the Pandora method and the CNN method are compared in Figure 6.8, which shows the fraction of each particle species classified as tracks and showers by Pandora and the CNN. We can see that CNN gives a stronger classification than Pandora for all particle species. Electron and photons are more often classified as showers, and pions, muons, and protons are more often classified as tracks.

Particle Species	Electron	Photon	Muon	Pion	Proton
Sample Size	10,581	694	1,728	1,240	2,049
Pandora FPR (%)	13.4	13.9	22.5	8.5	12.2
CNN FPR (%)	1.4	0.0	11.9	3.2	0.9
Pandora / CNN	9.53	–	1.88	2.62	13.9

Table 6.3: False positive rates for track shower classification with Pandora and the CNN.

The false positive rates (FPR) for Pandora and the CNN for each particle species are given in Table 6.3, as well as their ratio. The CNN only gives a modest improvement over Pandora for pions, and muons. However, there is a significant improvement in the false positive rate for electrons, photons, and protons. The biggest improvement is for photons, where, for a sample of 694 photons, the CNN has a false positive rate of zero.

The stronger classification from the CNN gives supplementary information to Pandora, which can be used in analyses to improve event selection or background rejection. This information is being utilised in a number of ProtoDUNE–SP analyses, some of these analyses are highlighted briefly in Section 6.4.

As well as providing supplementary information to Pandora for analysis, the CNN scores could be utilised during Pandora reconstruction as an additional guide for the reconstruction algorithms. This approach is being developed by Pandora for the DUNE far detector, based on a similar deep neural network for discriminating tracks and showers[110].

6.3 Validation on ProtoDUNE–SP Data

For validation on real ProtoDUNE–SP data three approaches were used: visual validation with event scans, comparisons of the overall score distributions, and the comparison of score distributions for different particle species. Data from ProtoDUNE–SP runs 5387 and 5809 were used for these validations; the data for these run was taken on under stable operating conditions, with an average beam energy of 1 GeV. The beam composition in run 5387 was tuned to contain mostly

hadrons, while in run 5809 it was tuned to have an enhanced electron component. The same operating conditions were used for the simulated data, in order to give the closest possible comparison between data and simulation.

As discussed in Section 6.2, the sample of Michel electron hits is orders of magnitude smaller than the other hits. In order to make a meaningful validation of the performance of the Michel electron classifier on data, the fraction of Michel electron hits in the sample needs to be increased. Therefore, discussion of the validation of the Michel electron classifier will be postponed until Chapter 7, which will discuss Michel electron event selection and reconstruction.

Hand scans of the events show qualitatively that the performance on the data is good. Figure 6.9 shows and example of the track score for hits in a reconstructed event from run 5387. The hits in this image are from the collection plane near the beam entry point, an electron shower tagged by the beam instrumentation (BI) can be seen near the centre of the image. In the event we can see that for hits in the tracks the CNN produces a large output score, and for the shower like activity in the event the score is low, as we expect. In addition, the classifier is able to identify that the hits adjacent to the track, which are from delta rays, are from electrons.

The shower score distribution for all hits gives an overall validation of the network performance between data and simulation. It is the only quantitative validation method for the CNN, which remains decoupled from the Pandora reconstruction algorithm. Therefore, it is independent of the differences between data and simulation, which impact the results of Pandora. The comparison is still dependent on the noise removal algorithm, the hit tagging algorithm, and the particle flux, which impact the input images to the CNN.

