Vol. 08, No. 01 (2024) 66 – 76
JIIF (Jurnal Ilmu dan Inovasi Fisika)
https://jurnal.unpad.ac.id/jiif/

CBDR-CNN APPROACH FOR RAPID IDENTIFICATION OF XRD DATA: A PRELIMINARY STUDY

JULIAN EVAN CHRISNANTO[1,*], NURFAUZI FADILLAH[2], FERRY FAIZAL[1]

[1]*Departemen Fisika, FMIPA, Universitas Padjadjaran*
Jl. Raya Bandung-Sumedang KM 21, Jatinangor 4563, Sumedang, Jawa Barat, Telp. 022-7796014
[2]*PT. Plabs. ID*
Jl. Setrasari Tengah No. 10, Sukarasa, Kec. Sukasari, Kota Bandung, Jawa Barat 40512

**Corresponding author*
Email: julian20001@mail.unpad.ac.id

Diserahkan : 24/10/2023
Diterima : 29/10/2023
Dipublikasikan : 02/02/2024

Abstract. In this study, we present a novel approach combining Content-Based Data Retrieval (CBDR) and 1-dimensional Convolutional Neural Networks (1D-CNN) for crystal structure analysis of powder materials by using X-Ray Diffraction (XRD) data. The introduction sets the background by highlighting the importance of X-ray diffraction analysis and the limitations of conventional approaches in dealing with complex crystal structures. To overcome this challenge, researchers have explored artificial intelligence techniques, specifically CNN for crystal structure classification based on XRD image graph represented as intensity values versus 2-theta (XRD pattern). The aims of this study are: implementing CBDR method on CNN model for crystal structure classification; simulating CBDR-CNN model for crystal structure classification; verifying CBDR-CNN model in crystal structure classification. Each class for CNN model such as crystal system, class material, sub-class material, and space-group achieved accuracies 99.86%, 99.99%, 99.95%, and 99.82% respectively. The results and discussion section presents the results of the CBDR-CNN model. The CBDR model effectively retrieved the most similar XRD spectrum data from the dataset based on the query properties, including Miller indices and peak position. The model effectively reduced the scope potential candidate materials, sub-materials, and space-groups. The 1D-CNN model showed high accuracy in predicting crystal properties such as material, sub-material, space-group, and crystal system. In conclusion, the CBDR-CNN approach potential revolutionizes XRD data analysis and crystal system prediction, which promotes progress in computer-aided materials study.

Keywords: convolutional neural networks (CNN), content-based data retrieval (CBDR), x-ray diffraction, crystal structure prediction, and crystal system

Abstrak. *M Dalam studi ini, kami menyajikan pendekatan baru yang menggabungkan Content-Based Data Retrieval (CBDR) dan 1-dimensional Convolutional Neural Networks (1D-CNN) untuk analisis struktur kristal bahan serbuk dengan menggunakan data X-Ray Diffraction (XRD). Pendahuluan menetapkan latar belakang dengan menyoroti pentingnya analisis difraksi sinar-X dan keterbatasan pendekatan konvensional dalam menangani struktur kristal yang kompleks. Untuk mengatasi tantangan ini, para peneliti telah mengeksplorasi teknik kecerdasan buatan, khususnya CNN untuk klasifikasi struktur kristal berdasa rkan grafik gambar XRD yang direpresentasikan sebagai nilai intensitas versus 2-theta (pola XRD). Tujuan dari penelitian ini adalah: mengimplementasikan metode CBDR pada model CNN untuk klasifikasi struktur kristal; mensimulasikan model CBDR-CNN untuk klasifikasi struktur kristal; melakukan verifikasi model CBDR-CNN dalam klasifikasi struktur kristal. Setiap kelas untuk model CNN seperti sistem kristal, kelas material, sub-kelas material, dan space-group mencapai akurasi 99,86%, 99,99%, 99,95%, dan 99,82% secara berurutan. Bagian hasil dan diskusi menyajikan hasil dari*

DOI: https://doi.org/10.24198/jiif.v8i1 *e-ISSN: 2549-7014*

model CBDR-CNN. Model CBDR secara efektif mengambil data spektrum XRD yang paling mirip dari kumpulan data berdasarkan properti kueri, termasuk indeks Miller dan posisi puncak. Model ini secara efektif mengurangi cakupan kandidat material, sub-material, dan space-group yang potensial. Model 1D-CNN menunjukkan akurasi yang tinggi dalam memprediksi sifat-sifat kristal seperti material, sub-material, space-group, dan sistem kristal. Kesimpulannya, pendekatan CBDR-CNN berpotensi merevolusi analisis data XRD dan prediksi sistem kristal, yang mendorong kemajuan dalam studi material berbantuan komputer.

***Kata kunci**: convolutional neural networks (CNN), content-based data retrieval (CBDR), difraksi sinar-x, prediksi struktur kristal, dan sistem kristal*

1. Introduction

This research is motivated by the author's curiosity about the phenomena involved in X-ray diffraction analysis. X-ray diffraction is a phenomenon in which X-rays is exposed/irradiated to a crystalline sample, and the rays are diffracted and refracted by the atoms in the crystal, creating an interference pattern on the screen [1]. This technique is essential for understanding and analyzing the arrangement of atoms and properties of various materials or crystals.

