Predicción de banda prohibida para óxidos de alta entropía (Fluorite y Mukesh) con machine learning y deep learning

Buitrago Garcia Wilhelm David[1], Rueda Mayorga Jose Luis[2], & Vasquez Rivera Héctor Daniel.[3]

[1]Estudiante de pregrado en Ingeniería Mecatrónica (e-mail: wibuitragog@unal.edu.co)
[2]Estudiante de pregrado en Ingeniería Mecatrónica (e-mail: jruedam@unal.edu.co)
[3]Estudiante de pregrado en Ingeniería Mecatrónica y Estadistica (e-mail: hevasquezr@unal.edu.co)

RESUMEN Este estudio se centra en la predicción del band gap en óxidos de alta entropía mediante inteligencia artificial, utilizando datos de espectroscopia UV-Vis transformados en gráficas Tauc's. Se desarrollaron modelos de Machine Learning (ML) y Deep Learning (DL). En ML, se entrenaron modelos de RandomForest, Linear-Regresor y DecisionTree de dos formas distintas: en el primer método, se utilizaron los algoritmos mencionados directamente para entrenar los modelos, mientras que en el segundo método se extrajeron características de los datos de absorbancia y se redujo la dimensionalidad utilizando PCA antes de entrenar los modelos. Aunque ambos métodos de ML mostraron precisión, el segundo método fue superior, con RandomForest obteniendo un RMSE de 0.39 eV y un R^2 de 65%. En DL, se implementaron dos enfoques diferentes: uno utilizando una red neuronal convolucional 1D y otro desarrollando una arquitectura de red convolucional personalizada. En el segundo enfoque de DL, se entrenó una red convolucional con imágenes de los diagramas Tauc's en formato PNG binarizado, el segundo método fue el mas destacable, obteniendo un RMSE de 0.36 eV con un R^2 de 68%. Estos resultados demuestran una precisión significativa en la predicción del band gap. Este estudio resalta la eficacia de la inteligencia artificial en la predicción de propiedades de materiales, contribuyendo al avance en la ciencia de materiales.

PALABRAS CLAVES Band gap prediction, High entropy oxides, Deep Learning, Machine Learning.

I. INTRODUCCIÓN

Desde los albores de la humanidad, los materiales han sido intrínsecos al desarrollo de la sociedad, ya sea obtenidos de la naturaleza o producidos por el ser humano. A lo largo de la historia, se han destacado periodos donde ciertos materiales han tenido una relevancia crucial, como en la Edad de Piedra, cobre, bronce y hierro.

El ser humano ha utilizado estos materiales de diversas maneras, desde la confección de vestimenta hasta la construcción de herramientas y tecnología. Sin embargo, en un mercado siempre cambiante, estos materiales pueden volverse obsoletos frente a nuevos desafíos, lo que impulsa a las industrias a buscar constantemente innovaciones.

En este sentido, el desarrollo e investigación de nuevos materiales es esencial para el avance de la sociedad. Aunque existen clasificaciones comunes como metales, polímeros, cerámicas y materiales híbridos, se ha reconocido la necesidad de explorar y desarrollar nuevas categorías de materiales capaces de sustituir con un mejor rendimiento materiales que usamos actualmente. Es por esto por lo que recientemente, ha surgido un creciente interés en una nueva clase de materiales denominados "materiales de alta entropía", los cuales pueden llegar a ofrecer propiedades optimizadas para diversas aplicaciones.

Los materiales de alta entropía, como los óxidos de alta entropía, son una clase única de materiales compuestos por una combinación equimolecular de al menos cinco elementos diferentes, que presentan una distribución aleatoria de átomos en su estructura cristalina. Los óxidos de alta entropía como su nombre indica, consiste principalmente en óxidos en lugar de metales u otros compuestos. Esta diferencia en la composición química puede resultar en propiedades únicas, por lo que es imperativo el estudio de sus características físicas, química, termodinámicas, ópticas, etc. En esta lista de características, el band gap (Energía de banda prohibida) juega un papel importante ya que, permite clasificar los materiales dentro de tres (3) grandes grupos: Conductores, Semiconductores y Aislantes.