In this comparison, and the following beam particle comparisons, a cut is made on the reconstructed charge of each hit. This cuts out the excess of low charge hits in data, as illustrated in Figure 6.10a. In addition, all hits in the first 45 cm of the APA closest to the beam entry point were removed. This region has a known issue in charge simulation, which can be seen as a discontinuity in the

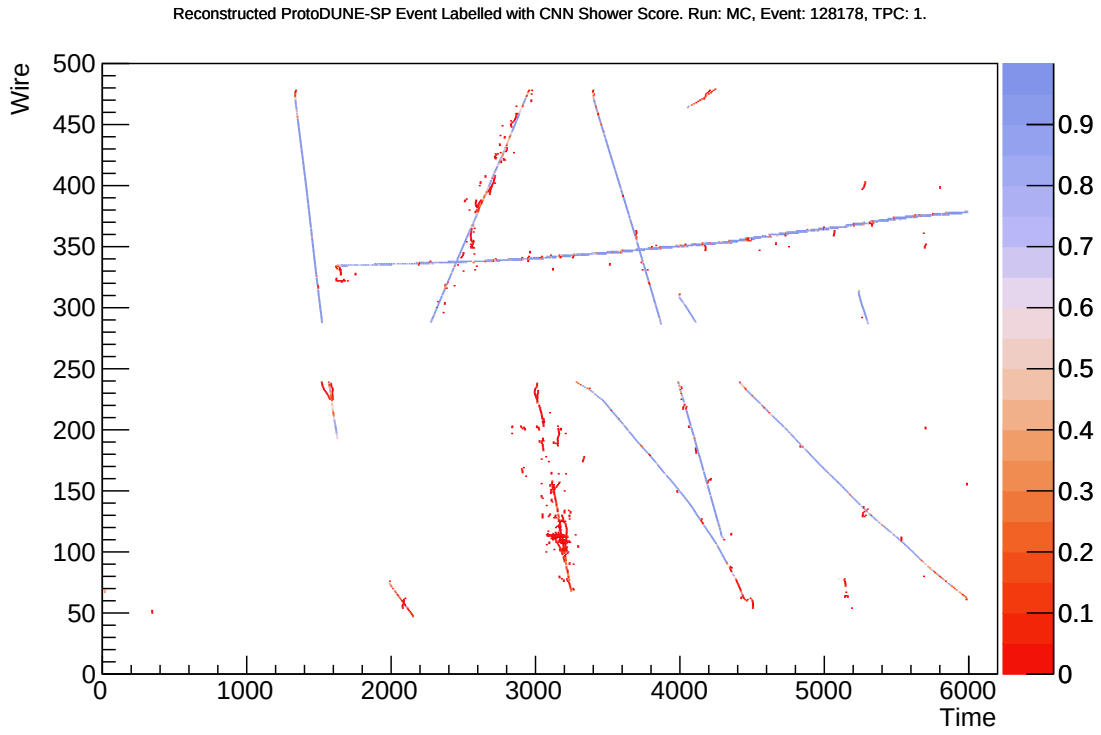
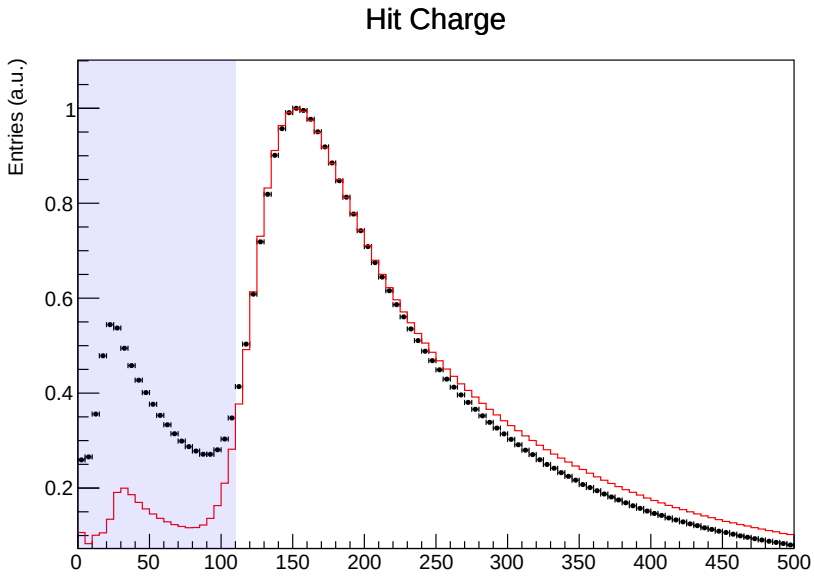


Figure 6.9: Track classifier scores for a reconstructed event in ProtoDUNE–SP data.