However, conventional analysis methods of X-ray diffraction are time-consuming and difficult when analyzing complex crystal structures [2]. The X-Ray Powder Diffraction (XRPD) method, a specialized application of XRD, has certain limitations in crystal structure analysis. These limitations include overlapping peaks of the interference pattern on the screen and noise from the background on the interference result screen, which can hinder crystal structure analysis [3].

To overcome the challenges of improving crystal structure analysis quality, researchers have explored the use of artificial intelligence. In particular, they have looked at the Convolutional Neural Networks (CNN) method, which is well-known for its success in diverse applications. In [2], CNN was used to classify crystal structures using a set of spectra from XRPD data in the form of images. In that study, three classifications were made: space-group, extinction group, and crystal system. The achieved accuracy for these classifications were 81.14%, 83.83%, and 94.99%, respectively.

The author proposes this study to improve the accuracy and performance of crystal structure analysis by including one of the methods of Content-Based Data Retrieval (CBDR) into the CNN model itself. This study differs from previous research in that it uses raw data from XRD spectra rather than images or XRD graphs themselves. Using the CBDR method provides advantages since it utilizes the features of the XRD data, such as the Miller indices, peak information, and other crystal parameters. The study introduces a new approach to crystal structure analysis by combining the CNN method and CBDR technique. The combination of XRD spectra instead of direct XRD graphs or images is anticipated to enhance the accuracy, efficiency, and performance in crystal structure analysis significantly. Moreover, by utilizing XRD spectra instead of direct XRD graphs or images, there is a change in data representation, reflecting both an increase in accuracy and more informative.

1.1 Related Studies

In study [2], discusses the use of convolutional neural networks (CNNs) for classifying crystal structures. In that study, researchers conducted a literature review on relevant research topics. The literature review provides supporting evidence that crystal structure classification is an important topic to be researched, particularly in the field of materials science. Previous relevant research on this topic has used machine learning techniques,

including conventional artificial neural networks [4], principal component analysis (PCA) [5], partial least-squares regression (PLSR) [6], and various specialized statistical approaches [7]. Certain approaches are constrained by the requirement of manual or traditional feature analysis techniques like the conventional artificial neural network techniques.

In [8], a new approach for crystal structure classification using artificial neural networks with machine learning methods was proposed. The approach proposed in the study demonstrates the capability to classify crystal structures based on crystal symmetry, even though the crystal symmetry exists in the form of defects or geometric deformations. This machine learning-based approach represents each crystal structure as a two-dimensional diffraction fingerprint and uses a subset of those structures to generate a classification model through a convolutional neural network (CNN) with convectional training data.

In reference to [9], the study presents an approach that utilizes machine learning to predict crystal dimensionality and space group from a variety of thin film XRD patterns. The proposed approach modifies XRD data using data augmentation techniques by employing simulated data acquired from the Inorganic Crystal Structure Database (ICSD) and experimental data. The proposed method for convolutional neural networks (CNN) achieved high accuracy for dimensionality classification and space group, achieving 93% and 89%, respectively. The study also revealed several causes of invalid classification, including phase mixtures in the sample, noise in the experimental data, and systematic errors in the experimental setup. The study discussed limitations of the proposed approach, including the need for a large and varied ICSD dataset and proper selection of experimental data.

2. Research Methods

This research was conducted in three general stages: pre-processing, processing, and post-processing. During the pre-processing stage, the dataset was prepared from the American Mineralogist Crystal Structure Database (AMCSD). This dataset will be used in the CBDR and CNN models. Once the data is processed into a dataset, a training model is prepared for each model. Next, in the processing stage, the given query XRD data is tested in the CBDR-CNN model using a csv data format. The data return stage is carried out based on similar values in the dataset, followed by a prediction of the query data in the context of crystal properties. Finally, in the post-processing stage, the prediction results of the CBDR-CNN model are analyzed. The simulation utilized Python Programming Language with version 3.10.7, Sklearn with version 1.0.0, and Tensorflow with version 2.11.0.

2.1 Content-Based Data Retrieval (CBDR)

Content-Based Data Retrieval (CBDR) is a recently developed technique that shares similarities with Content-Based Image Retrieval (CBIR), also known as Query By Image Content (QBIC). CBIR is an automated technique that retrieves similar images based on their visual content, utilizing low-level features such as color, texture, shape, and spatial location. Similarly, CBDR utilizes content-based features to retrieve similar data based on their property content from a database. When an image is used as a query in CBIR, it is converted into a feature vector through a feature extraction process. The similarity measure is applied to calculate the distance between feature vectors of a query image and a target image in a database. This enables retrieving similar images efficiently [10]. CBDR is inspired by CBIR and implements the same principle, focusing on searching data based on its properties. CBDR considers content-based characteristics, like attribute values, data patterns, and structural information, to represent and search for similar data

within a database. By utilizing these features, CBDR enables efficient retrieval of data that shares similar characteristics with the query data from the database.