En este contexto, en “Bandgap formation in graphene doped with BN, TiO2, Al2O3 and ZnO by sintering process” (Rajib et. al 2023), se define el band gap como la cantidad de energía necesaria para trasladar un electrón de la banda de valencia a la banda de conducción. Este principio se emplea ampliamente en el análisis de las propiedades fotocatalíticas de estos materiales. También se menciona que en 1966, Tauc propuso el uso de espectros de absorción óptica para determinar esta energía en semiconductores amorfos, método posteriormente refinado por Davis y Mott. Tauc sugirió una fórmula específica para calcular el coeficiente de absorción en relación con la energía, expresada como:

$$(\alpha.hv)^{\frac{1}{\gamma}} = B(hv - E_g)$$

Donde h representa la constante de Planck, v es la frecuencia de un fotón, E_g es la energía del band gap y B es una constante.

Se menciona que la forma de determinar el band gap a menudo se realiza mediante el análisis de espectros de reflectancia difusa. En los materiales semiconductores, existe una región donde la absorción de luz experimenta un aumento significativo y lineal conforme aumenta la energía. En el análisis gráfico de Tauc, la energía del band gap se calcula en el punto de intersección entre un ajuste lineal y el eje x, el cual representa el band gap en la región de linealidad mencionada previamente. [1].

Cabe mencionar que, el estudio de estas características puede ser un proceso lento y costoso, lo que ha llevado al uso creciente de técnicas de Inteligencia Artificial, como Machine Learning (ML) y Deep Learning (DL), para mejorar la eficiencia, el rendimiento y la precisión en la caracterización del band gap en materiales de alta entropía. En este contexto, el presente trabajo tiene como objetivo desarrollar modelos de ML y DL capaces de predecir el band gap en óxidos de alta entropía, con el fin de acelerar el proceso de obtención y caracterización de estos materiales.

Es por ello que en sintonía con lo anteriormente descrito, en “From prediction to design: Recent advances in machine learning for the study of 2D materials" (He et al., 2023). Se hace una revisión exhaustiva de las aplicaciones multifacéticas del ML en el estudio de materiales 2D, llenando el vacío en este campo. Se encuentra resumidos los últimos desarrollos de ML para la predicción de band gap, la clasificación magnética, el cribado de materiales catalizadores y el diseño de síntesis de materiales. En lo que respecta a la predicción de band gad se presentan los distintos tipos de algoritmos de ML que se han usado para este propósito obteniendo las mejores estimaciones con arboles de decisión y random forest, para el entrenamiento de los modelos en la mayoría de los casos se usa la base de datos “C2DB" que documenta las propiedades eléctricas y mecánicas de mas de 4000 materiales bi-dimensionales (2D) [2], [3].

Seguidamente, en el articulo “Machine-learning prediction of the computed band gaps of double perovskite materials” (Zhang et al., 2023). Se hace la aplicación de diferentes tecnicas como; Random Forest, Tuplewise Graph Neural Networks (TGNN), Alternating Conditional Expectations (ACE) y Kernel Ridge Regression (KRR). Para la predicción de band gaps en materiales en estado solido de “DOUBLE PEROVSKITE”, obteniendo una precisión de 85.6% con un error cuadrático medio de 0.64 eV. Cabe resaltar que en el artículo, de las 20 características suministradas se realiza el entrenamiento del mejor modelo con las 10 más importantes, es decir, que previamente a entrenar el modelo, se hace una inferencia entre las características para conocer la relevancia de cada una en la predicción del band gap [4].

Finalmente, en otro estudio titulado ’Bandgap prediction of two-dimensional materials using machine learning’ (Zhang et al. 2021), se aplicaron cuatro algoritmos de ML: árboles de decisión potenciados por gradiente, bosques aleatorios, máquinas de vectores de soporte (SVM) y deep learning con perceptrón multicapa. Los resultados mostraron que los árboles de decisión y los bosques aleatorios son los más precisos en la predicción de la brecha energética, con un coeficiente de determinación (R^2) superior al 90%, mientras que los SVM y el perceptrón multicapa alcanzaron un R^2 cercano al 70%. Además, se observó que al no utilizar la característica del acoplamiento espín-órbita (SOC), los cuatro modelos mejoraron su rendimiento, presentando un R^2 superior al 94%. En conclusión, los autores destacan que los algoritmos de ML logran una rápida predicción con una alta precisión [5].

II. METODOLOGÍA

Se presentan dos datasets en forma de .csv correspondientes a información obtenida de espectroscopia UV-Vis transformada a puntos de un diagrama Tauc’s (vease figura 1)realizada a dos óxidos de alta entropía: Fluorite con 530 datos y Mukesh con 2560 datos.