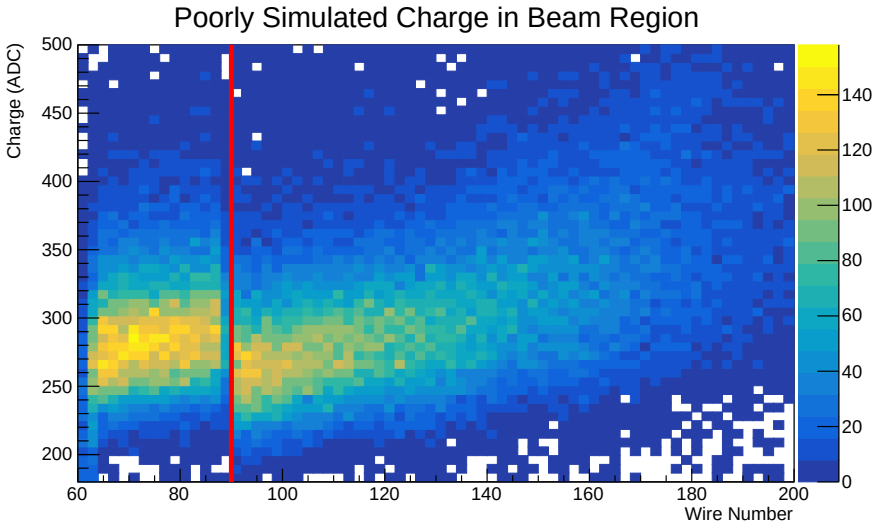
reconstructed charge. This charge discontinuity can be seen in in Figure 6.10b, which demonstrates the discontinuity for a simulated proton beam sample.

The overall shower score distribution for data and simulation is shown in Figure 6.11, as well as the fractional residuals in each bin. These distributions are normalised by the number of hits, allowing us to compare the shape of the score distribution in data and simulation. Overall there is a good agreement between the data and the simulation, however, the distribution in data has a slightly different shape at low CNN scores. This can be seen in the residuals, which have a negative slope for hits with shower score less than around 0.2.

We can also consider the comparison of the CNN to data for specific particle species, such as beam particles and cosmic–rays. The BI in ProtoDUNE–SP provides PID for beam particles with a very high purity. Therefore, in the case of particles originating from the charged particle beam, the BI can be used to provide an effective truth source. This allows the results of the CNN to be compared between data and simulation for different particle species. Any particles that arrive out–



(a) Charge Cut



(b) Charge Discontinuity

Figure 6.10: Quality cuts on reconstructed charge for CNN comparison.

of–time with the beam can be assumed to be cosmic–rays, and therefore form an additional truth source. The shower score distributions for electrons, pions, protons, and cosmic–rays are produced based on these samples.

It is important to note that the results for these true particle samples rely on both Pandora and the CNN scores. Therefore, phase space cuts were made, to select regions of the detector where Pandora has a good agreement between data and simulation. The aim of these cuts was to to minimise the impact of Pandora

