UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS

Máster en Física Teórica

TRABAJO DE FIN DE MÁSTER

Redes neuronales robustas frente al envenenamiento de datos aplicadas al problema de la discriminación de los eventos de cuello en DEAP-3600

Robust neural networks against data poisoning applied to the neck event discrimination problem in DEAP-3600

Clara Álvarez Rodríguez

Directores:

Miguel Cárdenas-Montes

Vicente Pesudo Fortes

Curso académico 2021-22

Redes neuronales robustas frente al envenenamiento de datos aplicadas al problema de la discriminación de los eventos de cuello en DEAP-3600
Robust neural networks against data poisoning applied to the neck event discrimination problem in DEAP-3600

Clara Álvarez Rodríguez[*]
Universidad Complutense de Madrid

Miguel Cárdenas-Montes[**] y Vicente Pesudo Fortes[***]
CIEMAT (Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas)
(Dated: 8 de junio de 2022)

En los experimentos de búsqueda directa de materia oscura se requiere poder rechazar las distintas contribuciones de fondo para maximizar su sensibilidad. En el experimento DEAP-3600 una de las contribuciones más relevantes al fondo son las partículas α producidas en el cuello del detector. Este tipo de eventos producen una señal lumínica que debido a la geometría del detector puede ser confundida con la señal que la interacción de un WIMP con un núcleo de argón produciría. Para reducir esta contribución de fondo se imponen determinados cortes en la selección de eventos, que, colateralmente, reducen la aceptancia, y por ende la sensibilidad, significativamente. Previamente, se ha propuesto la implementación de redes neuronales en el proceso de clasificación de estos eventos de cara a mejorar la aceptancia manteniendo un poder de rechazo alto. En este trabajo se propone el método de la destilación para robustecer estas redes neuronales frente al envenenamiento de los datos simulados en el conjunto de prueba. Como resultado se ha evaluado la precisión de dos modelos de destilado sobre dos tipos de perturbaciones distintas añadidas en distintas proporciones sobre los datos de prueba, y se ha observado una mejora en poderes de rechazo inferiores al 99 %.

CONTENIDO

I. INTRODUCCIÓN

Según el modelo cosmológico actual, la materia oscura y la energía oscura suponen la componente principal del Universo. De acuerdo con las mediciones del Fondo Cósmico de Microondas (CMB) llevadas acabo por el experimento Planck [1], las principales densidades en el actual modelo cosmológico ΛCDM son:

$$\Omega_\Lambda = 0.6897 \pm 0.0057 \qquad \Omega_m = 0.3103 \pm 0.0057$$

El primer parámetro describe la densidad de energía oscura y el segundo la densidad de materia, dentro del cual distinguimos entre materia bariónica $\Omega_b = 0.0492 \pm 0.001$ y materia oscura $\Omega_c = 0.261 \pm 0.004$, suponiendo esta última entorno al 84 % de la materia presente en el universo.

La existencia de la materia oscura está respaldada por distintas observaciones astrofísicas y cosmológicas [2]. Su existencia es necesaria para explicar fenómenos como las curvas de rotación de las galaxias o el ligamiento gravitacional de los cúmulos. En ambos fenómenos, la cantidad de materia bariónica presente no es suficiente para explicar su dinámica por lo que debe existir una materia no luminosa y no bariónica que aporte la masa necesaria para que esto ocurra. La cantidad total de materia oscura se puede inferir también a partir de las medidas asociadas a la nucleosíntesis de elementos ligeros tras el Big Bang, que da información sobre la cantidad de materia bariónica que contribuyó a esa nucleosíntesis, y mediante las observaciones de las oscilaciones acústicas de bariones medidas a partir de la anisotropía del CMB.

De este modo se introduce la hipótesis de la materia oscura, un nuevo tipo de partícula elemental no bariónica que no interacciona electromagnéticamente, y por lo tanto no emite luz. Interacciona gravitatoriamente con la materia ordinaria y los experimentos de detección se basan en la hipótesis de que interacciona también vía algún

* clalva02@ucm.es
** miguel.cardenas@ciemat.es
*** vicente.pesudo@ciemat.es

proceso similar a la interacción débil. Debe ser, además, suficientemente estable para estar presente a día de hoy en el universo, y fría, es decir, ser no relativista actualmente, así como durante la evolución del Universo para explicar la actual Estructura a Gran Escala del Universo.

Entre los candidatos viables se estudian los neutrinos estériles, los axiones, los agujeros negros primordiales o los WIMPs. Una de las opciones que encaja de manera natural dentro de las descripciones de ΛCDM son los WIMPs (*Weakly Interacting Massive Particle*), que provienen de una extensión del modelo estándar. La detección experimental de estos se puede llevar a cabo a través de experimentos de detección directa, como DEAP-3600.

A lo largo de las últimas décadas los algoritmos de Aprendizaje Automático se han utilizado como herramienta en distintos problemas de física, entre los cuales se encuentran las redes neuronales que se utilizan para la discriminación de señal y fondo, donde señal se refiere a aquellos eventos que se busca detectar en el experimento, y fondo al resto de eventos que ocurren en el experimento pero no se quieren estudiar. Un problema de estos algoritmos es el empeoramiento en su precisión al clasificar datos que difieren del conjunto de entrenamiento, sea porque se han generado artificialmente o por la introducción de perturbaciones o ruido en la medida experimental. Para estudiar este problema, en este trabajo se proponen dos formas de envenenamiento de los datos simulados para el experimento DEAP-3600 y se introduce un mecanismo de defensa, el destilado, como método para reducir el impacto de los datos envenenados en la precisión de una red neuronal.

II. WIMPS

Los WIMPs (χ), candidatos a materia oscura, son partículas estables y neutras, que habrían sido producidas en el Big Bang y que se encuentran presentes en el halo alrededor de las galaxias. Mientras la temperatura del Universo es mayor a su masa $k_B T > m_\chi c^2$, su densidad se mantiene en equilibrio mediante procesos de creación y aniquilación. Tras bajar la temperatura por debajo de esta, el equilibrio se ve suprimido, pues la tasa de aniquilación cae por debajo de la tasa de expansión del Universo, por lo que dejan de aniquilarse. Además se suprimen los procesos de creación en los que otra partícula y su antipartícula se aniquilan con suficiente energía como para crear un WIMP y su antipartícula, y pasa a tener una abundancia reliquia [2].

Existen tres tipos de experimentos a la hora de detectar WIMPs. De detección directa, a través de la dispersión elástica de los WIMPs del halo galáctico con los núcleos blanco; la detección indirecta a través de la detección del producto de aniquilaciones de pares WIMPs; y su búsqueda en colisionadores de alta energía como producto de las colisiones.

