Optimización de un modelo de red neuronal convolucional 1D para identificar el sistema cristalino de compuestos inorgánicos binarios

Germán Torres Arroyave,[a] Sebastian Diaz Granados,[b] and Juan Manuel Albarracín[c]
Instituto de física, Universidad de Antioquia, Medellín,
Colombia 050010

En este estudio, se creó un modelo de red neuronal convolucional 1D (CNN-1D) para clasificar patrones de difracción de rayos X de compuestos inorgánicos binarios en sistemas cristalinos. Se empleó un extenso conjunto de datos de *Materials Project*, con diversas técnicas de preprocesamiento para mejorar el rendimiento. La primera prueba, con menos datos y sin desplazamientos, logró una precisión del 74,5 %. En contraste, la segunda prueba, con más datos y desplazamientos, alcanzó un 97,3 %. La ampliación de datos demostró ser crucial. La matriz de confusión y métricas específicas respaldan la capacidad del modelo, destacando un rendimiento excepcional en la clase "Cubic". Estos resultados sugieren la aplicabilidad del modelo en la interpretación automatizada de datos de difracción de rayos X para la caracterización de materiales.

Palabras clave: Red Neuronal Convolucional, Difracción de Rayos X, Sistemas Cristalinos, Materials Project, Ampliación de Datos, Rendimiento Predictivo, Clasificación de Estructuras Cristalinas, Caracterización Automatizada de Materiales.

[a]Electronic mail: german.torres@udea.edu.co
[b]Electronic mail: sebastian.diazgranadosc@udea.edu.co
[c]Electronic mail: juan manuel.albarracin@udea.edu.co

I. INTRODUCCIÓN

La difracción de rayos X (XRD) ha sido crucial para desentrañar las propiedades cristalográficas de diversos materiales, proporcionando información valiosa sobre sus estructuras. Esta técnica clásica ha experimentado una transformación significativa en el ámbito de la ciencia de materiales, gracias a la convergencia de la computación de alto rendimiento y las avanzadas técnicas de aprendizaje automático. Esta sinergia ha potenciado la capacidad de interpretar grandes volúmenes de datos, llevando la investigación a nuevas fronteras y desbloqueando un potencial sin precedentes para el análisis de materiales a escalas microscópicas.

En el contexto del Aprendizaje Automático para la Interpretación de Enormes Datos de Difracción de Rayos X de Sincrotrón, se han logrado avances prometedores. Destacan la eficacia de los modelos de Redes Neuronales Profundas (DNN) en la interpretación de datos de mapeo de microdifracción de rayos X (μ-XRD)[1] y los enfoques basados en datos para analizar escaneos de μ-XRD Laue de sincrotrón.[2]

Ampliando el foco hacia la ciencia de materiales y la cristalografía, se emplea el aprendizaje automático para descubrir nuevos materiales y mejorar la eficiencia en experimentos cristalográficos. Investigaciones sugieren que el aprendizaje automático tiene el potencial de revolucionar la interpretación de datos de difracción de rayos X de sincrotrón, anticipando avances a medida que se acumulan más datos y se desarrollan algoritmos más sofisticados.[1]

En este contexto, los difractogramas son fundamentales para caracterizar materiales y determinar estructuras cristalinas. Estos patrones de difracción de rayos X detallan la disposición atómica, facilitando la identificación del sistema cristalino de un material, el cual se refiere a la disposición ordenada y repetitiva de los átomos que lo componen. Existen siete tipos conocidos de sistemas cristalinos: triclínico, monoclínico, ortorrómbico, tetragonal, trigonal, hexagonal y cúbico. Cada uno exhibe una geometría única en su celda unitaria, la unidad más pequeña que conserva las propiedades del cristal.[3]

Al explorar las propiedades del cristal, nos enfrentamos al desafío de clasificar sistemas cristalinos utilizando técnicas de aprendizaje automático. Este proceso implica la aplicación de redes neuronales profundas para resaltar características no visibles al ojo humano. Al tratar cada difractograma como una imagen 1D de intensidades, asignando el sistema cristalino como etiqueta, se abre la posibilidad de desarrollar un modelo capaz de identificar sistemas cristalinos a partir de difractogramas. En este artículo, nos centramos en el desarrollo de un

modelo de clasificación para difractogramas de rayos X de compuestos inorgánicos binarios mediante una red neuronal convolucional 1D, con el objetivo de discernir entre diferentes sistemas cristalinos.