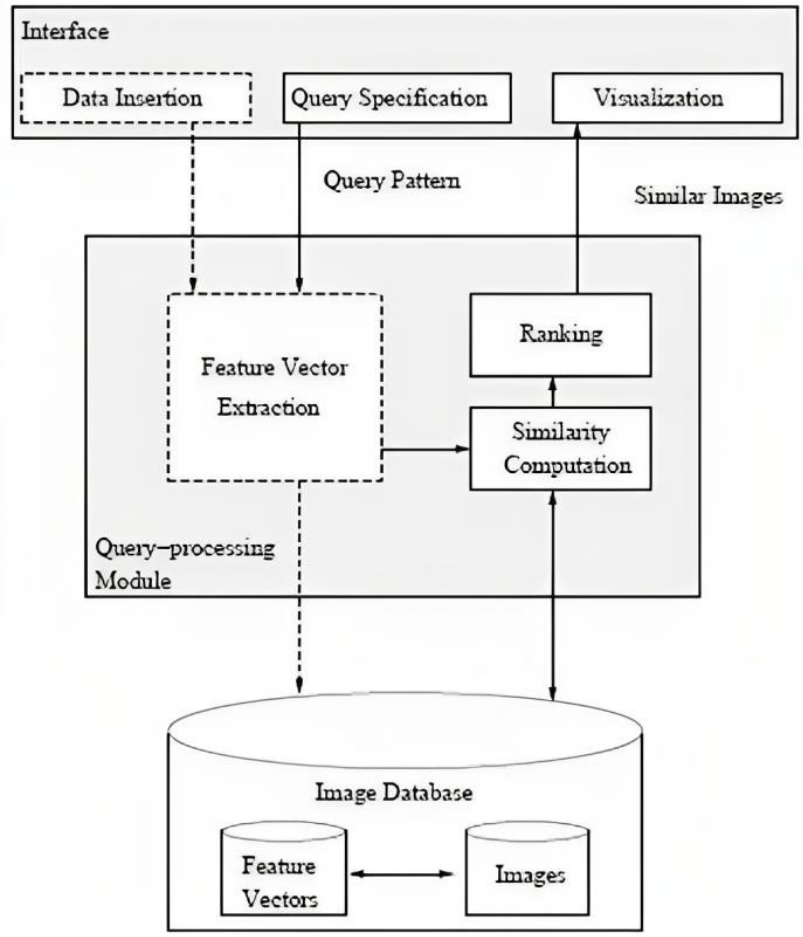

Figure 1. CBIR general architecture

In the context of CBDR, data is first transformed into a feature vector through the process of feature extraction, and subsequently, a similarity measure is computed. This similarity measure facilitates the calculation of the distance between the feature vectors of the query data and the data contained in the database. The current study employs Euclidean Distance as the similarity measure. In particular, Euclidean Distance determines the distance between points in a straight line. The distance calculation methodology relies on the Pythagorean theorem. The equation for Euclidean Distance can be described as follows:

$$D(q,\, d) = \left|\left|q - d\right|\right|^2 = \sqrt{\sum_{i=1}^{n}(q_i - d_i)^2} \tag{1}$$

The L-2 Norm is another term for Euclidean Distance. The L-2 Norm is a measure of distance that calculates the straight-line distance between two points. In the context of Euclidean Distance, the L-2 Norm is employed to determine the distance between two vectors by taking the square root of the sum of the squared differences between the components of the corresponding vectors [11].

2.2 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a type of artificial neural network that learn internal feature representations in a hierarchical structure, enabling them to generalize features in image-related tasks such as object recognition and other computer vision problems. CNNs leverage a convolutional layer to extract features from input images while preserving each pixel's spatial relationship [12].

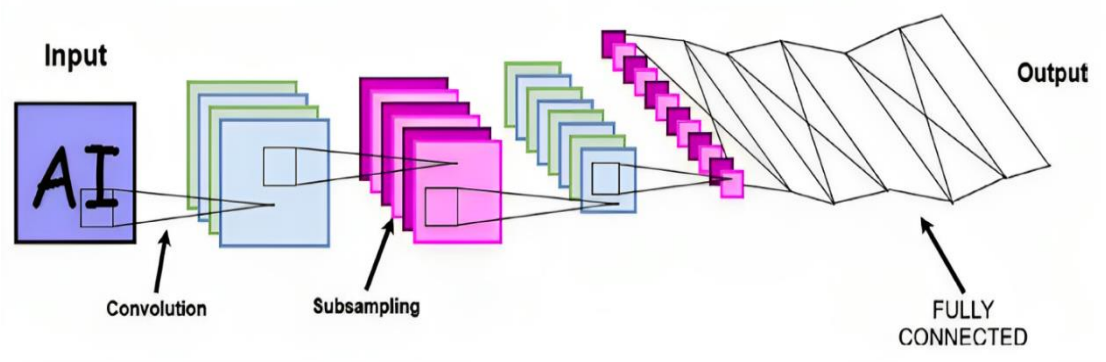

Figure 2. CNN layers

In mathematics, the convolution operation is a process that combines two functions f and g to produce a third function h. This process consists of two types of convolutions, continuous and discrete convolution. In numerical processing, the second type of convolution used is discrete convolution. It is commonly defined as follows:

$$(f * g)(x) = \sum_{i=1}^{m} f(i) \cdot g(x - i) \tag{2}$$

The function $f(x)$ is a function that represents the numerical input in the form of a vector, and the function $g(x)$ is the convolution kernel or filter. The kernel $g(x)$ is a window that operates on the input signal $f(x)$ shifted by a certain shift step, where the sum of the multiplication of the two functions at each discrete point is the result of the convolution, which is expressed as the output function $h(x)$.

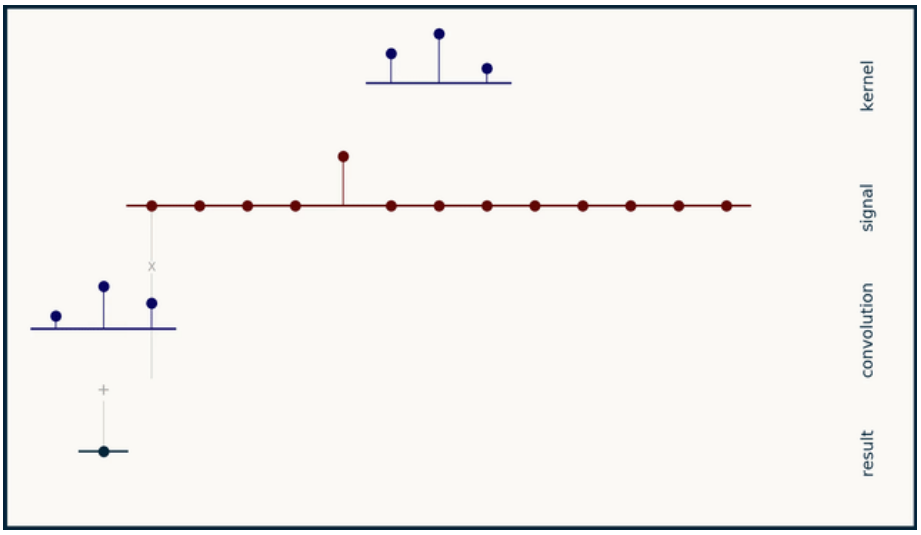

Figure 3. Example of signal convolution operation in one dimension

2.3 Prerequisites

These are tools were used in this research as follows:

1. American Mineralogist Crystal Structure Database (AMCSD).
2. Python v. 3.10.7.
3. Sklearn v.1.0.0.
4. Tensorflow v.2.11.0.

The American Mineralogist Crystal Structure Database (AMCSD) is a user interface for a crystal structure database. It comprises all the structures published in American Mineralogist, The Canadian Mineralogist, European Journal of Mineralogy, Physics and Chemistry of Minerals, and selected data sets from other journals. The database is funded by the National Science Foundation and is overseen by the Mineralogical Society of America and the Mineralogical Association of Canada. Python is a widely used high-level programming language. Python is capable in data analysis, machine learning, and scientific computations due to its extensive library and framework. Scikit-learn or sklearn is one of the prerequisite libraries. Scikit-learn is a machine learning library that offers multiple algorithms for classification, regression, clustering, and other purposes. Scikit-learn is incorporated into the CBDR model due to its Euclidean Distance calculation algorithm. Next, the library utilized in this simulation is Tensorflow, which is a Python-based library. Tensorflow is a highly robust open-source library for machine learning and deep learning applications. Tensorflow offers a flexible architecture that is particularly suitable for the CNN architecture utilized in this simulation.

The simulation scheme is as follows:

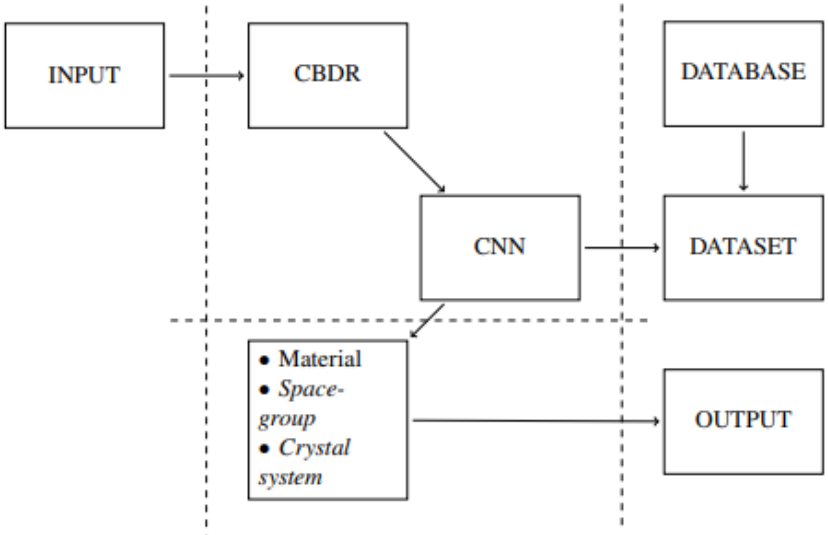

Figure 4. Simulation scheme

Figure 4 shows the construction of the simulation scheme. The scheme starts with retrieval of raw data from the AMCSD database, which undergoes data preparation based on the dataset diagram. Next, the input diagram shows how the query data is fed into the CBDR-CNN model, depicted in both the CBDR and CNN diagrams. The results of the CBDR and CNN models are depicted in the result diagram, showing properties such as material, space-group, and crystal system. The output is generated as an XRD graph visualization based on the query data entered.

Here it is the CNN architecture we used in this research, represented in Figure 5 below.

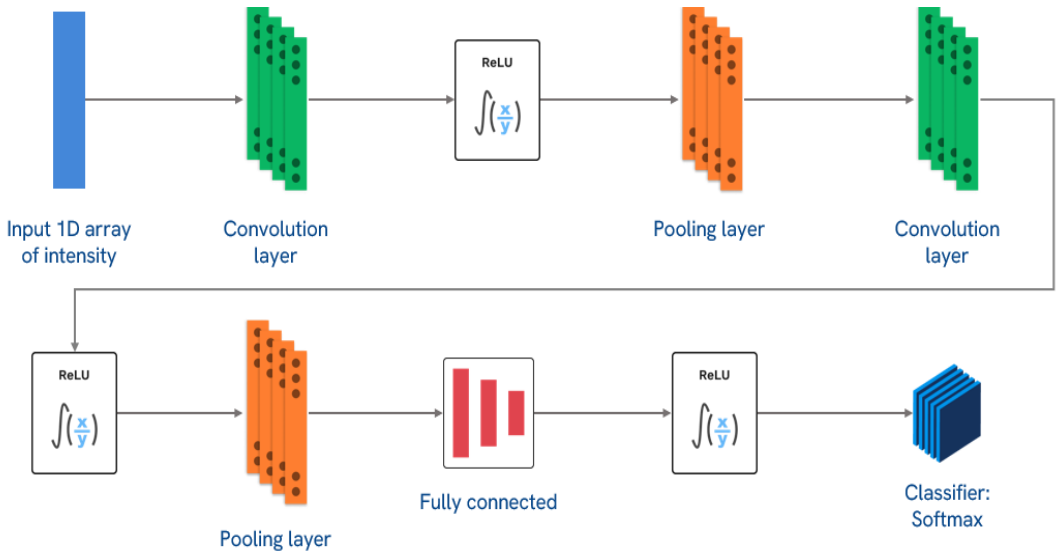

Figure 5. CNN architecture in this research

The CNN architecture utilized in this study plays a crucial role in classifying crystal structures based on XRD data. The architecture follows a sequential pattern of layers designed to extract and process relevant features from the input data. The initial layer of the model is a 1-dimensional convolutional layer. This layer is responsible for performing feature extraction on the input XRD data. It is composed of 32 filters, each with a size of 3, that slide across the input data to capture patterns and relationships within the intensity values. The Rectified Linear Unit (ReLU) is used as the activation function in this model, introducing non-linearity [13]. A max pooling layer is added after the convolutional layer. The max pooling layer reduces the dimensions of the feature maps obtained from the previous layer, retaining the most relevant information. Each feature map is subject to a pooling window of size 2 to capture the maximum value within that window. Another convolutional layer with 64 filters, each of size 3, is then added to the model. This layer extracts higher-level features from the down-sampled feature maps that were obtained from the previous pooling layer. The ReLU activation function is once again used to introduce non-linearity. Following the second convolutional layer, a second max pooling layer is introduced, similar to the first one. This further reduces the dimensions of the feature maps while preserving critical information. After the second pooling layer, the feature maps undergo a flattening operation. This transformation prepares the data for input into the fully connected layers. Two fully connected layers are added following the flattening layer. The initial fully connected layer consists of 128 neurons and employs the ReLU activation function. This layer functions as a feature extractor, introducing non-linearity into the model. The last fully-connected layer contains a size of neurons that is identical to the number of classes of crystal properties. The utilized activation function is the softmax function that creates probability distributions for each class [14].

3. Results and Discussion

We evaluated the performance of the developed 1D-CNN model for classifying crystal structures across multiple categories. As far as the crystal system is concerned, the model demonstrated exceptional accuracy, achieving a test accuracy of 99.86% with a minimal test loss of 0.0026. The material classification was equally impressive, with a test

accuracy of 98.99% and a test loss of 0.0193. The sub-material classification exhibited exceptional accuracy, with a test accuracy of 99.95% and a corresponding test loss of 0.0016.

