Universidad Popular Autónoma del Estado de Puebla

Vicerrectoría Académica

Decanato de Ingenierías

Implementación de algoritmo de aprendizaje, para mejorar la separación de datos en el observatorio HAWC

Tesis para obtener el Grado de Maestro en Ciencia de Datos e Inteligencia de Negocios

Presentado por:
Julián Federico Orea Díaz

Director de tesis
Dr. Ibrahim Daniel Torres Aguilar

H. Puebla de Zaragoza, México. (Junio 2021)

Universidad Popular Autónoma del Estado de Puebla

Vicerrectoría Académica

Decanato de Ingenierías

Ciencia de Datos e Inteligencia de Negocios

Se aprueba la Tesis:

“Implementación de algoritmos de aprendizaje, para mejorar la separación de datos en el observatorio HAWC”

Nombre del alumno:
Julián Federico Orea Díaz

Comité Asesor

_____

Dr. Ibrahim Daniel Torres Aguilar
Director de Tesis

Dra. Rosa María Cantón Croda Asesora	Mtro. Charles Galindo Jr. Asesor
H. Puebla de Zaragoza, México	(Junio 2021)

Índice general

1. Introducción 6
1.1 Antecedentes 6
1.2 Planteamiento del problema 8
1.3 Justificación 9
1.4 Objetivo general 10
1.5 Objetivos particulares 10
1.6 Detección de radiación Cherenkov en agua 10
2. Marco teórico 14
2.1 Cascadas atmosféricas extendidas 14
2.1.1 Características de las EAS 16
2.2 Rayos gamma 16
2.3 Rayos cósmicos 18
2.4 Efecto Cherenkov 18
2.5 Observatorio HAWC 19
2.5.1 Tanques de agua Cherenkov 20
2.5.2 Detección de partículas 21
2.5.3 Outriggers 22
2.5.4 Separación de rayos gamma / rayos cósmicos en HAWC 23
2.5.5 Resolución angular 25
2.5.6 AERIE 26
2.5.7 XCDF 26
2.5.8 Datos simulados 27
2.5.9 Variables candidatas 28
2.5.10 Binning 30
2.5.11 Método estándar de separación de partículas 32
2.6 Aprendizaje automático. 32
2.6.1 Redes Neuronales 34
2.6.2 MLP 34
2.7 Evaluación. 36

2.7.1 Factor Q.....36
2.7.2 Curvas ROC:.....37
3. Metodología.....40
3.1 Datos.....40
3.1.1 Datos simulados.....40
3.2 Entrenamientos de red neuronal.....43
3.2.1 Entrenamientos NN10.....43
3.2.2 Entrenamientos NN4, NN7, NN8.....45
3.3.3 Configuración de red neuronal.....50
3.3.4 Generación de salida red neuronal.....52
3.4 Cortes Óptimos.....53
3.5 Comparación con modelo estándar.....54
4. Análisis y resultados.....56
4.1 Entrenamientos de la red neuronal.....56
4.1.2 Entrenamientos.....56
4.2 Comparaciones de Métodos de separación.....60
4.2.2 Eficiencia en gamma.....60
4.2.3 Eficiencia en hadrón.....65
4.2.4 Factor Q.....69
4.3 Conclusiones.....79
Referencias.....81

1. Introducción

1.1 Antecedentes

La astrofísica, es una rama de la ciencia que estudia los fenómenos del universo a partir de la energía emitida por los objetos celestes, esta energía se le conoce como radiación, la forma en que se divide la radiación dependiendo su frecuencia o energía se le conoce como espectro electromagnético. Las diferentes divisiones del espectro electromagnético son radio, infrarrojo, visible, ultravioleta, rayos x y rayos gamma (Ilustración 1).

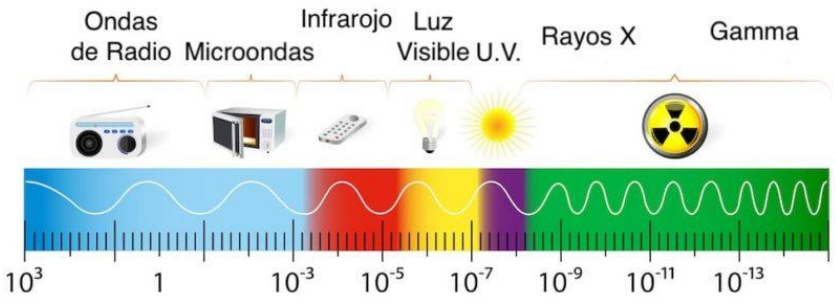


Ilustración 1. Se muestra el espectro electromagnético, que es la forma en que se divide la radiación dependiendo su frecuencia (Raffino, 2020).

Debido a que la radiación está formada por partículas elementales desprovistas de masa nombradas fotones, cuando estos fotones alcanzan energías sobre los 10^5 eV[1] se les considera rayos gamma. Cuando un rayo gamma alcanza energías sobre los 10^{14} eV y entra en contacto con la atmósfera del planeta, se generan nuevas partículas y conforme penetra más en la atmósfera, se repite

[1] Electronvoltio es energía cinética que adquiere un electrón al atravesar en el vacío una diferencia de potencial de un voltio. Equivale, aproximadamente, a 1.602 19 X 10^{-19} joules.

este proceso, creando un efecto cascada, a este efecto se le conoce como cascadas electromagnéticas (Longair, 2011).

Por lo general hay dos formas mediante la que los astrofísicos de altas energías detectan los rayos gamma, la primera es de forma directa a través de telescopios satelitales y la segunda es de forma indirecta mediante observatorios que se encuentran en la tierra, los cuales son capaces de detectar las cascadas electromagnéticas.

Uno de los observatorios capaces de detectar de forma indirecta los rayos gamma es el observatorio HAWC, este observatorio ha sido capaz de generar mapas del cielo de fuentes de rayos gamma, estos mapas actualmente ayudan a comprender los eventos del Universo que generan estas partículas como lo son galaxias, pulsares, cuásares (Ilustración 2), remanentes de supernova, nebulosas, cúmulos de estrellas (Ilustración 3), blazares y otras fuentes aun no identificadas (Abeysekara, y otros, 2017).

Ilustración 2. Impresión artística del cuásar 3C 279. Un quásar es un faro brillante de luz intensa procedente del centro de una galaxia distante, cuya emisión es tan potente que puede eclipsar a toda la galaxia. Su fuente de energía proviene de un agujero negro supermasivo que se alimenta vorazmente de materia, desatando un torrente de radiación. Crédito ESO/M. Kornmesser

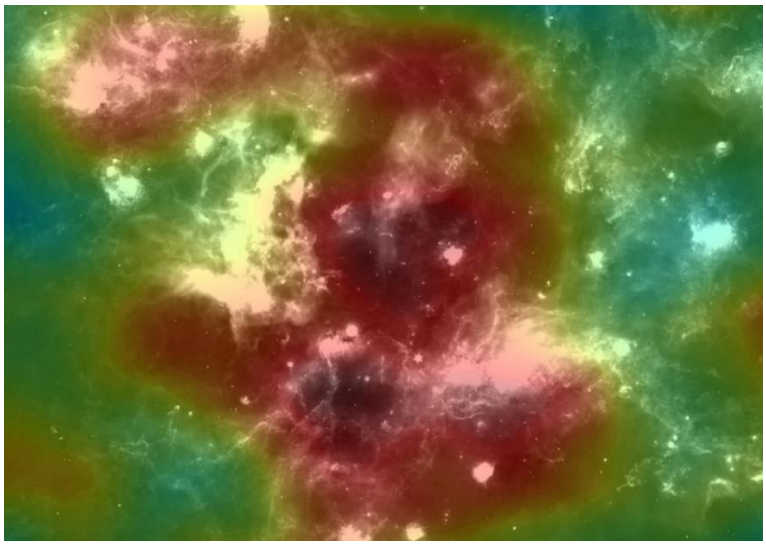


Ilustración 3. HAWC ha detectado rayos gamma con energías por encima de 200 TeV provenientes de Cygnus. En Cygnus se encuentra el "Capullo", una superburbuja alrededor de estrellas de alta masa recién nacidas. Crédito (HAWC collaboration, 2021)

1.2 Planteamiento del problema

Dentro de la astrofísica una de las áreas más importantes es la de altas energías, la cual se plantea algunos de los problemas más desafiantes de la ciencia en la actualidad, por ejemplo, explicar los agujeros negros súper masivos presentes en el núcleo de galaxias activas, o cómo las partículas cargadas son aceleradas a energías extremadamente altas en ambientes astronómicos, el origen de enormes flujos de partículas de alta energía en galaxias activas, etc. Por lo que la astrofísica de altas energías hace factible el estudio de las propiedades de la materia bajo condiciones que aún no pueden ser reproducidas en un laboratorio. Esto hace que en muchos casos los problemas que se plantean solo puedan ser explorados de forma indirecta. Para intentar dar respuesta a estos problemas se construyen observatorios de altas energías y con los datos que obtienen se pretende comprender más sobre los diversos fenómenos del universo.

Los rayos cósmicos los cuales son detectados por estos observatorios, son partículas provenientes del espacio interestelar, que tienen un rango de energía de 10^9 eV hasta 10^{21} eV, estos representan el 99.9999% de la radiación que llega a la atmósfera, el resto que es un .00001% son rayos gamma, por lo que para identificarlos es necesario separarlos mediante algoritmos. La importancia de

identificar los rayos gamma reside en que estos permiten generar mejores mapas del cielo, los cuales ayudan a estudiar los fenómenos que provocan este tipo de radiación (Ilustración 4).

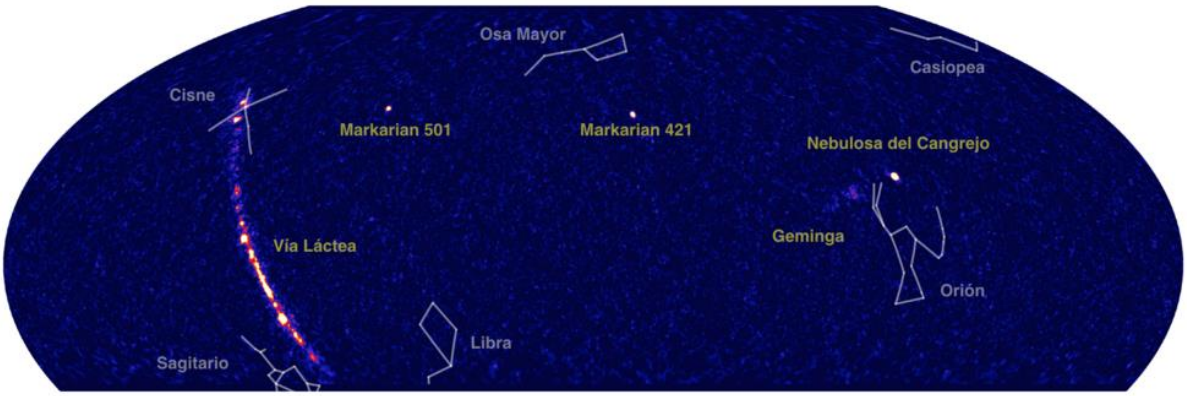


Ilustración 4. Mapa del cielo generado por HAWC, en donde se muestran las fuentes de rayos gamma (color amarillo y rojo) y su ubicación respecto a las constelaciones. Crédito (HAWC collaboration, 2021)

1.3 Justificación

Los rayos cósmicos (RC) llegan a la atmósfera en mayor cantidad que los rayos gamma (RG), por lo que es necesario desarrollar métodos que nos permitan separar uno de otro, el poder diferenciarlos ayudará a conocer la dirección de un rayo gamma con mayor exactitud y por lo tanto identificar la fuente de origen.

El rendimiento de un buen método de separación de RC y RG es aquel que es capaz de disminuir la cantidad de eventos que son clasificados como RG cuando en realidad son RC y aumentar la correcta selección de eventos producidos por RG.

Debido a lo anterior se propone implementar el modelo basado en una red neuronal para diferenciar los rayos gamma y rayos cósmicos. Se ha escogido este modelo debido a que anteriormente ya se ha trabajado con este método de clasificación en el observatorio HAWC (Capistrán, Torres, Altamirano, & Collaboration, 2015) (Capistrán, 2020), logrando buenos

resultados, por lo que al seguir explorando este método se podrían obtener más resultados en esta área.

1.4 Objetivo general

Mejorar la separación de RG y RC con el fin de aumentar la sensibilidad del observatorio HAWC en la detección de fuentes que generan RG, a través de la aplicación de redes neuronales y el uso de datos simulados de los eventos generados por los RG y RC en el entrenamiento, verificación y prueba de las redes neuronales, finalmente se comparará los resultados con el modelo de separación estándar que usa actualmente HAWC, lo cual ayudará a determinar si mejoró la detección de partículas.

1.5 Objetivos particulares

- Establecer uno o más modelos de datos que ayude en la separación de RC y RG.
- Seleccionar las variables de entrada que mejoren el desempeño en la separación de datos.
- Comparar los resultados arrojados por la red neuronal con los resultados de separación del modelo estándar usado por HAWC, para conocer si existe alguna mejora en la separación de RG y RC.

1.6 Detección de radiación Cherenkov en agua

Los observatorios Cherenkov en agua están compuestos por arreglos de detectores llenos de agua (WCDs) y dentro de ellos contienen tubos fotomultiplicadores (PMTs) que son dispositivos muy sensibles a la luz ultravioleta, visible e infrarrojo, que convierten a estas en señal eléctrica, utilizando el agua como medio dialectico para la detección de rayos gamma y rayos cósmicos (Ilustración 5).

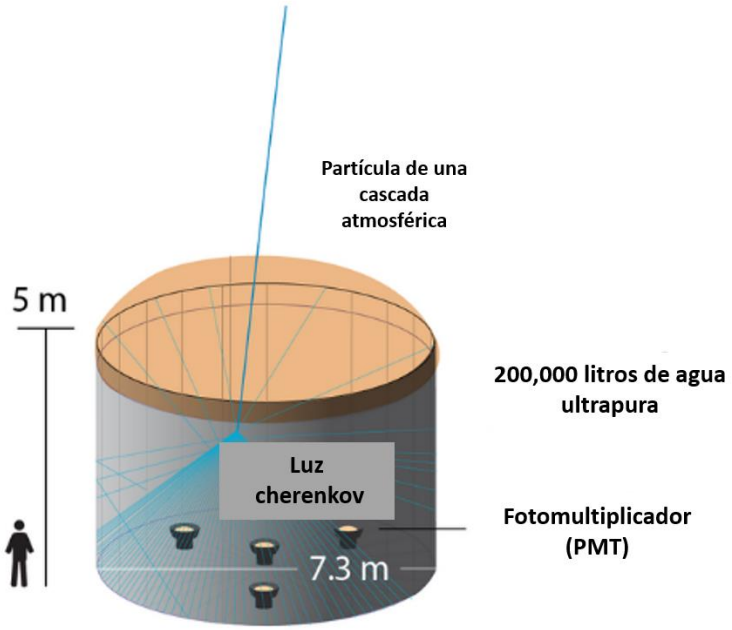


Ilustración 5. Ejemplo de un detector del observatorio HAWC, el cual contiene tubos fotomultiplicadores (PMTs). Crédito (HAWC collaboration, 2021)

El observatorio Haverah Park, fue el primer observatorio en utilizar la técnica de detección Cherenkov en agua, este observatorio se enfocaba en detectar rayos cósmicos y estaba compuesto por 15 WCDs y cinco piscinas, se considera el antecesor del observatorio MILAGRO.

El observatorio MILAGRO (Ilustración 6) consistía en una piscina llena de agua ultrapura y dentro de la piscina contenía PMTs, lo cual le permitía detectar rayos gamma (Ilustración 7), sin embargo la gran diferencia en la cantidad de rayos cósmicos respecto a los rayos gamma representaba un gran problema, para intentar mejorar la sensibilidad de detección de rayos gamma, desarrollaron un método de separación, el cual consistía en establecer un corte a la variable compactness = NB/PEMax (Atkins, 2003) donde NB es el número de PMTs con un pulso mayor a un umbral en fotoelectrones (PE) durante un evento y PEMax es la carga máxima en PE de un PMT en el evento, fuera de un radio de 40 m a partir del núcleo de los eventos. El valor de esta variable cuando se detecta un rayo gamma es grande, y el valor para un rayo cósmico es pequeño. La eficiencia de detección de rayos gamma usando este método es superior al 50% y su rechazo es del 20% según lo reportado por (Atkins, 2003).

Ilustración 6. Vista aérea del observatorio Milagro. Crédito Universidad de Maryland.

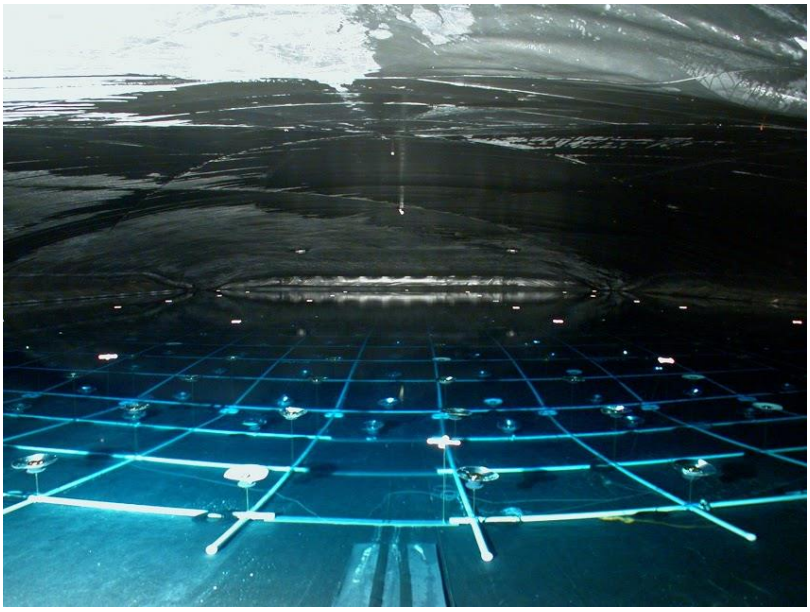


Ilustración 7. Vista del interior del observatorio milagro, en donde se observa la red de detectores. Crédito Universidad de Maryland.

El observatorio LHAA-SO construido a una altitud de 4410 metros en Daocheng, provincia de Sichuan, China (Ilustración 8), se centra en estudiar el cielo del norte en busca de fuentes de rayos gamma, en el trabajo realizado por (Wang, Liao, Zha, & Cao, 2019), se reporta que a través del uso de análisis multivariado separan las partículas de rayos gamma y rayos cósmicos, después sus resultados los comparan con el método estándar usado por el observatorio LHAA-SO, llegando a concluir que existe un incremento en la separación conforme se aumenta el número de variables usadas en los métodos multivariados, igualmente la energía es importante para la separación, ya que es más fácil separar partículas con mayor energía.

Ilustración 8. Vista aérea del observatorio LHAA-SO el cual cuenta con 3,120 detectores. Crédito IHEP

2. Marco teórico

2.1 Cascadas atmosféricas extendidas

Las cascadas atmosféricas extendidas (EAS) pueden ser clasificadas mediante la partícula que la originó: si es originada por un rayo gamma se le considera una cascada electromagnética, mientras que si es iniciada por un rayo cósmico es una cascada hadrónica.

En las cascadas electromagnéticas cuando un RG entra en contacto con la atmósfera se produce un par electrón-positrón, conforme aumenta la distancia, el electrón emite un fotón por medio de radiación Bremsstrahlung, la cual es un tipo de radiación producida por la desaceleración de una partícula cargada, este proceso sigue repitiéndose hasta que la energía de la partícula secundaria sea menor a la energía crítica la cual es de 85 MeV. (Ilustración 9) El número de partículas llega a un máximo y a partir de ese momento dejan de producirse partículas y comienza un proceso de absorción de partículas por la atmósfera (Matthews, 2005).

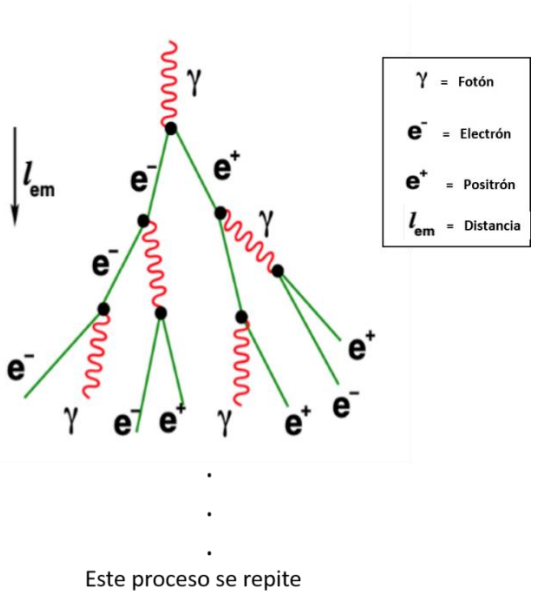


Ilustración 9. Representación de una cascada electromagnética, en donde los puntos negros representan la interacción entre el núcleo de aire y la partícula. En este tipo de cascada un fotón produce partículas secundarias como electrones, positrones y más

fotones. Crédito (Capistrán, Implementación de algoritmos para la optimización de detección de fuentes en el observatorio HAWC, 2020)

Por otro lado, las cascadas hadrónicas se forman cuando un RC colisiona con los átomos de la atmósfera, según (Matthews, 2005) en la primera interacción entre un RC y los átomos de la atmósfera se crean nuevas partículas; una parte de estas partículas creadas rápidamente decaen en dos fotones y estos fotones a su vez generan pequeñas cascadas electromagnéticas. Mientras que la otra parte de las partículas creadas, después de viajar cierta longitud en el aire, decaen y producen muones, que son partículas que tienen la misma carga eléctrica que los electrones, pero su masa es aproximadamente 200 veces mayor, igualmente se producen más partículas elementales llamadas hadrones, este proceso es continuo hasta que la energía de estas partículas sea menor a la energía crítica (85 Mev), una vez llegado a este punto continua la producción de partículas hadrónicas, pero en menor cantidad (Ilustración 10).

