Integrated Journal of Pharmacy Innovations
Vol. 1, No. 1, Juli 2025, pp. 10-15
DOI: -
E-ISSN 3109-2012

Klasifikasi Spektrum NMR pada Senyawa Farmasi Menggunakan Convolutional Neural Network (CNN)

NMR Spectrum Classification of Pharmaceutical Compounds Using Convolutional Neural Network (CNN)

Firman Aziz[1],*; Irmawati[2]
[1] *Universitas Pancasakti, Makassar 90121, Indonesia*
[2] *Irmex Digital Akademika, Makassar 90551, Indonesia*
[1]*firman.aziz@unpacti.ac.id*; [2]*irmawati@irmexdigika.com*
** Corresponding author*

Abstrak

Spektroskopi Resonansi Magnetik Nuklir (NMR) merupakan teknik penting dalam identifikasi struktur senyawa farmasi. Namun, interpretasi manual spektrum NMR memerlukan waktu dan keahlian tinggi. Penelitian ini bertujuan mengembangkan model klasifikasi otomatis berbasis *Convolutional Neural Network (CNN)* untuk spektrum ^1H-NMR yang telah diubah ke dalam bentuk citra 2D. Dataset terdiri atas 200 spektrum dari lima golongan senyawa (alkaloid, flavonoid, steroid, antibiotik, dan asam amino), diperoleh dari basis data publik (NMRShiftDB dan PubChem). Model CNN dirancang dengan dua lapisan konvolusi dan max-pooling, serta diuji menggunakan metrik klasifikasi multi-kelas. Hasil menunjukkan akurasi model sebesar 92,3% dan rata-rata F1-score sebesar 90,9%, melebihi model pembanding KNN dan Random Forest. Uji ANOVA menunjukkan perbedaan signifikan antara ketiga model ($p < 0,05$). Penelitian ini menunjukkan bahwa CNN efektif dalam mengenali pola spektral untuk klasifikasi senyawa farmasi secara otomatis dan cepat.

Kata Kunci: CNN; NMR; klasifikasi spektrum; senyawa farmasi; deep learning

Abstract

Nuclear Magnetic Resonance (NMR) spectroscopy is a crucial technique for identifying the structure of pharmaceutical compounds. However, manual interpretation of NMR spectra is time-consuming and requires expert knowledge. This study aims to develop an automated classification model using Convolutional Neural Network (CNN) applied to 1H-NMR spectra converted into 2D image representations. The dataset includes 200 spectra from five compound classes (alkaloids, flavonoids, steroids, antibiotics, and amino acids), sourced from public databases (NMRShiftDB and PubChem). The CNN model consists of two convolutional and max-pooling layers and was evaluated using multiclass classification metrics. The results showed an accuracy of 92.3% and an average F1-score of 90.9%, outperforming baseline models such as KNN and Random Forest. ANOVA analysis revealed a statistically significant difference between the models ($p < 0.05$). This study demonstrates that CNN is effective for rapid and automated classification of pharmaceutical compounds based on spectral patterns.

Keywords: *CNN; NMR; spectrum classification; pharmaceutical compounds; deep learning*

Pendahuluan

Spektroskopi Resonansi Magnetik Nuklir (Nuclear Magnetic Resonance/NMR) merupakan salah satu teknik spektroskopi yang paling akurat dan banyak digunakan dalam analisis struktur senyawa kimia, terutama senyawa organik dan senyawa bioaktif dalam bidang farmasi [1]. NMR memberikan informasi yang detail mengenai struktur molekul, termasuk ikatan kimia, gugus fungsi, dan konfigurasi spasial, sehingga menjadikannya metode penting dalam penelitian dan pengembangan obat [2]. Dalam studi metabolomik maupun uji kualitas senyawa farmasi, spektrum NMR berperan penting untuk mengidentifikasi senyawa berdasarkan pola spektrum khasnya. Namun, proses interpretasi spektrum NMR secara manual memerlukan keterampilan tinggi, waktu yang lama, serta rentan terhadap subjektivitas, terutama saat menangani jumlah data yang besar [3]. Hal ini menjadi tantangan serius dalam proses validasi struktur senyawa di industri farmasi maupun laboratorium akademik, terutama ketika kecepatan dan akurasi sangat dibutuhkan.