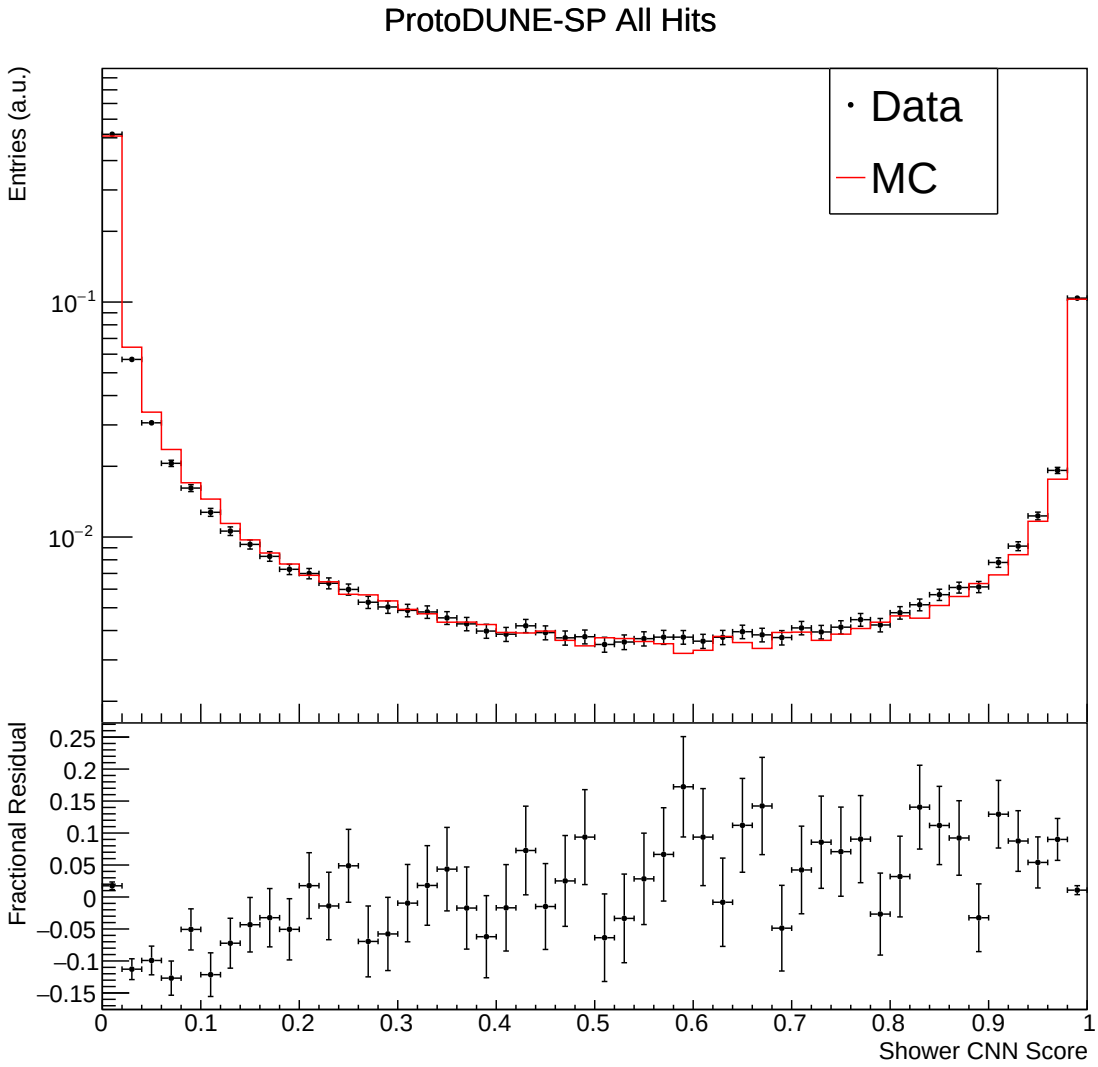
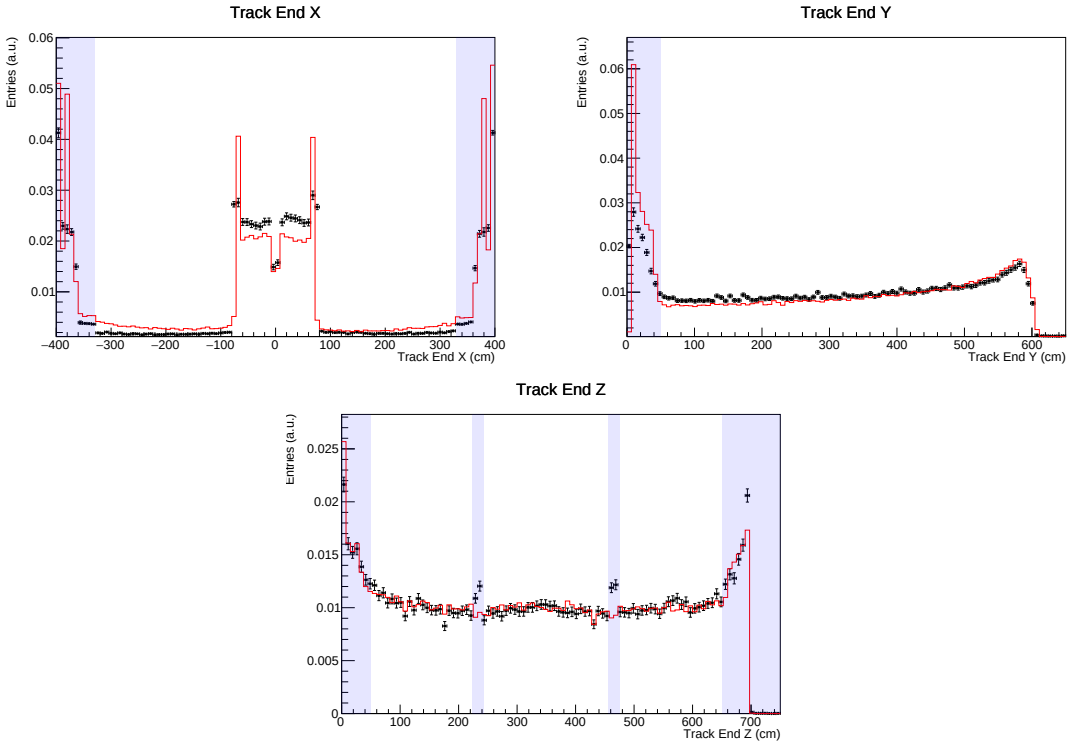