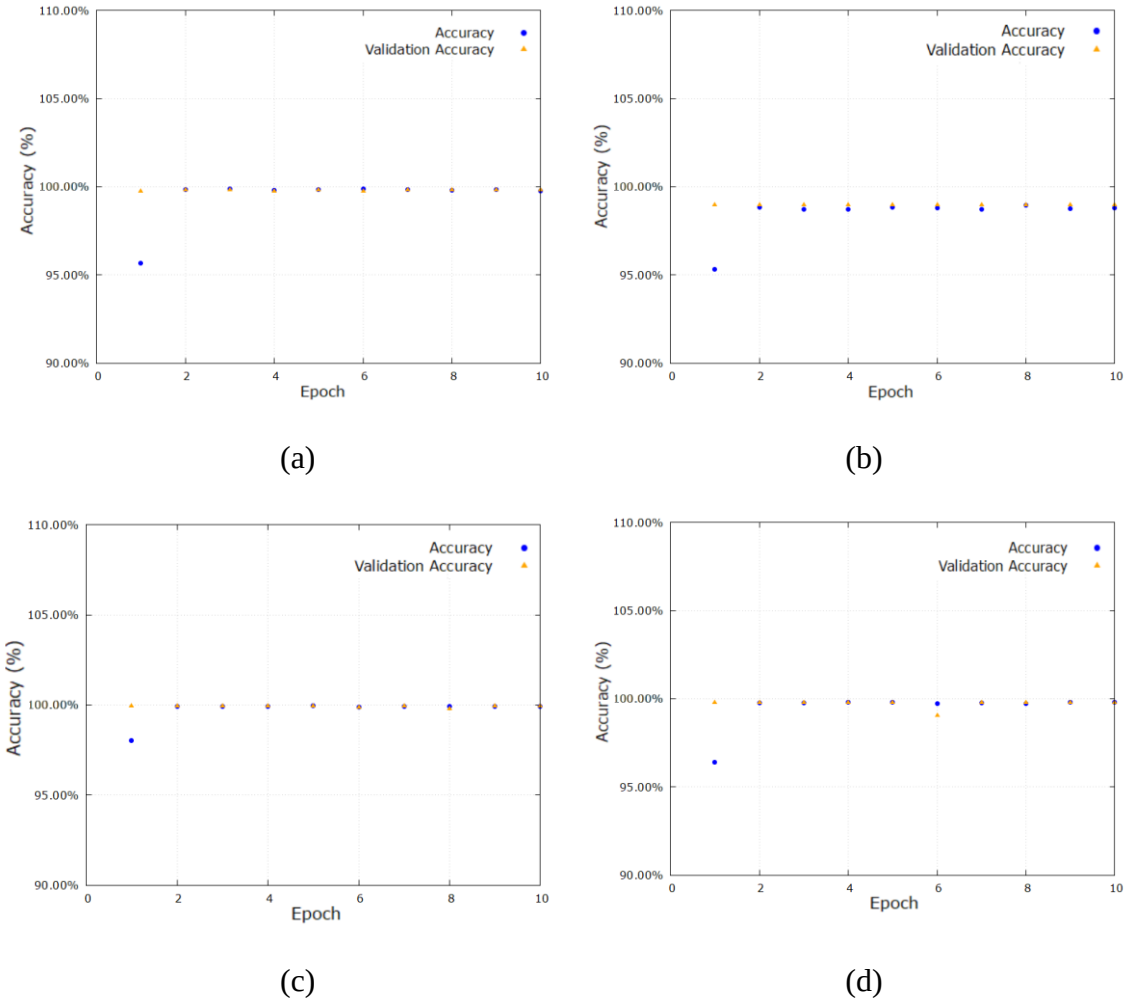

Figure 6. Accuracy graph of CNN training on: (a) crystal system, (b) material, (c) sub-material, (d) space-group

Furthermore, the space group classification exhibited a high accuracy of 99.82% with a test loss of 0.0072. All the training and validation results were obtained after completing 10 epochs, underscoring the robustness and effectiveness of the CNN architecture in processing XRD data to classify crystal structures. The ability of the CBDR-CNN model to retrieve relevant data based on query properties and predict crystal properties was thoroughly evaluated. The CBDR model effectively retrieved the most similar spectra from the dataset by inputting XRD data with specific Miller indices and peak positions. This retrieval narrowed down potential candidate materials, sub-materials, and space groups, and offered insightful predictions. For instance, the CBDR-CNN model successfully predicted the crystal properties to be Beryl with hexagonal symmetry (space group P6) by using the input data for a specific query, demonstrating the model's capability of rapid material identification. In addition to textual outputs, the results were also supported by visual representations. The XRD data graph corresponding to the query input is provided, representing peak intensities at Miller indices. This visual representation improves the interpretability of the results and assists researchers in