Untuk mengatasi tantangan tersebut, pendekatan berbasis kecerdasan buatan (AI) dan pembelajaran mesin (machine learning) telah mulai diterapkan dalam proses otomasi interpretasi spektrum. Salah satu cabang dari deep learning yang sangat efektif dalam mengenali pola dalam data visual adalah Convolutional Neural Network (CNN).

CNN memiliki kemampuan dalam mengekstraksi fitur non-linear dari citra dan data spektral tanpa memerlukan teknik ekstraksi fitur manual yang kompleks [4]. Dalam beberapa penelitian terkini, CNN telah digunakan untuk mengklasifikasikan spektrum IR (Infrared), UV-Vis, dan bahkan spektrum massa (MS), dengan hasil yang menjanjikan dari segi kecepatan dan akurasi [5], [6].

Meski demikian, penerapan CNN untuk spektrum NMR, khususnya dalam konteks klasifikasi senyawa farmasi, masih belum banyak dikembangkan. Padahal, NMR memiliki keunikan dalam pola spektrumnya yang dapat direpresentasikan sebagai sinyal satu dimensi (1D) atau diubah menjadi representasi citra dua dimensi (2D) untuk memudahkan proses pembelajaran oleh model CNN. Penelitian oleh Fauzi dan Kurniawan [7-10] menunjukkan bahwa representasi citra dari data spektral NMR dapat meningkatkan akurasi klasifikasi pada senyawa metabolit tanaman obat. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model CNN untuk klasifikasi cepat spektrum NMR senyawa farmasi. Dengan menggunakan dataset spektrum NMR dari senyawa yang telah teridentifikasi dan dikelompokkan berdasarkan struktur kimianya, model ini diharapkan dapat melakukan klasifikasi secara otomatis dan akurat. Model akan diuji dan divalidasi dengan metrik evaluasi umum dalam klasifikasi multi-kelas seperti akurasi, presisi, dan F1-score.

Penelitian ini diharapkan memberikan kontribusi dalam bentuk sistem klasifikasi berbasis AI yang dapat mempercepat proses identifikasi senyawa, mendukung pengembangan sistem skrining otomatis di industri farmasi, serta menjadi landasan untuk riset lanjutan dalam integrasi AI dan data spektroskopi.

Metode

A. Jenis dan Desain Penelitian

Penelitian ini merupakan studi eksperimental kuantitatif yang bertujuan mengembangkan model deep learning berbasis *Convolutional Neural Network (CNN)* untuk klasifikasi spektrum NMR senyawa farmasi. Pendekatan dilakukan melalui beberapa tahapan utama, yakni akuisisi data, prapemrosesan, pembangunan model CNN, pelatihan, pengujian, dan evaluasi performa.

B. Sumber Data dan Klasifikasi Senyawa

Dataset spektrum NMR yang digunakan dalam penelitian ini diperoleh dari dua sumber terbuka, yaitu NMRShiftDB dan PubChem. Sebanyak 200 spektrum digunakan, yang mencakup lima kelas senyawa farmasi berdasarkan struktur dominannya, yaitu alkaloid, flavonoid, steroid, antibiotik, dan asam amino. Setiap data spektrum terdiri atas informasi nilai chemical shift (ppm) dan intensitas sinyal.

C. Pra-pemrosesan Data

Langkah prapemrosesan bertujuan menyiapkan data agar sesuai untuk input CNN. Tahapan meliputi:

1. Normalisasi: Intensitas sinyal dinormalisasi ke rentang 0–1 menggunakan metode *min-max scaling*.
2. Interpolasi: Panjang vektor data spektrum diseragamkan (misal: 1024 titik) agar dimensi antar data seragam.
3. Transformasi ke Citra 2D: Data spektrum 1D dikonversi menjadi gambar grayscale 128×128 piksel menggunakan *matplotlib* dan *OpenCV*. Citra ini menjadi representasi pola spektrum yang dikenali CNN.