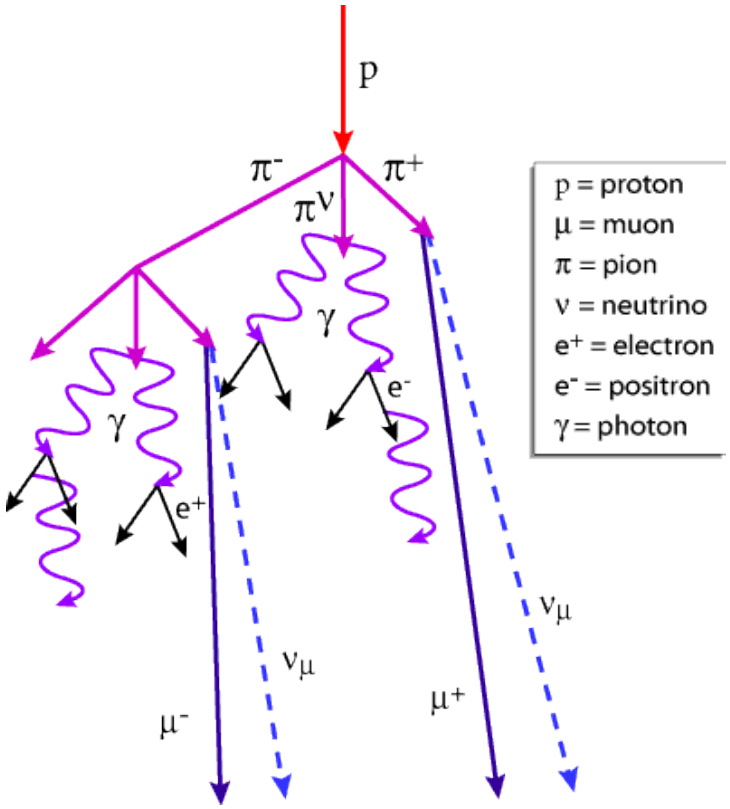


Ilustración 10. Representación de una cascada hadrónica, la cual se forma cuando un RC o protón colisiona con los átomos de la atmósfera, en la primera interacción entre un RC y los átomos de la atmósfera se crean nuevas partículas; una parte de estas partículas creadas rápidamente decaen en dos fotones y estos fotones a su vez generan pequeñas cascadas electromagnéticas.

Mientras que la otra parte de las partículas creadas, después de viajar cierta longitud en el aire, decaen y producen muones junto con más partículas elementales. Crédito (HAWC collaboration, 2021)

2.1.1 Características de las EAS

Las características de las cascadas atmosféricas (EAS) son importantes para su estudio, ya que nos permiten identificar con mayor claridad la partícula fuente que las ocasiona, algunas de las características más importantes son:

- *Eje de la cascada*

  Al momento de colisionar una partícula en la atmósfera este tiene una dirección, la prolongación de esta dirección hasta un plano se le conoce como eje de la cascada.

- *Núcleo de la cascada*

  Por lo general una cascada atmosférica tiene un centro, que es donde existe un máximo del número de partículas, a este máximo se le conoce como núcleo y las cargas van disminuyendo gradualmente conforme se alejan del núcleo.

- *Plano de la cascada*

  En la Ilustración 11 se muestra el plano de la cascada la cual tiene forma de un disco, las puntas de este disco se curvean y ensanchan ligeramente.

2.2 Rayos gamma

Los rayos gamma (RG) es radiación electromagnética, los cuales se pueden comportar como partículas o como ondas, los RG no pueden penetrar la atmósfera debido a que es muy opaca, para poder detectarlos directamente solo puede ser a través de satélites especializados fuera de la atmósfera o de forma indirecta a través de observatorios en la tierra que son capaces de

detectar las interacciones que producen los rayos gamma con la atmósfera (Flynn, 2010) (Ilustración 12).

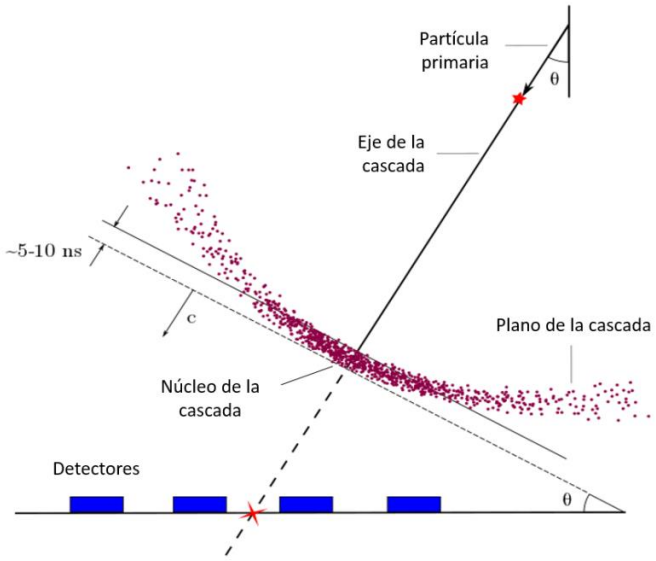


Ilustración 11. Se muestra una cascada atmosférica extendida (EAS) con algunas de las características más importantes como los son el eje, núcleo y plano de la cascada. Crédito (HAWC collaboration, 2021).

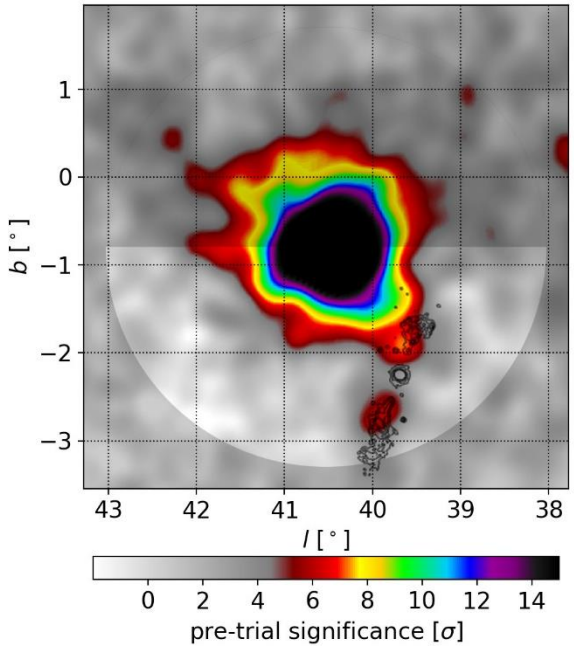


Ilustración 12. Rayos gamma registrados por HAWC desde la nebulosa MGRO J1908+06. Crédito (HAWC collaboration, 2021).

2.3 Rayos cósmicos

Actualmente se sabe que los rayos cósmicos están compuestos por todos los tipos de núcleos atómicos, desde el núcleo de hidrógeno hasta núcleos más pesados como el hierro, el espectro de energía de los rayos cósmicos ha sido medido hasta 10^{21} eV, a las más altas energías los rayos cósmicos tienen aproximadamente la misma energía que una pelota de tenis golpeada con fuerza, con la diferencia de que esta energía se encuentra contenida en un solo núcleo atómico. Los modelos actuales predicen que los rayos cósmicos con energía mayor a 10^{15} son acelerados en fuentes extra galácticas como los núcleos activos de galaxias (HAWC collaboration, 2021).

2.4 Efecto Cherenkov

El efecto Cherenkov se genera cuando una partícula cargada viaja más rápido que la velocidad de la luz en un medio, como lo es el agua o el aire. Esta velocidad está determinada por

el índice de refracción del medio, en el aire a una presión y temperatura estándar el índice es de 1.0003 y en el agua es de 1.33 (Stanev, 2004).

Al iniciarse el efecto Cherenkov se emite un cono de luz alrededor de la trayectoria de la partícula. En donde el ángulo de apertura del cono depende del índice de refracción del medio, en el agua el ángulo es de 41° y en el aire de 1° (Ilustración 12).

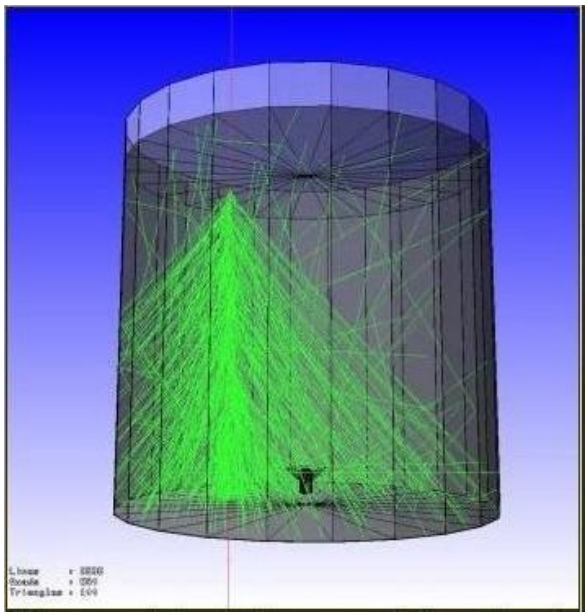


Ilustración 13. Simulación del paso de una partícula cargada a través de un tanque de agua y el cono de luz alrededor de la trayectoria (HAWC collaboration, 2021).

2.5 Observatorio HAWC

El observatorio a gran altura Cherenkov en agua, o HAWC por sus siglas en inglés, es un laboratorio diseñado para detectar rayos gamma y rayos cósmicos, cuenta con una apertura que cubre más del 15% del cielo. Con su amplio campo de visión, el observatorio está expuesto a dos terceras partes del cielo durante cada ciclo de 24 horas. El rango de detección de la energía en HAWC es de 300 GeV[2] y más de 300 TeV[3] (Collaboration H. , 2020) (Ilustración 14).

[2] Gigaelectronvolt
[3] Teraelectronvolt

Ilustración 14. Vista de observatorio HAWC ubicado dentro del parque nacional Pico de Orizaba, México. Crédito (HAWC collaboration, 2021).

2.5.1 Tanques de agua Cherenkov

Para registrar el paso de las partículas creadas en cascadas atmosféricas producidas por RC y RG, el detector HAWC usa el método Cherenkov en agua, con esta técnica, el detector es usado para muestrear las partículas de la cascada atmosférica al nivel de la superficie de la tierra, detectando la luz Cherenkov producida cuando las partículas de las cascadas atmosféricas pasan a través de los tanques llenos con agua ultrapura.

Los detectores Cherenkov en agua de HAWC están hechos de tanques de láminas de acero corrugado con una altura de 4 metros y con un diámetro de 7.3 metros. Cada tanque tiene por dentro una bolsa que contiene el agua y cuatro tubos fotomultiplicadores (PMTs) que son sensibles a las longitudes de onda en el rango ultravioleta (Ilustración 15).

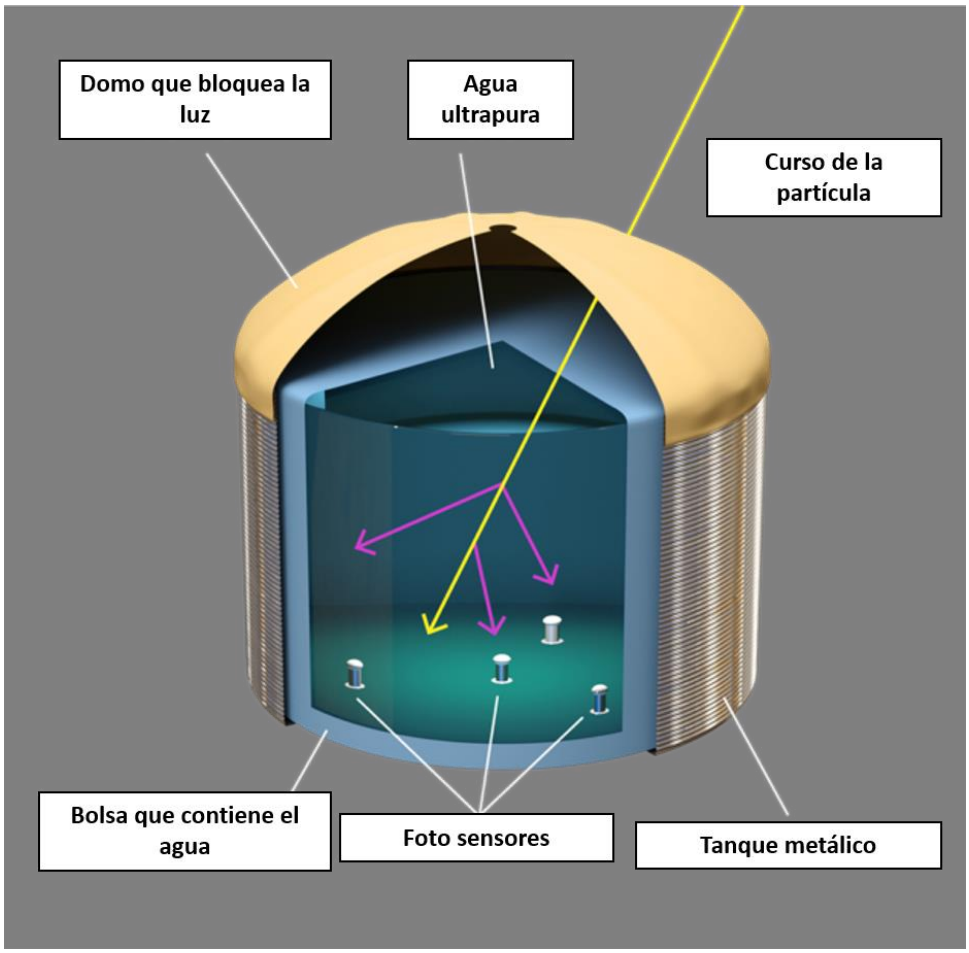


Ilustración 15. Se muestra un tanque Cherenkov en agua, el cual tiene el objetivo de registrar el paso de las partículas creadas en cascadas atmosféricas producidas por RC y RG. Crédito Universidad de Rochester.

2.5.2 Detección de partículas

La producción de luz Cherenkov es extremadamente eficiente dentro del agua debido a su alto índice de refracción. La luz Cherenkov se emite en un cono frontal que rodea la dirección de movimiento de la partícula cargada. El ángulo de apertura del cono depende del índice de refracción del medio. Debido a que el cono de Cherenkov en el agua es tan grande, casi todas las partículas cargadas que entran en el tanque deben de ser observadas por lo menos por uno de los cuatro PMTs.

Debe notarse que los tanques Cherenkov en agua pueden usarse para detectar cascadas producidas por rayos gamma y rayos cósmicos. Dado que el agua en los tanques es densa, un rayo gamma producirá un par electrón/positrón una vez entre en el tanque. Estas partículas cargadas emitirán entonces radiación Cherenkov en su paso por el agua la cual será detectada por los PMTs (HAWC collaboration, 2021).

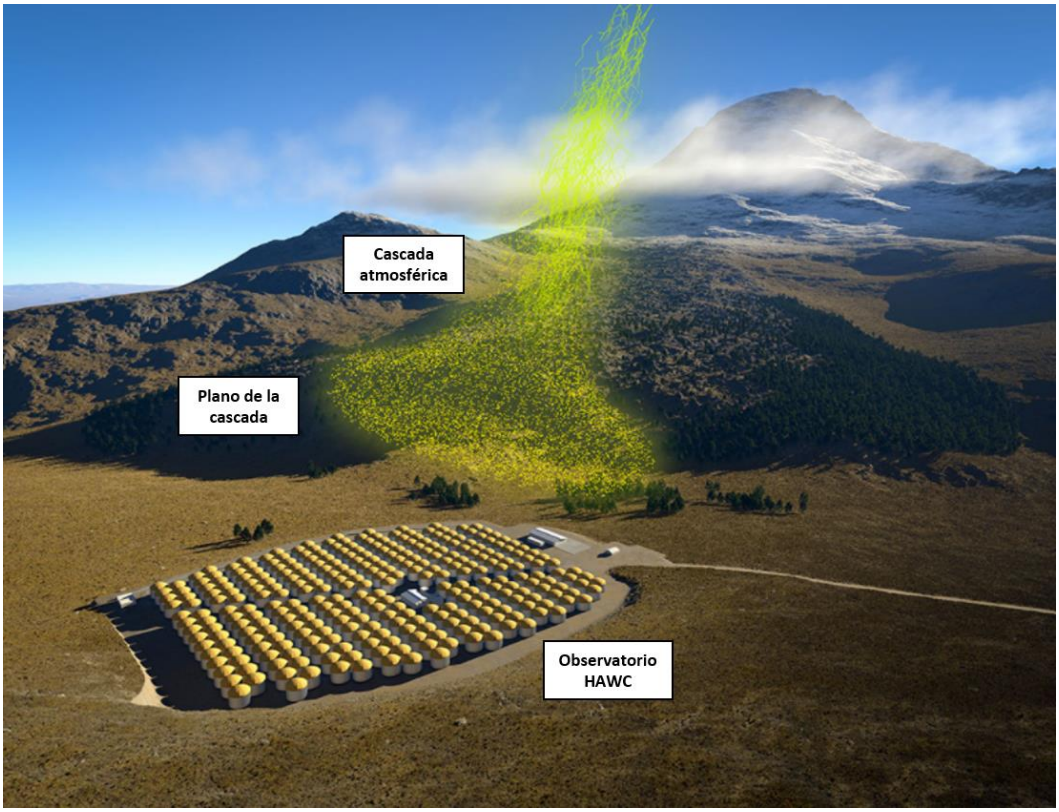


Ilustración 16. Simulación de una cascada atmosférica cayendo sobre el observatorio HAWC. Crédito Universidad de Rochester.

2.5.3 Outriggers

Los outriggers es una serie de detectores más pequeños que se encuentran alrededor de los detectores principales de HAWC, compuestos por tanques de polietileno de 2 metros de diámetro, cada uno de estos tanques está lleno de agua ultrapura y equipado con un PMT en la parte inferior (Ilustración 17), se implementaron con el objetivo de detectar partículas secundarias que llegan

fuera del arreglo principal de detectores de HAWC, se estima que aumenta diez veces la sensibilidad en la detección de rayos gamma por encima de 10 TeV (Capistrán, Torres, Moreno, & Collaborator, 2017).

Ilustración 17. Se muestran los outriggers instalados alrededor del arreglo principal de detectores de HAWC. Crédito (HAWC collaboration, 2021).

2.5.4 Separación de rayos gamma / rayos cósmicos en HAWC

Para poder ser capaces de observar rayos gamma con alta sensibilidad, los datos obtenidos por HAWC deben ser filtrados de eventos debidos a rayos cósmicos. Los rayos cósmicos pueden ser discriminados de los rayos gamma al observar el patrón de PMTs con señales en el detector (es decir, el "perfil" de la cascada atmosférica). Las cascadas de rayos gamma tienden a tener un perfil que disminuye radialmente desde el centro de la cascada, en contraste con los perfiles de las cascadas iniciadas por rayos cósmicos que son relativamente más aleatorias y presentarán grumos en el patrón de PMTs con señales (HAWC collaboration, 2021) (Ilustración 18).

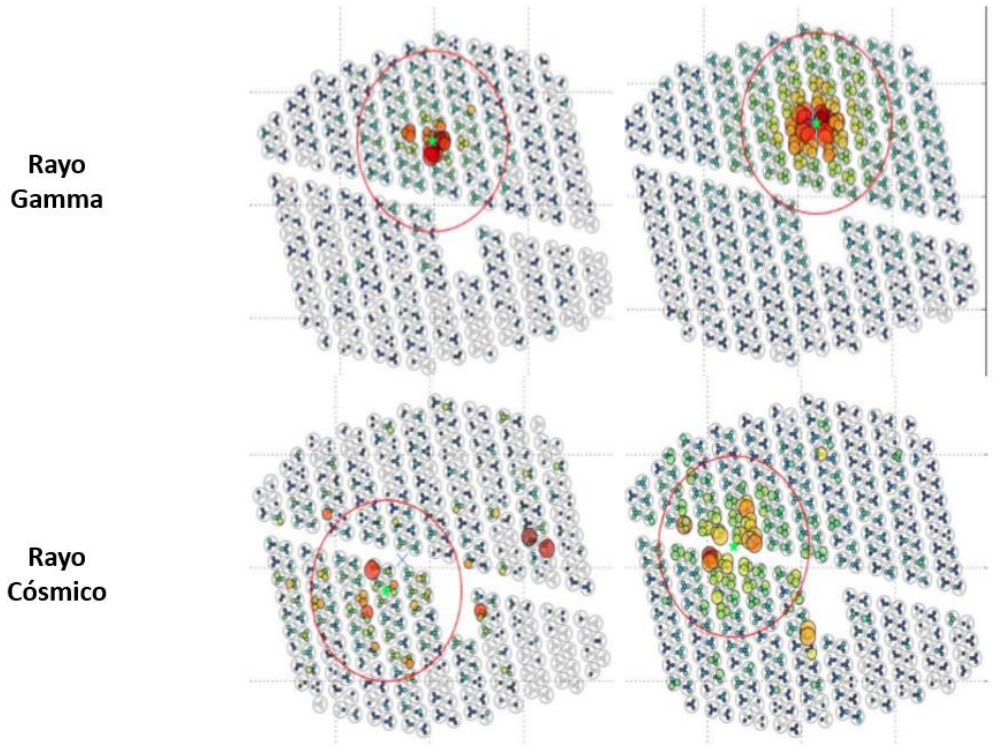


Ilustración 18. Vista del perfil de la cascada de un rayo gamma (superior) comparada con el perfil de un rayo cósmico(inferior). Crédito HAWC

Usando simulaciones de cascadas producidas por rayos gamma y rayos cósmicos, se ha estimado la habilidad del observatorio para rechazar cascadas iniciadas por rayos cósmicos en los datos de HAWC. En base a los patrones de PMTs con señales que se observan en las simulaciones, se ha encontrado que se puede rechazar >99% de las cascadas de rayos cósmicos con energías ligeramente superiores a 3 TeV usando una selección en el perfil de las cascadas. También se ha encontrado que el desempeño de la discriminación mejore con la energía, conforme más y más información (PMTs con señales) está presente en los datos con mayores energías (Ilustración 19).