Puesto que los WIMPs no interaccionan electromagnéticamente, en los experimentos de detección directa, se espera que produzcan una dispersión elástica en el núcleo, dando lugar a retrocesos nucleares (NR). Al contrario betas y gammas, que son uno de los principales fondos en estos experimentos, interactúan con los electrones produciendo retrocesos electrónicos (ER). Para recoger la energía producida por el retroceso nuclear según el experimento se utiliza: el centelleo, en el que fotones son producidos por desexcitaciones en el blanco; la ionización, cuando se mide la carga que el blanco libera a lo largo del retroceso; y el calor generado por fonones o vibraciones de red cristalina. Entre los materiales que se utilizan para estas funciones están los cristales, como el NaI (ANAIS [3]) y los elementos nobles, entre los que destaca el uso del xenón (XENON1T [4]) y el argón (DEAP-3600 [5]).

A. Detectores de argón

Los procesos de dispersión entre núcleos y WIMPs son no relativistas, pues los WIMPs son materia oscura fría. La energía de retroceso asociada a esta interacción depende de la masa del WIMP, de la masa del blanco y de la energía y ángulo de incidencia [6],

$$E_R = \frac{4M_W M_N}{(M_N + M_W)^2} \frac{(1-\cos\theta)}{2} E_0. \tag{1}$$

La interacción de los WIMPs con la materia ordinaria podría ocurrir mediante un canal dependiente o independiente de spin. El argón solo es sensible al canal independiente del spin, pues tiene spin nulo (A=18, Z=40). La sección eficaz para el canal independiente del spin viene descrita por [2]

$$\sigma_0 = \frac{4\mu^2}{\pi} \left[Z f_p + (A-Z) f_n\right]^2, \tag{2}$$

donde f_i son las funciones de acoplamiento de los WIMPs a los nucleones y μ la masa reducida del sistema WIMP-blanco.

El estado actual en la búsqueda mediante dispersión de nucleón-WIMP mediante el canal independiente spín se puede observar en la figura 1. Estos experimentos buscan explorar el espacio de parámetros de los WIMPs y delimitar su sección eficaz. El Suelo de Neutrinos es un límite experimental a la detección de materia oscura, y se produce cuando las señales de materia oscura se ocultan bajo la dispersión elástica coherente de neutrinos de diferente origen con los núcleos del blanco. Estas curvas experimentales excluyen los parámetros por encima de ellas, pues no se ha observado ninguna señal a pesar

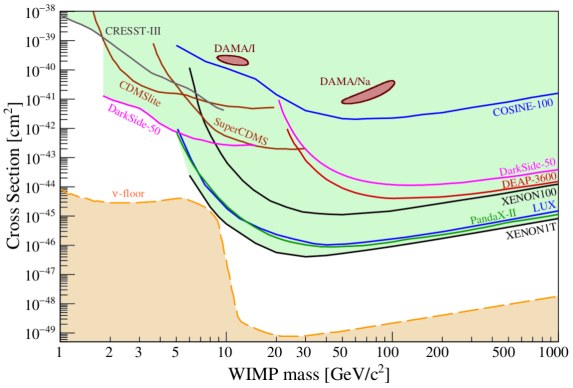


Figura 1: Límites superiores para la sección eficaz en el canal independiente de spin obtenidos a través de distintos experimentos. Figura obtenida de [7].

de que los experimentos tienen sensibilidad para medirla.

Las partículas cargadas que cruzan el medio de argón pierden energía principalmente al interactuar con los electrones de los átomos. Sin embargo, los neutrones y los WIMPs interactúan principalmente con los núcleos del medio, produciendo un retroceso nuclear que debido al movimiento puede excitar e ionizar los átomos en su camino. En ambos casos, estos átomos excitados o ionizados se adhieren a otros átomos en su estado fundamental, produciendo estados moleculares. Estos excímeros emiten luz de centelleo, desde un estado singlete o triplete, cuando la molécula excitada decae a dos átomos. Se emite un fotón con una energía del orden de 9.8 eV, que es menor que la energía requerida para excitar un átomo de argón a su primer estado excitado. Esto hace que el fotón no pueda ser reabsorbido y se propaga libremente dentro del detector. Por esta razón el argón es transparente a su propia luz de centelleo.

Los eventos de retroceso electrónico originados por betas, gammas y muones, dan lugar de manera dominante al excímero triplete, que tiene una vida media de (1.4−1.6) μs, mientras que los retrocesos nucleares producidos por neutrones o WIMPs dan lugar principalmente a estados singlete con una vida más corta ($\approx$ 6 ns). Esta diferencia de tiempos en el Ar es mucho mayor que en el Xe, y es uno de los mayores argumentos para su uso en la detección de materia oscura a través de la discriminación por forma de pulso (PSD). Esta discriminación se hace a través del parámetro F_{prompt}, que se define como la fracción de carga de un evento registrada en una ventana de tiempo más pequeña sobre el total de la carga del evento [5],

$$\mathrm{F_{prompt}} = \frac{\sum_{-28\ \mathrm{ns}}^{60\ \mathrm{ns}} \mathrm{PE(t)}}{\sum_{-28\ \mathrm{ns}}^{10\,\mu\mathrm{s}} \mathrm{PE(t)}}, \qquad (3)$$

y permite identificar qué excímero se ha poblado principalmente. Los NR suelen presentar valores altos por encima de 0.6, mientras que los ER tienen valores por debajo de 0.3. Este parámetro resulta eficiente para establecer un corte que elimine el fondo generado por el decaimiento β del isótopo radiactivo argón-39, que se trata de la principal contribución de contaminación en el argón.

III. EXPERIMENTO DEAP-3600

El detector DEAP-3600 (*Dark Matter Experiment using Argon Pulseshape Discrimination*) se encuentra en SNOLAB en Sudbury, Canadá. Está localizado a 2 km de profundidad bajo tierra, el equivalente a un aislamiento de 6 km de agua.

En la figura 2 se muestra un corte transversal del detector que consiste en un recipiente esférico de acrílico (AV) de 1.7 m de diámetro lleno de LAr ultra-puro. Su superficie interna está cubierta de tetrafenil butadieno (TPB, $C_{28}H_{22}$) que actúa como transformador de longitud de onda, absorbiendo la luz de centelleo del LAr a 128 nm y emitiendo aproximadamente a 420 nm, longitud de onda que puede ser detectada por los fotomultiplicadores [8]. Los 30 cm de la parte superior del AV están llenos con argón gaseoso (GAr), de modo que la interfaz liquido/gas se encuentra a 55 cm sobre el ecuador. El AV está rodeado por un conjunto de 255 tubos fotomultiplicadores (PMTs) orientados hacia el interior. Toda la información sobre el evento, principalmente su energía, su posición y el tipo de partícula que lo ha generado, se extrae a partir de la cantidad de fotones detectados por estos PMTs y de su patrón espacial y temporal.