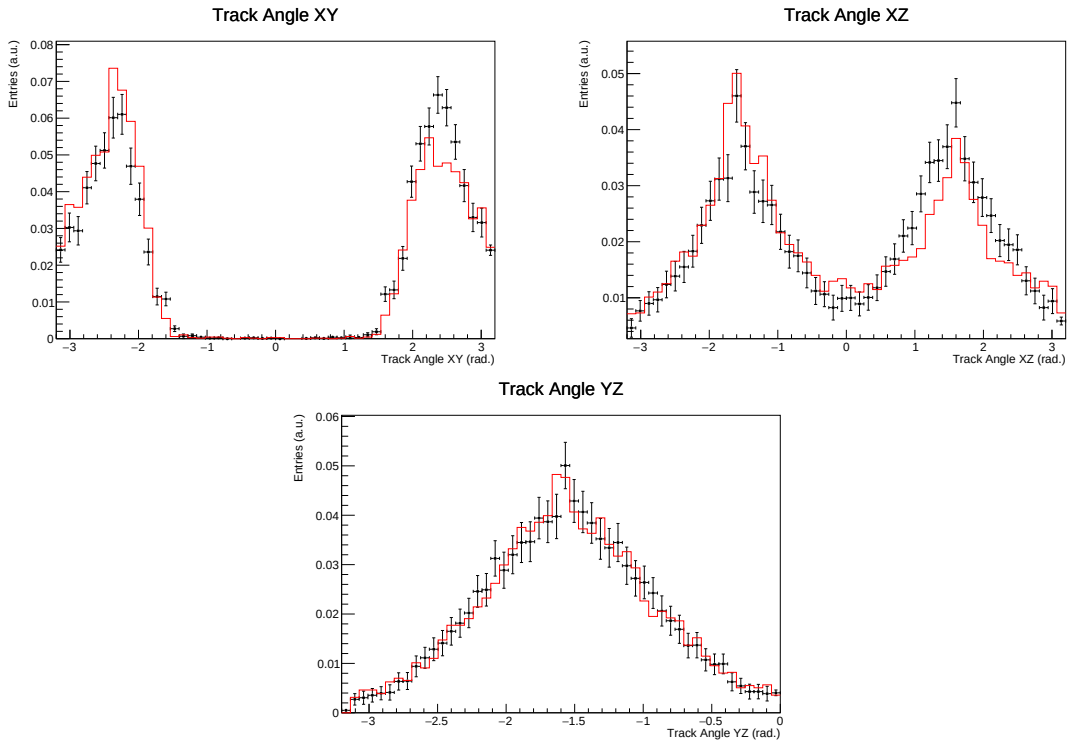
Figure 6.11: Overall shower score distribution from the CNN for all hits in data and simulation.