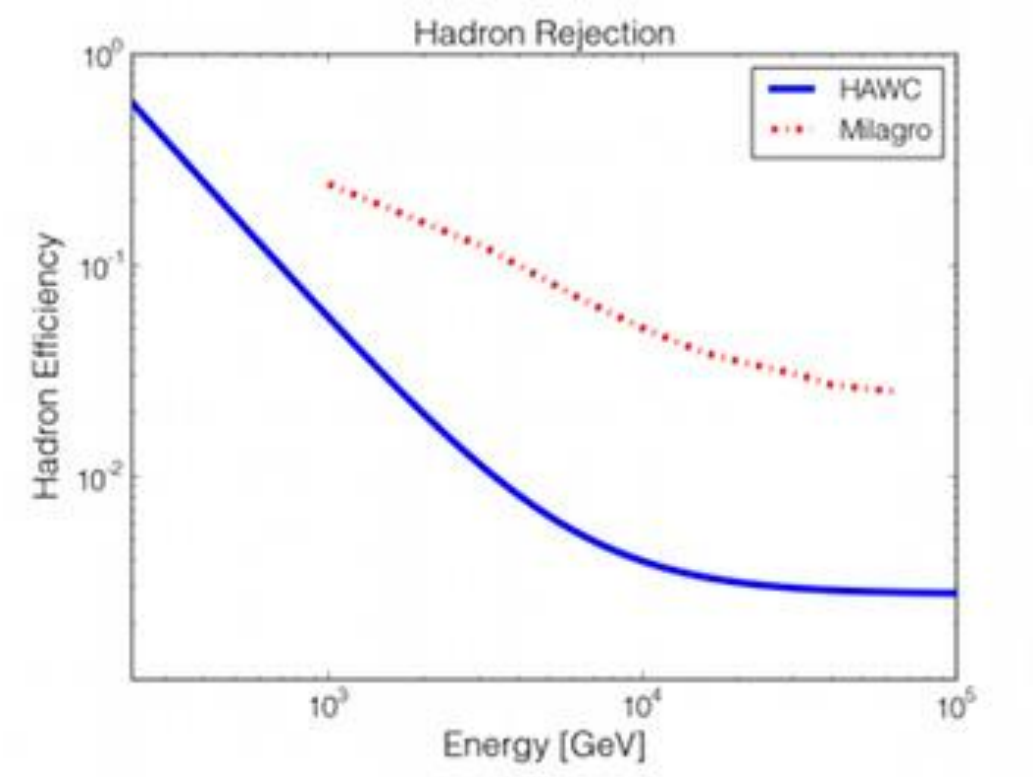


Ilustración 19. Se muestra la comparación en el poder de discriminación de rayos cósmicos entre los observatorios HAWC y Milagro, en donde a mayores energías la capacidad de discriminación es mayor. En línea azul se observa que la eficiencia en hadrón, la cual es el número de rayos cósmicos que se clasificaron como rayos gamma, en HAWC es mejor que la de Milagro, pues se busca que esta cada vez sea más baja.

2.5.5 Resolución angular

Casi tan importante como la separación gamma/hadrón es la resolución angular del detector. La resolución angular se define como la incertidumbre típica hecha al reconstruir la dirección de llegada de una cascada atmosférica. Todos los detectores tienen una resolución angular finita, que tiene como efecto el que las características de las fuentes no se puedan conocer con precisión infinita.

Lo que se desea es mantener a la resolución angular tan pequeña como sea posible. Como es de esperarse la ventaja es que de este modo es posible observar fuentes astronómicas pequeñas; además la resolución también afecta la sensibilidad del detector a fuentes puntuales. Esto se debe

a que las fuentes puntuales que se observan con cualquier experimento contendrán una mezcla de "señal" (eventos de la fuente) y "fondo o ruido" (eventos que no son de la fuente). Conforme la resolución angular disminuye, el cociente entre el fondo y la señal disminuirá. Esencialmente la señal permanecerá constante, porque se trata de una fuente puntual, mientras que el fondo disminuirá (HAWC collaboration, 2021).

2.5.6 AERIE

Aerie es una plataforma que ayuda a procesar, manejar y trabajar los datos en HAWC, cuenta con las herramientas básicas para el manejo de las simulaciones y la creación de mapas del cielo mediante el uso de los datos reales generados por el observatorio, igualmente se utiliza para generar simulaciones de eventos producidos por rayos gamma y rayos cósmicos. El software está estructurado como un conjunto de proyectos interdependientes en C++ y unidos por un núcleo central. El núcleo permite un ciclo de ejecución para analizar conjuntos de datos, clases para almacenar datos simulados y bibliotecas para manejar tareas de geometría, coordenadas astronómicas o conversiones de tiempo. Igualmente contiene otros proyectos para realizar tareas más especializadas, como la reconstrucción de mapas, aerie se puede ejecutar con C++ o con scripts de Python.

2.5.7 XCDF

El formato de datos explícitamente compactado, en inglés eXplicitly-Compacted Data Format (XCDF), es un formato de datos binarios diseñado para guardar los campos con una exactitud específica definida por el usuario. Utiliza un empaquetamiento de bit para almacenar los campos con una precisión dada, para el conjunto de valores dados y proporciona una compresión sustancial.

El empaquetamiento de bit es el proceso de compresión de datos para reducir el número de bits necesarios para representar los valores de los datos, es decir, empaqueta estos bits en serie por cada dato, ignorando el límite de la palabra digital.

Los archivos XCDF almacenan un conjunto de campos que tienen un índice en común, cada índice se refiere a un evento, cada campo tiene: un nombre, una resolución y el tipo de la variable (enteros positivos, enteros o puntos flotantes). Los datos son escritos dentro de cada campo y los eventos son escritos en el archivo (HAWC Collaboration, 2018).

2.5.8 Datos simulados

La simulación de los eventos en HAWC se realiza mediante el software CORSIKA (Cosmic Ray Simulation Kascade), el cual simula el proceso de creación de cascadas atmosféricas, en donde las partículas de RG Y RC al pasar a través de la atmósfera y al colisionar con los núcleos de aire producen otras partículas, este proceso se repite de forma continua, creando el efecto cascada. (Ilustración 20)

Para que el programa simule las cascadas se debe establecer la configuración inicial, en donde se especifican el número de cascadas a simular, cual es el tipo de partícula primaria, en que rango de energías trabajara, entre otras opciones más. Al terminar, el programa generará la simulación indicada.

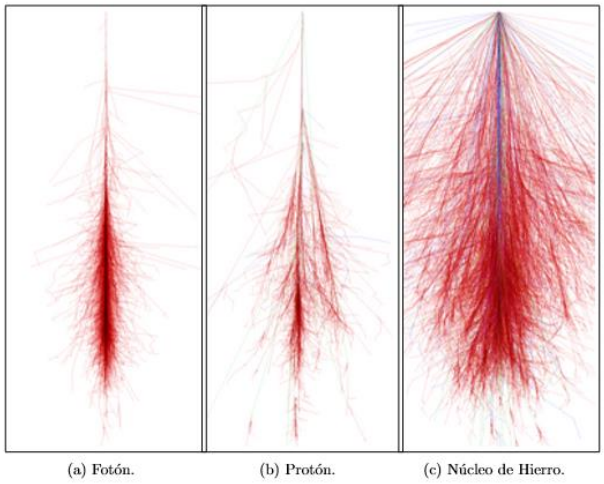


Ilustración 20. Representación de las simulaciones de cascadas atmosféricas producidas por un fotón, protón y un núcleo de Hierro. Crédito HAWC

Finalizada esta primera etapa, ahora con el uso del software GEANT4 se simula el proceso a través del cual las partículas secundarias pasan a través de la materia, en el caso de HAWC, es a través de los WCDs hasta llegar a los PMTs, al terminar esta segunda simulación, se procede a simular la respuesta del PMT, por último, esta respuesta del PMT se reconstruye, para así poder obtener las variables de los eventos.

2.5.9 Variables candidatas

De las variables que se obtienen de las simulaciones de eventos, algunas son relevantes dado que pueden ayudar en la separación de RG y RC, a continuación, se describirá cada una de estas variables candidatas.

- LIC

  Dado que los muones se generan mayormente en las cascadas hadrónicas, la variable LIC se encarga de buscar estas partículas, para ello debe localizar la carga máxima depositada en un PMT fuera de un círculo de 40 m a partir del núcleo, este valor es dado por la variable CxPE40. Si el valor de esta variable es alto se puede considerar como un muon, debido a que los muones por lo general depositan su carga a cierta distancia del núcleo, si se cumple esta condición se le puede considerar un RC, de lo contrario se consideraría que es un RG. Igualmente, LIC usa otra variable la cual es nHitSP20 esta representa el número de PMTs con señal en un rango de 20 nanosegundos del plano de la cascada. Se utiliza una escala Log_{10} en ambas variables para restringir el valor de la variable LIC.

$$LIC = Log_{10}\left(\frac{CxPE_{40}}{nHitSP20}\right) \quad (1)$$

- PINC

Esta variable mide la distribución de la carga alrededor del núcleo de la cascada, por lo que, en un RG, las cargas por lo general se concentran en el núcleo y disminuyen de forma suave, mientras que en los RC las concentraciones de cargas son irregulares.

- LogNNEnergy

  Es una variable que usando el método de redes neuronales permite estimar la energía de la partícula primaria.

- dismax

  Con frecuencia los muones en una cascada hadrónica se producen lejos del núcleo en comparación con la cascada electromagnética por lo que esta variable permite localizar los PMTs con las cargas más altas y calcula la distancia entre estos PMTs, finalmente la distancia máxima será el valor de dismax.

- LDFChi2

  Para entender LDFChi2 primero se debe definir lo que es distribución lateral, esta fue derivada por Nishimura, Kamata y Greisen (NKG) y es una aproximación teórica que proporciona la densidad de la partícula cargada como función de la distancia r dependiendo de la edad s de la cascada. La expresión está dada por:

$$p(r) = \frac{N(S)}{R_M^2}\, f(r) \left(\frac{r}{R_M}\right)^{S^{-2}} \left(1 + \frac{r}{R_M}\right)^{S^{-4,5}}, f(r) = \frac{\Gamma(4{,}5 - S)}{2\pi\Gamma(S) - \Gamma(4{,}5 - 2_S)} \qquad (2)$$

donde N(s) es el número total de partículas cargadas (Atreidis, 2017).

En las cascadas producidas por RG su distribución lateral es homogénea, mientras que en las cascadas producidas por RC la distribución es caótica, la variable LDFChi2 permite saber el grado de ajuste de la distribución lateral de los eventos.

- LDFamp

  La variable LDFamp permite medir la amplitud del ajuste de la distribución lateral.

- fhit

  La variable fhit se compone de dos variables la primera es nHitSP20 la cual define el número de PMTs con señal durante el evento, la cual se divide por nChavail que es el número de PMTs en funcionamiento durante la detección. Mientras mayor sea el valor de esta variable, esto nos indicara que más PMTs fueron activados durante el evento.

$$fhit = nHitSP20/nChavail \quad (3)$$

- LDFAge

  Es un parámetro que provee el mejor ajuste de la distribución lateral del evento.

- sos

  Esta variable nos permite diferenciar un RG de un RC a través de la medición de las cargas de los PMTs vecinos del núcleo de la cascada.

- LLH

  Es un estimador de energía que se ajusta a un modelo de RC y RG para cada evento.

2.5.10 Binning

La forma en que el observatorio HAWC divide los datos es mediante 10 grupos también llamados Binning, los cuales representan la fracción de PMTs activados en cada evento, esto se representa a través de la variable fhit, en donde fhit también se puede entender como el porcentaje de PMTs que detecta señal durante el evento.

Igualmente tomando en cuenta el núcleo de la cascada sección 2.1.1 se define si el evento cayó dentro de los WCDs, para ello se usa la variable rec.coreFiduscale, si el valor de esta variable es menor a 100 entonces el núcleo cayó dentro del arreglo de WCDs y si es mayor a 100 y menor a 150, entonces se considera que el núcleo cayó fuera del arreglo.

Tabla 1.

Rangos de fhit

Dentro-Fuera			
fhit	Rango		
0	2.7%	-	4.7%
1	4.7%	-	6.8%
2	6.8%	-	10.4%
3	10.4%	-	16.1%
4	16.1%	-	24.5%
5	24.5%	-	35.1%
6	35.1%	-	47.2%
7	47.2%	-	59.9%
8	59.9%	-	72.2%
9	72.2%	-	82.8%
10	82.8%	-	100.0%

La tabla mostrada representan la forma en que se dividen los datos en HAWC, utilizando las variables fhit y rec.coreFiduscale, en donde fhit define el porcentaje de detectores activados por el evento y rec.coreFiduscale indica si el evento fue dentro o fuera del arreglo de detectores.

2.5.11 Método estándar de separación de partículas

La manera en que normalmente HAWC separa las partículas de rayos gamma y rayos cósmicos, es mediante el uso de dos variables: LIC y PINC. En donde LIC es una variable enfocada en buscar la presencia de muones en las cascadas atmosféricas y PINC es una variable que cuantifica la suavidad de la función de distribución lateral, es decir en un rayo gamma las cargas se concentran alrededor del núcleo y conforme se desarrolla la cascada, las cargas van disminuyendo gradualmente, mientras que los rayos cósmicos tienen diferentes concentraciones de cargas, lo que genera una disminución de cargas más irregular desde el centro de la cascada (Ilustración 18).

Entonces para que una partícula sea reconocida como rayo gamma estas dos variables deben cumplir la siguiente condición:

$$LIC < CutLIC \;\&\&\; PINC < CutPINC \qquad (4)$$

LIC: variable enfocada en buscar la presencia de muones

CutLic: Número óptimo para LIC

PINC: variable que cuantifica la suavidad de la función de distribución lateral.

CutPINC: Número óptimo para PINC

Si la condición se cumple entonces es catalogado como rayo gamma en caso contrario como rayo cósmico.

2.6 Aprendizaje automático.

El aprendizaje automático es una rama de estudio que le da la habilidad a las computadoras de aprender sin ser explícitamente programadas. Arthur Samuel, 1959.

Otra definición más técnica puede ser:

Se dice que un programa de computadora aprende de experiencia E, con respecto a alguna tarea T y alguna medida de desempeño P, mejora con experiencia E, si su desempeño en T, como fue medido con P, mejora con experiencia E. Tom Mitchell, Carnegie Mellon University.

Un filtro de spam en los correos se puede considerar un programa de aprendizaje automático ya que puede aprender de los ejemplos de spam y de email que se le da. En este caso la tarea T es reconocer el spam, la experiencia E son los datos de entrenamiento, y la medida de desempeño P se especifica, por ejemplo, podría ser la taza de spam correctamente clasificados. Esta medida de desempeño se le llama exactitud.

Por lo que usar aprendizaje automático es bueno para resolver problemas que requieren una larga lista de reglas, entonces el aprendizaje automático usualmente simplifica el código y el desempeño mejora, igualmente es bueno usar aprendizaje automático en ambientes fluctuantes, ya que un buen sistema de aprendizaje automático puede adaptar los datos, por último, también podemos obtener descubrimientos de un volumen grande de datos.

Un método de aprendizaje automático construye un modelo el cual ayudará en la tarea de clasificación que se le indique, para la construcción del modelo, se realizan tres etapas las cuales son entrenamiento, verificación y prueba.

En la primera etapa, la de entrenamiento, al algoritmo se le da un conjunto de datos llamado set de entrenamiento, el cual contiene un número de atributos con sus respectivas instancias, así como la clase a la que pertenece, entonces el algoritmo estudia los atributos con el fin de reconocer y predecir el tipo de clase a la que pertenece, dado que no es una rutina programada, se puede decir que el algoritmo aprende de los datos. En la etapa de verificación se usa un set de datos diferente al de entrenamiento para verificar su desempeño, así como identificar si el algoritmo no está cayendo en un sobre entrenamiento, el cual significa que el algoritmo en lugar de aprender sobre los datos, está memorizando los datos, y esto generaría que, al ingresar nuevos datos, la clasificación se realice de forma errónea. Finalmente, en la etapa de verificación se usa un set de datos diferente a los anteriores el cual tiene el objetivo de evaluar el modelo final respecto a los resultados esperados.

Una de las herramientas más utilizadas en el aprendizaje automático son las redes neuronales, dado que son modelos matemáticos que intentan reproducir las neuronas del cerebro,

se busca que el modelo aprenda de los datos que se le dan y al usar un nuevo conjunto de datos pueda predecir la clase a la que pertenecen.

2.6.1 Redes Neuronales

Las redes neuronales son una de las herramientas más importantes del aprendizaje automático, ya que son versátiles, escalables y potentes, lo que las hace ideales para abordar problemas altamente complejos, ya sea para clasificar miles de millones de imágenes, potenciar servicios de reconocimiento de voz o recomendar los mejores videos a los usuarios de Youtube.

Warren McCulloch y Walter Pitts propusieron un modelo muy simple de la neurona biológica, que más tarde se conoció como neurona artificial: Tiene una o más entradas binarias (on/off) y una salida binaria. La neurona artificial activa su salida cuando más de un cierto número de entradas están activas. Bajo este modelo simplificado es posible construir una red de neuronas artificiales que computen cualquier proposición lógica. (Ilustración 21)

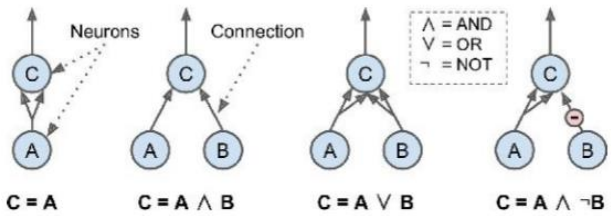


Ilustración 21. Se muestran cuatro modelos simples de neurona artificial, en donde A y B representan las neuronas de entrada y C representa la neurona de salida, cada modelo necesita de diferentes opciones para activar la neurona C. Fuente. (Géron, 2017).

2.6.2 MLP

MLP es un tipo de arquitectura de red neuronal que se compone de una capa de entrada (de paso), una o más capas ocultas, y una capa final llamada capa de salida (Ilustración 22). Cada capa, excepto la capa de salida, incluye una neurona de polarización y está completamente conectada a

la siguiente capa. Cuando una red neuronal tiene dos o más capas ocultas, se denomina red neuronal profunda (Géron, 2017).

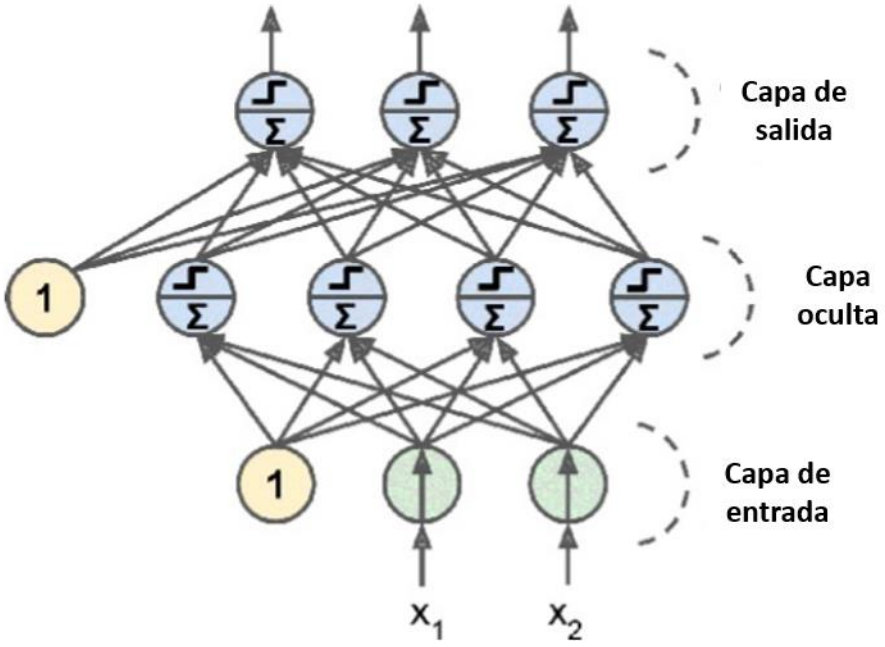


Ilustración 22. Ejemplo de arquitectura MLP, en donde se muestra la capa de entrada, la capa oculta y una capa de salida, el 1 representa la neurona de polarización, x1 y x2 representan los datos de entrada. Fuente. (Géron, 2017)

El proceso de aprendizaje de una MLP es: para cada instancia de entrenamiento, el algoritmo primero hace una predicción, mide el error, luego pasa por cada capa en reversa para medir la contribución de error de cada conexión y finalmente ajusta ligeramente los pesos de conexión para reducir el error (Géron, 2017).