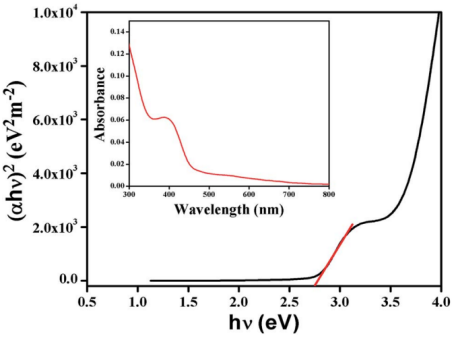


FIGURE 1: Ejemplo de UV-Vis (Wavelength vs Absorbance) visto como un gráfico Tauc’s, Tomada de Photosensitive Schottky barrier diode behavior of a semiconducting Co(III)–Na complex with a compartmental Schiff base ligan (2019)

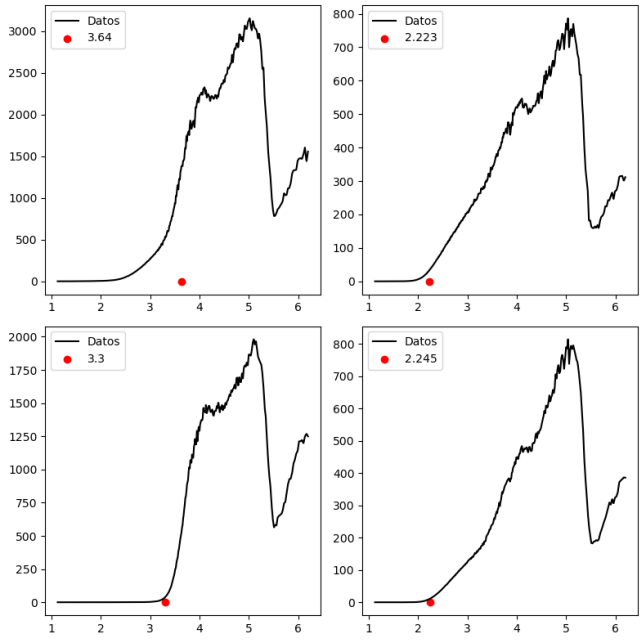


FIGURE 2: Visualización de los datos

Cada conjunto de datos presenta dos columnas: x1,x2. Donde x1 es una lista unidimensional (1D) que representa la energía de los fotones incidentes (eV) y x2 es una lista unidimensional (1D) de datos correspondientes al absorbencia del material para fotones incidentes. Ambas en forma de cadena de texto (string).

	x1	x2
0	"[1.12726309532197, ⋯ , 1.1293180115103607]"	"[1.12726309532197, ⋯ , 1.1282891794029504]"

TABLE 1: Ejemplo de un posible dato

Cabe aclarar que adicionalmente, el dataset de Flourite presenta una columna denominada "band_gap_std", sin embargo debido a que solamente se encuentra en este dataset y es una medida hallada a partir de la predicción a realizar, se ha decidido por la eliminación de la misma.

Dado que la columna x1 esta presente en ambos datasets con el mismo rango y forma se decidió eliminarla, sin embargo x2 se encuentra relacionado con x1 y este ultimo no presenta un comportamiento normalmente espaciado y eliminarlo representaría perdida de información en x2, por tanto se decide hacer una interpolación a los datos, haciendo que queden equidistantes

A. MACHINE LEARNING

La predicción de bandgap se abordo con algoritmos de ML, este se realizo desde dos enfoques con el fin de explorar las diferentes formas con las cuales se podría realizar la estimación, de los cuales surgieron los siguientes métodos:

1) Método 1

Este método consiste en entrenar algoritmos de Machine Learning usando la totalidad de los datos 1D de x2 (901 elementos) de cada dato, es decir que, se tratara cada punto o valor de este arreglo 1D como una característica o predictor del problema, a priori se espera entonces que, este método tenga un buen rendimiento pero, sea costoso computacionalmente.

Se dividió el dataset en train/test(20%) y se entrenaron distintos modelos de machine learning de regresión como lo fueron: LinearRegresor, DecisionTree y RandomForest. Para la búsqueda de mejores hiperparametros se utilizo el método RandomizedGridCV de sklearn utilizando crossvalidation 5-fold.