Los PMTs están acoplados a unas guías de luz (LGs) de 45 cm, que transportan los fotones del interior del AV a los PMTs. Las LGs se introducen por la diferencia de temperaturas entre el LAr, que se encuentra a 85 K, y los PMTs que operan al rededor de 240-290 K, y para minimizar el número de rayos gamma que llegan al LAr, causados por la contaminación radiológica de los PMTs. El volumen entre las diferentes LGs está relleno de polietileno de alta densidad y Styrofoam, que actúa como blindaje de neutrones procedentes de los distintos componentes, así como de aislante térmico.

La simetría esférica del volumen del detector se rompe con una abertura en la parte superior del AV, que conduce a un cuello y un reborde acrílico. El cuello contiene un serpentín de enfriamiento de acero inoxidable lleno de N_2 líquido (LN_2). El LAr condensado entra al AV, dirigido por un conjunto de guías de flujo (FGs) de acrílico ubicadas en la abertura del cuello.

Todo este aparato se encuentra en una vasija esférica de acero inoxidable sumergido en un tanque de agua de 7.8 m de alto por 7.8 m de diámetro. La vasija contiene 48 PMTs orientados hacia afuera en su superficie exterior.

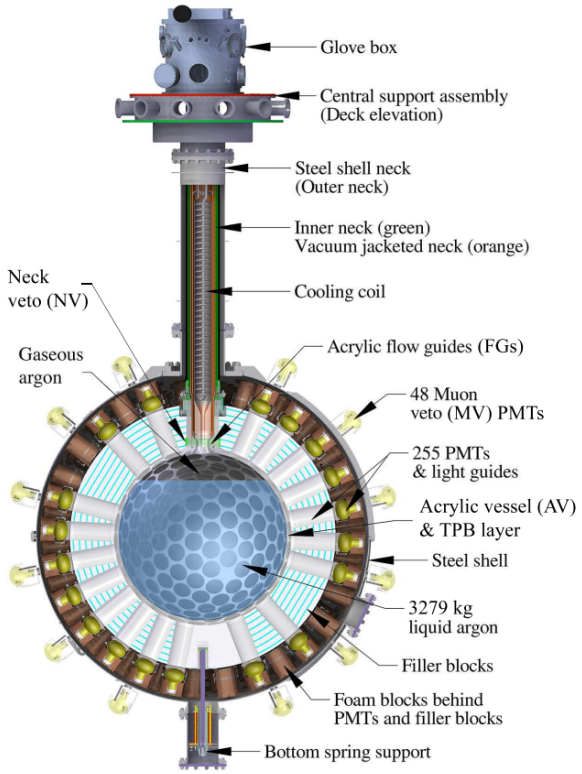


Figura 2: Corte transversal del detector DEAP-3600. Figura obtenida de [5].

Juntos, el tanque de agua y los PMTs sirven como un veto de muones sensible a radiación Cherenkov. Además el agua suprime parte del fondo de neutrones y gamma provenientes de la caverna en la que se encuentra.

A. Fondo y cortes

Una parte fundamental de un experimento de búsqueda de materia oscura es la caracterización de los fondos presentes en el detector y la discriminación de estos respecto a una potencial señal de interés. Esta discriminación se hace mediante cortes que, idealmente, rechacen eficientemente los eventos de fondos y afecten en pequeña medida a la señal buscada. Por ejemplo, la radiación externa deja preferentemente señales en las capas más externas del Ar por lo que se introducen diversos cortes fiduciales que descartan los eventos en los últimos cm de Ar adyacentes a la AV. Asimismo, solo se aceptan los eventos que se reconstruyen por debajo del nivel de llenado del LAr (z $<$ 550 mm) y dentro de un radio de 630 mm.

La región de interés (ROI) de WIMPs es una región en el espacio F_{prompt} vs. fotoelectrones (PEs) detectados, diseñada para la búsqueda de los retrocesos nucleares de baja energía que abarca el rango de 95-200 PEs, correspondiente a 48-100 keV, que se muestra en la figura 3.

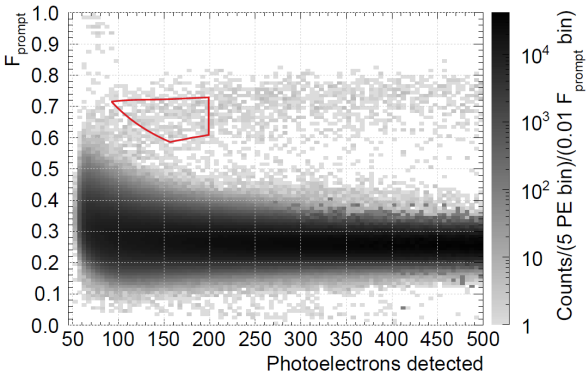


Figura 3: Distribución de F_{prompt} vs. PE para datos tomados con una fuente de neutrones AmBe. La ROI de búsqueda de WIMPs se muestra en rojo. Se puede ver la separación entre la banda NR ($F_{prompt} \sim 0.7$) y la banda ER ($F_{prompt} \sim 0.3$). Figura obtenida de [5].

En la ROI de interés definida, después de los cortes mencionados, los eventos de fondo más significativos se deben a neutrones y a eventos α degradados. El fondo de neutrones está dominado por reacciones (α, n) inducidas por decaimientos α en el detector. La tasa de neutrones se estima in situ contando los retrocesos nucleares seguidos por la captura de neutrones de rayos γ de alta energía de ^{1}H y ^{40}Ar dentro de una ventana de tiempo de coincidencia de 1 ms.

El fondo α se origina tanto en la masa de LAr, como en la superficie interna del acrílico. En el caso de las desintegraciones α en la masa de LAr, toda la energía se deposita en el LAr y da como resultado más luz de centelleo en comparación con la que se espera que se produzca en las interacciones con los WIMPs, lo que sirve como parámetro de discriminación para este fondo. Por el contrario, las desintegraciones α en la superficies del AV internas se eliminan de forma eficiente mediante la reconstrucción de la posición y seleccionando un volumen fiduciario en el centro del detector. Para mitigar eventos de tipo decaimientos α dentro del AV se aplica un corte adicional para eliminar eventos donde el 3.5 % o más de la carga total del evento se detecta en un solo PMT.