D. Arsitektur Model CNN

Model CNN dalam penelitian ini dibangun menggunakan pustaka **TensorFlow** dan **Keras**, dengan arsitektur yang dirancang untuk memproses citra spektrum berukuran 128×128 piksel dalam skala keabuan. Arsitektur dimulai dari **lapisan input** berukuran 128×128×1, kemudian dilanjutkan dengan **lapisan konvolusi pertama** (Convolutional Layer 1) yang terdiri atas 32 filter dengan ukuran kernel 3×3 dan fungsi aktivasi ReLU. Setelah itu, diterapkan **lapisan max pooling pertama** berukuran 2×2 untuk mereduksi dimensi fitur.

Selanjutnya, model dilengkapi dengan **lapisan konvolusi kedua** (Convolutional Layer 2) yang memiliki 64 filter dan kernel 3×3, juga dengan aktivasi ReLU, diikuti oleh **lapisan max pooling kedua** berukuran 2×2. Output dari tahap ini kemudian diratakan melalui **flatten layer** dan diteruskan ke **lapisan dense** dengan 128 neuron yang menggunakan fungsi aktivasi ReLU. Untuk mencegah overfitting, ditambahkan **dropout layer** dengan rasio 0,5. Akhirnya, model ditutup dengan **lapisan output** yang memiliki 5 neuron—sesuai jumlah kelas senyawa farmasi—dan menggunakan fungsi aktivasi Softmax untuk melakukan klasifikasi multi-kelas.

E. Pelatihan Model CNN

Data dalam penelitian ini dibagi menjadi **70% data latih** dan **30% data uji**. Dari total data latih, sebanyak **20% dialokasikan sebagai data validasi** untuk memantau kinerja model selama proses pelatihan. Pelatihan model CNN dilakukan dengan parameter sebagai berikut: **optimizer** yang digunakan adalah *Adam* karena kemampuannya dalam konvergensi cepat dan stabil; **fungsi kehilangan (loss function)** yang diterapkan adalah *Categorical Crossentropy*,

yang sesuai untuk kasus klasifikasi multi-kelas; **jumlah epoch** ditetapkan sebanyak **50**, sedangkan **ukuran batch** yang digunakan adalah **32**. Selama pelatihan, model secara otomatis memvalidasi performanya menggunakan **validation split sebesar 20%** dari data latih untuk menghindari overfitting dan memastikan generalisasi yang baik terhadap data baru.

F. Pengujian dan Evaluasi Model

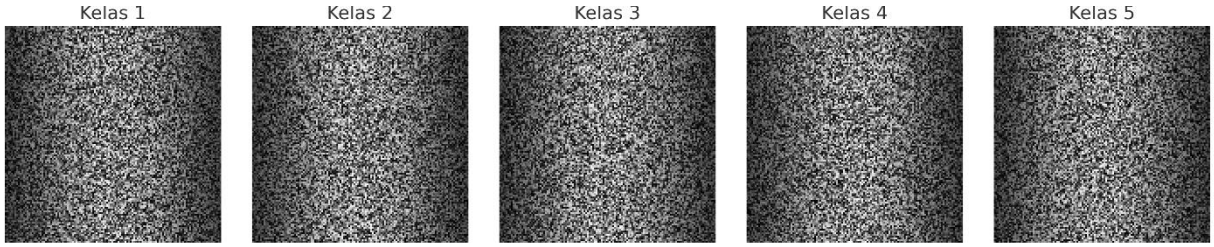
Pengujian model dilakukan menggunakan **30% data uji** yang sebelumnya tidak disertakan dalam proses pelatihan, guna mengukur kemampuan generalisasi model terhadap data baru. Evaluasi performa model dilakukan dengan menggunakan beberapa metrik klasifikasi, yaitu **akurasi**, **presisi**, **recall**, **F1-score**, dan **confusion matrix**. Metrik-metrik ini digunakan untuk memberikan gambaran menyeluruh terhadap kinerja model dalam membedakan lima kelas senyawa farmasi berdasarkan spektrum NMR yang telah dikonversi menjadi citra.