II. METODOLOGÍA

A. Plataforma de computación y librerías de software

Para la adquisición y preprocesamiento de datos, se optó por *Google Colab* como entorno de desarrollo, aprovechando la potencia de la API de *Materials Project*. A través de esta interfaz, se accedió a información detallada sobre las estructuras de diversos materiales. Posteriormente, haciendo uso de las bibliotecas `pandas`, `numpy`, `json`, `random`, y `pymatgen` en Python, se extrajeron y procesaron los difractogramas correspondientes. Los datos procesados fueron guardados como archivos con extensión `.JSON`.

En el proceso de compilación, entrenamiento y almacenamiento del modelo, se emplearon las bibliotecas `Keras`, `tensorflow` y `sklearn`. Dada la intensidad computacional del modelo, la ejecución se llevó a cabo en la *GPU T4* de *Google Colab*, con una capacidad de RAM de 12,7 GB y un almacenamiento en disco de 78,2 GB.

B. Adquisición de los datos de difracción de rayos X (Difractogramas) y preprocesamiento

Con el propósito de simplificar la cantidad de datos y la complejidad del modelo, se optó por concentrarse exclusivamente en compuestos inorgánicos binarios. La construcción de los conjuntos de datos de entrenamiento y validación se llevó a cabo mediante una combinación de datos experimentales y teóricos. Todas las muestras utilizadas provienen de la base de datos de *Materials Project*.[4]

Para evitar sesgos en el proceso de entrenamiento, se extrajo el mismo número de difractogramas en un rango 2θ de $(0°, 135°)$ para cada uno de los 7 tipos de sistemas cristalinos. De cada difractograma se generaron dos arrays: uno para la intensidad de pico y otro para su posición angular. Además, se extrajo la clase asociada, actuando como etiqueta para cada par de arrays y correspondiendo al sistema cristalino específico del material.

Para abordar las diferencias en las longitudes de los arrays de posición e intensidad entre

los difractogramas, se implementó un proceso de normalización. Este método implicó la creación de un nuevo array normalizado en el rango de 0° a 135° con un paso de 0,01° (es decir, 13500 puntos). Posteriormente, se ajustaron y asignaron las intensidades de los picos a este nuevo array, garantizando así una representación uniforme y coherente de los difractogramas, independientemente de las diferencias en las longitudes originales de los arrays.

Con los datos normalizados, se crearon dataframes individuales para cada sistema cristalino, conteniendo dos columnas: una correspondiente a los arrays de intensidades normalizadas y la otra a las etiquetas del sistema cristalino. Estos datos fueron almacenados en archivos con formato `.JSON`, generándose un archivo por cada sistema cristalino.

C. Aumento, diversificación y procesamiento de datos

Luego de recopilar los datos, los importamos en un nuevo notebook de Colab. Es relevante señalar que, en un primer intento, se seleccionaron 950 datos por sistema cristalino, totalizando 6650 datos, antes de intentar diversificar el conjunto mediante aumentos.

La estrategia empleada para diversificar el conjunto de datos implicó la generación de conjuntos sintéticos mediante desplazamientos de cada difractograma en variaciones de $-0{,}01°$ y $0{,}01°$. No obstante, las limitaciones de capacidad de almacenamiento temporal del notebook generaron preocupaciones, lo que condujo a la selección, en una segunda prueba, de 600 datos de cada sistema cristalino y la creación de 1200 datos en total a través de conjuntos sintéticos. En total, se utilizaron 12600 datos para representar los 7 sistemas cristalinos, concatenándolos en dos arrays: uno para las intensidades y otro para las etiquetas, donde cada array simboliza una imagen 1D de intensidades.

Posteriormente, se llevó a cabo la codificación numérica de las etiquetas y se procedió al desorden aleatorio de los datos. Este paso se implementó para mitigar sesgos inherentes al orden original después de la concatenación, asegurando así una distribución homogénea y aleatoria de los datos.

En un esfuerzo adicional por optimizar la adaptación de los datos al modelo, se reformó la estructura de los arrays de intensidad a $(12600, 13500, 1)$ y se normalizaron los datos al rango de 0 a 1. Adicionalmente, las etiquetas se reescribieron en formato "one-hot".

Finalmente, los datos se dividieron en conjuntos de entrenamiento y validación, siguiendo

una proporción del 90 % para el conjunto de entrenamiento y del 10 % para el conjunto de validación.