La mayor contribución al fondo después de aplicar los cortes anteriores proviene de las desintegraciones α de ^{210}Po en las superficies de las FGs acrílicas en el cuello de la AV. Como se observa en la figura 4, en la base del cuello del detector, por encima de la esfera, se encuentras dos FGs separados denominados FG interno y externo (IFG y OFG). De modo que hay tres superficies donde ocurren estos decaimientos, las superficies interior y exterior del IFG (IFG-IS e IFG-OS) y la superficie interior del OFG (OFG-IS). Las partículas α emitidas por estas desintegraciones tienen

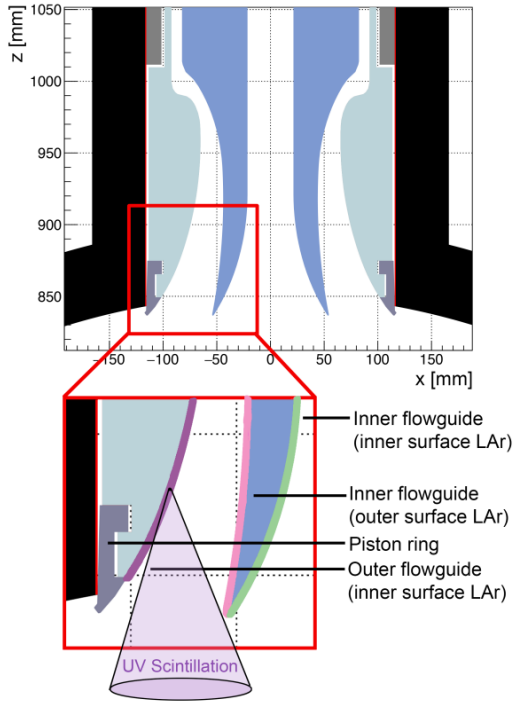


Figura 4: Corte transversal del cuello del detector DEAP-3600. Se observa las dos guías de luz internas donde se originan los decaimientos α problemáticos. Figura obtenida de [5].

lugar por encima de los PMTs, por eso los fotones que van hacia arriba se pierden y los PMTs solo detectan una fracción de los fotones totales producidos. Esto hace que tanto su energía como la reconstrucción de la posición puedan ser confundidos con los de la señal producida por un WIMP interactuando en el hemisferio sur del detector.

De cara a eliminar parte de este fondo, los eventos se rechazan si alguno de los 3 primeros pulsos observados se registra en la región GAr. Además debido al reflejo de los fotones UV en la interfaz GAr/LAr, los PMTs por encima del nivel de llenado ven en promedio una fracción mayor de PE por decaimiento en el cuello que por retrocesos similares a WIMP, de modo que se eliminan los eventos que tienen el 4 % de la carga total en los 2 anillos superiores de PMTs (10 PMTs en total).

La eficacia de los cortes es evaluada mediante el poder de rechazo, definido como la unidad menos la ratio entre el número de eventos de fondo que pasan los cortes ($N_{bg,s}$), sobre el total de eventos de fondo (N_{bg}):

$$R_f = 1 - \frac{N_{bg;s}}{N_{bg}}. \tag{4}$$

Por otro lado, el impacto de los cortes en la búsqueda de materia oscura se cuantifica a través de la aceptancia, la cual se define como la ratio de eventos de señal seleccionados:

$$A = \frac{N_{sg,s}}{N_{sg}} \tag{5}$$

donde N_{sg} es el número eventos señal y $N_{sg,s}$ corresponde al número de estos eventos que pasa el corte. Un problema asociado a los cortes introducidos para eliminar el fondo de partículas α del cuello reside en que para obtener un poder de rechazo alto, la aceptancia de eventos de retrocesos nucleares se reduce simultáneamente. Esto ha implicado que la colaboración buscase otras alternativas basadas en Aprendizaje Automático.

IV. REDES NEURONALES

Una Red Neuronal Artificial es un algoritmo de Aprendizaje Automático supervisado formado por neuronas o nodos, que son pequeñas unidades de computación sencilla. Su característica más distintiva es su arquitectura, que consta de una capa de entrada, una o varias capas ocultas y una capa final de salida, unidas cada una de las neuronas de una capa con las neuronas de las capas contiguas, como se representa en la figura 5. Dentro de cada neurona se realiza un ajuste lineal y se aplica sobre el resultado una función de carácter no lineal. La salida, a_j^ℓ, de la neurona j en la capa ℓ es

$$a_j^\ell = f_\ell\left(z_j^\ell\right) \qquad z_j^\ell = w_{jk}^\ell a_k^{\ell-1} + b_j^\ell \tag{6}$$

donde w_{jk}^ℓ y b^ℓ son los pesos y el sesgo asociado a esta neurona, los cuales se aplican sobre los resultados obtenidos en las neuronas de la capa anterior $(\ell - 1)$, $a_k^{\ell-1}$. La función f_ℓ se llama función de activación (o transferencia) y puede ser una función umbral, sigmoide, tanh, $ReLU$, $Softmax$ u otra en función de la red [9]. La finalidad de la aplicación de una función de activación sobre estos resultados es, además de introducir la no linealidad en la red neuronal, acotar el valor de la neurona para que la red neuronal no se paralice por neuronas divergentes. Se ha demostrado que una red neuronal construida con al menos una capa oculta puede aproximar cualquier función computable con una precisión arbitraria [10].

Para entrenar una red neuronal se inicia con un conjunto de pesos y sesgos aleatorios, y se entrena presentándole un conjunto de datos de entrada llamado conjunto de entrenamiento X, del cual se conocen las salidas (etiquetas) reales Y. Durante el entrenamiento se busca minimizar una función de error, ajustando los pesos entre las neuronas conectadas. En el caso de una clasificación se puede utilizar la función de entropía cruzada categórica,

$$C = -\sum_x y(x) \log\left(F(x)\right) \tag{7}$$

donde se suma sobre todos los datos de entrenamiento. $F(x)$ se refiere al resultado de aplicar la red sobre el