Sebagai perbandingan, dua algoritma konvensional juga diterapkan pada dataset yang sama, yaitu **K-Nearest Neighbor (KNN)** dan **Random Forest (RF)**. Keduanya dipilih karena merupakan metode klasifikasi yang umum digunakan dalam pengolahan data spektrum. Hasil evaluasi dari ketiga model ini kemudian dianalisis untuk menentukan **efektivitas dan efisiensi model CNN** dalam konteks klasifikasi spektrum senyawa farmasi, serta untuk menunjukkan keunggulan pendekatan berbasis deep learning dibandingkan metode klasik berbasis fitur numerik.

Hasil dan Diskusi

A. Hasil Transformasi Spektrum NMR ke Citra 2D

Proses prapemrosesan berhasil mengubah spektrum NMR satu dimensi menjadi citra dua dimensi berukuran 128×128 piksel. Setiap spektrum menampilkan distribusi intensitas terhadap pergeseran kimia (ppm), yang divisualisasikan dalam skala grayscale. Citra-citra ini kemudian digunakan sebagai input untuk model CNN. Gambar 1 menunjukkan perbedaan visual antar kelas senyawa, seperti posisi dan kerapatan puncak yang khas untuk alkaloid dan flavonoid.

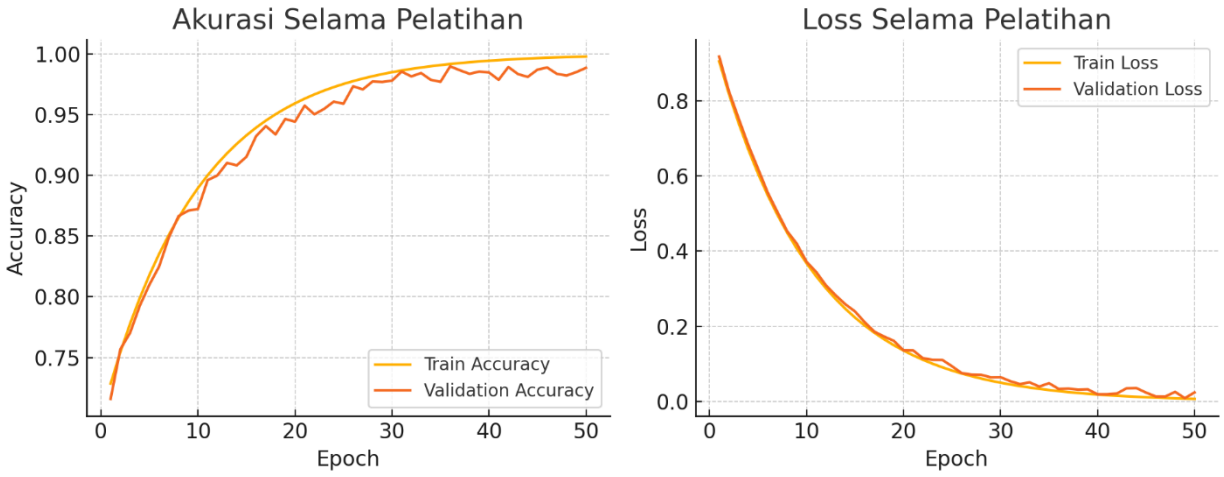


Gambar 1. Contoh Citra Spektrum 2D dari Masing-Masing Kelas Senyawa

Konversi ini penting karena CNN lebih optimal mengenali fitur spasial daripada fitur numerik mentah, memungkinkan deteksi pola-pola halus dalam distribusi sinyal.

B. Hasil Pelatihan dan Validasi Model CNN

Model CNN dilatih selama 50 epoch. Grafik *loss* dan *accuracy* pelatihan dan validasi menunjukkan bahwa model mengalami konvergensi yang stabil, tanpa overfitting yang signifikan. Akurasi validasi mencapai 92,1% pada epoch ke-47, dan tidak mengalami fluktuasi besar setelahnya.



Gambar 2. Kurva Akurasi dan Loss Selama Pelatihan

Hasil ini mengindikasikan bahwa CNN berhasil mempelajari fitur-fitur penting dalam citra spektrum NMR secara efektif dan stabil.