evaluating the precision and accuracy of predictions. The CNN model's exceptional accuracy in classifying different crystal properties highlights its potential as a useful tool in crystallography. The model's robustness across different classes and its capability to learn complex patterns from XRD data offer potential for the improvement in speed and accuracy of material characterization. The incorporation of CBDR into the CNN architecture introduces a novel approach to retrieve and predict data, particularly in the field of crystallography. The capability to narrow down probable candidates for material properties, based on query data can significantly expedite the process of material discovery and characterization.

```
------
CBDR Model
------
Query Properties:
Filename: Beryl_0001040.csv
Miller Indices: 1 0 0, 2 1 1, 1 1 2, 0 0 2, 1
Peak Position: 11.10, 31.23, 27.43, 19.32, 22
Peak Intensities: 100.00, 57.16, 55.45, 24.23,

Most similar files:
1.
File: Beryl_0001604.csv
Material: Beryl
Sub Material: Beryl
Space Group: P6
Score: 1.0
Path: CBDR Datasets\Beryl\Beryl_0001604.csv

2.
File: Beryl_0001040.csv
Material: Beryl
Sub Material: Beryl
Space Group: P6
Score: 1.0
Path: CBDR Datasets\Beryl\Beryl_0001040.csv
```

```
Narrowed Material:
['Beryl', 'Biotite', 'Adamite']
Narrowed Sub Material:
['Beryl', 'Biotite', 'Adamite']
Narrowed Space-Group:
['Pnnm', 'P6', 'C2']
```

```
-----
CNN Model
-----
1/1 [==============================] - 0s 206ms/step
1/1 [==============================] - 0s 246ms/step
1/1 [==============================] - 0s 213ms/step
1/1 [==============================] - 0s 228ms/step
Predicted Material: Beryl
Predicted Sub Material: Beryl
Predicted Space Group: P6
Predicted Crystal System: Hexagonal
```