D. Arquitectura del modelo

Se construyó una arquitectura de red neuronal convolucional 1D (Conv1D) utilizando Keras. La capa de entrada de la red esperaba vectores de forma (13500,1). A continuación, se implementaron tres capas convolucionales, cada una con 60 filtros y tamaños de kernel de 100, 50 y 25 respectivamente. Cada capa convolucional incluía una capa de dropout con una tasa de 0.3 y todas eran seguidas por una capa de AveragePooling1D. Después de las capas convolucionales, los datos se aplanaron para prepararlos para las capas densas. La primera capa densa tenía 700 neuronas y la segunda tenía 70 neuronas, ambas incluían dropout y utilizaban la función de activación ReLU. Finalmente, la capa de salida era una capa densa con un número de neuronas igual a 7 (número de clases) y utilizaba la función de activación softmax, lo que la hacía adecuada para la clasificación multiclase. La normalización por lotes no se aplicó en ninguna capa.[5]

E. Entrenamiento y evaluación del modelo.

En la implementación del modelo, se utilizó la función de pérdida `categorical_crossentropy` y se evaluó el rendimiento mediante la métrica de `accuracy`. En cuanto al optimizador, se optó por `Adam`, ajustando su tasa de aprendizaje (`learning rate`) a 0,001.

En el proceso de entrenamiento, se estableció una proporción de validación del 20 %, con un tamaño de lote de 64 datos por iteración durante 70 iteraciones.

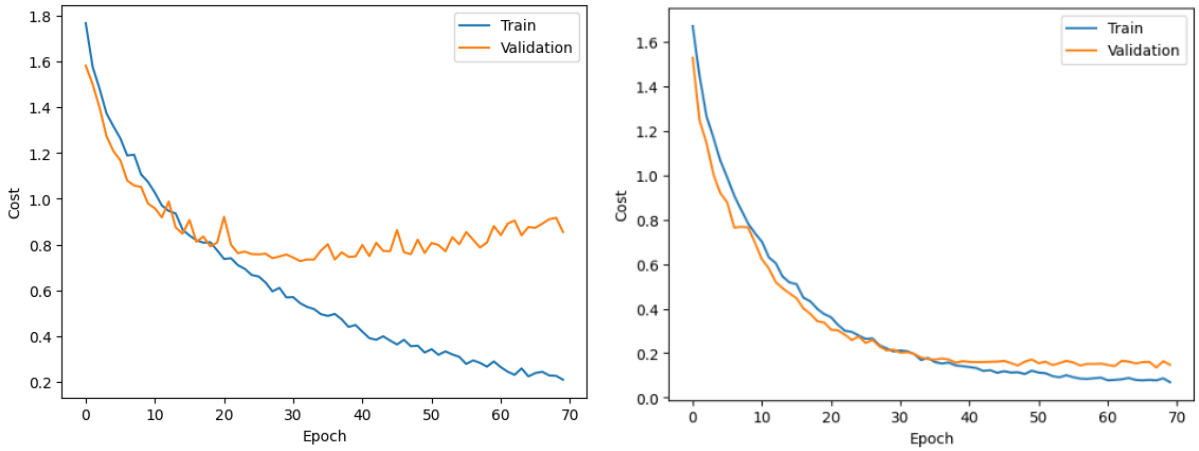
Al concluir, se generaron gráficos para visualizar la función de coste y la precisión del modelo en cada iteración para los conjuntos de datos de entrenamiento y validación en ambas pruebas realizadas. Además, se obtuvo la matriz de confusión para evaluar el rendimiento del modelo en la clasificación de las diversas clases.

III. RESULTADOS Y ANÁLISIS

A continuación se presentan los resultados obtenidos mediante la evaluación del modelo para los dos diferentes conjuntos de datos descritos en la metodología. La evaluación se basa

en la observación de las métricas clave, como la función de pérdida y la precisión, en los conjuntos de datos de entrenamiento y validación:

Según se observa en la Figura 1, en la primera prueba, los datos de entrenamiento convergen de manera suave y decreciente, lo que sugiere una fase efectiva de aprendizaje por parte del modelo. Sin embargo, los datos de validación muestran una tendencia diferente, con un aumento leve a partir de la iteración 20. Este comportamiento podría indicar un sobreajuste de los datos, sugiriendo la necesidad de ajustar la complejidad del modelo o reconsiderar el procesamiento de los datos. En esta prueba, el modelo logró una precisión de $74,5\,\%$ y una pérdida de prueba de 1,097.