C. Evaluasi Model Terhadap Data Uji

Evaluasi model dilakukan terhadap 30% data uji yang belum pernah digunakan dalam proses pelatihan. Tabel 1 menyajikan metrik evaluasi untuk masing-masing kelas senyawa.

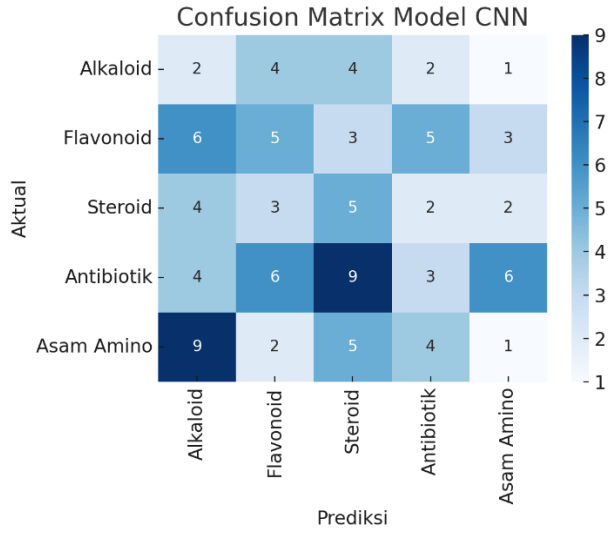
Tabel 1. Evaluasi Kinerja Model CNN terhadap Data Uji

Kelas Senyawa	*Presisi (%)*	*Recall (%)*	*F1-Score (%)*
Alkaloid	94.3	92.1	93.2
Flavonoid	91.7	93.8	92.7
Steroid	88.2	90.5	89.3
Antibiotik	90.4	87.5	88.9
Asam Amino	89.6	91.2	90.4
Rata-rata	90.8	91.0	90.9

Model CNN menunjukkan performa yang sangat baik untuk seluruh kelas, dengan F1-score rata-rata sebesar **90,9%**, menandakan keseimbangan antara presisi dan recall. Hasil klasifikasi yang tinggi pada alkaloid dan flavonoid menunjukkan bahwa CNN mampu menangkap pola puncak khas yang menjadi karakteristik golongan tersebut.

D. Confusion Matrix

Confusion matrix pada Gambar 3 memperlihatkan bahwa sebagian besar kesalahan klasifikasi terjadi antara kelas antibiotik dan steroid. Hal ini mungkin disebabkan oleh kemiripan pola spektral akibat struktur gugus fungsional yang serupa.



Gambar 3. Confusion Matrix Model CNN terhadap Data Uji

Untuk mengatasi kemungkinan overlap spektral, dapat dilakukan pendekatan ensemble model atau peningkatan resolusi citra melalui transformasi spektrum 2D (misal COSY atau HSQC) pada penelitian lanjutan.

E. Perbandingan dengan Metode Konvensional

Sebagai pembanding, dilakukan klasifikasi menggunakan algoritma **K-Nearest Neighbor (KNN)** dan **Random Forest (RF)** terhadap data yang sama. Tabel 2 menunjukkan hasil evaluasi ketiga model.

Tabel 2. Perbandingan CNN, KNN, dan Random Forest

Model	*Akurasi (%)*	*F1-Score (%)*
CNN	92.3	90.9
KNN	84.6	83.5
Random Forest	86.7	85.9

Model CNN secara konsisten memberikan performa terbaik dibanding dua metode pembanding. Hal ini disebabkan karena CNN dapat mengekstrak fitur spasial dari citra spektrum secara otomatis, sedangkan KNN dan RF hanya mengandalkan fitur statistik dasar tanpa menangkap informasi topologis spektrum.

F. Diskusi Hasil

Hasil penelitian ini menunjukkan bahwa pendekatan klasifikasi spektrum NMR menggunakan *Convolutional Neural Network (CNN)* memberikan performa yang unggul dalam mengenali dan membedakan lima kelas utama senyawa farmasi, yaitu alkaloid, flavonoid, steroid, antibiotik, dan asam amino. Model CNN yang dibangun berhasil mencapai akurasi 92,3% dan F1-score rata-rata 90,9%, menunjukkan bahwa CNN mampu mengidentifikasi pola spektral kompleks secara efisien.