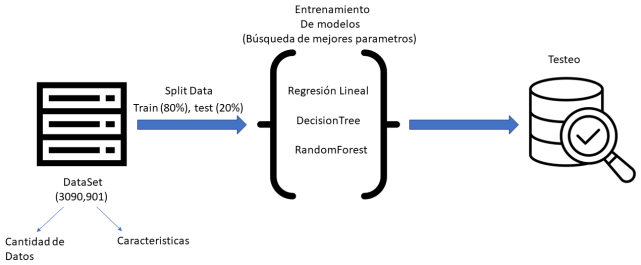


FIGURE 3: Flujo de trabajo Método 1 ML

2) Método 2

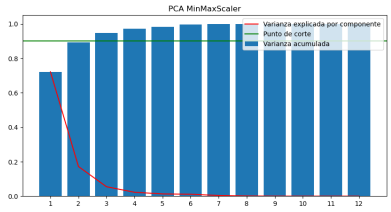
En este método se exploro una posible extracción de características a los arreglos 1D, se decidió extraer características básicas como lo son:

- Valor Máximo
- Valor Medio
- Desviación Estándar
- Asimetría
- Curtosis
- Pendiente Mínima
- Pendiente Máxima
- Pendiente Media
- Área bajo la curva
- Cantidad de puntos hasta alcanzar valor máximo (Subida)
- Cantidad de puntos restantes después de alcanzar valor máximo (Bajada)
- Puntos medios entre Subida y Bajada

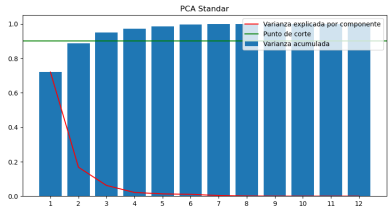
Posteriormente, se escalan los datos usando MinMaxScaler (MMScaler), StandarScaler (StadScaler) y RobustScaler (RobScaler) para luego ser normalizandos por medio de PowerTransform con el método yeo-johnson. En este caso, se tiene todavía una moderada dimensionalidad de los datos, por lo que se realiza PCA para la reducción de la misma.

Se aplico reducción dimensionalidad a los datos utilizando el método de PCA y se realizo el análisis de varianza en las componentes, cabe aclarar que se tiene como objetivo tomar

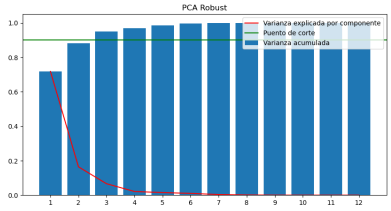
las componente que aporten una representación del 90% de la varianza de los datos.



(a) PCA para MinMaxScaler



(b) PCA para StandarScaler



(a) PCA para Robust Scaler

FIGURE 5: Representación visual de Sesgo y Varianza dependiendo de la complejidad del modelo

Se puede observar de la figura 5 una gran variabilidad de los datos es encontrada en las primeras dos componentes las cuales aportan aproximadamente 90% de la varianza de los datos, por tanto, se decide trabajar en este nuevo espacio de representación (2 dimensiones).

Luego, se divide el conjunto de datos entre train/test(20%) para el posterior entrenamiento de modelos.

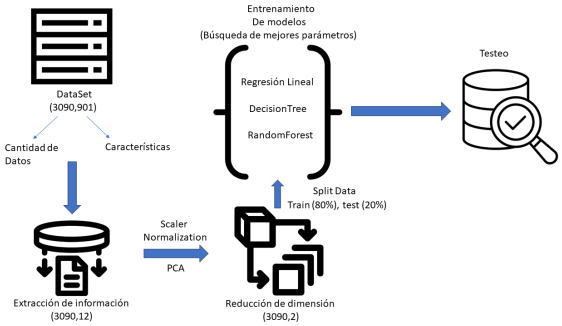


FIGURE 6: Flujo de trabajo Método 2 ML

B. DEEP LEARNING

Por otro lado, se busco predecir el bandgap implementando técnicas de aprendizaje profundo, desde dos enfoques distinto. Primeramente se preprocesaron los dato y luego se aplicaron los siguientes métodos.

1) Preprocesamiento de los datos

Es importante intentar eliminar la mayor cantidad de ruido de los datos para evitar que el modelo relacione el ruido como parte de las muestras, por lo que se aplico el filtro de Savitzky-Golay, el cual permite suavizar los datos y eliminar el ruido de alta frecuencia.