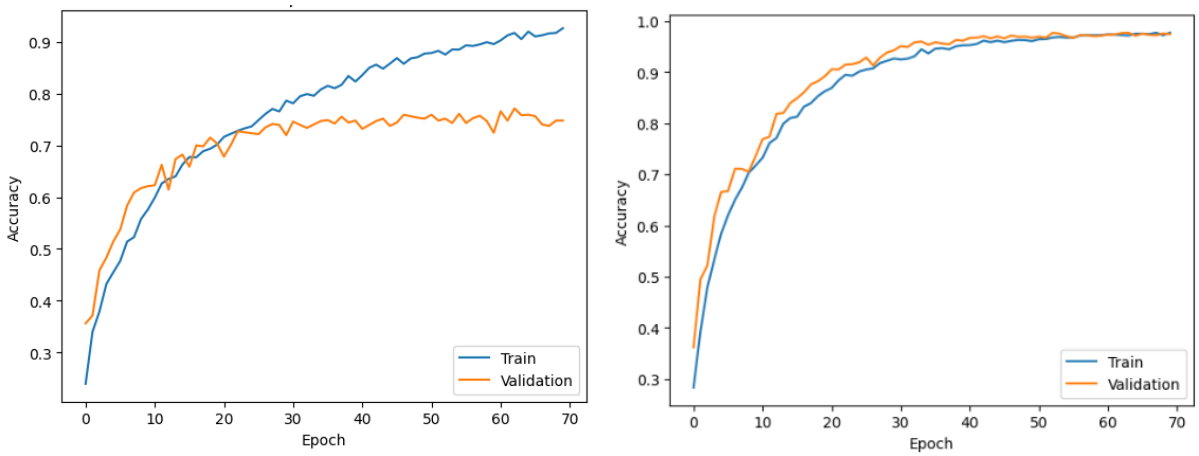
(a) Primera prueba

(b) Segunda prueba

Figura 1: Gráfico que representa la función de coste en función de cada época (o iteración) del entrenamiento y la validación

Por otro lado, en la segunda prueba, la gráfica muestra una convergencia suave y una disminución constante en ambos conjuntos de datos. En este caso, el modelo logró una precisión notable del $97{,}3\,\%$ y una pérdida de prueba de 0,1904. Este último valor, considerablemente más cercano a 0 en comparación con la primera prueba, nos indica una mejora significativa en la capacidad del modelo para generalizar y hacer predicciones precisas en el conjunto de prueba.

De forma análoga, al examinar la Figura 2, se nota un aumento progresivo de la precisión por iteración en ambas pruebas. Sin embargo, en la primera prueba, se destaca que la precisión del conjunto de validación no sigue el mismo patrón de crecimiento que la de los datos de entrenamiento, indicando la posibilidad de que el modelo no esté generalizando bien



(a) Primera prueba

(b) Segunda prueba

Figura 2: Gráfico que representa la evolución de la precisión del modelo a lo largo de cada época durante el proceso de entrenamiento y la validación

los datos nuevos. Este contraste no se evidencia en la segunda prueba, donde la precisión del conjunto de validación sigue la tendencia positiva de los datos de entrenamiento, sugiriendo una mejor capacidad de generalización del modelo en ese caso.

Cuadro I: Informe de clasificación para los datos de prueba

Classification Report				
	precision	recall	f1-score	support
triclinic	0.97	0.99	0.98	174
monoclinic	0.98	0.94	0.96	202
orthorhombic	0.96	0.98	0.97	169
tetragonal	0.98	0.98	0.98	190
trigonal	0.94	0.96	0.95	181
hexagonal	0.98	0.99	0.98	178
cubic	1	0.99	0.99	166

Los resultados revelan una influencia directa de la cantidad de datos de entrenamiento en la calidad predictiva del modelo. Este hallazgo subraya la importancia crucial de estrategias como los desplazamientos aplicados a los difractogramas para aumentar la cantidad de datos. Estas estrategias no solo amplían la diversidad del conjunto de datos, sino que también

contribuyen significativamente a mejorar la capacidad predictiva del modelo.

En el Cuadro I destaca el sólido rendimiento del modelo de clasificación al anticipar diversas estructuras cristalinas. Las métricas de precisión, recall y F1-score para cada clase (triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal y cubic) revelan un equilibrio eficaz en la identificación de instancias positivas y negativas. El rendimiento excepcional en la clase "Cubic" con una precisión del 100 % y un F1-score del 99 % se atribuye a la prevalencia de la estructura cristalina cúbica, la más simple y común. Estos resultados sugieren que el modelo puede hacer predicciones precisas y capturar variaciones en las estructuras cristalinas del conjunto de prueba, destacando en la matriz de confusión (Figura 3).