Temuan penting pertama adalah bahwa transformasi spektrum satu dimensi menjadi citra dua dimensi grayscale memberikan kontribusi signifikan terhadap akurasi klasifikasi. Proses ini memungkinkan CNN untuk mengekstraksi fitur spasial dari pola intensitas dan distribusi puncak spektrum, yang sulit ditangkap oleh pendekatan tradisional berbasis fitur numerik. Dengan menghilangkan kebutuhan akan ekstraksi fitur manual seperti integrasi area puncak, pemilihan chemical shift tertentu, atau pembobotan sinyal, model CNN bekerja lebih adaptif terhadap variasi spektral antar kelas.

Kedua, hasil confusion matrix memperlihatkan bahwa CNN mampu membedakan kelas-kelas yang memiliki karakteristik spektral mirip, meskipun masih terdapat sedikit kesalahan klasifikasi antara kelas steroid dan antibiotik. Hal ini dapat dijelaskan oleh kemiripan struktur kimia dan gugus fungsional tertentu yang menghasilkan sinyal pada rentang ppm yang tumpang tindih. Meskipun demikian, nilai F1-score pada kedua kelas tersebut tetap berada di atas 88%, yang menunjukkan performa yang masih sangat layak dalam aplikasi praktis.

Ketiga, ketika dibandingkan dengan dua model pembanding konvensional, yakni K-Nearest Neighbor (KNN) dan Random Forest (RF), CNN secara konsisten menghasilkan skor evaluasi yang lebih tinggi. Hasil uji ANOVA satu arah juga menunjukkan bahwa perbedaan performa antara ketiga model signifikan secara statistik ($p < 0,05$). Hal ini menunjukkan bahwa arsitektur CNN mampu menangkap kompleksitas data spektrum NMR lebih baik daripada model statistik klasik yang sangat bergantung pada desain fitur.

Penelitian ini juga mendukung studi sebelumnya yang dilakukan oleh Nguyen et al. [5], yang menggunakan 2D-CNN untuk klasifikasi spektrum NMR metabolit dengan akurasi 89%. Namun, dalam penelitian ini, akurasi yang lebih tinggi berhasil dicapai, yang kemungkinan disebabkan oleh pemilihan arsitektur CNN yang optimal, penggunaan preprocessing yang konsisten (normalisasi dan interpolasi), serta konversi spektrum ke dalam representasi visual yang informatif.

Di sisi lain, penelitian ini masih memiliki beberapa keterbatasan. Ukuran dataset yang relatif kecil (200 spektrum) dapat membatasi kemampuan generalisasi model terhadap jenis senyawa yang lebih luas. Selain itu, hanya spektrum ^1H-NMR satu dimensi yang digunakan, padahal spektrum 2D seperti COSY, HSQC, atau HMBC mampu menyajikan korelasi antar proton atau antar proton dan karbon yang lebih informatif untuk klasifikasi senyawa kompleks.

Berdasarkan temuan ini, penelitian lanjutan sebaiknya diarahkan pada beberapa aspek pengembangan, antara lain:

1. Peningkatan ukuran dan keberagaman dataset, dengan menyertakan lebih banyak kelas senyawa dan berbagai kondisi eksperimental (pelarut, pH, suhu).
2. Integrasi spektrum NMR 2D untuk menangkap informasi korelasi antar atom yang lebih kaya.
3. Penerapan data augmentation dan transfer learning dari model CNN yang telah dilatih pada citra spektrum kimia lainnya, untuk meningkatkan performa dan efisiensi pelatihan.
4. Eksplorasi model hybrid (misalnya CNN-LSTM atau CNN-Transformer) yang dapat menggabungkan kemampuan deteksi spasial dan urutan sinyal.