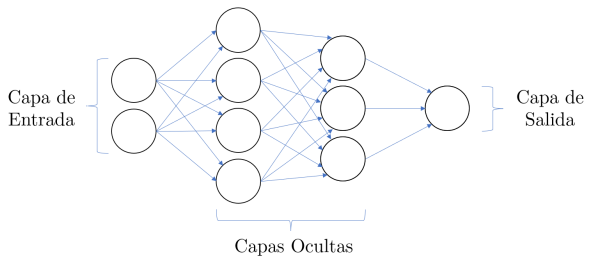


Figura 5: Red Neuronal Artificial prealimentada, con dos capas ocultas de 4 y 3 neuronas, una capa de entrada de 2 neuronas y una sola neurona en la de salida.

dato x e $y(x)$ a su etiqueta real. A partir de los errores medidos, se aplica un algoritmo de propagación del error hacia atrás con el que se varían los parámetros de las neuronas. Los hiperparámetros de la red son parámetros ajustables que permiten controlar el proceso de entrenamiento de un modelo, como el número de capas y número de neuronas por capa o el tamaño de los lotes de datos que se utilizan para entrenarla. De cara a la elección de estos se puede utilizar un conjunto de validación, para poder comparar el desempeño de cada una de las elecciones y utilizar la mejor. Una vez entrenada la red neuronal para conocer su capacidad de generalización se utiliza un conjunto de prueba para determinar el nivel de rendimiento de la red neuronal sobre datos con los que no ha sido previamente entrenada [10].

Para evaluar el funcionamiento de una red neuronal, se utilizan las funciones denominadas métricas. Estas se obtienen a través de la comparación de los resultados obtenidos al aplicar el modelo sobre el conjunto de prueba con su etiqueta real. La precisión mide la frecuencia con la que las predicciones equivalen a las etiquetas reales

$$\text{Precisión} = \frac{F(X) - Y}{N} \tag{8}$$

donde N se refiere al número de datos comparado. Entre estas métricas, para problemas de clasificación como el que estamos estudiando, destaca la matriz de confusión, en la que se representan en el eje horizontal las etiquetas reales y en el vertical las predichas por el modelo, como se representa en la tabla I. En el caso de una clasificación en dos clases, la diagonal contiene los valores de los verdaderos positivos (TP) y verdaderos negativos (TN), que son aquellos datos que la red ha predicho de forma correcta, mientras que los valores fuera de la diagonal son los falsos positivos (FP) y falsos negativos (FN), que son aquellos que la red no ha sido capaz de clasificar correctamente.

Además se suele trabajar con la ratio de verdaderos positivos (TPR) y la ratio de falsos positivos (FPR),

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{9}$$

Tabla I: Matriz de confusión.

		Etiqueta	
		Positivo	Negativo
Predicción	Positivo	TP	FP
	Negativo	FN	TN

que representan la ratio de verdaderos positivos frente al total de positivos reales del conjunto inicial y la ratio de falsos positivos frente al total de negativos reales del conjunto inicial. Estos parámetros se pueden relacionar con la aceptancia y el poder de rechazo a través de

$$A \equiv TPR \qquad R_f \equiv 1 - FPR. \tag{10}$$

Teniendo en cuenta que la salida de la red neuronal en una clasificación binaria se trata de un valor entre 0 y 1, como se representa en la figura 6, en el caso de una neurona de salida sigmoide, se debe imponer un umbral a partir del cual se clasifica como positivo o negativo a la salida. Para buscar el parámetro óptimo, de cara a maximizar la aceptancia para un poder de rechazo predeterminado, se utiliza la curva ROC (Característica Operativa del Receptor), donde se representa la aceptancia frente al poder de rechazo, según se varía el umbral de discriminación, como se representa en la figura 7.

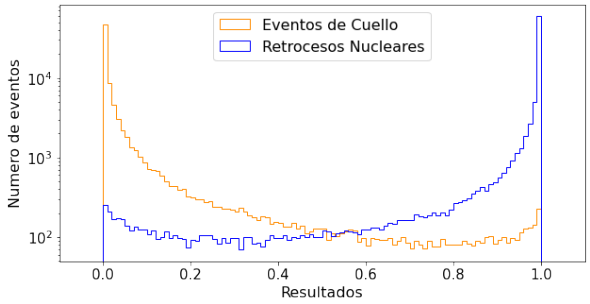


Figura 6: Resultado de la clasificación de retrocesos nucleares y eventos de cuello que forman el conjunto de test.

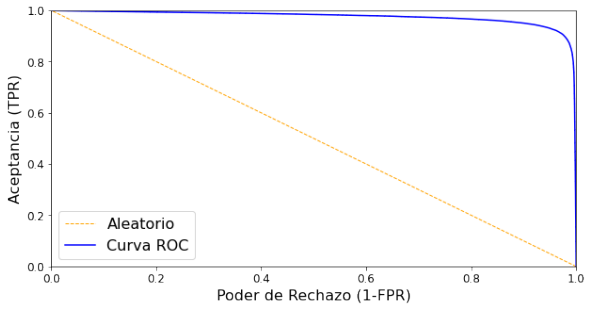


Figura 7: En azul se observa una curva ROC para un clasificador binario y en naranja un clasificador que funciona de forma aleatoria.

A. Mecanismo de Defensa

El funcionamiento de una red neuronal se puede ver afectado por la presencia de muestras adversarias entre los datos. Estas afectan a la capacidad de actuación de la red a través de su presencia durante su uso. Estos datos se pueden generar agregando perturbaciones δX a las entradas de la red X, que deben ser lo suficientemente pequeñas para no ser detectadas a través de otros mecanismos, pero lo suficientemente notables como para afectar en el funcionamiento de la red. El uso de muestras adversarias, está relacionado con que los datos reales nunca son ideales, pues existen limitaciones experimentales que deterioran el dato, por lo que hacer las redes neuronales robustas es de gran interés para su aplicación en física de partículas. En DEAP-3600, existen diversos motivos por los cuales los datos experimentales se pueden ver desplazados de los datos generados a través de las simulaciones utilizadas para entrenar los algoritmos de Aprendizaje Automático utilizados para el rechazo de eventos de fondo. La simulación de los eventos que tienen lugar en el detector es una operación compleja que abarca física de partículas, física nuclear, física química y óptica. En particular, el modelo óptico del detector tiene muchos parámetros y no es posible la caracterización independiente de todos ellos. Esto se traduce en posibles desviaciones entre los datos simulados y los datos experimentales. Los errores asociados a los distintos parámetros que se utilizan para generar estas simulaciones, como pueden ser el asociado al índice de refracción del argón o el asociado a la longitud de dispersión del TPB, pueden provocar que el entrenamiento con estos datos no sirva para generalizar la información a los posteriormente recogidos por el experimento. Además, los datos experimentales, se pueden ver afectados por variaciones asociadas a los distintos componentes del experimento, introduciéndose así pequeñas modificaciones que contribuyen a una degradación en la capacidad de clasificación de los eventos de señal y fondo.