Seguidamente para proporcionarle mas información al modelo sobre los datos se opto por calcular la derivada de x2 para usarla como un segundo canal en el entrenamiento del modelo, en este momento tomo vital importancia suavizar los datos y reducir el ruido:

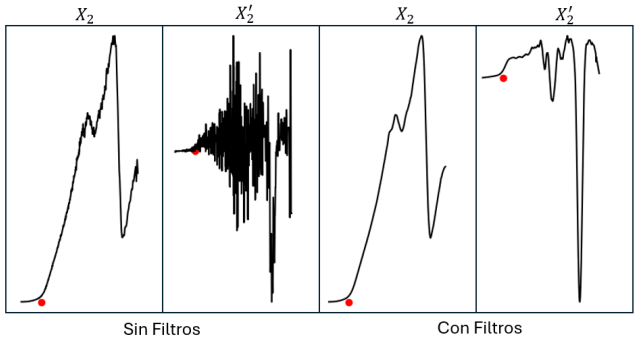


FIGURE 7: Comparación de visualización de los datos sin filtros vs con filtros

2) Método 1

Este método consiste en implementar un red neuronal convolucional de 1 dimensión, con dos canales ($X2$ y $X2'$), la arquitectura usada se baso en la propuesta por Alex Shenfield y Martin Howarth en "A Novel Deep Learning Model for the Detection and Identification of Rolling Element-Bearing Faults" donde usan redes convolucionales 1D y RNN para detectar anomalías en el funcionamiento de motores a partir de las señales generadas por la vibración de los mismos. la arquitectura para este método fue la siguiente: [6]

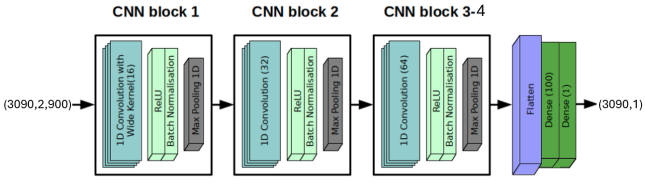


FIGURE 8: Arquitectura Conv 1D

3) Método 2

Este método implica la generación de imágenes a partir de los datos interpolados, filtrados y barajados tanto para imágenes normales como para sus derivadas. Estas imágenes se guardan con un tamaño de 270x270 píxeles.

Se le aplicaron las siguiente transformaciones:

- Redimensionar a 64x64 píxeles
- Normalización de las imágenes
- Se aplican dos técnicas de binarización:
  - Los píxeles de la imagen que superan el umbral de 0.64 se establecen en 1.0 , mientras que el resto se establecen en 0.0.
  - Los píxeles de la imagen que están por debajo del umbral de 0.64 se establecen en 1.0, y el resto se establecen en 0.0.

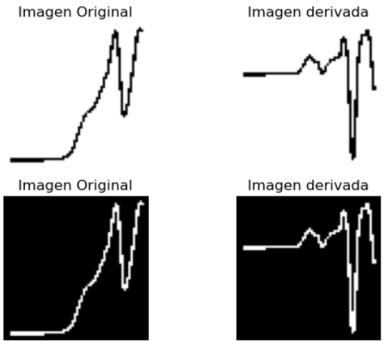


FIGURE 9: Imagenes Binarizadas

La arquitectura del modelo usada en este método fue construida y propuesta por los autores, el modelo resultante contó con 2116513 parámetros y tiene la siguiente estructura:

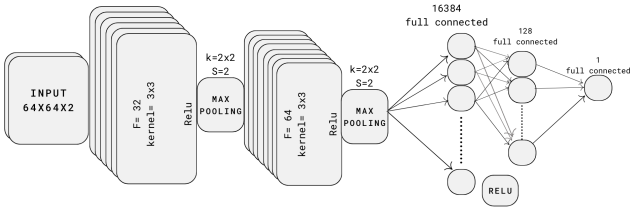


FIGURE 10: Arquitectura de CNN

Durante el entrenamiento, se utiliza la función de pérdida Mean Squared Error (MSE) y Mean Absolute Error, se utiliza el optimizador SGD y ADAM para ajustar los pesos del modelo. También se utilizan métricas como el Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) y el coeficiente de determinación R^2.

III. RESULTADOS

A. MACHINE LEARNING - MÉTODO 1

Los resultado al realizar la búsqueda aleatoria en malla de los mejores hiperparametros para los modelos son:

- **Regresión Lineal (LinearRegressor)**
  - **n_jobs:** -1
- **Arbol de desición (DesicionTreeRegressor)**
  - **Max_depth:** 5
  - **Min_samples_split:** 5
- **Bosques aleatorios (RandomForestRegressor)**
  - **n_estimators:** 800
  - **Max_depth:** 50
  - **min_samples_leaf:** 2

MSE	RMSE	MAE
RandomForest (0.159)	RandomForest (0.399)	RandomForest (0.254)
DecisionTree (0.163)	DecisionTree (0.404)	DecisionTree (0.258)
LinearRegression (23.905)	LinearRegression (4.889)	LinearRegression (0.761)