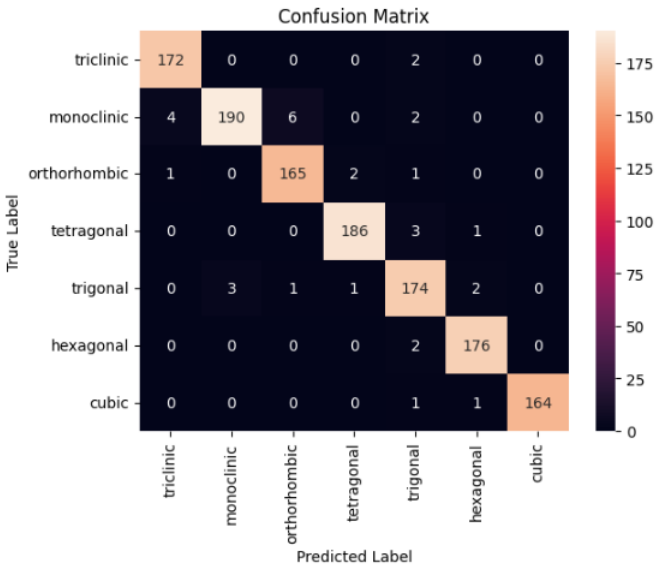


Figura 3: Matriz de confusión para la evaluación de la precisión del modelo en la segunda prueba

IV. CONCLUSIONES

Este estudio proporciona una evaluación detallada de la aplicación de un modelo de Red Neuronal Convolucional 1D (CNN-1D) en la identificación de sistemas cristalinos a partir de patrones de Difracción de Rayos X (XRD). Destaca la importancia crucial de estrategias de preprocesamiento y la inclusión de datos experimentales en el entrenamiento del modelo para mejorar su robustez y precisión. La comparación entre instancias sin y

con desplazamiento proporciona una valiosa perspectiva sobre las posibles variaciones en el rendimiento del modelo, ofreciendo una comprensión más completa de su aplicabilidad en escenarios específicos.

El estudio demuestra de manera ilustrativa cómo los modelos de Redes Neuronales Convolucionales 1D pueden desempeñar un papel fundamental en la identificación de sistemas cristalinos a partir de patrones de difracción de rayos X. Los tratamientos aplicados a los datos, incluida la aleatorización, han sido efectivos para reducir posibles sesgos y mejorar la capacidad del modelo para generalizar a nuevos datos.

A pesar de las limitaciones computacionales, el uso eficiente de la GPU en Google Colab ha permitido la ejecución rápida del modelo, destacando la viabilidad de implementaciones prácticas en entornos de recursos de libre acceso.

Es relevante señalar que, aunque este estudio se enfoca específicamente en compuestos inorgánicos binarios, los métodos empleados poseen un potencial de aplicabilidad más amplio en otros ámbitos de la cristalografía. La adaptabilidad del modelo y las estrategias empleadas sugieren que estas técnicas podrían ser exploradas y extendidas a problemas similares en la caracterización de materiales..

V. DESCARGO DE RESPONSABILIDADES

ChatGPT4 fue empleado con el fin único de mejorar la sintaxis y ortografía del texto. Toda referencia e investigación fue realizada por los integrantes del grupo.

REFERENCIAS

[1]Zhao X., Luo Y., Liu J., Liu W., Rosso K. M., Guo X., Geng T., Li A., Zhang X. (2023). Machine Learning Automated Approach for Enormous Synchrotron X-Ray Diffraction Data Interpretation. *ArXiv*. `https://arxiv.org/abs/2303.10881`

[2]Song Y., Tamura N., Zhang C., Karami M., Chen X. (2019). Data-driven approach for synchrotron X-ray Laue microdiffraction scan analysis. *ArXiv*. `https://arxiv.org/abs/1909.06572v1`

[3]Cruz-Gandarilla F., Calyco C. (2005). Aplicaciones de la Difracción de Rayos-X a Materiales Policristalinos. *ResearchGate*.

https://www.researchgate.net/profile/Francisco-Cruz-Gandarilla/publication/273458745_Aplicaciones_de_la_Difraccion_de_Rayos-X_a_Materiales_Policristalinos/links/550340b30cf24cee39fd6d77/Aplicaciones-de-la-Difraccion-de-Rayos-X-a-Materiales-Policristalinos.pdf

[4]Materials Project. https://next-gen.materialsproject.org

[5]Clase 11 parte 3.1- Implementación de una CNN 1D," *TSFCIIIIAyaplicaciones*, YouTube, 2021. https://www.youtube.com/watch?v=7k5fzoqkkCQ&ab_channel=TSFCIIIIAyaplicaciones