El envenenamiento de datos es una técnica que, mediante un empeoramiento controlado de la calidad de los datos simulados, permite verificar cuan robusta es una red. El objetivo es introducir mecanismos de defensa que no generen un gran impacto en la arquitectura de la red, mantengan la precisión y la velocidad de la red inicial y sean capaces de defenderla contra muestras adversarias que no se alejen mucho de los datos reales. Un mecanismo de defensa contra estas muestras adversarias es el uso de la destilación durante el entrenamiento de las redes [11].

La red neuronal utilizada en este trabajo tiene dos neuronas en la capa de salida, correspondientes a las dos categorías posibles: señal y fondo. Por lo tanto las etiquetas asociadas a los distintos conjuntos de datos se encuentran codificadas en *one-hot encoding*, un vector con un uno en la posición asociada a la clase a la que pertenece y un cero en la que no. Además las neuronas de salida tienen como función de activación la función softmax,

$$\sigma(\mathbf{z})_j^\ell = \frac{e^{z_j^\ell}}{\sum_{k=1}^{K} e^{z_k^\ell}} \tag{11}$$

donde K es el numero total de clases y los exponentes han sido definidos en la ecuación 6. Esta función devuelve la probabilidad de pertenencia de los datos de entrada X a cada una de las clases, de modo que los datos de salida $F(X)$ son vectores de probabilidad. En los datos de entrenamiento, cada vector de entrada está asociado a una sola clase, con probabilidad total, a este etiquetado Y se le denomina *hard label*. El mecanismo de destilado se aprovecha de los vectores de probabilidad que genera la red $Y' \equiv F(X)$ y que se denominan *soft label* [11].

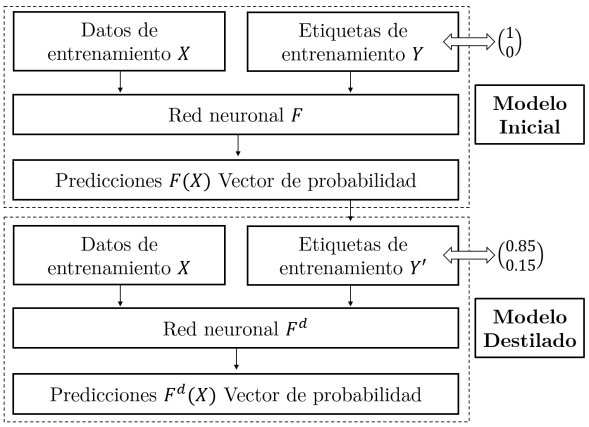


Figura 8: Destilado en dos modelos.

Este algoritmo, descrito por el esquema de la figura 8, consiste en el entrenamiento de una red con el conjunto de datos de entrenamiento X y su etiquetado *hard label* Y. Una vez entrenada, se utiliza para generar los vectores de probabilidad Y' (*soft label*) asociados a cada uno de los datos de entrenamiento. Se resetea el modelo, generándose de nuevo sus parámetros de forma aleatoria y se entrena con las etiquetas *soft level* Y'. Este segundo modelo es el denominado modelo destilado. El beneficio de usar *soft label* como etiquetas de entrenamiento radica en el conocimiento adicional que se encuentra en los vectores de probabilidad en comparación con las etiquetas *hard label*.

B. Ruido y modelo

Para el entrenamiento y evaluación de la red neuronal, se han utilizado un total de 585963 datos Monte Carlo generados en DEAP-3600 usando el código Geant4. De los cuales 292963 provienen de simulaciones de eventos de retroceso nuclear y 293000 de eventos de decaimientos α

del cuello del detector. Cada uno de estos datos consiste en un vector de longitud 255 correspondiente a la carga registrada (proporcional al número de fotones) por cada PMT según un índice correlacionado con la altura del PMT en el detector, estando el PMT_0 cercano al cuello del detector y el PMT_{254} en su polo sur. Esta carga es normalizada a la carga total del evento,

$$\frac{q_{PE;i}}{\sum_i q_{PE;i}}, \tag{12}$$

donde $q_{PE;i}$ es la carga recogida por el PMT i. Esta normalización se realiza para conseguir que la red aprenda patrones de luz. Este conjunto de datos se divide en un conjunto de entrenamiento con el 75 % de los datos y un conjunto de prueba con el 25 %. Sobre el conjunto de prueba se han introducido dos perturbaciones distintas. El "ruido a" (R_a) consiste en una distribución uniforme entre -1 y 1. El "ruido b" consiste en uno de los tres valores $[-1, 0, 1]$ con la misma probabilidad (R_b), multiplicado por cada uno de los valores del conjunto de prueba. Ambos ruidos se suman al conjunto de prueba en proporciones (M) del 1 %, 5 %, 10 % y 20 %.

$$X_a = X + \delta X_a \qquad \delta X_a = M \cdot R_a \tag{13}$$

$$X_b = X + \delta X_b \qquad \delta X_b = M \cdot R_b \cdot X \tag{14}$$

El nuevo dato se normaliza a través de la ecuación 12.

El modelo de red neuronal ha sido generado con la librería especializada en redes neuronales `keras` [12] y consta de una capa de entrada con 255 neuronas de lectura, una por cada PMT del detector. Dos capas ocultas de 64 y 32 neuronas, con función de activación *ReLU*. Y una capa de salida con dos neuronas, una por cada clase, señal y fondo, con función de activación *Softmax*. El entrenamiento se ha llevado a cabo con una función de coste de entropía cruzada categórica, y mediante el optimizador *Adam*. A partir de este modelo de referencia, se han construido dos modelos de destilado, el primero con un solo destilado y el segundo con dos, aplicando una segunda vez el mecanismo de destilado sobre el modelo destilado. Cada uno de los modelos se ha implementado sobre los distintos conjuntos de prueba un total de 25 veces por caso.