Para entrenar una MLP se usan diferentes parámetros que a continuación se enlistan con más detalle:

- Número de ciclos: es el número de veces que se usará el set de datos de entrenamiento, para que este sea adaptado a los parámetros indicados y se pueda reducir el error.
- Número de capas ocultas: Es el número de capas a usar, en cada capa se tienen las neuronas artificiales.

- Función de transferencia: Representa simultáneamente la salida de la neurona y su estado de activación.
- Función sináptica: Representa la combinación de la información de entrada de las neuronas
- Método de entrenamiento: Este método o algoritmo tiene el objetivo de minimizar el error mediante la modificación de los pesos de la neurona.
- Taza de aprendizaje: Se utiliza para poder obtener una solución óptima en un tiempo de cómputo ideal, si este indicador es alto eso significa que el entrenamiento requiere menos tiempo de cómputo y si es bajo significa que el tiempo de cómputo será muy elevado.

2.7 Evaluación.

Para poder evaluar la correcta separación de rayos gamma y rayos cósmicos se pueden utilizar diferentes parámetros para conocer el resultado de los modelos utilizados:

- Eficiencia en gamma: Es el número de eventos tipo gamma que se clasificaron correctamente como gamma.
- Eficiencia en hadrón: Es el número de rayos cósmicos que se clasificaron como rayos gamma o el número de rayos cósmicos que no se clasificaron como rayos cósmicos.
- Especificidad: Es el número de rayos cósmicos que se clasificaron como rayos cósmicos
- Exactitud: es el porcentaje que se clasificó correctamente de la señal y el ruido.

2.7.1 Factor Q

Es un parámetro que nos indica la relación entre RG y RC, que viene dada por la fórmula

$$Q = \frac{\epsilon gam}{\sqrt{\epsilon had}} \qquad (5)$$

En donde

ϵgam: es la eficiencia en RG

ϵhad: es la eficiencia en RC

Por lo general se usa este parámetro para mostrar el punto en donde mejor se está separando los RG de los RC, pues al comparar distintos factores Q, aquel con el valor más alto, indica que ahí se tiene una mejor separación.

2.7.2 Curvas ROC:

Es una técnica de visualización, organización y selección de clasificadores basándose en su rendimiento (costo-beneficio), muy usado en medicina y en el diagnóstico de sistemas, últimamente se ha incrementado su uso en machine learning (Fawcett, 2006).

Dado un clasificador y una instancia, hay cuatro posibles resultados. Si la instancia es positiva y se clasifica como positiva, se cuenta como un verdadero positiva; si está clasificado como negativa, se cuenta como falso negativa. Si la instancia es negativa y se clasifica como negativa, se cuenta como verdadera negativa; si se clasifica como positiva, se cuenta como falsa positiva (Ilustración 23).

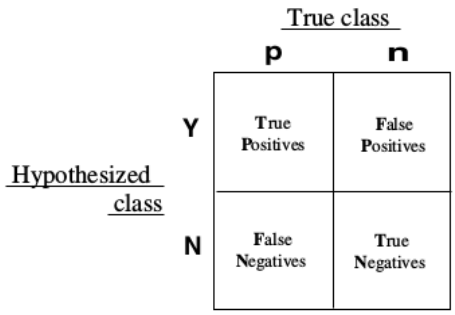


Ilustración 23. Tabla de confusión. Fuente (Fawcett, 2006).

En una gráfica de curvas ROC (Ilustración 24) el eje X representa la eficiencia en hadrones (falsos positivos) y el eje Y, la eficiencia en gamma (verdaderos positivos). Los puntos representan el resultado de cinco diferentes clasificadores, donde el ideal es el punto D.

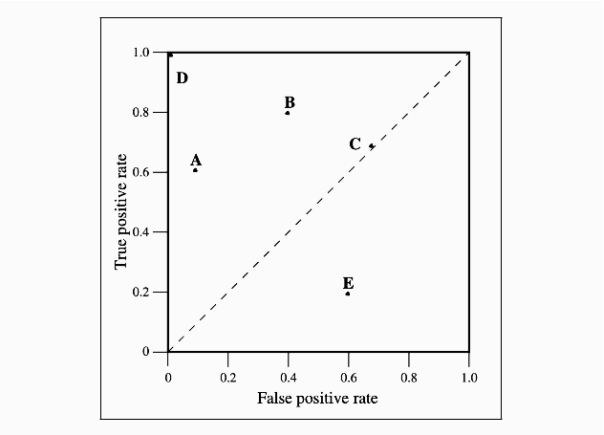


Ilustración 24. Resultado de cinco diferentes clasificadores, en donde la línea punteada representa un clasificador con resultados al azar, es decir, dado que los resultados son al azar se esperaría obtener la mitad de verdaderos positivos y la mitad de falsos positivos, por lo que esta línea nos permite tener una comparación para nuestros clasificadores, si realmente serían mejor que un clasificador aleatorio. En este ejemplo el clasificador D tiene el mejor desempeño pues tiene un alto porcentaje de verdaderos positivos y al mismo tiempo un bajo porcentaje de falsos positivos por otro lado, el clasificador E, no tiene un buen desempeño pues el porcentaje de verdaderos positivos es muy bajo, mientras que el porcentaje de falsos positivos es alto. Fuente. (Fawcett, 2006)

Un clasificador necesita de un umbral para producir un clasificador discreto (binario): si la salida del clasificador está por encima del umbral, el clasificador produce un valor positivo, de lo contrario un valor negativo. Cada valor umbral produce un diferente punto en el espacio ROC (Ilustración 25).

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

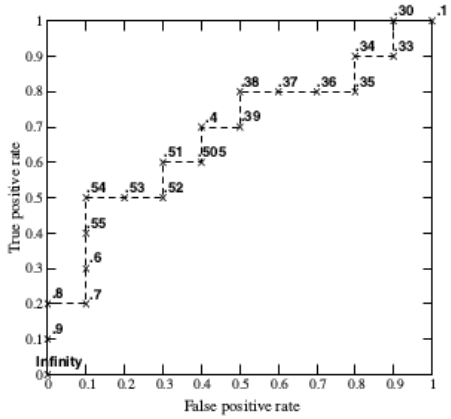


Ilustración 25. Curva ROC creada por las puntuaciones dadas por un clasificador en un set de datos. La tabla a la izquierda muestra 20 datos y la puntuación asignado por el clasificador. La grafica a la derecha muestra la curva ROC generada por cada puntuación conforme el umbral disminuye de .9 a .1 (Fawcett, 2006).

La Ilustración 26 muestra las áreas bajo dos curvas ROC, A y B. El clasificador B tiene mayor área y, por lo tanto, mejor promedio de actuación.

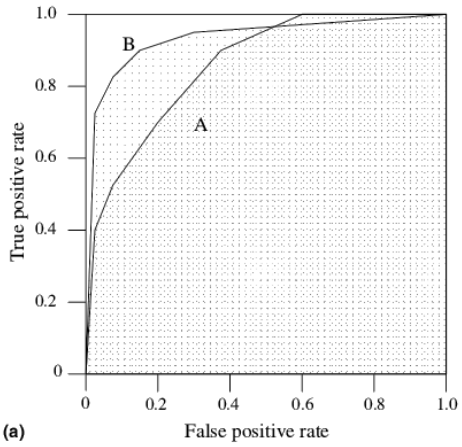


Ilustración 26. Se muestra el área bajo dos curvas ROC, en donde el eje X representa los falsos positivos y el eje Y los verdaderos positivos. Fuente (Fawcett, 2006).

3. Metodología

Para la realización del presente trabajo se emplearon datos de las simulaciones de las cascadas de rayos cósmicos (RC) y rayos gamma (RG), como segundo paso se realizaron los códigos para la correcta selección de eventos a utilizar, después se seleccionaron las variables que más ayudaron en la separación de datos. Una vez teniendo los datos necesarios se realizaron los entrenamientos de la red neuronal con el fin de poder separar los eventos, finalmente se evaluó el rendimiento de la red neuronal.

3.1 Datos

En el presente trabajo solo se usaron datos simulados de RG y RC para los entrenamientos de la redes neuronales, dado que al usar datos reales el resultado podría ser inferior según lo analizado en el trabajo de: (Capistrán, 2020), igualmente el obtener un set de datos reales de RG implica más tiempo de cómputo dado que la cantidad es muy pequeña comparada con los RC.

3.1.1 Datos simulados

Como se menciona en la sección 2.5.8 los datos simulados de las cascadas atmosféricas se generan mediante software especializados, al terminar el proceso de simulación, se crean los archivos con los parámetros de los eventos en un formato XCD.

Para la obtención de datos simulados en el observatorio HAWC se siguen los siguientes pasos:

1. Se debe acceder al clúster de HAWC, mediante las credenciales proporcionadas por el observatorio.
2. Una vez dentro del clúster mediante una ruta específica, se localizan los archivos de la simulación, los cuales se encuentran en formato XCD.
3. Estos archivos XCD contienen el número de eventos, así como las variables correspondientes a cada evento (Ilustración 27).

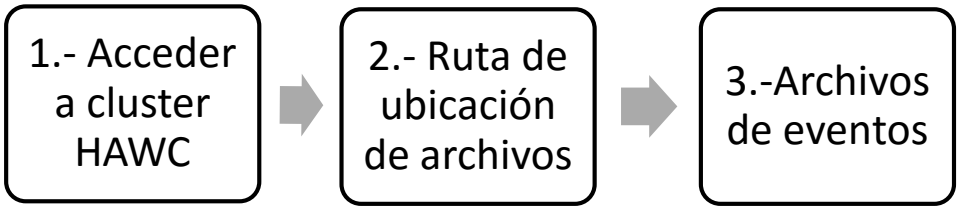


Ilustración 27. Pasos para obtener archivos con los eventos simulados de las cascadas atmosféricas de RG y RC.

Por lo general el observatorio HAWC simula los eventos de nueve partículas. En el presente trabajo se unieron los datos de protón, silicio, neón, helio, carbón, hierro, magnesio y oxígeno en un solo archivo y se le considera como RC, este archivo con alrededor de 21 millones de eventos representa el 71% del total de datos (Ilustración 28), mientras que el archivo simulado de gamma se le considera RG, con alrededor de 8 millones de eventos representa el 29% del total de eventos. Tomando en cuenta esto se puede entender que la cantidad de datos de RC es mayor que la de RG debido a que se unieron los datos de ocho partículas, mientras que los datos de RG solo representan las simulaciones de rayos gamma.

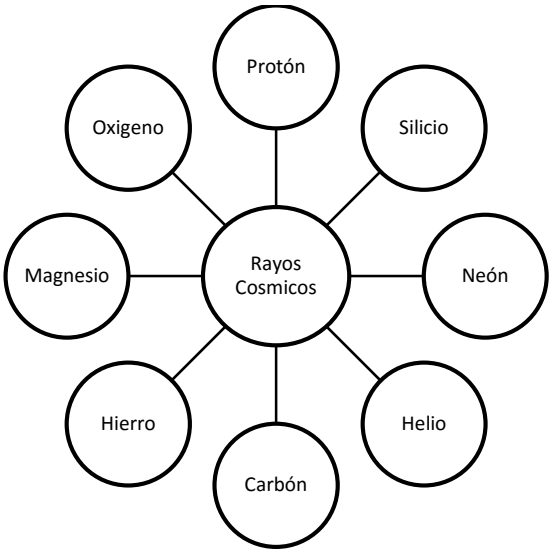


Ilustración 28. Se muestran las 8 partículas que se toman como rayos cósmicos.

Para la red neuronal se utilizaron los datos simulados de RG y RC, los cuales se dividieron para las diferentes etapas: entrenamiento, verificación y prueba: 25% para entrenamiento, 25% para verificación y 50% para prueba, estos últimos se usaron en la comparación con el método estándar de HAWC (Ilustración 29).

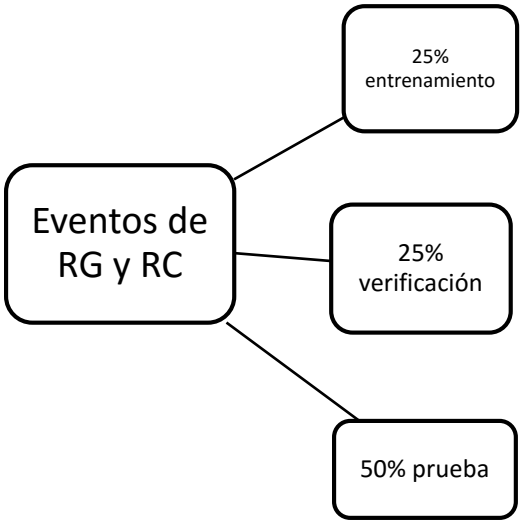


Ilustración 29. Se muestra la forma en que se separan los datos para las diferentes etapas del método de red neuronal.

3.2 Entrenamientos de red neuronal

3.2.1 Entrenamientos NN10

Los entrenamientos NN10 son nombrados así debido a que usan 10 variables de entrada, las cuales son seleccionadas debido a su importancia en la separación de datos, algunas de estas variables son variables que anteriormente no se han analizado usando redes neuronales.

- *Configuración de variables de entrada*

Para mejorar la separación de datos es importante seleccionar las variables que más ayuden en esta separación, para ello se eligen 10 variables para entrenar todos los fhit, de las cuales 7 (LIC, PINC, LogNNEnergy, dismax, LDFamp, LDFChi2, nfit20) son elegidas después de un análisis comparativo usando curvas ROC, aquellas con una mejor zona de clasificación fueron seleccionadas, según lo reportado por (Capistrán, 2020), igualmente se eligen 3 variables mas (LDFAge, sos, LLH) debido a que son nuevas variables agregadas en la actualización Pass5 del software de HAWC y podrían ayudar en la separación de datos (Ilustración 30).

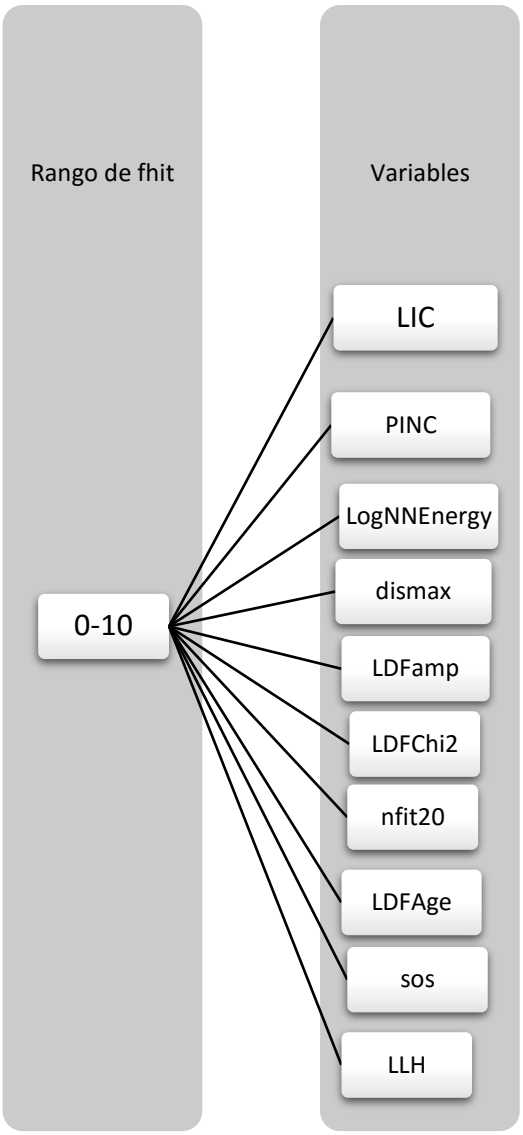


Ilustración 30. Configuración de variables de entrada de entrenamientos NN10

- *Modelo de entrenamiento*

En el modelo que se propone para los entrenamientos NN10, se dividen los datos simulados en dos grupos: In-HAWC y Out-HAWC, aquellos eventos en donde el núcleo se detecta en el arreglo principal de WCDs de HAWC se consideran In-HAWC, mientras que aquellos en donde el núcleo del evento se detecta fuera del arreglo principal de WCDs se consideran Out-HAWC, después, en cada grupo, se dividen nuevamente los datos en 3 grupos de fhit, esto con el fin de

reconocer con mayor facilidad en que rangos de fhit las redes neuronales tienen un mejor desempeño, el primer grupo abarca del fhit 0 al 2, el segundo grupo del fhit 3 al 7 y el tercer grupo del fhit 8 al 10 (Ilustración 31).

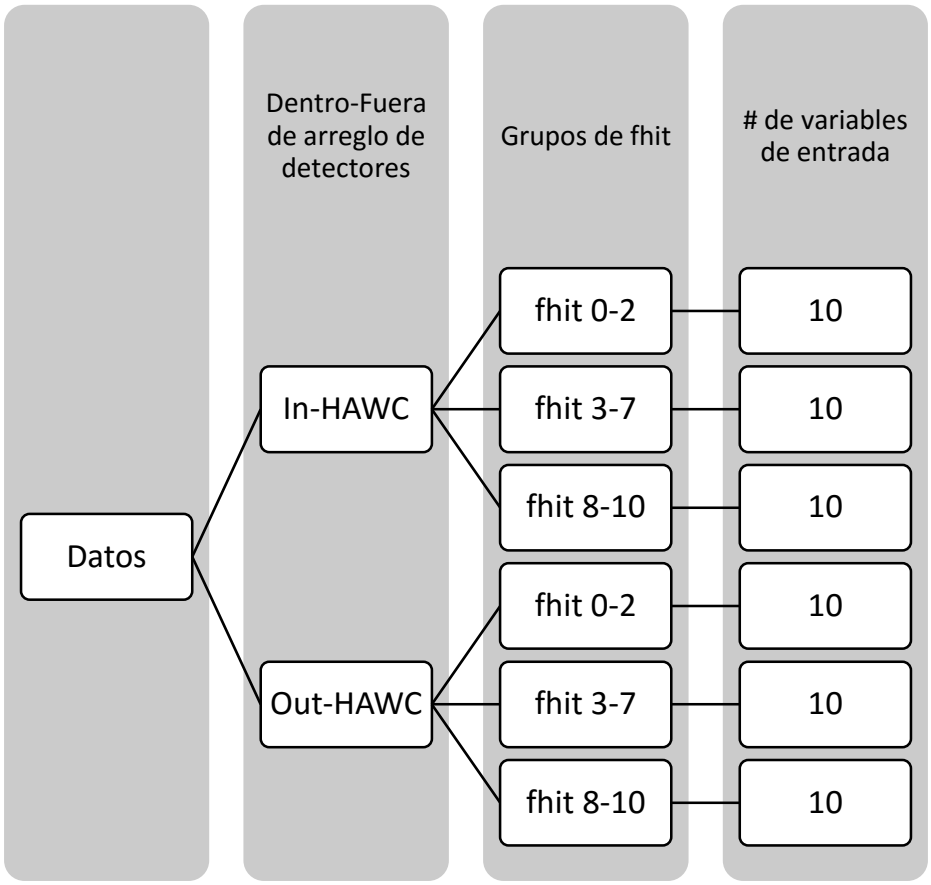


Ilustración 31. En el modelo de los entrenamientos NN10 se dividen los datos por In-Out HAWC, a su vez estos se dividen en 3 grupos de fhit y cada grupo de fhit es entrenado con 10 variables de entrada.

3.2.2 Entrenamientos NN4, NN7, NN8

El entrenamiento NN4 es nombrado así debido a que usa 4 variables de entrada para los fhit 0-2, mientras que el NN7 usa 7 variables para los fhit 3-7 y el NN8 usa 8 variables para los fhit 8-10, la selección de estas variables por grupo de fhit se debe a que se eligen las que tuvieron

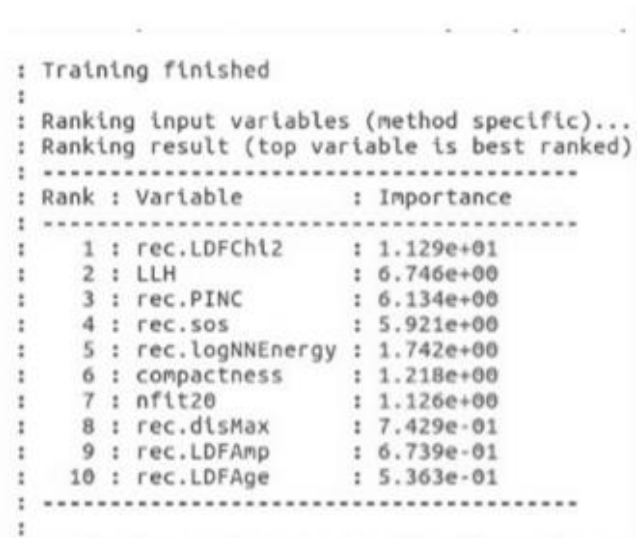
el mejor desempeño en los entrenamientos NN10, en la sección 3.2.2 se explica a detalle esta selección de variables.