R^2	Max_Error
RandomForest (0.646)	DecisionTree (2.616)
DecisionTree (0.637)	RandomForest (2.616)
LinearRegression (-52.293)	LinearRegression (114.744)

TABLE 2: Resultados obtenidos utilizando todos los puntos como características del modelo

B. MACHINE LEARNING - MÉTODO 2

Los resultado al realizar la búsqueda aleatoria en malla de los mejores hiperparametros para los modelos son:

- **Regresión lineal (LinearRegressor)**

  Para cada normalización de datos se obtuvieron los resultados:
  - **n_jobs:** -1
- **Arbol de desición (DesicionTreeRegressor)**
  - **Random_State:** 42

Normalización	max_depth	max_features	min_samples_leaf	min_samples_split
MinMaxScaler	10	None	10	2
StandarScaler	10	sqrt	8	2
RobustScaler	None	None	20	2

- **Bosques Aleatorios (RandomForestRegressor)**
  - **Random_state:** 42

Normalización	n_estimator	max_depth	max_features	min_samples_leaf	min_samples_split
MinMaxScaler	300	None	sqrt	1	2
StandarScaler	300	30	sqrt	1	2
RobustScaler	300	30	sqrt	1	2

Obteniendo:

- Regresión Lineal

MSE	RMSE	MAE
RobScaler (0.284)	RobScaler (0.532)	RobScaler (0.413)
StadScaler (0.29)	StadScaler (0.538)	StadScaler (0.415)
MMScaler (0.294)	MMScaler (0.542)	MMScaler (0.415)

R^2	Max_Error
RobScaler (0.368)	RobScaler (2.247)
StadScaler (0.354)	StadScaler (2.366)
MMScaler (0.345)	MMScaler (2.422)

TABLE 3: Resultados obtenidos del mejor scaler para Regresión lineal

- Arboles de desiciones

MSE	RMSE	MAE
StadScaler (0.224)	StadScaler (0.473)	MMScaler (0.315)
MMScaler (0.23)	MMScaler (0.479)	StadScaler (0.334)
RobScaler (0.253)	RobScaler (0.503)	RobScaler (0.348)

R^2	Max_Error
StadScaler (0.501)	StadScaler (1.919)
MMScaler (0.488)	RobScaler (2.225)
RobScaler (0.436)	MMScaler (3.128)

TABLE 4: Resultados obtenidos del mejor scaler para Arboles de desiciones

- RandomForest

MSE	RMSE	MAE
MMScaler (0.229)	MMScaler (0.479)	StadScaler (0.338)
StadScaler (0.23)	StadScaler (0.48)	MMScaler (0.342)
RobScaler (0.253)	RobScaler (0.503)	RobScaler (0.348)

R^2	Max_Error
MMScaler (0.488)	RobScaler (2.225)
StadScaler (0.487)	MMScaler (2.248)
RobScaler (0.436)	StadScaler (2.315)

TABLE 5: Resultados obtenidos del mejor scaler para RandomForest

Entonces, sabiendo que el mejor rendimiento lo tuvo DecisionTree con el StandarScaler, entonces al aplicar este escalado de datos a los modelos con sus mejores hiperparametros se obtuvieron:

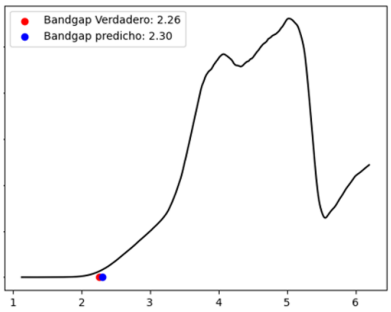
MSE	RMSE	MAE
RandomForest (0.157)	RandomForest (0.396)	RandomForest (0.294)
DecisionTree (0.224)	DecisionTree (0.473)	DecisionTree (0.334)
LinearRegression (0.29)	LinearRegression (0.538)	LinearRegression (0.415)

R^2	Max_Error
RandomForest (0.65)	RandomForest (1.769)
DecisionTree (0.501)	DecisionTree (1.919)
LinearRegression (0.354)	LinearRegression (2.366)

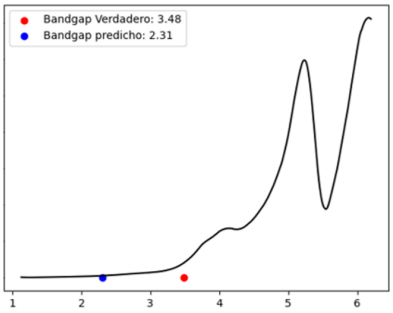
TABLE 6: Resultado de los finales para modelos con StandarScaler(Escalador escogido como mejor)