V. RESULTADOS Y ANÁLISIS

En la figura 9 se observa el efecto sobre la precisión del modelo de referencia que generan las dos perturbaciones introducidas sobre el conjunto de prueba en distintas proporciones. Se observa como la precisión disminuye a medida que aumenta el porcentaje de ruido introducido para ambos ruidos, destacando la degradación del rendimiento entre el "ruido a" y el "ruido b" para los ruidos de más del 10 %. El "ruido a" se trata de un

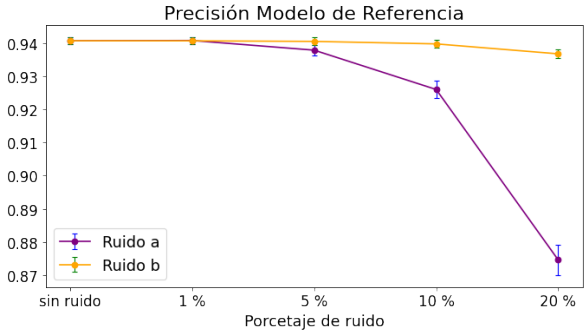
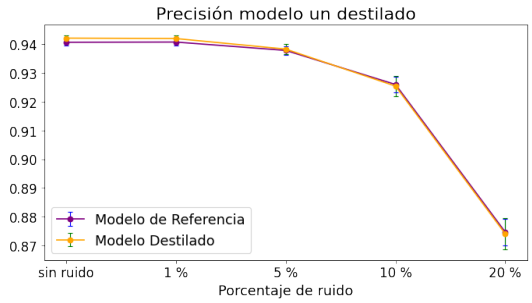
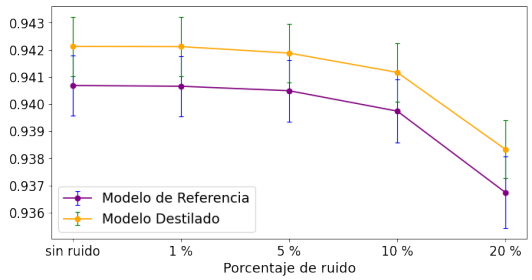


Figura 9: Precisión del modelo de referencia sobre el conjunto de prueba añadiendo los ruidos a y b en distintos porcentajes.

ruido más agresivo, pues cambia todos los porcentajes de luz recogidos por los PMTs, sin tener en cuenta el valor recogido por estos. Es por esto que este ruido degrada tanto el rendimiento, pues es capaz de introducir variaciones que afecten de forma significativa el patrón de luz. El "ruido b" es un ruido más conservador, pues solo afecta al 66 % de los PMTs, y como en cada evento parte de los PMTs no recoge carga, al ser proporcional al valor inicial, este porcentaje se reduce. Además, al estar relacionado con el valor inicial de los datos, afecta de forma menos significativa al patrón de luz, por lo que el degradamiento es menor que para el "ruido a".



(a) "ruido a"



(b) "ruido b"

Figura 10: Comparación de la precisión del modelo de un destilado, con umbral a 0.5, con el de referencia sobre el conjunto de prueba con ruidos a y b en distintos porcentajes.

Con el umbral de clasificación en 0.5, hemos obtenido

los valores de la precisión del modelo de referencia y de los modelos con uno y dos destilados al introducir el conjunto de prueba con los distintos ruidos. En la figura 10, se observa la precisión con la que el modelo de un destilado actúa sobre los distintos porcentajes de cada uno de los ruidos. Para el "ruido a", se aprecia la mejora en la precisión de los datos sin ruido y con ruido al 1 % y para el "ruido b" esta mejora se da de manera más notable y sobre todos los porcentajes, como podemos observar en los valores tabulados de la media y desviación estándar en la tabla II. Para observar si existe una mejora en el funcionamiento del modelo sobre los datos alterados con ruido, se realiza la prueba de los rangos con signo de Wilcoxon [13], un test no paramétrico que se utiliza para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas. Al realizar este test obtenemos significancia estadística para todas las mejoras mencionadas.

Tabla II: Precisión (media y desviación estándar) para un umbral de clasificación en 0.5. En negrita las comparaciones de los modelos destilados con la referencia con un test de significancia de 95 %.

		Referencia	Destilado 1	Destilado 2
sin ruido		0.9407(11)	**0.9421(11)**	**0.9414(16)**
a	1 %	0.9408(11)	**0.9420(11)**	**0.9414(16)**
	5 %	0.9379(15)	0.9383(18)	0.9381(19)
	10 %	0.926(3)	0.9254(11)	0.926(3)
	20 %	0.875(5)	0.8740(11)	0.875(6)
b	1 %	0.9407(11)	**0.9421(11)**	**0.9414(16)**
	5 %	0.9405(11)	**0.9419(11)**	**0.9412(16)**
	10 %	0.9397(12)	**0.9412(11)**	**0.9405(15)**
	20 %	0.9367(13)	**0.9383(11)**	**0.9377(16)**

Los resultados de aplicar el modelo de dos destilados sobre estos conjuntos de prueba alterados están tabulados en la tabla II. Para el "ruido a" se obtiene mejora en la precisión con significancia estadística para los datos sin ruido y con ruido al 1 %, pero de manera menos significativa que con el modelo de un destilado. Para el "ruido b" esta mejora se da sobre todos los porcentajes con significancia estadística, pero también con una mejora más pequeña que para el modelo con un solo destilado.

Desplazando el umbral de clasificación, hemos recogido los valores de la precisión de ambos modelos para un poder de rechazo de fondo (BRP) del 99 % y del 99.9 %. En la tabla III se recoge la media y desviación estándar de la precisión de los distintos modelos con un poder de rechazo del 99 %. Aplicando el test de Wilcoxon, se observa significancia estadística de mejora para todos los porcentajes del "ruido b" en los modelos de uno y dos destilados. En cuanto al "ruido a", la mejora en la precisión es menos significativa, pero tiene significancia estadística en ambos modelos para los datos sin ruido y con ruido al 1 % y al 20 %, además de con el modelo de un destilado para los datos con ruido al 5 %. Idealmente para la búsqueda de materia oscura se busca un poder de rechazo tan grande como sea posible. El mayor valor de rechazo estudiado en este trabajo es 99.9 %. En la tabla IV se observa que no se obtienen mejoras para ninguno de los dos destilados. Para el doble destilado, se obtiene significancia estadística de un empeoramiento general para la precisión sobre los datos con ambos ruidos.

Tabla III: Precisión (media y desviación estándar) para $R_f = 99\,\%$. En negrita las comparaciones de los modelos destilados con la referencia con un test de significancia de 95 %.

		Referencia	Destilado 1	Destilado 2
sin ruido		0.819(5)	**0.827(3)**	**0.821(6)**
a	1 %	0.820(5)	**0.827(4)**	**0.822(5)**
	5 %	0.811(6)	**0.816(7)**	0.812(6)
	10 %	0.768(9)	0.77(1)	0.76(1)
	20 %	0.61(1)	**0.62(1)**	**0.622(8)**
b	1 %	0.819(5)	**0.827(4)**	**0.821(6)**
	5 %	0.818(5)	**0.826(4)**	**0.820(6)**
	10 %	0.816(5)	**0.823(4)**	**0.817(6)**
	20 %	0.804(5)	**0.812(4)**	**0.806(6)**

Tabla IV: Precisión (media y desviación estándar) para $R_f = 99.9\,\%$. En negrita las comparaciones de los modelos destilados con la referencia con un test de significancia de 95 %.