- *Configuración de variables de entrada*

Al terminar los entrenamientos NN10, los cuales se mencionan en la sección 3.2.1, la red neuronal arroja un ranking de aquellas variables con el mejor desempeño por cada grupo de fhit, por lo que tomando en cuenta estos resultados se decide elegir aquellas variables con el mayor nivel de importancia, para realizar otros entrenamiento, así en los fhit 0-2 se eligen las primeras 4 variables (Ilustración 32) en los fhit 3-7 se eligen las primeras 7 variables (Ilustración 33), por último para los fhit 8-10 se eligen las primeras 8 variables (Ilustración 34), en la (Ilustración 35) se muestra la configuración con todas las variables elegidas por grupos de fhit.

```
: Training finished
:
: Ranking input variables (method specific).
: Ranking result (top variable is best ranke
: ---------------------------------------
: Rank : Variable        : Importance
: ---------------------------------------
:    1 : rec.PINC        : 2.146e+01
:    2 : LLH             : 1.393e+01
:    3 : rec.LDFChi2     : 1.244e+01
:    4 : rec.sos         : 1.219e+01
:    5 : nfit20          : 2.364e+00
:    6 : rec.dlsMax      : 2.066e+00
:    7 : rec.logNNEnergy : 2.037e+00
:    8 : compactness     : 1.084e+00
:    9 : rec.LDFAmp      : 2.601e-01
:   10 : rec.LDFAge      : 2.067e-01
: ---------------------------------------
:
```

Ilustración 32. Se muestra el ranking de variables que arroja la red neuronal al terminar el entrenamiento NN10 en los fhit 0-2, de las cuales se eligen las primeras cuatro para realizar otro entrenamiento llamado NN4.



```
: Training finished
:
: Ranking input variables (method specific)...
: Ranking result (top variable is best ranked)
: ----------------------------------------
: Rank : Variable        : Importance
: ----------------------------------------
:    1 : rec.LDFChi2     : 1.129e+01
:    2 : LLH             : 6.746e+00
:    3 : rec.PINC        : 6.134e+00
:    4 : rec.sos         : 5.921e+00
:    5 : rec.logNNEnergy : 1.742e+00
:    6 : compactness     : 1.218e+00
:    7 : nfit20          : 1.126e+00
:    8 : rec.disMax      : 7.429e-01
:    9 : rec.LDFAmp      : 6.739e-01
:   10 : rec.LDFAge      : 5.363e-01
: ----------------------------------------
:
```

Ilustración 33. Se muestra el ranking de variables que arroja la red neuronal al terminar el entrenamiento NN10 en los fhit 3-7, de las cuales se eligen las primeras siete para realizar otro entrenamiento llamado NN7.

```
: Training finished
:
: Ranking input variables (method specific)...
: Ranking result (top variable is best ranked)
: ----------------------------------------
: Rank : Variable        : Importance
: ----------------------------------------
:    1 : rec.LDFChi2     : 1.401e+01
:    2 : LLH             : 1.123e+01
:    3 : rec.PINC        : 5.072e+00
:    4 : rec.sos         : 3.960e+00
:    5 : nfit20          : 2.950e+00
:    6 : rec.LDFAmp      : 2.778e+00
:    7 : compactness     : 1.776e+00
:    8 : rec.LDFAge      : 1.151e+00
:    9 : rec.disMax      : 6.299e-01
:   10 : rec.logNNEnergy : 5.002e-01
: ----------------------------------------
:
```

Ilustración 34. Se muestra el ranking de variables que arroja la red neuronal al terminar el entrenamiento NN10 en los fhit 8-10, de las cuales se eligen las primeras ocho para realizar otro entrenamiento llamado NN8.

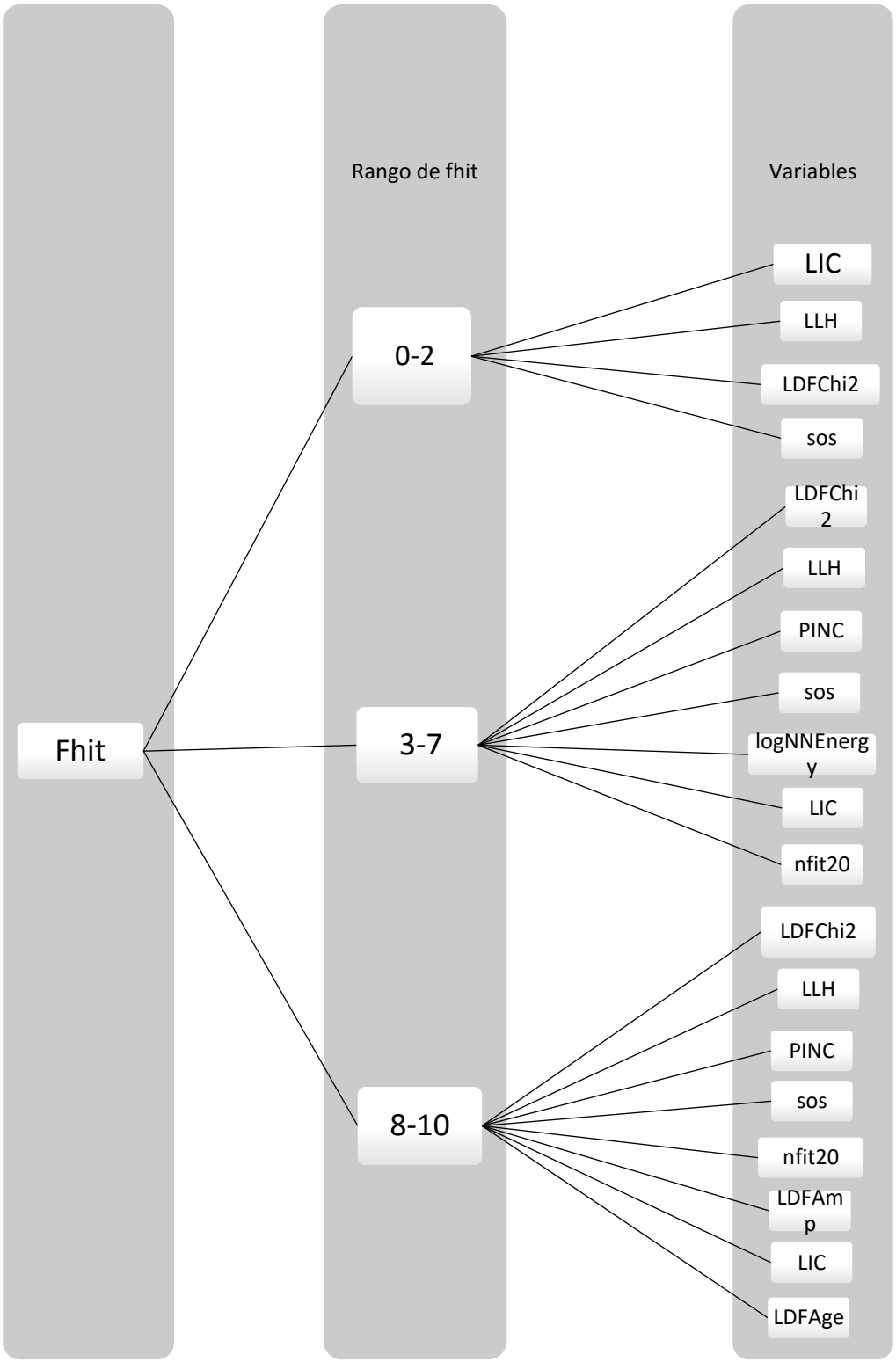


Ilustración 35. La configuración de variables de entrada de los entrenamientos NN4, NN7, NN8 en donde se usan 4 variables en los fhit 0-2, 7 variables en los fhit 3-7 y 8 variables en los fhit 8-10.

- *Modelo de entrenamiento*

El cambio principal que se hace en el modelo para los entrenamientos NN4, NN7, NN8 respecto al modelo de los entrenamientos NN10 (sección 3.2.1), es en la configuración de variables de entrada, así en los entrenamientos NN4 se usan 4 variables, en los entrenamientos NN7 se usan 7 variables, mientras que en los entrenamientos NN8 se usan 8 variables (Ilustración 36).

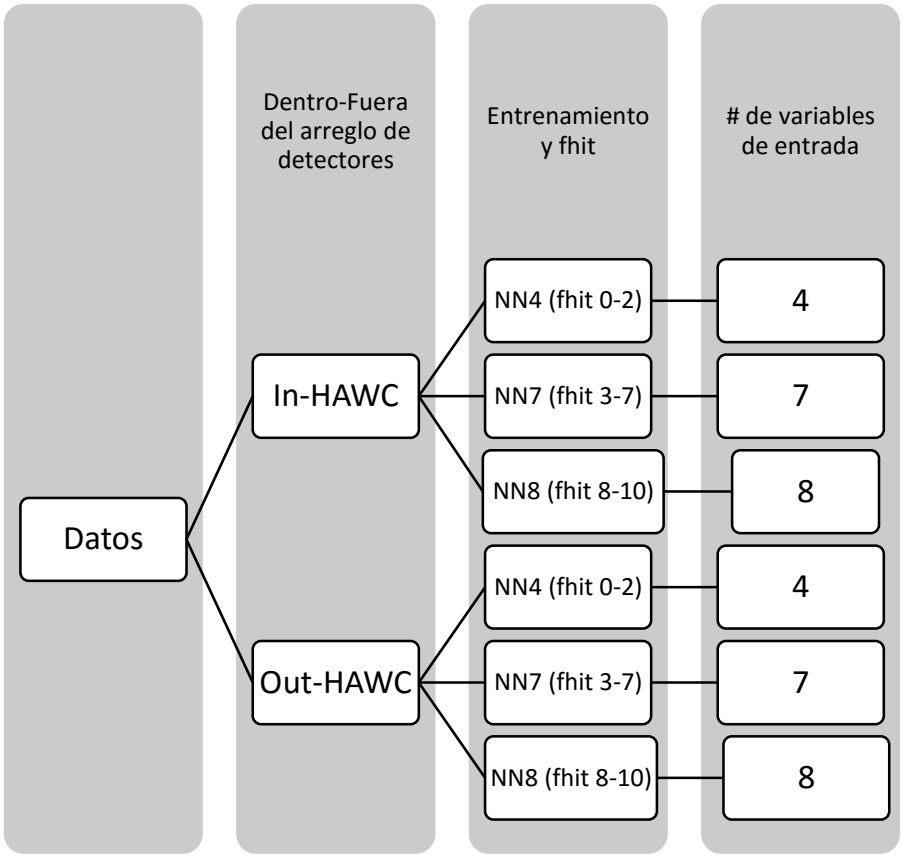


Ilustración 36. En el modelo para los entrenamientos NN4, NN7, NN8, se dividen los datos por In-Out HAWC, a su vez estos se dividen en 3 grupos de fhit, el grupo fhit 0-2 se entrena con 4 variables, el grupo fhit 3-7 se entrena con 7 variables y el grupo fhit 8-10 se entrena con 8 variables.

En la Tabla 2 se resumen todos los entrenamientos realizados, los entrenamientos NN10 es la primera propuesta de entrenamientos para la red neuronal, por otro lado, los entrenamientos NN4, NN7 y NN8 se proponen después de analizar las variables con mejor desempeño en los entrenamientos NN10 por cada grupo de fhit. El objetivo que se busca al realizar estos entrenamientos es compararlos con el método estándar de HAWC para identificar si existe una variabilidad en el desempeño de clasificación de datos al cambiar el número de variables en los entrenamientos.

Tabla 2.

Entrenamientos totales.

Método	Red Neuronal											
Entrenamientos	**NN10**						**NN4**		**NN7**		**NN8**	
Grupos In-Out HAWC	In			Out			In	Out	In	Out	In	Out
Grupos fhit	0-2	3-7	8-10	0-2	3-7	8-10	0-2	0-2	3-7	3-7	8-10	8-10
# Variables	10	10	10	10	10	10	4	4	7	7	8	8

En esta tabla se muestran los diferentes entrenamientos del método de red neuronal y la forma en cómo se dividieron los datos, por In-Out HAWC, grupos de fhit, así como el número de variables que se utilizan en cada uno.

3.3.3 Configuración de red neuronal

La arquitectura de la red es x:10:10:1 el primer valor x es el número de neuronas de la capa de entrada el cual depende de la configuración de variables de entrada, los siguientes valores (10:10) corresponden a la capa oculta con 10 neuronas cada una y el último valor es la capa de la neurona de salida, si el valor de salida es 1 se considera RG y si es cero entonces es RC.

La función sináptica es tipo suma, la cual es la suma de las entradas multiplicada por sus pesos (Ecuación 6).

$$f = \sum_{i=0}^{n} x_i w_i \qquad (6)$$

Donde:

x= representa las entradas de la red neuronal

w=los pesos de las entradas

La función de transferencia para las neuronas de entrada y salida, es lineal, mientras que para la oculta es sigmoide, la cual determina la salida de las neuronas tomando en cuenta un umbral dado. (Ecuación 7)

$$o(f) = \frac{1}{1+e^{-f}} \qquad (7)$$

El método de entrenamiento es BFGS (Broyden-Fletcher-Goldfarb-Shannon), el cual utiliza las segundas derivadas de la función del error.

El número de ciclos definidos son 500.

Para el entrenamiento de la red neuronal se utiliza la biblioteca de TMVA, por lo que los archivos de entrada deben ser en formato ROOT. El código que se utiliza para configurar la red neuronal se toma del trabajo realizado por (Capistrán, 2020) y se adapta para poder utilizar los modelos de entrenamiento propuestos en este trabajo, este código está escrito principalmente en lenguaje de programación C++ y Python. Básicamente lo que hace el código es agregar a los archivos de entrada una variable que especifica los eventos utilizados en la etapa de entrenamiento, verificación y prueba, después para la etapa de entrenamiento establece la configuración de la red neuronal (variables, datos de entrenamiento y modelo) para después iniciar con el entrenamiento, por último se evalúa el modelo construido y almacena el resumen en un archivo tipo ROOT, así como los pesos óptimos del modelo en formato XML (Ilustración 37).

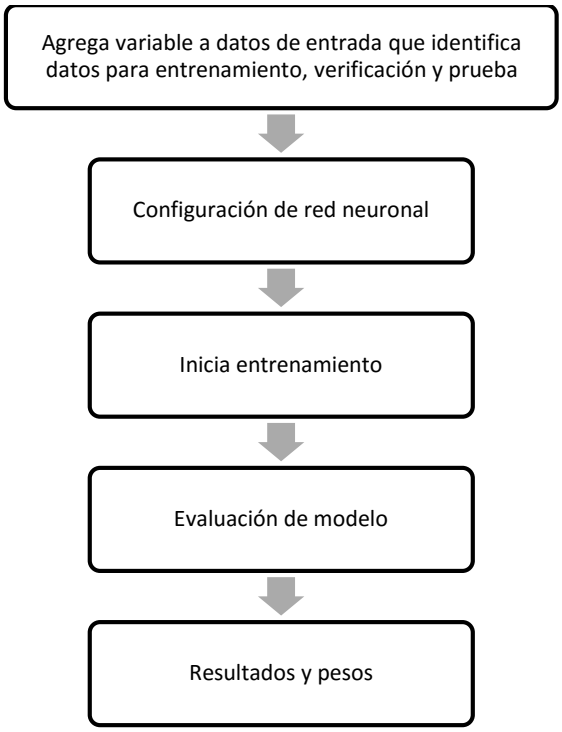


Ilustración 37. Pasos en el código para realizar entrenamiento de red neuronal.

3.3.4 Generación de salida red neuronal.

Para la generación de la salida de la red neuronal se toma el archivo XML que contiene los pesos óptimos, entonces el código lee los pesos, así como el valor de las variables para evaluarlos y finalmente dar el valor de salida (Ilustración 38).

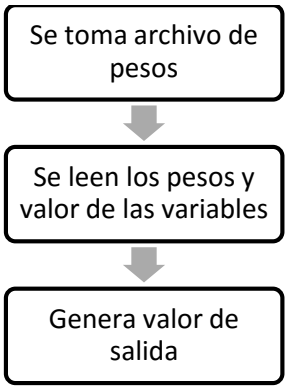


Ilustración 38. Pasos para generar la salida de la red neuronal

3.4 Cortes Óptimos

Ya que la salida de la red neuronal ($NNout$) es un número continuo con valores entre ~0 y ~1, los cuales están relacionados con la probabilidad de ser RG y RC, se debe definir un corte ($\emptyset NN$), el cual ayudará a determinar si cada salida es RG o RC, si la condición $NNout > \emptyset NN$ es verdadera se considera RG y si es falsa RC (Ilustración 39). Dado que se busca optimizar el valor de este corte ($\emptyset NN$), la forma de lograrlo es a través de la generación de histogramas de RG y RC utilizando la variable de salida $NNout$ de los datos de entrenamiento y verificación para cada grupo de fhit. Una vez teniendo el histograma de un grupo de fhit, se calcula el factor Q en función del corte $NNout > \emptyset NN$, esto nos generara los factor Q asociados a cada corte, en donde se considera un corte óptimo aquel que maximice el factor Q, este proceso se repite para los demás grupos de fhit con el fin de determinar el corte óptimo en estos grupos de fhit.

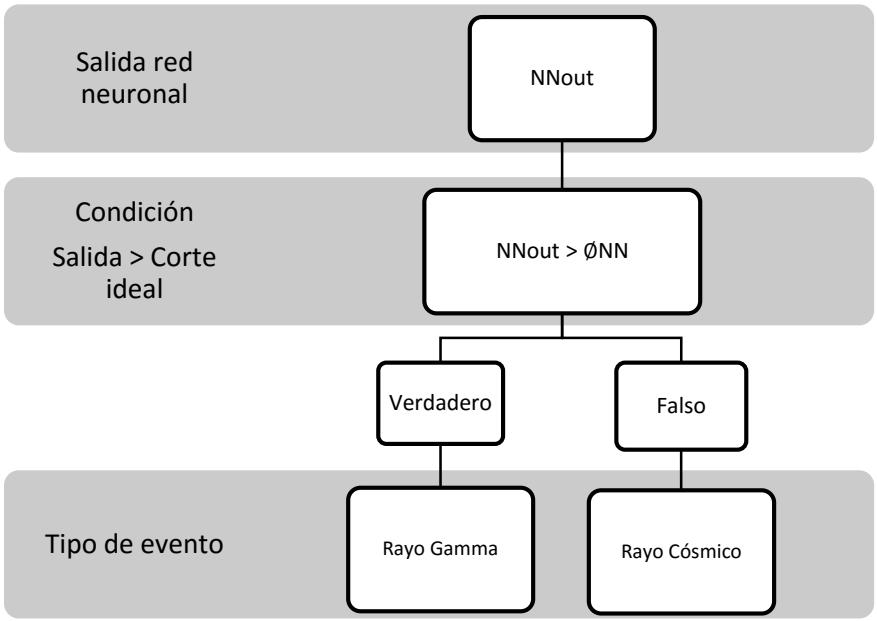


Ilustración 39. Una vez teniendo la salida de la red neuronal se evalúa a partir de la función NNout > ØNN, si es verdadero se considera rayo gamma de lo contrario es rayo cósmico.

3.5 Comparación con modelo estándar

Se realiza la comparación de los resultados de este trabajo contra el método estándar de separación utilizado en HAWC, para ello se compara la eficiencia en gamma y hadrones, así como el factor Q, de los métodos de entrenamiento propuestos, para cada uno de los fhit, lo cual ayuda a conocer en que rangos de fhit la separación de RG y RC es mejor (Ilustración 40).

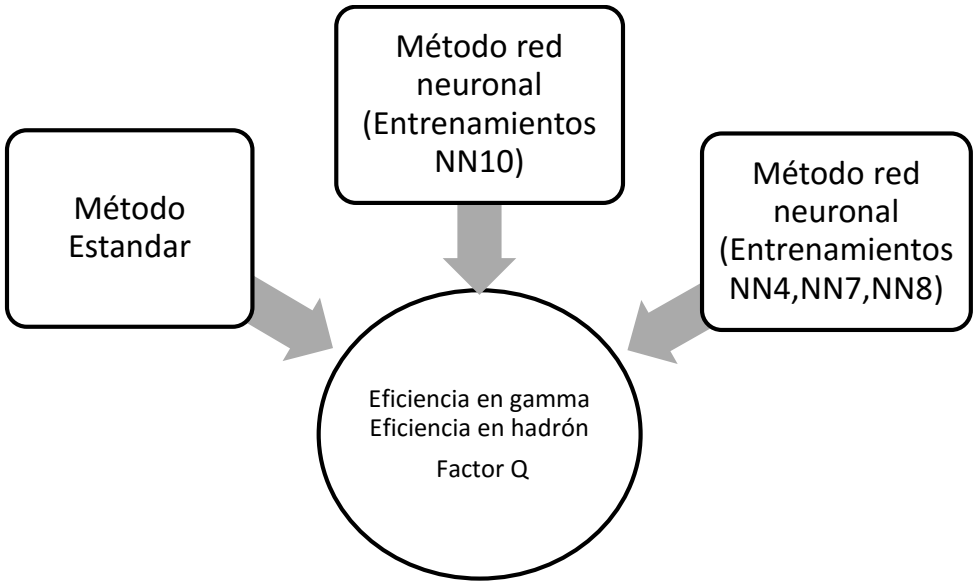


Ilustración 40. Se busca comparar los resultados de la eficiencia en gamma, hadrón y factor Q del método estándar y los 2 modelos de entrenamientos del método de red neuronal.

4. Análisis y resultados

El objetivo del presente capitulo es describir los diferentes entrenamientos del método red neuronal y comparar los resultados de la separación de partículas, con los del método estándar, por lo que en la primera sección se describe la configuración de los entrenamientos, así como las variables utilizadas en cada uno de ellos, las cuales ayudan al entrenamiento de la red neuronal para la correcta separación de datos. Finalmente se presentan los resultados de cada uno de los entrenamientos y se comparan con los resultados del método de separación de partículas usado en el observatorio HAWC, en donde métricas como eficiencia en gamma, eficiencia en hadrón y factor Q, determinan aquel con la mejor clasificación de partículas.