C. DEEP LEARNING - MÉTODO 1

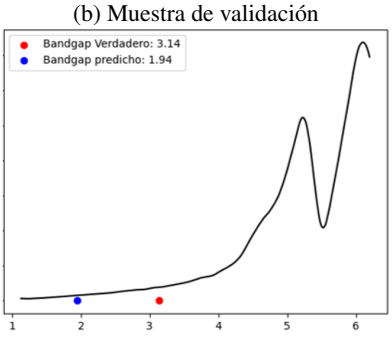
El método arrojo un modelo capaz de predecir el bandgap a partir de los datos de x2 debidamente pre-procesados, una muestra de las predicciones son las siguientes:



(a) Muestra de validación



(b) Muestra de validación



(c) Muestra de validación

De las representaciones anteriores podemos apreciar de que manera el modelo esta prediciendo el bandgap. En la gráfica (a) realizo una buena predicción, sin embargo existieron casos como los de la gráfica (b) y (c) donde la predicción estuvo lejos del objetivo, esto se presento en gráficas con un incremento temprano y una tasa de cambio en principio constante.

El desempeño del modelo esta representado en las siguientes metricas:

MAE	MSE
0.5854554	0.6722501
RMSE	R^2
0.8199086	-0.356794

TABLE 7: Mejores Resultados de Método Conv1D

Las métricas arrojaron que este método no consigue realizar unas predicciones adecuadas, dado que el R^2 no se encuentra cerca al 1, esto podría deberse a lo anteriormente mencionado de lo sensible que es este modelo a pendientes no pronunciadas.

D. DEEP LEARNING - MÉTODO 2

Se probaron diferentes configuraciones cambiando parámetros clave. Con el optimizador SGD, se variaron el learning rate, las épocas y el tamaño del batch, manteniendo las demás instancias fijas en valores específicos: 0.001 para el learning rate, 50 para las épocas, y 500 para el tamaño del batch. También se experimentó con distintas funciones de error, tales como MAE y MSE. Asimismo, se llevó a cabo una evaluación con el optimizador ADAM, con la única exclusión de la variación de las épocas.

Fondo Blanco & MAE & SGD						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	0.61	0.46	0.68	-0.36
50	0.01	500	0.56	0.43	0.68	-0.36
50	0.001	500	0.54	0.57	0.75	-0.3

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.39	0.46	0.68	0.3
50	0.001	128	0.43	0.43	0.68	0.24
50	0.001	600	0.54	0.57	0.75	-0.3

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
60	0.001	500	0.54	0.56	0.68	-0.31
100	0.001	500	0.56	0.6	0.68	-0.39
600	0.001	500	0.45	0.357	0.75	0.16

TABLE 8: Resultados obtenidos con SGD y MAE con Fondo blanco

Fondo Negro & MAE & SGD						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	0.3	0.2	0.45	-0.49
50	0.01	500	0.42	0.33	0.57	-0.22
50	0.001	500	0.5	0.47	0.69	-0.11

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.36	0.28	0.53	0.34
50	0.001	128	0.41	0.32	0.57	0.24
50	0.001	600	0.51	0.49	0.7	-0.14

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
60	0.001	500	0.49	0.44	0.66	-0.02
100	0.001	500	0.47	0.39	0.62	0.08
600	0.001	500	0.4	0.32	0.56	0.24

TABLE 9: Resultados obtenidos con SGD y MAE con Fondo Negro

Fondo Blanco & MSE & SGD						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	0.39	0.25	0.5	0.4
50	0.01	500	0.38	0.24	0.498	0.42
50	0.001	500	0.56	0.38	0.62	0.1

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.4	0.26	0.5	0.38
50	0.001	128	0.44	0.3	0.55	0.29
50	0.001	600	0.55	0.39	0.62	0.08

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
60	0.001	500	0.56	0.4	0.63	0.05
100	0.001	500	0.53	0.37	0.61	0.12
600	0.001	500	0.45	0.31	0.55	0.26

TABLE 10: Resultados obtenidos con SGD y MSE con Fondo blanco

Fondo Negro & MSE & SGD						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	Nan	Nan	Nan	Nan
50	0.01	500	0.44	0.3	0.54	0.29
50	0.001	500	0.47	0.35	0.57	0.21

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.35	0.21	0.46	0.502
50	0.001	128	0.39	0.25	0.5	0.4
50	0.001	600	0.47	0.33	0.57	0.21