		Referencia	Destilado 1	Destilado 2
sin ruido		0.55(2)	0.543(19)	**0.523(16)**
a	1 %	0.55(2)	0.55(2)	**0.524(16)**
	5 %	0.54(2)	**0.53(2)**	**0.51(2)**
	10 %	0.500(19)	**0.488(19)**	**0.47(2)**
	20 %	0.354(14)	0.360(14)	0.351(18)
b	1 %	0.54(2)	0.543(19)	**0.524(16)**
	5 %	0.54(2)	0.542(18)	**0.524(17)**
	10 %	0.542(18)	0.539(18)	**0.520(16)**
	20 %	0.53(2)	0.529(16)	**0.507(12)**

Por lo tanto, hemos observado como el degradamiento en la precisión del modelo de clasificación se ve aumentado por perturbaciones que no mantengan el patrón de luz de los datos, el porcentaje en el que se introduzcan y

el número de datos que se ven afectados. Además, hemos visto como para poderes de rechazo del 99 % y menores, se mantiene o mejora la precisión a través del destilado, obteniéndose mejores resultados para el modelo con un destilado que para el de dos destilados, mientras que para BRP del 99.9 % se empeora la capacidad de predicción del modelo.

VI. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se presenta un mecanismo para generar redes robustas para la clasificación de eventos en el experimento de búsqueda directa de materia oscura DEAP-3600. Concretamente, en la clasificación de eventos de señal de retrocesos nucleares, frente al fondo de eventos de decaimientos α del cuello del detector, que es uno de los fondos dominantes. Los cortes actuales para mitigar este fondo suponen una gran reducción en la aceptancia y el uso de redes neuronales se está estudiando para mejorar la sensibilidad del experimento. Este trabajo se engloba dentro de una de las líneas de trabajo de la colaboración, enfocada a reducir los errores sistemáticos asociados al uso de simulaciones Monte Carlo para definir algoritmos y umbrales de corte. Se propone un mecanismo de destilado de una red neuronal, para mejorar su precisión al enfrentarse a datos que contienen ruido, de cara a aumentar su robustez.

Como resultado de este trabajo se ha evaluado el degradamiento producido sobre un modelo de referencia de una red neuronal de clasificación al añadir dos perturbaciones distintas, una más agresiva y una más conservadora, sobre el conjunto de prueba. Se ha evaluado la precisión obtenida al introducir el mecanismo de defensa del destilado una y dos veces sobre este modelo de referencia, observando su rendimiento sobre los conjuntos de prueba perturbados. Se ha observado como para una clasificación con umbral de clasificación en 0.5 se obtienen mejoras con ambos modelos sobre todos los porcentajes del ruido menos agresivo, y para los más bajos del más agresivo, de forma más significa con un solo destilado. También se han observado estas mejoras para un BRP de 99 %. Sin embargo, ninguno de los modelos obtiene mejoras en la precisión para un BRP de 99.9 %, y en algunos casos se obtiene un empeoramiento en el desempeño de la clasificación.

Las ventajas asociadas a este método de defensa son que el modelo destilado mantiene la arquitectura del modelo inicial, mantiene o mejora la precisión para BRPs menores o iguales al 99 % y es igual de veloz que el modelo inicial, pese a un aumento en el tiempo de entrenamiento, debido a que el modelo se debe entrenar dos o tres veces en función del numero de destilados. El mecanismo del destilado para redes neuronales ha sido evaluado en visión por computación, pero no se ha implementado en física de partículas previamente, por lo que se presenta como un posible campo de estudio la implantación de esta técnica en otros problemas de clasificación en física de partículas.

[1] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro, N. Bartolo, S. Basak, *et al.*, Planck 2018 results-vi. cosmological parameters, Astronomy & Astrophysics **641**, A6 (2020).

[2] B.-R. Montes Núñez, Analysis of the first underground run and background studies of the argon dark matter experiment, (2016).

[3] J. Amaré, S. Cebrián, C. Cuesta, E. García, M. Martínez, M. Á. Oliván, Y. Ortigoza, A. O. de Solórzano, C. Pobes, J. Puimedón, *et al.*, The anais dark matter project: status and prospects, in *The Fourteenth Marcel Grossmann Meeting On Recent Developments in Theoretical and Experimental General Relativity, Astrophysics, and Relativistic Field Theories: Proceedings of the MG14 Meeting on General Relativity, University of Rome "La Sapienza", Italy, 12–18 July 2015* (World Scientific, 2018) pp. 2414–2419.

[4] E. Aprile, J. Aalbers, F. Agostini, S. A. Maouloud, M. Alfonsi, L. Althueser, F. Amaro, S. Andaloro, V. C. Antochi, E. Angelino, *et al.*, Search for coherent elastic scattering of solar b 8 neutrinos in the xenon1t dark matter experiment, Physical review letters **126**, 091301 (2021).

[5] R. Ajaj, P.-A. Amaudruz, G. Araujo, M. Baldwin, M. Batygov, B. Beltran, C. Bina, J. Bonatt, M. Boulay, B. Broerman, *et al.*, Search for dark matter with a 231-day exposure of liquid argon using deap-3600 at snolab, Physical Review D **100**, 022004 (2019).

[6] J. Lewin and P. Smith, Review of mathematics, numerical factors, and corrections for dark matter experiments based on elastic nuclear recoil, Astroparticle Physics **6**, 87 (1996).

[7] M. Schumann, Direct detection of wimp dark matter: concepts and status, Journal of Physics G: Nuclear and Particle Physics **46**, 103003 (2019).

[8] M. Kuźniak, D.-. collaboration, *et al.*, Status of the deap-3600 experiment, in *Journal of Physics: Conference Series*, Vol. 2156 (IOP Publishing, 2021) p. 012070.

[9] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms* (Cambridge university press, 2014).

[10] S.-C. Wang, Artificial neural network, in *Interdisciplinary computing in java programming* (Springer, 2003) pp. 81–100.

[11] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in *2016 IEEE symposium on security and privacy (SP)* (IEEE, 2016) pp. 582–597.

[12] F. Chollet *et al.*, Keras, `https://keras.io` (2015).

[13] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers* (John Wiley & Sons, 2010).