4.1 Entrenamientos de la red neuronal

En este capítulo se describen los diferentes entrenamientos que se usaron para aplicar las variables candidatas que se mencionan en la sección 2.5.9, igualmente se describirá la configuración de la red neuronal, la cual será comparada con sus diferentes entrenamientos contra el método estándar.

4.1.2 Entrenamientos

- *Entrenamientos NN10*

Los entrenamientos son nombrados de acuerdo al número de variables de entrada, para así poder diferenciarlos, en este capítulo se describirá los entrenamientos realizados con 10 variables los cuales se llamaran NN10, para realizar estos entrenamientos se dividieron los datos simulados en dos grupos: In-HAWC y Out-HAWC, después para uno de estos dos grupos se dividen nuevamente los datos por fhit en 3 grupos, el primer grupo abarca del fhit 0 al 2, el segundo grupo

del fhit 3 al 7 y el tercer grupo del fhit 8 al 10, cada uno de estos grupos de fhit fue entrenado de forma separada en donde se usó la configuración de 10 variables de entrada, el modelo de entrenamiento se describe en la sección 3.2.1. La configuración de la arquitectura, número de ciclos, función de transferencia, función sináptica, etc, de cada una de los entrenamientos NN10 fue la descrita en la sección 3.3.3, en la Ilustración 41 se muestra la arquitectura de uno de los entrenamientos NN10 en donde se observan las 10 variables como neuronas de entrada. Una buena práctica en la utilización de las redes neuronales es verificar que estas no caigan en un sobre entrenamiento, para ello se usan los histogramas de la salida de la red neuronal entrenada y se comparan con la salida de la red neuronal pero ahora con datos de verificación, en la Ilustración 42 se puede observar que la distribución de la salida de ambos grupos de datos es muy similar, lo cual confirma que el entrenamiento fue realizado correctamente pues se usó un set de datos de entrada diferente, para evitar el sobre entrenamiento estas dos salidas deben ser similares sin llegar a ser idénticas para ello se usa el Kolmogorov-Smirnov (KS) test, el cual es un procedimiento de bondad de ajuste que permite medir el grado de concordancia entre dos distribuciones, en la ilustración 43 se muestra el KS test tanto para RG (Señal) y RC (Ruido), con un valor de .766 nos indica que existen ligeras diferencias entre las distribuciones de entrenamiento y verificación para rayos gamma y con un valor de .053 igualmente nos indica que existen ligeras diferencias para rayos cósmicos, si el valor fuera 0 esto indicaría que las distribuciones son idénticas y si el valor fuera 1 entonces habría una alta probabilidad de inconsistencia en los datos.

- *Entrenamientos NN4, NN7, NN8*

La diferencia entre los entrenamientos NN10 y los entrenamientos NN4, NN7, NN8 es en la configuración de variables de entrada para cada grupo de fhit: asi en los entrenamientos NN4 para el grupo fhit 0-2 se usan 4 variables , en los NN7 para el grupo de fhit 3-7 se usan 7 variables, mientras que en los entrenamientos NN8 para el grupo de fhit 8-10 se usan 8 variables, como se describió en la sección 3.2.2, estos entrenamientos se proponen después de analizar el comportamiento de las 10 variables usadas en los entrenamientos NN10. La configuración y estructura de la red neuronal (Ilustración 43) para estos entrenamientos fue la mencionada en la

sección 3.3.3, mientras que para verificar el sobre entrenamiento se realizó el KS test, entre los datos de entrenamiento y verificación (Ilustración 44).

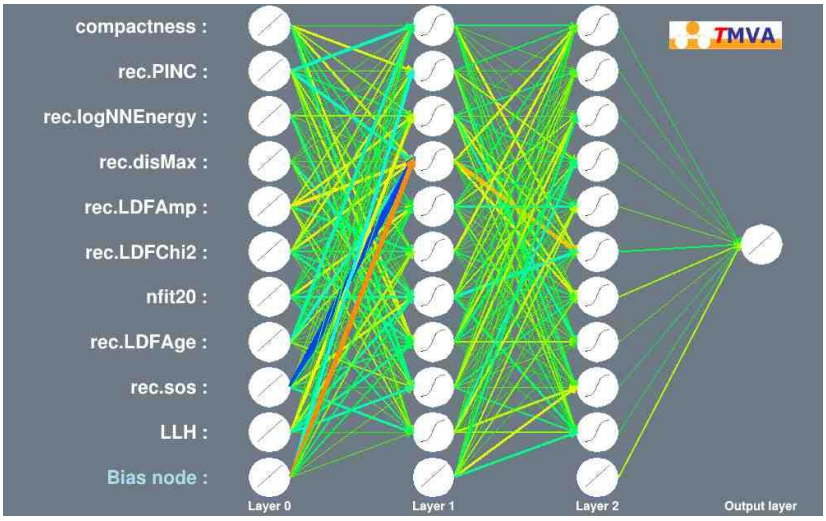


Ilustración 41. Arquitectura del entrenamiento NN10 en el fhit 0-2, en el lado izquierdo se muestran las variables usadas en este entrenamiento, los círculos blancos indican las neuronas, las líneas de colores las conexiones entre las neuronas, se puede observar que se usan diez neuronas para la capa de entrada, diez neuronas para las dos capas ocultas y una neurona para la capa de salida.

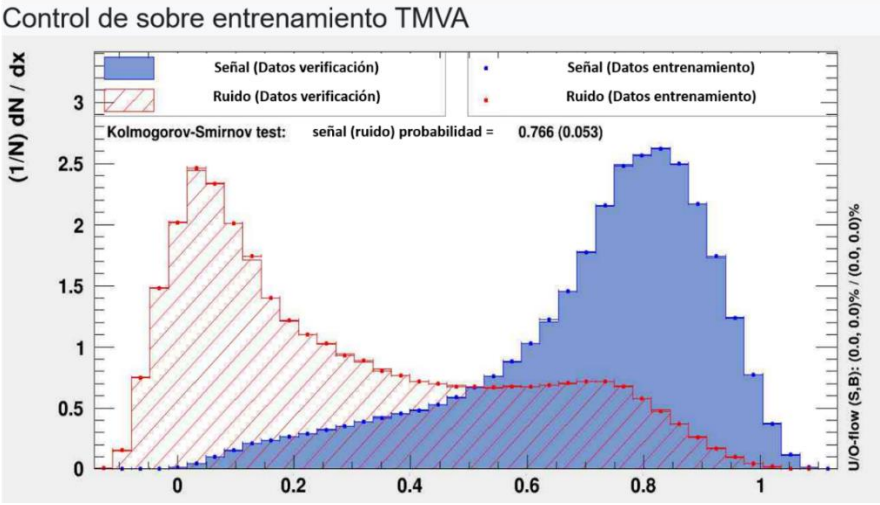


Ilustración 42. Se puede observar la distribución de la salida de la red neuronal con datos de entrenamiento (puntos), y se compara con la salida de la red neuronal, pero usando datos de verificación (líneas), de color rojo se resaltan los datos de Ruido (RC) y de color azul los datos de Señal (RG), igualmente se muestra el KS test entre las dos distribuciones de señal (.766) y entre las dos distribuciones de ruido (.053).

En las secciones 3.2.1 y 3.2.2 se describe a detalle la configuración de los entrenamientos del método de red neuronal, mientras que la tabla 2 resume todos los entrenamientos realizados.

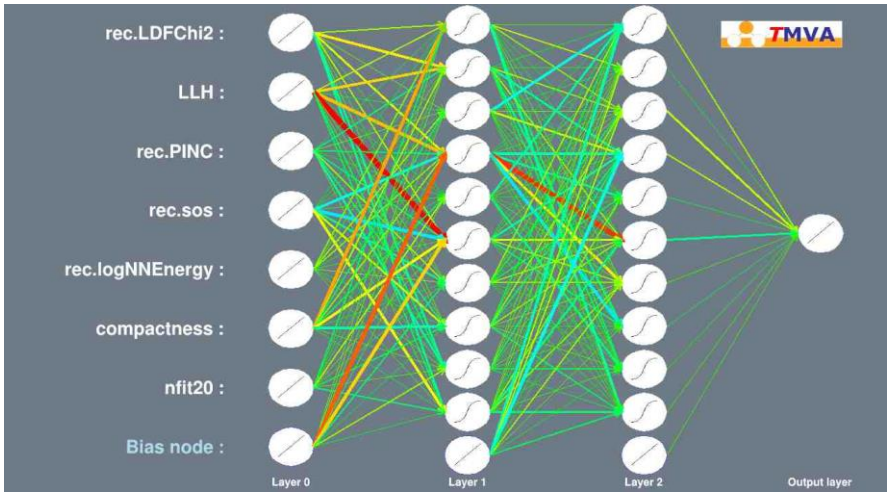


Ilustración 43. Arquitectura del entrenamiento NN7 en el fhit 3-7, en el lado izquierdo se muestran las variables usadas en este entrenamiento, los círculos blancos indican las neuronas, las líneas de colores las conexiones entre las neuronas, se puede observar que se usan siete neuronas para la capa de entrada, diez neuronas para las dos capas ocultas y una neurona para la capa de salida.

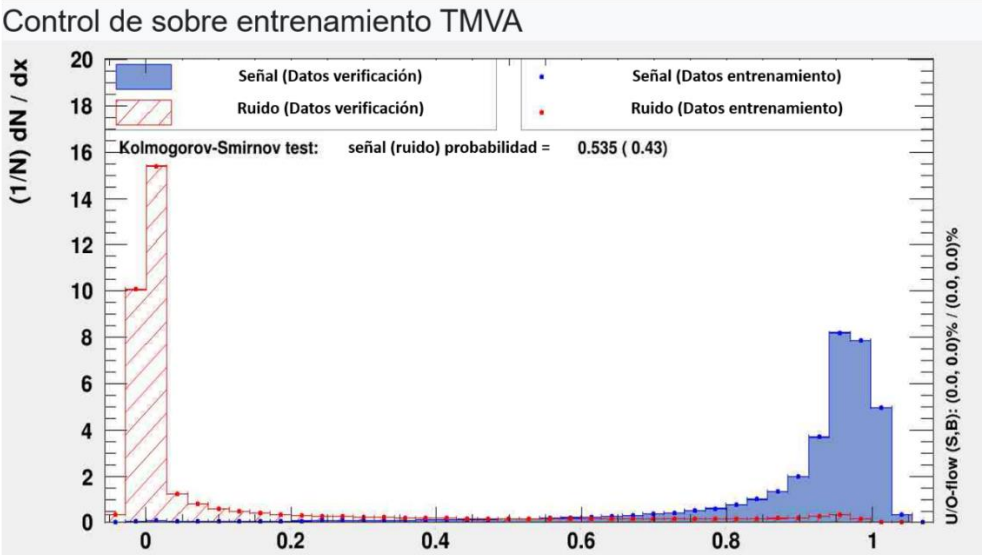


Ilustración 44. Se puede observar la distribución de la salida de la red neuronal con datos de entrenamiento (puntos), y se compara con la salida de la red neuronal, pero usando datos de verificación (líneas), de color rojo se resaltan los datos de Ruido (RC) y de color azul los datos de Señal (RG), igualmente se muestra el KS test entre las dos distribuciones de señal (.535) y las dos distribuciones de ruido (.43)

4.2 Comparaciones de Métodos de separación

En esta sección se realiza la comparación entre el método de red neuronal y el método estándar, a través de parámetros que indiquen el rendimiento de la separación de partículas como es la eficiencia en gamma, eficiencia en hadrón y el factor Q.

4.2.2 Eficiencia en gamma

Como se menciona en la sección 2.7 una de las formas para evaluar el rendimiento de los métodos es a través de la eficiencia en gamma, la cual cuantifica el porcentaje de eventos tipo gamma, correctamente clasificados.

En la ilustración 45 se muestran los resultados de eficiencia en gamma In-HAWC, de los entrenamientos del método red neuronal: NN10, NN4, NN7, NN8 y el método estándar, de igual forma en la tabla 3 se compara de forma cuantitativa las diferencias entre estos métodos, se puede notar que para los fhit 0,1 y 3, el método estándar mantiene una mejor eficiencia en gamma, superado por el entrenamiento NN10 en el fhit 2, respecto a los fhit 4 al 7, NN7 es superior, destacando el fhit 6 en donde la eficiencia es un 65% más alta que el método estándar y un 59% mejor en el fhit 7, mientras que en los fhit 8 al 10 el entrenamiento NN10 tuvo el mejor desempeño ya que en el fhit 9 fue 56% más eficiente y en el fhit 10 un 80% mejor respecto al método estándar.

En la Ilustración 46, se muestran los resultados de eficiencia en gamma Out-HAWC; es decir para aquellos eventos en donde el núcleo fue localizado fuera del arreglo principal de detectores de HAWC. Se puede observar en la tabla 4 que el entrenamiento NN7 obtuvo un mejor desempeño en los fhit 4 al 7, se destaca que en el fhit 6 la eficiencia fue un 67% más alta respecto al método estándar y un 70% en el fhit 7, por otro lado, el entrenamiento NN10 tiene el mejor desempeño en los fhit 2, 8 y 9 ya que en el fhit 8 fue un 63% más eficiente, y en el fhit 9 un 80%.

En general se puede observar que tanto para In-HAWC como para Out-HAWC, la eficiencia en gamma de los entrenamientos de red neuronal fueron superiores al método estándar, en los fhit 4 al 10.

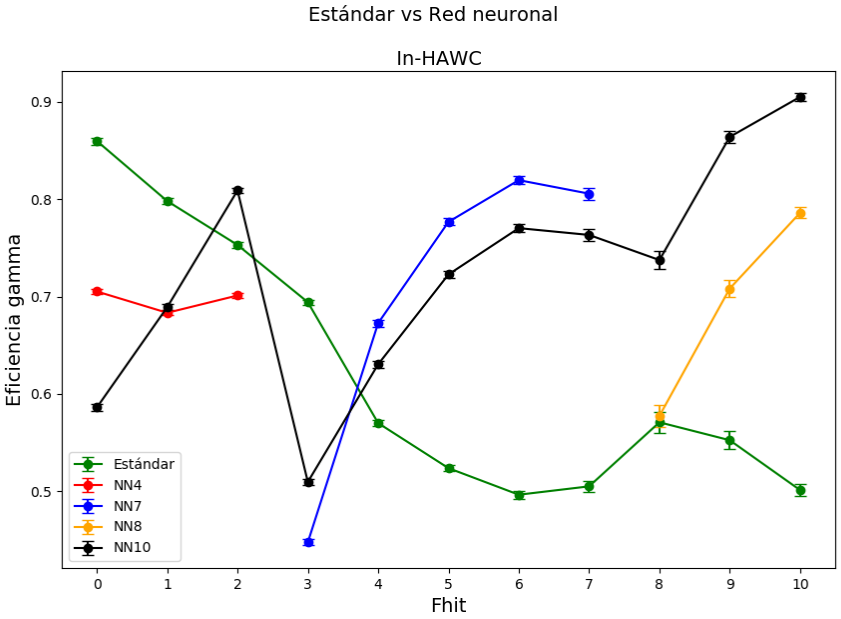


Ilustración 45. Se reporta la eficiencia en gamma In-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7(azul), NN8(amarillo) y el método estándar(verde). En el eje X se muestran los fhit en los que fueron entrenados, mientras que el eje Y el porcentaje de eficiencia en gamma. El entrenamiento NN7 reporta en general un mejor desempeño en los fhit 4 al 7, mientras que el NN10 tiene el mejor desempeño en los fhit 8 al 10.

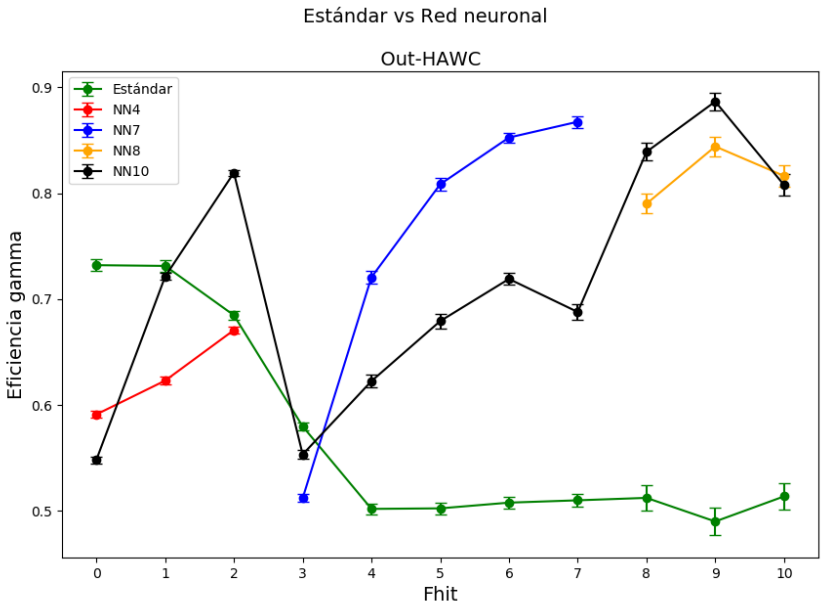


Ilustración 46. Se reporta la eficiencia en gamma Out-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7(azul), NN8(amarillo) y el método estándar(verde). En el eje X se muestran los fhit en los que fueron entrenados, mientras que en el eje Y el porcentaje de eficiencia en gamma. El entrenamiento NN7 reporta en general un mejor desempeño en los fhit 4 al 7, mientras que el entrenamiento NN10 tiene el mejor desempeño en los fhit 2, 8 y 9.

Tabla 3.

Eficiencia en gamma In-HAWC

	In-HAWC				
Fhit	**Eficiencia en gamma Método-Estándar**	**Eficiencia en gamma entrenamientos NN4, NN7,NN8**	**% de mejora NN4, NN7, NN8 vs Estándar**	**Eficiencia en gamma entrenamientos NN10**	**% de mejora NN10 vs Estándar**
0	0.86	0.71	-18.0	0.59	-31.8
1	0.80	0.68	-14.4	0.69	-13.7
2	0.75	0.70	-6.9	0.81	7.4
3	0.69	0.45	-35.5	0.51	-26.6
4	0.57	0.67	18.0	0.63	10.6
5	0.52	0.78	48.3	0.72	38.0
6	0.50	0.82	65.1	0.77	55.2
7	0.51	0.81	59.5	0.76	51.1
8	0.57	0.58	1.2	0.74	29.2
9	0.55	0.71	28.2	0.86	56.4
10	0.50	0.79	56.8	0.91	80.6

Se muestran los resultados de las eficiencias en gamma In-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron entrenados, en la segunda columna la eficiencia en gamma del método estándar, en la tercer columna la eficiencia de los entrenamientos NN4, NN7, NN8, en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4, NN7 y NN8 respecto al método estándar, en la quinta columna la eficiencia lograda por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar. En color verde se destaca el % de mejora logrado en los fhit 5, 6, 7, 9 y 10 en ambos modelos de entrenamiento.

Tabla 4.

Eficiencia en gamma Out-HAWC.

Out-HAWC					
Fhit	**Eficiencia en gamma-Método-Estándar**	**Eficiencia en gamma entrenamientos NN4, NN7,NN8**	**% de mejora NN4, NN7, NN8 vs Estándar**	**Eficiencia en gamma entrenamiento NN10**	**% de mejora NN10 vs Estándar**
0	0.732	0.5911	-19.2	0.5482	-25.1
1	0.7313	0.623	-14.8	0.7212	-1.4
2	0.6847	0.6707	-2.0	0.8193	19.7
3	0.5798	0.512	-11.7	0.5533	-4.6
4	0.5019	0.7206	43.6	0.6225	24.0
5	0.5024	0.8084	60.9	0.6792	35.2
6	0.5078	0.8525	67.9	0.719	41.6
7	0.51	0.8674	70.1	0.6878	34.9
8	0.5123	0.79	54.2	0.839	63.8
9	0.49	0.8441	72.3	0.8865	80.9
10	0.5137	0.8163	58.9	0.8077	57.2

Se muestran los resultados de las eficiencias en gamma Out-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron entrenados, en la segunda columna la eficiencia en gamma del método estándar, en la tercer columna la eficiencia de los entrenamientos NN4, NN7, NN8 en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4,NN7 y NN8 respecto al método estándar, en la quinta columna la eficiencia lograda por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar. Se puede notar que a partir del fhit 4 los entrenamientos de red neuronal mejoran la eficiencia en gamma.

4.2.3 Eficiencia en hadrón

Dado que la eficiencia en hadrón es el porcentaje de eventos tipo hadrón clasificados incorrectamente, se busca que un buen método de separación tenga valores bajos para determinar su desempeño. En la Ilustración 47, se muestra el desempeño de la eficiencia en hadrón In-HAWC del método estándar y de los entrenamientos del método de red neuronal; se puede notar que en los fhit 0 y 1 el entrenamiento NN10 tiene buen desempeño, ya que logró ser un 60% y 42% mejor respecto al método estándar, como se muestra en la tabla 5, mientras que en el fhit 2 el mejor valor lo obtuvo el entrenamiento NN4, logrando ser un 26% mejor que el método estándar, en el fhit 3 los entrenamientos NN10 y NN7, obtuvieron igualmente resultados favorables, logrando mejorar un 60%, del fhit 4 al fhit 10 el desempeño de los métodos es muy similar, logrando obtener valores de eficiencia en hadrón cercanos a 0.