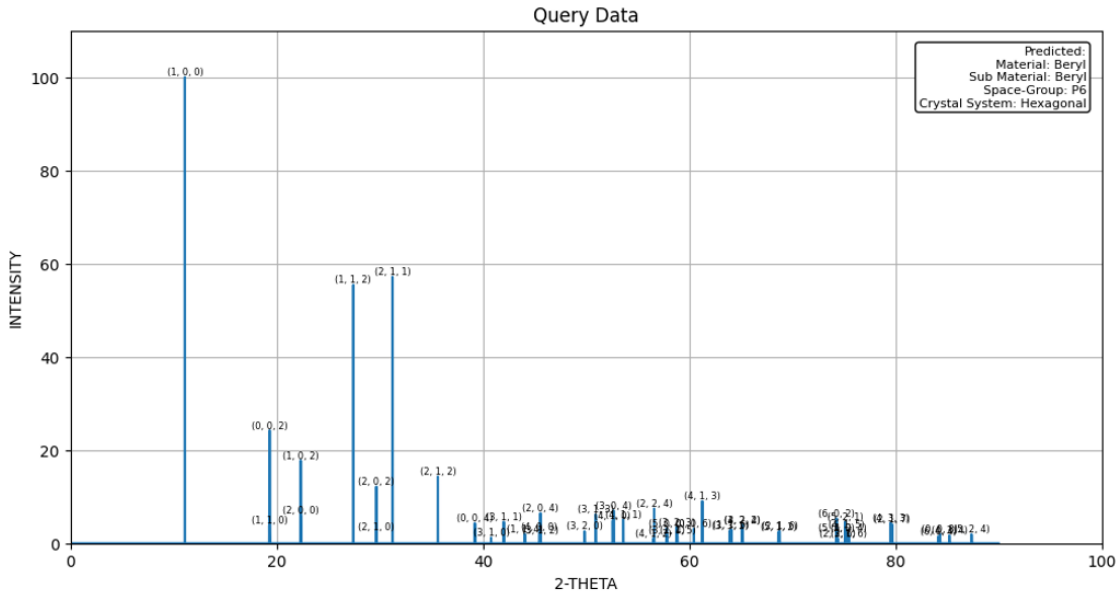

Figure 6. CBDR-CNN results for Beryl raw material; Simulation result

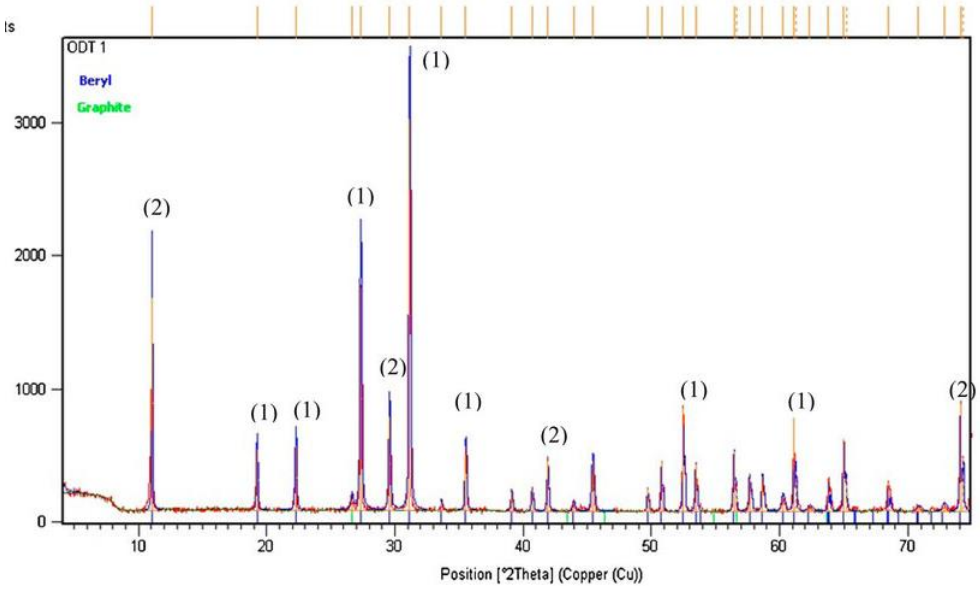

Figure 7. CBDR-CNN results for Beryl raw material: Experimental result

4. Conclusions

The CBDR-CNN method was used to analyze XRD data for crystal structure analysis, resulting in accurate prediction of essential crystal properties. The CBDR-CNN method represents notable progress in powder material research and provides potent research for precise material characterization. Integrating CBDR with deep learning techniques presents new possibilities for discovering materials and making scientific advances in various domains, including material science. The primary data source for this study was the American Mineralogist Crystal Structure Database. In the future, researchers could broaden their investigations by incorporating data from other databases. This approach could provide a wider range of materials and crystal structures for analysis, thereby enabling the creation of more robust and versatile models. Although this research primarily focuses on computational data, integrating experimental XRD data could be an opportunity to verify and enhance the model's accuracy.

Bibliography

1. Y. Waseda, E. Matsubara, and K. Shinoda, *X-{Ray}{Diffraction}{Crystallography}: {Introduction}, {Examples} and {Solved} {Problems}*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
2. W. B. Park *et al.*, “Classification of crystal structure using a convolutional neural network,” *IUCrJ*, vol. 4, no. 4, pp. 486–494, Jul. 2017, doi: 10.1107/S205225251700714X.
3. A. Altomare, C. Cuocci, G. D. Gatta, A. Moliterni, and R. Rizzi, “Methods of crystallography: powder {X}-ray diffraction,” in *Mineralogical {Crystallography}*, 1st ed., J. Plášil, J. Majzlan, and S. Krivovichev, Eds. Mineralogical Society of Great Britain \& Ireland, 2017, pp. 79–138.

4. M. Tatlier, “Artificial neural network methods for the prediction of framework crystal structures of zeolites from {XRD} data,” *Neural Comput. Appl.*, vol. 20, no. 3, pp. 365–371, Apr. 2011, doi: 10.1007/s00521-010-0386-4.
5. S. M. Obeidat, I. Al-Momani, A. Haddad, and M. Bani Yasein, “Combination of {ICP}-{OES}, {XRF} and {XRD} techniques for analysis of several dental ceramics and their identification using chemometrics,” *Spectroscopy*, vol. 26, no. 2, pp. 141–149, 2011, doi: 10.1155/2011/894143.
6. D. Lee, H. Lee, C.-H. Jun, and C. H. Chang, “A {Variable} {Selection} {Procedure} for {X}-ray {Diffraction} {Phase} {Analysis},” *Appl. Spectrosc.*, vol. 61, no. 12, pp. 1398–1403, Dec. 2007, doi: 10.1366/000370207783292127.
7. C. J. Gilmore, G. Barr, and J. Paisley, “High-throughput powder diffraction. {I}. {A} new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles,” *J. Appl. Crystallogr.*, vol. 37, no. 2, pp. 231–242, Apr. 2004, doi: 10.1107/S002188980400038X.
8. A. Ziletti, D. Kumar, M. Scheffler, and L. M. Ghiringhelli, “Insightful classification of crystal structures using deep learning,” *Nat. Commun.*, vol. 9, no. 1, p. 2775, Jul. 2018, doi: 10.1038/s41467-018-05169-6.
9. F. Oviedo, “Fast and interpretable classification of small {X}-ray diffraction datasets using data augmentation and deep neural networks,” *npj Comput. Mater.*, 2019.
10. V. Tyagi, *Content-{Based} {Image} {Retrieval}*. Singapore: Springer Singapore, 2017.
11. R. Suwanda, Z. Syahputra, and E. M. Zamzami, “Analysis of {Euclidean} {Distance} and {Manhattan} {Distance} in the {K}-{Means} {Algorithm} for {Variations} {Number} of {Centroid} {K},” *J. Phys. Conf. Ser.*, vol. 1566, no. 1, p. 12058, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012058.
12. N. K. Manaswi, *Deep {Learning} with {Applications} {Using} {Python}*. Berkeley, CA: Apress, 2018.
13. M. Zufar and B. Setiyono, “Convolutional Neural Networks Untuk Pengenalan Wajah Secara Real-Time,” *J. Sains dan Seni ITS*, vol. 5, no. 2, p. 128862, 2016.
14. C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation {Functions}: {Comparison} of trends in {Practice} and {Research} for {Deep} {Learning}.” arXiv, Nov. 2018, Accessed: Jul. 12, 2023. [Online]. Available: http://arxiv.org/abs/1811.03378