on the observed CNN score distributions. Cuts on the start and end points of the reconstructed tracks were made, with the same cuts being applied to both the start and end point of each track. These cuts are illustrated for the end point of the tracks in Figure 6.12a. The reconstructed angular distributions for tracks after these cuts are shown in Figure 6.12b.

The shower classifier scores for the pion, proton, and electron test beam samples are shown in Figure 6.13. The data in the pion and proton distributions was taken from ProtoDUNE–SP run 5387, and the data in the electron distribution was from run 5809. The data in all of these beam particle distributions are normalised by



(a) Spatial cuts.



(b) Angular distributions.

Figure 6.12: Data Quality Cuts for CNN Comparison.

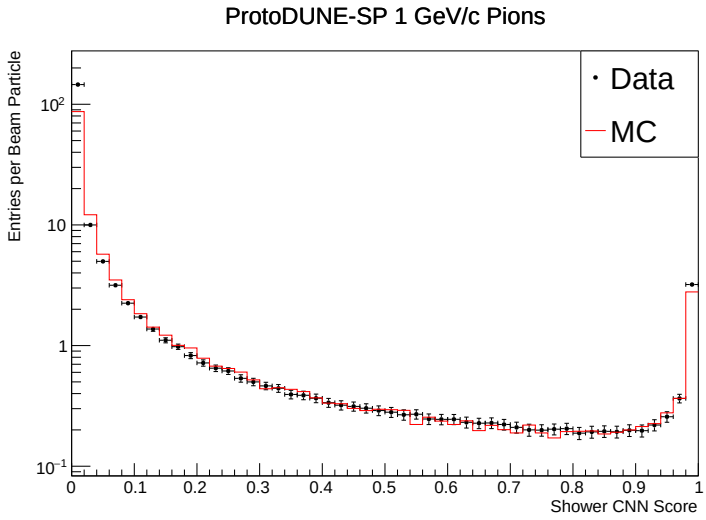
the number of triggered beam particles of the given flavour.

Overall, there is a good agreement between the data and simulation in terms of the shower score distributions for each particle species. However, there are still some discrepancies, which highlights the differences between the data and the simulation. The difference is most pronounced for the electron sample, which gets consistently more entries at low CNN scores. One consistent difference between data and simulation for all particle species, is the tendency for Pandora to cluster more hits into each reconstructed particle in data than in simulation. This can be seen in the bins at the two ends of the distributions, which contain the majority of the hits, and consistently have more entries in data than in simulation.