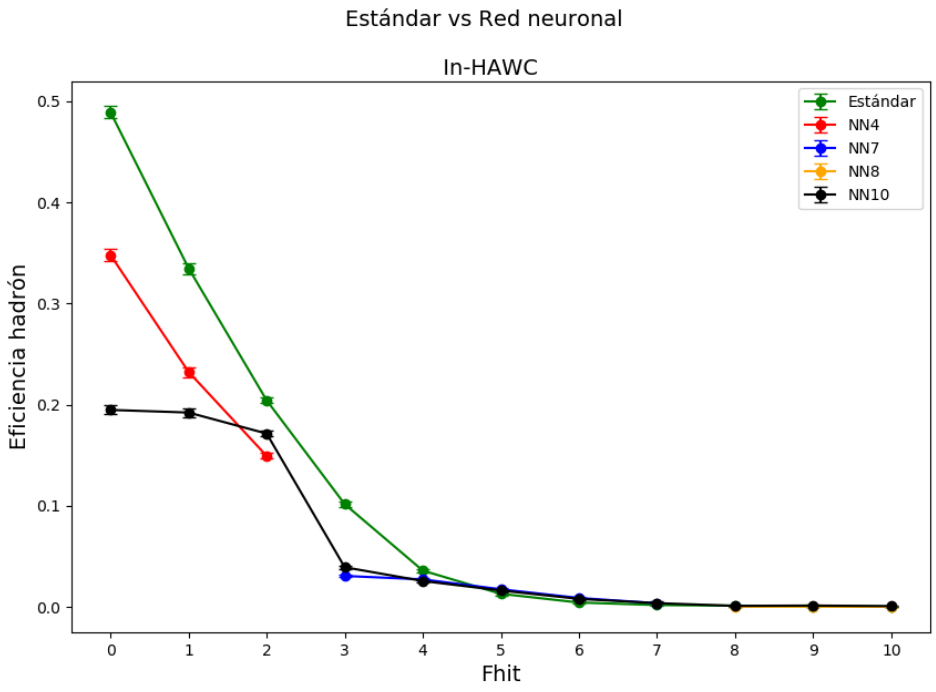


Ilustración 47. Se reporta la eficiencia en hadrón In-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7 (azul), NN8 (amarillo) y el método estándar (verde). En el eje X se muestran los fhit en los que fueron entrenados, mientras que en el eje Y el porcentaje de eficiencia en hadrón. Se puede notar que en los fhit 0 al 4 los entrenamientos propuestos logran un mejor desempeño respecto al método estándar.

Tabla 5.

Eficiencia en hadrón In-HAWC.

	In-HAWC				
Fhit	**Eficiencia en hadrón-Método-Estándar**	**Eficiencia en hadrón entrenamientos NN4, NN7,NN8**	**% de mejora Estándar vs NN4, NN7, NN8**	**Eficiencia en hadrón entrenamiento NN10**	**% de mejora Estándar vs NN10**
0	0.4893	0.348	28.0	0.1948	60.2
1	0.3349	0.2324	30.0	0.1922	42.6
2	0.2043	0.1493	26.0	0.1714	16.1
3	0.1016	0.0306	69.0	0.0392	61.4
4	0.0359	0.0272	24.0	0.0256	28.7
5	0.0129	0.0174	-34.9	0.0162	-25.6
6	0.0043	0.0088	-104.0	0.0079	-83.7
7	0.0018	0.0036	-100.0	0.0037	-105.6
8	0.0012	0.0002	83.3	0.001	16.7
9	0.0003	0.0003	0.0	0.0013	-333.0
10	0.0001	0.0001	0.0	0.0008	-700.0

Se muestran los resultados de las eficiencias en hadrón In-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron entrenados, en la segunda columna la eficiencia en hadrón del método estándar, en la tercer columna la eficiencia de los entrenamientos NN4, NN7, NN8 en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4, NN7 y NN8 respecto al método estándar, en la quinta columna la eficiencia lograda por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar. Se destaca que en los fhit 0 al 4 los entrenamientos logran mejorar la eficiencia respecto al método estándar. En los fhit 5 al 10 los dos métodos tienen un buen desempeño pues sus eficiencias son cercanas a 0, lo cual indica que muy pocos eventos tipo hadrón están siendo incorrectamente clasificados.

En la Ilustración 48 se analiza la eficiencia en hadrón Out-HAWC, en esta se puede notar que en el fhit 0 el entrenamiento NN10 tiene la mejor eficiencia logrando ser un 56% mejor que

el método estándar como también se observa en la Tabla 6, en el fhit 1 y 2 el entrenamiento NN4 tiene el mejor desempeño, logrando ser un 42% mejor en el fhit 1 y 29% en el fhit 2, mientras que en el fhit 3 NN7 obtuvo el mejor desempeño, finalmente en los fhit 4 al 10 tanto la red neuronal como el estándar logran obtener eficiencias cercanas a 0, lo cual indica una muy buena eficiencia en hadrón.

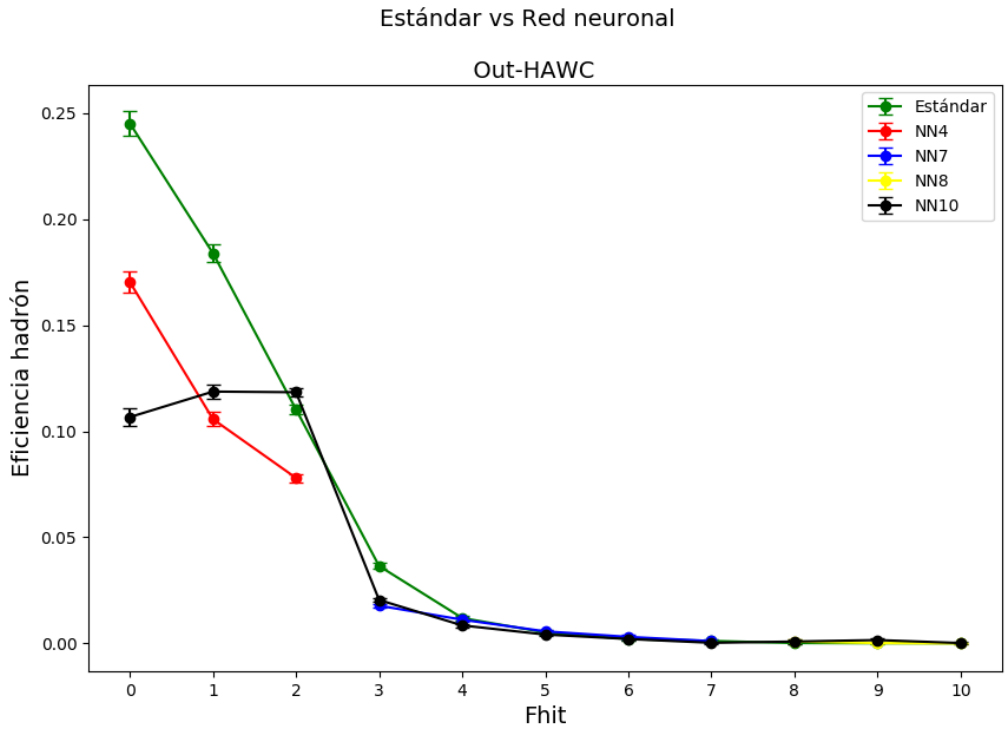


Ilustración 48. Se reporta la eficiencia en hadrón Out-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7 (azul), NN8 (amarillo) y el método estándar (verde). En el eje X se muestran los fhit en los que fueron entrenados, mientras que el eje Y el porcentaje de eficiencia en hadrón. Se puede observar que el entrenamiento NN4 en general tiene un mejor desempeño respecto al método estándar en los fhit 0 al 2.

Se puede concluir respecto a la eficiencia en hadrón que en los fhit 0 al 3 los entrenamientos basados en el método de red neuronal tienen una mejor eficiencia y en los fhit 4 al 10 la eficiencia tanto de la red neuronal como del método estándar es muy buena pues esta llega a ser cercana 0, lo cual implica que muy pocos eventos de tipo hadrón están siendo clasificados como gamma en estos fhit.

Tabla 6

Eficiencia en hadrón Out-HAWC.

Out-HAWC					
Fhit	**Eficiencia en hadrón-Método-Estándar**	**Eficiencia en hadrón entrenamientos NN4, NN7,NN8**	**% de mejora Estándar vs NN4, NN7, NN8**	**Eficiencia en hadrón entrenamiento NN10**	**% de mejora Estándar vs NN10**
0	0.2451	0.1703	30.5	0.1067	56.5
1	0.1839	0.1057	42.5	0.1187	35.5
2	0.1103	0.078	29.3	0.1184	-7.3
3	0.0365	0.0178	51.2	0.0205	43.8
4	0.012	0.0112	6.7	0.0085	29.2
5	0.0052	0.0057	-9.6	0.0042	19.2
6	0.0021	0.0031	-47.6	0.0021	0.0
7	0.0012	0.0011	8.3	0.0003	75.0
8	0.0002	0.0009	-350.0	0.0009	-350.0
9	0.0001	0.0002	-100.0	0.0016	-1500.0
10	0.0001	0	100.0	0.0002	-100.0

Se muestran los resultados de las eficiencias en hadrón Out-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron entrenados, en la segunda columna la eficiencia en hadrón del método estándar, en la tercer columna la eficiencia de los entrenamientos NN4, NN7, NN8, en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4, NN7 y NN8 respecto al método estándar, en la quinta columna la eficiencia lograda por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar, se destaca la eficiencia lograda en los fhit 0, 1 y 3 por la red neuronal, en donde logra mejorar la eficiencia del método estándar.

4.2.4 Factor Q

Una vez se han calculado las eficiencias se procede a calcular el factor Q para cada fhit, dado lo planteado en la sección 2.7.1 este parámetro nos permite saber el modelo que tiene un mejor desempeño en la separación de partículas.

En las Ilustraciones 49 y 50 , se comparan los métodos de separación para In-Out HAWC a través del factor Q, tomando la escala de los resultados, no es posible apreciar la diferencia entre los métodos, por lo que en las siguientes secciones se analiza a detalle las diferencias de los factor Q para cada fhit.

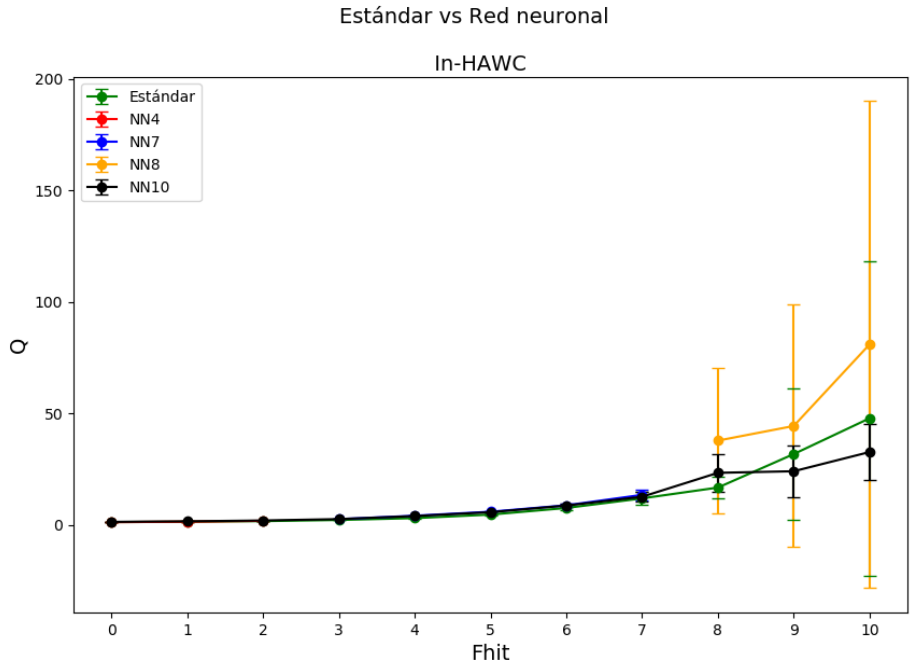


Ilustración 49. Se reporta el factor Q In-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7 (azul), NN8 (amarillo) y el método estándar (verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se puede observar que en los fhit 8, 9 y 10 los factores Q arrojan errores estadísticos muy altos, debido a la baja estadística en estos fhit.

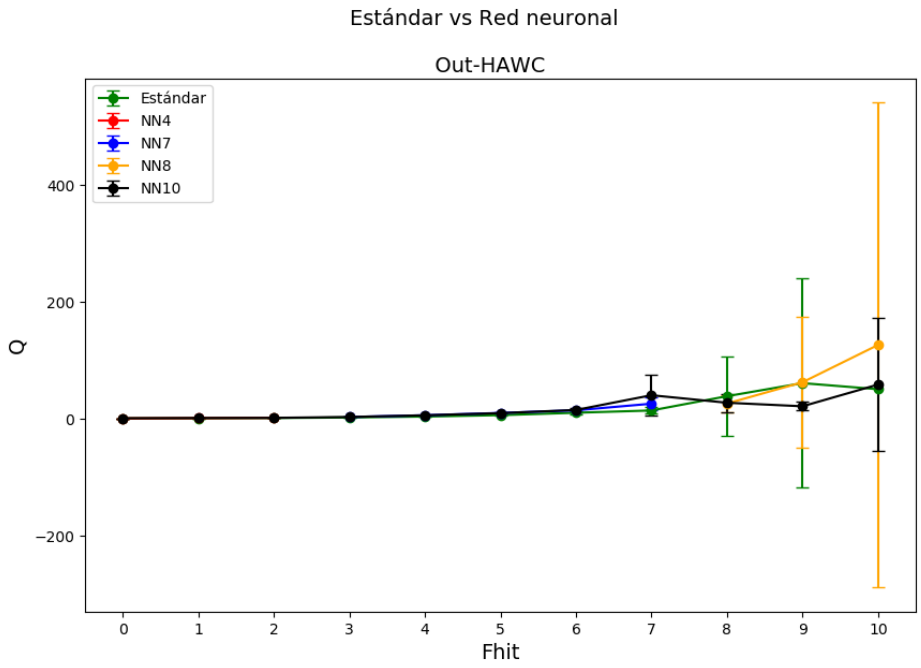


Ilustración 50. Se reporta el factor Q Out-HAWC, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo), NN7 (azul), NN8 (amarillo) y el método estándar(verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se puede observar que en los fhit 7, 8, 9 y 10 los factores Q arrojan errores estadísticos muy altos, debido a la baja estadística en estos fhit.

- *Factor Q en fhit 0-2*

En la Ilustración 51, se muestra las diferencias del factor Q de los métodos de separación en los fhit 0 al 2 para In-HAWC, en donde el entrenamiento NN10 es superior, ya que como se muestra en la tabla 7, en el fhit 0 alcanza a ser un 8% mejor que el factor Q del método estándar, mientras que en el fhit 1 un 13% y en el fhit 2 un 17% más alto. Lo cual nos indica que en general la separación de partículas en estos fhit es mejor con el entrenamiento NN10.

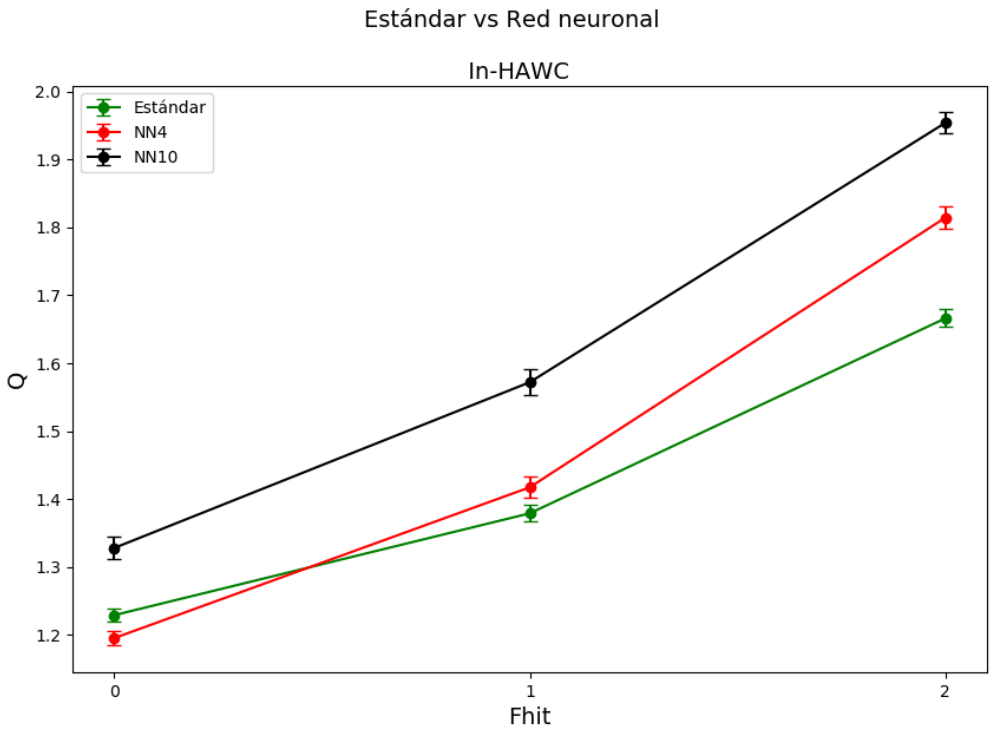


Ilustración 51. Se reporta el factor Q In-HAWC en los fhit 0-2, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo) y el método estándar (verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se puede observar que el entrenamiento NN10 tiene un factor Q superior en los fhit 0 al 2.

En la Ilustración 52, para Out-HAWC se puede observar que el factor Q del método de red neuronal utilizando el entrenamiento NN10 es superior al factor Q del método estándar en el fhit 0 al 2, logrando ser un 13% mejor en el fhit 0, 22% en fhit 1 y un 15% en fhit 2 como lo muestra la tabla 8. Igualmente se puede observar que el entrenamiento NN4 obtiene buenos resultados en el fhit 1 y 2.

Tabla 7.

Factor Q In-HAWC.

	In-HAWC				
fhit	**Factor Q método estándar**	**Factor Q entrenamientos NN4, NN7, NN8**	**% de mejora NN4, NN7, NN8 vs Estándar**	**Factor Q entrenamientos NN10**	**% de mejora NN10 vs Estándar**
0	1.2292	1.1954	-2.7	1.328	8.0
1	1.3791	1.4174	2.7	1.572	13.9
2	1.6661	1.8143	8.8	1.9538	17.2
3	2.1779	2.5576	17.4	2.574	18.1
4	3.0067	4.0779	35.6	3.9422	31.1
5	4.6033	5.8811	27.7	5.6713	23.2
6	7.5678	8.7504	15.6	8.6523	14.3
7	11.878	13.4498	13.2	12.5935	6.0
8	16.7198	37.7965	126.0	23.3376	39.5
9	31.7469	44.3532	39.7	24.0519	-24.2
10	47.7949	80.9304	69.3	32.6614	-31.6

Se muestran los resultados de los factores Q In-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron calculados, en la segunda columna el factor Q del método estándar, en la tercer columna el facto Q de los entrenamientos NN4, NN7 y NN8, en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4, NN7 y NN8 respecto al método estándar, en la quinta columna el factor Q logrado por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar, Se resalta el porcentaje de mejora logrado en los fhit 1 al 5, ya que en estos fhit los resultados no son afectados por el error estadístico.

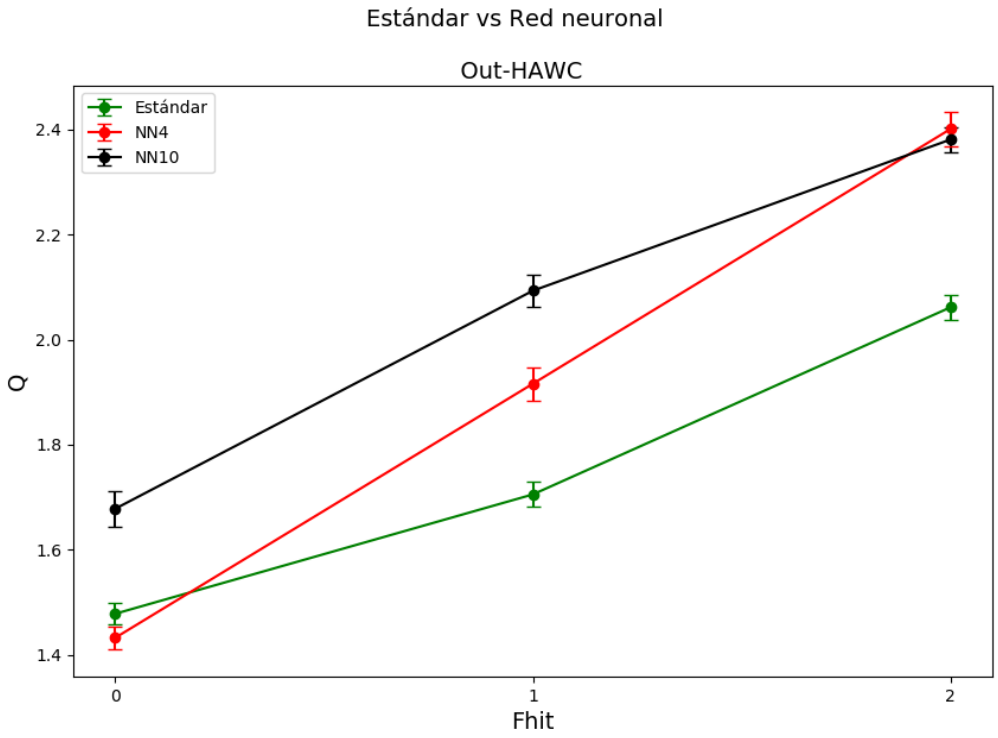


Ilustración 52. Se reporta el factor Q Out-HAWC en los fhit 0-2, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN4 (rojo) y el método estándar(verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se nota que el entrenamiento NN10 tiene un desempeño superior al estándar en los fhit 0 al 2, igualmente se puede notar que el entrenamiento NN4 es superior en los fhit 1 y 2.