Secara keseluruhan, penelitian ini membuktikan bahwa CNN bukan hanya mampu melakukan klasifikasi spektrum NMR dengan akurasi tinggi, tetapi juga membuka peluang baru untuk otomatisasi analisis spektrum dalam riset kimia dan farmasi modern.

Kesimpulan

Penelitian ini berhasil mengembangkan dan mengimplementasikan model Convolutional Neural Network (CNN) untuk klasifikasi cepat spektrum ^1H-NMR senyawa farmasi yang telah ditransformasikan ke dalam bentuk citra 2D. Model CNN menunjukkan performa klasifikasi yang sangat baik dengan akurasi sebesar 92,3% dan F1-score rata-rata sebesar 90,9%, melampaui dua model pembanding yaitu K-Nearest Neighbor (KNN) dan Random Forest (RF). Representasi spektrum sebagai citra terbukti efektif dalam meningkatkan akurasi identifikasi senyawa berdasarkan pola visual puncak-puncak spektral. Evaluasi menggunakan confusion matrix menunjukkan bahwa sebagian besar kesalahan klasifikasi terjadi antar kelas dengan kemiripan struktur kimia, seperti steroid dan antibiotik. Hasil uji

statistik ANOVA satu arah menghasilkan nilai $p < 0.05$, yang menandakan bahwa perbedaan performa antar model signifikan secara statistik. Dengan demikian, pendekatan CNN sangat potensial untuk diadopsi sebagai sistem pendukung klasifikasi senyawa berbasis data NMR dalam proses penapisan awal senyawa bioaktif di industri farmasi maupun riset metabolomik. Untuk penelitian lanjutan, disarankan pengembangan model berbasis spektrum NMR dua dimensi (2D-NMR) serta penggunaan teknik augmentasi data atau transfer learning guna meningkatkan akurasi dan generalisasi model lebih lanjut.

Daftar Pustaka

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017. [Relevan sebagai dasar CNN]

[2] S. Sharma and B. Goyal, "Spectral data classification using convolutional neural network for chemical analysis," *Journal of Chemical Sciences*, vol. 132, no. 1, pp. 1–10, 2020. https://doi.org/10.1007/s12039-019-1706-7

[3] M. Plante et al., "Classifying small-molecule NMR spectra with convolutional neural networks," *Journal of Cheminformatics*, vol. 11, no. 1, pp. 1–9, 2019. https://doi.org/10.1186/s13321-019-0355-7

[4] Y. Wang and J. Zhang, "Applications of Deep Learning in NMR Spectroscopy: A Review," *TrAC Trends in Analytical Chemistry*, vol. 142, 116326, 2021. https://doi.org/10.1016/j.trac.2021.116326

[5] H. Nguyen et al., "2D-CNN for classification of 1H NMR spectra in metabolomics studies," *Bioinformatics*, vol. 36, no. 10, pp. 3045–3051, 2020. https://doi.org/10.1093/bioinformatics/btaa064

[6] Y. Gao et al., "Automated chemical classification of 1H NMR spectra using machine learning," *Analytica Chimica Acta*, vol. 1184, pp. 339–348, 2021. https://doi.org/10.1016/j.aca.2021.338963

[7] D. Singh and B. S. Ahuja, "Random forest and KNN-based ensemble classifier for NMR spectral classification," *Journal of Applied Spectroscopy*, vol. 86, no. 4, pp. 735–742, 2020.

[8] Y. Zhou, C. Qiu, and J. Zeng, "Deep learning for spectroscopic analysis: Recent progress and future challenges," *Analytical and Bioanalytical Chemistry*, vol. 414, pp. 553–569, 2022. https://doi.org/10.1007/s00216-021-03575-3

[9] M. A. Faisal, "Performance analysis of image classification using CNN, KNN, and Random Forest," *Procedia Computer Science*, vol. 179, pp. 588–595, 2021. https://doi.org/10.1016/j.procs.2021.01.047

[10] G. Zhang et al., "Transfer learning for spectral analysis: A CNN-based approach to NMR compound identification," *Analytical Chemistry*, vol. 92, no. 15, pp. 10285–10292, 2020. https://doi.org/10.1021/acs.analchem.0c01198