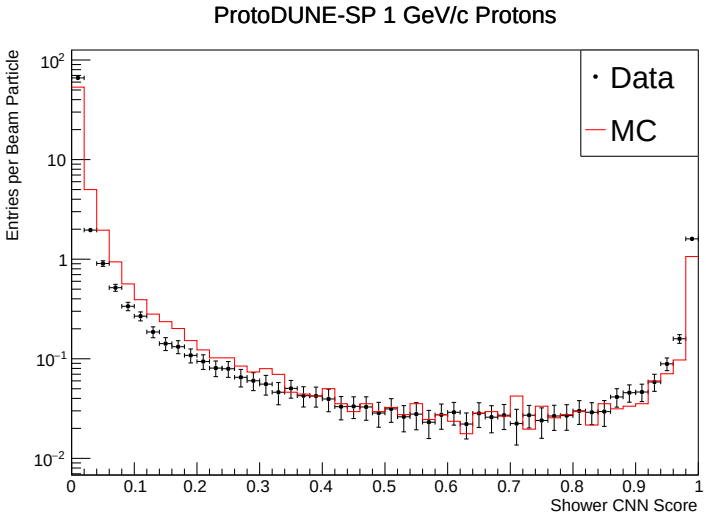
The cosmic–ray sample was selected based on the cosmic–ray part of the Pandora reconstruction chain, discussed in Chapter 3. Any reconstructed particles, which have been labelled as both clear cosmic–rays and tracks by Pandora are included in the sample. The data in this sample was taken from run 5387; while a cosmic only run would be preferable, all of the ProtoDUNE–SP simulations include test beam particles. Therefore, a test beam run was chosen in order to match as closely as possible to the simulation.

The shower score distribution for cosmic–ray tracks is shown in Figure 6.14. These distributions are normalised by the total number of hits in each distribution, this is because there is a large difference in the average number of reconstructed hits in clear cosmic–rays from Pandora between data and simulation. Again a good agreement is seen over the majority of the range, but in data there is a prominent excess of hits with a high shower score. The cause of this excess is not currently understood, one possibility is that delta–rays are more often reconstructed as part of cosmic–ray tracks in data than in simulation.

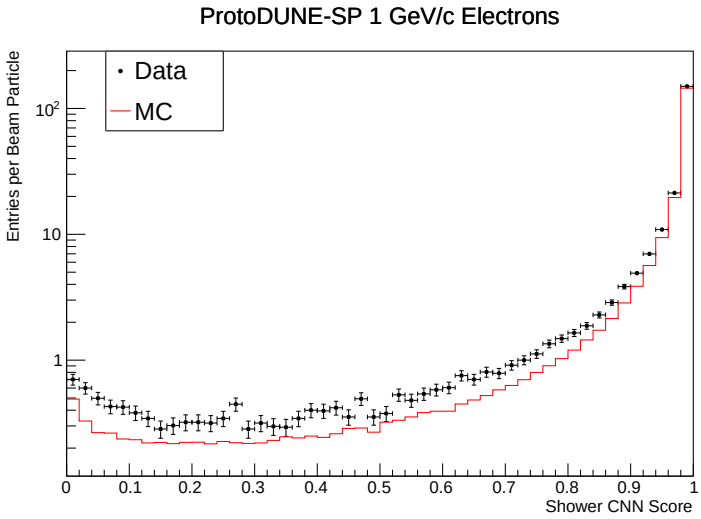
In practice, the CNN will be used to select events based on choosing hits with CNN scores above some threshold. This could be based on an average over a number of hits, on a hit by hit basis, or something more complicated. For the purposes of understanding the approximate errors involved in selecting hits with the CNN, we will consider selecting hits on a hit–by–hit basis here. The uncertainties involved



(a) pions



(b) protons



(c) electrons

Figure 6.13: Shower classifier scores for particles from the ProtoDUNE–SP beam.