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
60	0.001	500	0.49	0.34	0.588	0.19
100	0.001	500	0.41	0.28	0.532	0.33
600	0.001	500	0.47	0.335	0.57	0.37

TABLE 11: Resultados Obtenidos con SGD y MSE con Fondo Negro

Fondo Blanco & MSE & ADAM						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	0.58	0.42	0.65	-0.0006
50	0.01	500	0.455	0.32	0.57	0.23
50	0.001	500	0.27	0.13	0.37	0.67

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.27	0.14	0.38	0.65
50	0.001	128	0.29	0.157	0.39	0.633
50	0.001	600	0.47	0.148	0.38	0.65

TABLE 12: Resultados obtenidos con ADAM y MSE con Fondo Blanco

Fondo Negro & MSE & ADAM						
Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.1	500	0.58	0.42	0.65	-0.005
50	0.01	500	0.55	0.4	0.6	-0.002
50	0.001	500	0.26	0.13	0.36	0.685

Epocas	Lr	Bacth	MAE	MSE	RMSE	R^2
50	0.001	64	0.28	0.16	0.41	0.6
50	0.001	128	0.37	0.23	0.48	0.59
50	0.001	600	0.3	0.16	0.4	0.61

TABLE 13: Resultados obtenidos con ADAM y MSE con Fondo Negro

IV. CONCLUSIONES

En conclusión, el Random Forest se destacó como el mejor modelo en ambos experimentos, lo que está respaldado por la literatura existente. Aunque el método 2, que implicaba la extracción de características básicas de los datos, mostró una mejora la cual si bien no fue significativa en términos de métricas de rendimiento, sí resultó en tiempos de computación más bajos.

Esto se debe a que el método 2 presentaba una dimensionalidad de datos significativamente más baja en comparación con el método 1. Lo que significa que este acercamiento puede llegar a ser acertado para la predicción de band gap a través de lista 1D de puntos de diagrama Tauc's.

En resumen, se elige como mejor modelo el RandomForest del método 2 con un R^2 de 65%, con un error aproximado de 0.39 eV y presentando un error máximo de 1.77 eV.

Por otro lado, en la predicción con deep learning la investigación encontró que al tratar x2 como una señal procesada con convoluciones 1d, el modelo resultante es muy sensible a las pequeñas variaciones de la pendiente por lo que este realiza una predicción prematura en comparación a la etiqueta real.

Ahora bien, se obtuvieron mejores resultados al entrenar al modelo con las imágenes de las gráficas de discretización de fondo negro alcanzando un R^2 de 0.68 utilizando Adam como optimizador y MSE como función de perdida. Sin embargo, con este método aparentemente se realizaba la predicción a partir de la pendiente mas pronunciada, por lo que en algunos casos difería en gran medida de la etiqueta real.

REFERENCIAS

[1] R. Nandee, M. Asaduzzaman Chowdhury, M. D. Arefin Kowser, S. Kumer Nondy, N. Hossain, M. Rasadujjaman, A. Al Mostazi, M. Baizid Molla, S. Barua, M. Masud Rana, and M. Sherajul Islam, "Bandgap formation in graphene doped with bn, tio2, al2o3 and zno by sintering process," Results in Chemistry, vol. 6, DOI 10.1016/j.rechem.2023.101229, Dec. 2023.

[2] H. He, Y. Wang, Y. Qi, Z. Xu, Y. Li, and Y. Wang, "From prediction to design: Recent advances in machine learning for the study of 2d materials," Nano Energy, vol. 118, DOI 10.1016/j.nanoen.2023.108965, Dec. 2023.

[3] M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "Recent progress of the computational 2d materials database (c2db)," 2D Materials, vol. 8, DOI 10.1088/2053-1583/ac1059, no. 4, Oct. 2021.

[4] J. Zhang, Y. Li, and X. Zhou, "Machine-learning prediction of the computed band gaps of double perovskite materials," DOI 10.48550/arXiv.2301.03372, p. 15–27. Academy and Industry Research Collaboration Center (AIRCC), Jan. 2023.

[5] Y. Zhang, W. Xu, G. Liu, Z. Zhang, J. Zhu, and M. Li, "Bandgap prediction of two-dimensional materials using machine learning," PLoS ONE, vol. 16, DOI 10.1371/journal.pone.0255637, no. 8 August, Aug. 2021.

[6] A. Shenfield and M. Howarth, "A novel deep learning model for the detection and identification of rolling element-bearing faults," Sensors, vol. 20, DOI 10.3390/s20185112, no. 18, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5112