- *Factor Q en fhit 3-7*

La Ilustración 53 muestra que en los fhit 3 al 5 para In-HAWC, los entrenamientos NN7 y NN10 son superiores en el factor Q al método estándar, mientras que en los fhit 6 al 7 el factor Q es estadísticamente igual debido a los márgenes de error. El error estadístico alto en estos fhit se debe a que como se muestra en la ilustración 47 la eficiencia en hadrón es muy baja en estos fhit, lo cual genera que la cantidad de eventos resultantes tipo hadrón sea igualmente muy baja con valores cercanos a 0, entonces si $Q = \frac{\epsilon gam}{\sqrt{\epsilon had}}$, al obtener valores cercanos a cero en la eficiencia en hadrón genera que el error estadístico sea alto.

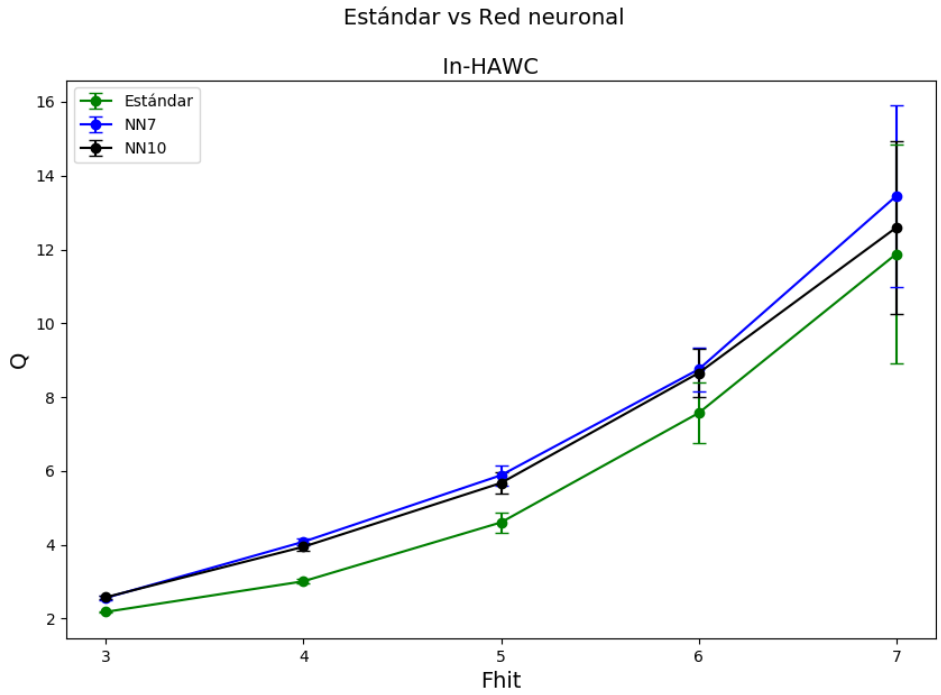


Ilustración 53. Se reporta el factor Q In-HAWC en los fhit 3-7, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN7 (azul) y el método estándar(verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se nota que los entrenamientos NN7 y NN10 son superiores al método estándar, mientras que en los fhit 6 y 7 no es posible determinar cuál es mejor debido al error estadístico.

En la ilustración 54 para Out-HAWC se observa que el comportamiento es similar que en In-HAWC ya que en los fhit 3 al 5, los entrenamiento NN10 y NN7 son superiores al método estándar, sin embargo en los fhit 6 al 7 no se puede determinar que método es el mejor debido al error estadístico, ya que como se menciona en la sección anterior, esto se debe a que la eficiencia en hadrón tiene valores cercanos a 0.

Tabla 8.

Factor Q Out-HAWC.

			Out-HAWC		
fhit	**Factor Q método estándar**	**Factor Q entrenamientos NN4, NN7, NN8**	**% de mejora NN4, NN7, NN8 vs Estándar**	**Factor Q entrenamientos NN10**	**% de mejora NN10 vs Estándar**
0	1.4784	1.4326	-3.0	1.6781	13.5
1	1.7055	1.916	12.3	2.0929	22.7
2	2.0615	2.4007	16.4	2.3806	15.4
3	3.0338	3.8325	26.3	3.8626	27.3
4	4.5884	6.8198	48.6	6.768	47.5
5	6.9898	10.7299	53.5	10.4369	49.3
6	11.1138	15.2721	37.4	15.8346	42.4
7	14.9849	26.5849	77.4	41.0841	174.1
8	39.4343	26.9347	-31.6	28.2374	-28.3
9	61.7548	63.2878	2.4	22.1675	-64.1
10	51.3682	127.0548	147.3	59.307	15.4

Se muestran los resultados de los factores Q Out-HAWC por fhit de los entrenamientos del método red neuronal NN4 (Rojo), NN7 (Azul), NN8 (amarillo) y el método estándar, en la primer columna se muestran los fhit en los que fueron calculados, en la segunda columna el factor Q del método estándar, en la tercer columna el facto Q de los entrenamientos NN4, NN7 y NN8, en la cuarta columna el porcentaje de mejora logrado de la red neuronal con los entrenamientos NN4,NN7 y NN8 respecto al método estándar, en la quinta columna el factor Q logrado por el entrenamiento NN10 y en la sexta columna el porcentaje de mejora logrado del entrenamiento NN10 respecto al método estándar, Se resalta el porcentaje de mejora logrado en los fhit 1 al 5, ya que en estos fhit los resultados no son afectados por el error estadístico.

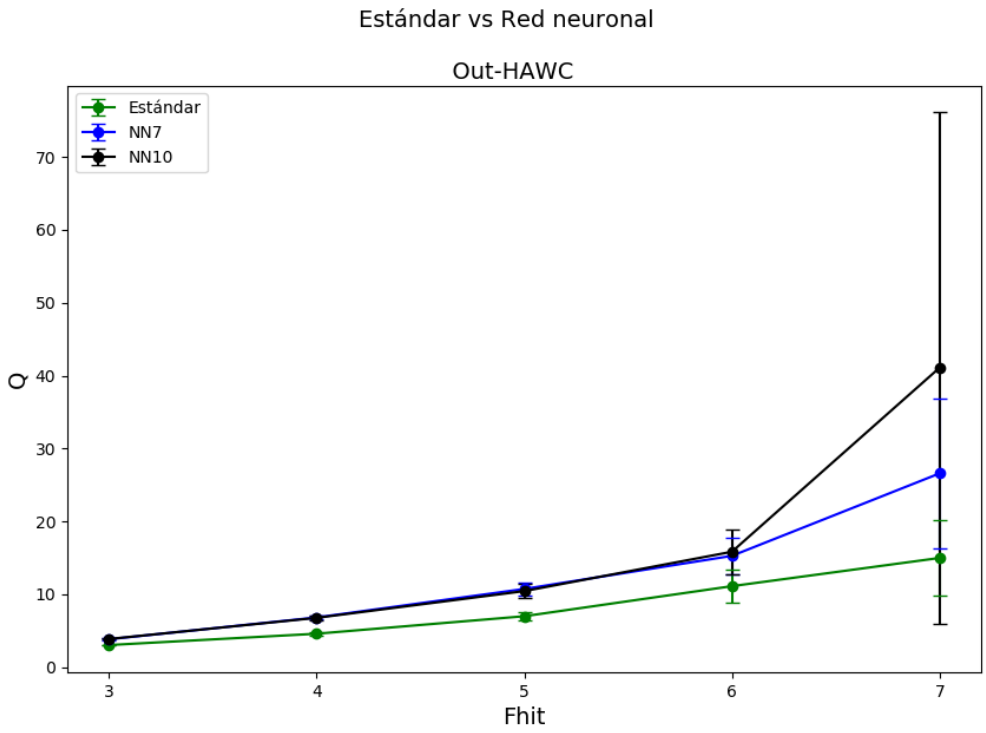


Ilustración 54. Se reporta el factor Q Out-HAWC en los fhit 3-7, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN7 (azul) y el método estándar(verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se observa que el factor Q de NN7 y NN10 es superior al estándar en los fhit 4 y 5.

- *Factor Q en fhit 8-10*

En las ilustraciones 55 y 56 no se puede determinar el método que tiene el mejor desempeño, debido a que los errores estadísticos del factor Q en estos fhit son altos, estos errores estadísticos se deben a la baja estadística y a que la eficiencia en hadrón en estos fhit tiene valores cercanos a 0, como se observa en las Ilustraciones 47 y 48.

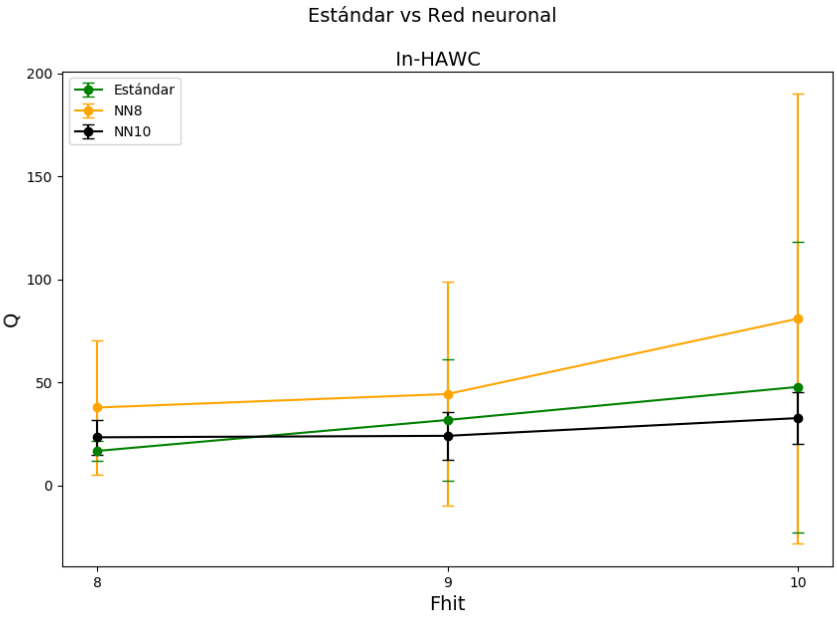


Ilustración 55. Se reporta el factor Q In-HAWC en los fhit 8-10, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN8 (amarillo) y el método estándar (verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se observa que no es posible determinar cuál método tiene el mejor desempeño debido a los errores estadísticos altos. Estos errores estadísticos son resultado de la baja estadística en estos fhit.

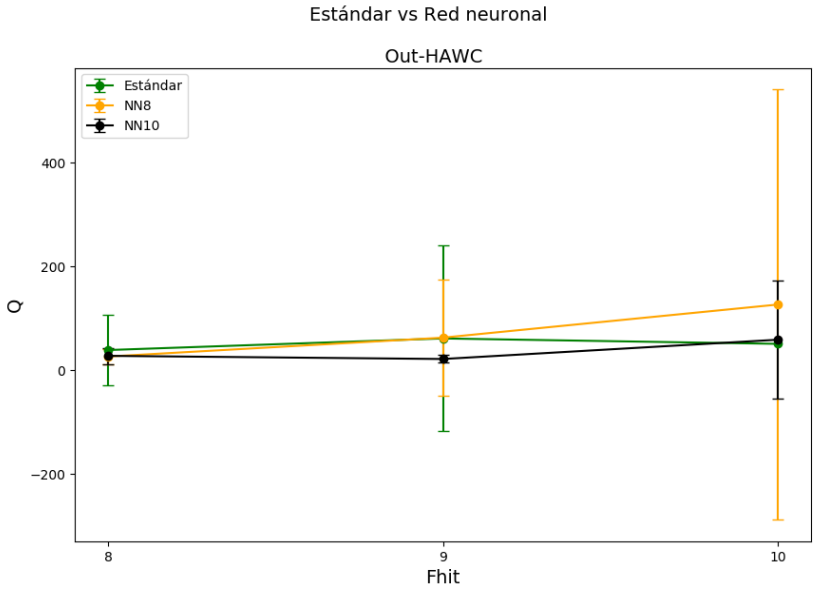


Ilustración 56. Se reporta el factor Q Out-HAWC en los fhit 8-10, de los métodos de separación; red neuronal con los entrenamientos NN10 (negro), NN8 (amarillo) y el método estándar (verde). El eje X muestra los fhit en los que fue calculado, mientras que el eje Y el valor de cada factor Q. Se observa que no es posible determinar cuál método tiene el mejor desempeño debido a los errores estadísticos altos. Estos errores estadísticos son resultado de la baja estadística en estos fhit.

Al analizar los factor Q de los métodos de separación se puede concluir que el modelo de entrenamiento NN10 basado en el método de red neuronal tiene un buen desempeño en los fhit 0 al 2 tanto para In-HAWC como en Out-HAWC, igualmente se puede decir que en los fhit 3 al 5 el factor Q de los entrenamientos NN10 y NN7 es más alto que el método estándar, por otro lado, en los fhit 6 al 10 no se puede concluir que modelo tiene el factor Q más alto debido a que los errores estadísticos en estos fhit son altos, por lo que para determinar que método de separación es el mejor en estos fhit, se debe recurrir a otra métrica como lo es la eficiencia en gamma.

4.3 Conclusiones

A lo largo de este trabajo se mostraron los resultados de eficiencia en gamma, hadrón y factor Q, de los entrenamientos del método de red neuronal, y se compararon con el método estándar de separación de partículas utilizado en el observatorio HAWC.

Para ello se propusieron diversos entrenamientos basados en el método de red neuronal, el primer grupo de entrenamientos, llamado NN10, utiliza 10 variables de entrada (LIC, PINC, LogNNEnergy, dismax, LDFamp, LDFChi2, nfit20, LDFAge, sos, LLH), después de analizar el desempeño de las variables en estos entrenamientos, se proponen nuevos entrenamientos, en donde las variables se depuran por grupos de fhit según su desempeño, así en los fhit 0-2 se utilizan 4 variables (PIN,LLH, LDFChi2,sos) para los entrenamientos y se nombran NN4, en los fhit 3-7 se usan 7 variables (LDFChi2, LLH, PINC, sos, LogNNEnergy, LIC, nfit20) y se nombran NN7, finalmente en los fhit 8-10 se usan 8 variables (LDFChi2, LLH, PINC, sos, nfit20, LDFamp, LIC, LDFAge) nombrándolos entrenamientos NN8. La arquitectura de la red neuronal que se usa en todos los entrenamientos es X:10:10:1 en donde X es el número de neuronas de la capa de entrada, 10:10 es la capa oculta con 10 neuronas y el último valor es la capa de la neurona de salida; si el valor de salida es 1 se considera un rayo gamma y si es 0 un rayo cósmico. Como datos de entrada se utilizaron datos simulados de las cascadas atmosféricas de las partículas de interés, los cuales se dividieron para las diferentes etapas de entrenamiento (25%), verificación (25%) y prueba (50%). Al finalizar la etapa de entrenamiento se verificó que no se cayera en un sobre entrenamiento, esto se logró, al comparar la salida de la red neuronal usando datos de entrenamiento contra la salida, pero ahora con datos de verificación, en la Ilustración 45 se muestra un ejemplo de cómo la distribución de las dos salidas es similar pero no idéntica, lo cual indica que no hay sobre entrenamiento. Después usando los datos de prueba se calcula la eficiencia en gamma y eficiencia en hadrón, finalmente con los resultados de las eficiencias se calcula el factor Q de los entrenamientos.

Al obtener las eficiencias y el factor Q de los métodos de separación, se compararon para poder medir su desempeño en la separación de partículas. Se puede concluir que para In-HAWC en los fhit 0, 1, 3 el método estándar sigue teniendo una mejor eficiencia en gamma, sin embargo el entrenamiento NN7 muestra el mejor desempeño en los fhit 4-7, y el NN10 en los fhit 8 al 10,

por lo que se puede notar que los entrenamientos propuestos basados en el método de red neuronal tienen en general un mejor desempeño en los fhit 4 al 10. Para Out-HAWC también se puede concluir que la eficiencia en gamma es mejor en los entrenamientos propuestos respecto al método estándar en los fhit 4-10, se destaca que el entrenamiento NN10 alcanza una eficiencia de casi un 90% en el fhit 9, logrando así ser un 80% mejor que el método estándar. Por otro lado, al analizar la eficiencia en hadrón para In-HAWC, se observa que en los fhit 0-3 en general los entrenamientos propuestos tienen un mejor desempeño respecto al método estándar, y en los fhit 4-10 tanto el método de red neuronal como el método estándar tienen un buen desempeño pues la eficiencia en hadrón logra valores cercanos 0. Para Out-HAWC en la Ilustración 48 se puede concluir que el entrenamiento NN4 es el que destaca en el desempeño en los fhit 0-2, y en los fhit 3 al 10 los dos métodos de separación tienen un buen desempeño en la eficiencia.

En la comparación de los métodos usando el factor Q, se analiza a detalle el comportamiento por cada fhit, así en las tablas 7 y 8 se reporta que los entrenamientos propuestos tienen un mejor factor Q en los fhit 1 y 2 tanto en In como en Out HAWC, también se observa un factor Q más alto en los fhit 3 al 5, sin embargo, en los fhit 6 al 10, no se puede determinar que método tiene un mejor desempeño debido a los errores estadísticos.

Dado que el nivel de estadística en la eficiencia en hadrón necesaria para calcular el factor Q en los fhit 6-10 es muy poca, por el bajo número de eventos resultante al realizar el corte de separación de partículas, lo ideal es que las conclusiones de desempeño se realicen tomando en cuenta el factor Q para los fhit 0 al 5, mientras que para los fhit 6 al 10 se comparará solo la eficiencia en gamma, debido a que en esta si se cuenta con la estadística suficiente, a partir de esto, se puede concluir que en los fhit 0 al 2 el modelo que mejor se comportó fue el de red neuronal usando el entrenamiento NN10, en los fhit 3 al 5 los entrenamientos NN7 y NN10 tuvieron un mejor desempeño sobre el estándar, en los fhit 6 y 7 el entrenamiento NN7 fue superior en la eficiencia en gamma, por último en los fhit 8 al 10, el entrenamiento NN10 obtuvo el mejor desempeño.

Referencias

Abeysekara, A., Albert, A., Alfaro, R., Alvarez, C., Álvarez, J., Arceo, R., . . . Brisbois, C. (2017). Observation of the Crab Nebula with the HAWC Gamma-Ray Observatory. *THE ASTROPHYSICAL JOURNAL 843-1*.

Atkins, R. e. (2003). Observation of TeV Gamma Rays from the Crab Nebula with Milagro Using a New Background Rejection Technique. *The Astrophysical Journal* , 595:803–811.

Atreidis, G. (2017). Numerical study of the electron and muon lateral distribution in atmospheric showers of high energy cosmic rays. *XIIth Quark Confinement & the Hadron Spectrum.*

Capistrán, T. (2020). Implementación de algoritmos para la optimización de detección de fuentes en el observatorio HAWC. Tonantzintla, Puebla, México: Tesis Instituto Nacional de Astrofísica Óptica y Electrónica.

Capistrán, T., Torres, I., Altamirano, L., & Collaboration, H. (August de 2015). New method for Gamma/Hadron separation in HAWC using neural networks. *The 34th International Cosmic Ray Conference,* (pág. 8). The Hague, The Netherlands: PoS. Obtenido de ResearchGate.

Capistrán, T., Torres, I., Moreno, E., & Collaborator, H. (2017). Characterization of a outer detector (outriggers) for HAWC. *Journal of Physics: Conference Series*, 792.

Collaboration, H. (2020). Constraints on Lorentz invariance violation from HAWC observations of gamma rays above 100 TeV. *Physical Review Letters 131101*, 124.

Collaboration, T. V. (2011). *Very Energetic Radiation Imaging Telescope*. Obtenido de Very Energetic Radiation Imaging Telescope: http://veritas.sao.arizona.edu/

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*.

Flynn, S. (2010). Gamma-ray/Hadron Separation Techniques for the HAWC. University thesis, University of Wisconsin - Madison.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. United States of America: O'Reilly Media, Inc.

Gussert, M. (2016). *A spectral analysis of the crab nebula and other sources with HAWC.* Colorado: Colorado State University.

H.E.S.S, C. (2004). *High Energy Stereoscopic System*. Obtenido de High Energy Stereoscopic System: https://www.mpi-hd.mpg.de/hfm/HESS/.

HAWC Collaboration. (2018). *Explicitly Compacted Data Format*. Obtenido de Explicitly Compacted Data Format: https://private.hawc-observatory.org/wiki/index.php/XCDF

HAWC collaboration. (2021). *HAWC HIgh Altitud Water Cherenkov Gamma-Ray Observatory*. Obtenido de HAWC HIgh Altitud Water Cherenkov Gamma-Ray Observatory: https://www.hawc-observatory.org/science/cosmicrays.php

INAOE. (2011). *High Altitude Water Cherenkov*. Obtenido de High Altitude Water Cherenkov: http://www.inaoep.mx/~hawc/.

Longair, M. (2011). *High energy astrophysics.* New York: Cambridge University Press.

Matthews, J. (2005). A Heitler model of extensive air showers. *Astroparticle Physics*, 22:387–397.

Stanev, T. (2004). High Energy Cosmic Rays. *Springer-Praxis Series in Astronomy and Astrophysics Series*.

Tsunesada, Y. (2010). The extreme energy cosmic rays. *AIP Conf.Proc*, 1238:82– 89.

Wang, X., Liao, W., Zha, M., & Cao, Z. (2019). Gamma hadron separation using single parameter method and multivariate algorithms with lhaaso-wcda experiment. *P o s (i c r c 2 0 1 9).*