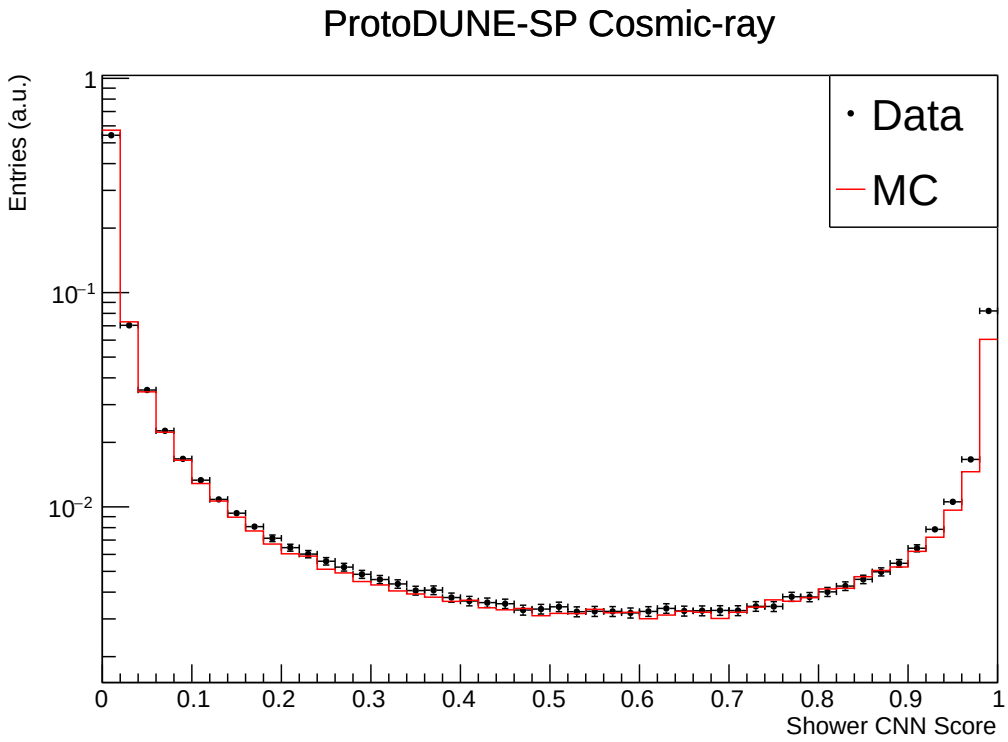


Figure 6.14: Shower classifier score for hits from cosmic–rays.

with selecting hits in this way can be evaluated by considering the fraction of hits selected into the sample given a selection threshold.

In Table 6.4, we list the fraction of individual hits selected into the appropriate category for each sample in data and simulation. The fractional difference between the numbers in each case is an estimate of the percentage uncertainty associated with this simple selection algorithm. For the sample of all hits, the track category was used, and a fractional difference of 1% was seen. This varies depending on the particle species, with the largest difference being 3% for the cosmic ray sample.

Overall, the CNN results have been shown to agree well with data across a range of particle types. The discrepancies seen are also influenced by the difference between data and simulation for the Pandora reconstruction framework, which makes the cause of the discrepancies difficult to disentangle. However, the general agreement between data and simulation is a sign that the results from the CNN are sensible, and the additional classification strength of the CNN over Pandora makes it a useful tool in analyses of the ProtoDUNE–SP data. The discrepancies

Hit Source	Selected Fraction MC (%)	Selected Fraction Data (%)	Data / MC
All	77.1	76.2	0.99
Pion	93.8	95.4	1.02
Proton	97.1	96.6	0.99
Electron	96.9	95.8	0.99
Cosmic–ray	92.9	90.3	0.97

Table 6.4: Fraction of hits selected as showers for reconstructed particles in ProtoDUNE–SP data and simulation.

will impact each analysis differently, therefore the uncertainties involved with using the CNN classifier should evaluated on a case–by–case basis; for hit selection with a simple hit–by–hit algorithm the uncertainties are on the order of 1-3% depending on the particle species.

6.4 Application in ProtoDUNE–SP Analyses

The output of the CNN classifier has been applied in a number of ProtoDUNE–SP analyses since it was developed. Chapter 7 will detail one use of the network in a Michel electron analysis. In addition, the scores from the network are being incorporated into analyses by other members of the ProtoDUNE–SP experiment, including:

- Selecting shower candidates for neutral–pion event selection[111].
- Identifying charge exchange candidates in charged–pion cross section analyses [112].
- Identifying Michel electron contaminated tracks for stopping muon calibration[113].