Available online at:
Al-Fiziya: Journal of Materials Science, Geophysics, Instrumentation and Theoretical Physics
Website: http://journal.uinjkt.ac.id/index.php/al-fiziya
Vol.3 No.I Tahun 2020, 9 - 14

Evaluasi Implementasi Algoritma *Machine Learning K-Nearest Neighbors* (kNN) pada Data Spektroskopi Gamma Resolusi Rendah

Muhammad Sholih Fajri†, Nizar Septian, Edi Sanjaya

Program Studi Fisika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta, Jalan Ir. H. Djuanda No.95, Ciputat, Kota Tangerang Selatan, Banten, 15412, Indonesia

†Sholih.fajri15@mhs.uinjkt.ac.id

Abstrak

Pada artikel ini kami mengevaluasi bagaimana implementasi algoritma *machine learning k-Nearest Neighbors* (kNN) pada data spektroskopi gamma beresolusi rendah. Penelitian ini bertujuan untuk mengetahui bagaimana performa kNN dalam mempelajari data tersebut. Kami melakukan berbagai variasi, yaitu: jumlah data *training*, jumlah data tes, jenis *metric*, dan nilai *k* untuk memperoleh performa terbaik dari algoritma ini. Data spektroskopi gamma diambil menggunakan sintilator NaI(Tl) Leybold Didactic dengan resolusi energi sebesar 10.9 keV per *channel*. Hasil variasi menunjukkan bahwa algoritma kNN memberikan hasil prediksi klasifikasi radioisotop yang sangat fluktuatif.

Kata Kunci: Akurasi, *Euclidean*, Gamma, *k-Nearest Neighbors*, *Manhattan*, *Minkowski*, Radioisotop.

Abstract

In this paper we evaluate the implementation of a machine learning algorithm namely k-Nearest Neighbors (kNN) on low resolution gamma spectroscopy data. The aim is to provide the information of how well the algorithm performs on learning the data. We did the variation of number of training and test data, type of metric used, and values of k in order to see the best performance of the algorithm. The gamma spectroscopy data were taken using NaI(Tl) scintillator made by Leybold Didactic with resolution of 10.9 keV per channel. The variations show that the kNN algorithm produce significantly fluctuating accuracy to the prediction of radioisotope class.

Keywords: *Accuracy, Euclidean, Gamma, k-Nearest Neighbors, Manhattan, Minkowski, Radioisotope.*

DOI: *10.15408/fiziya.v3i1.16180*

PENDAHULUAN

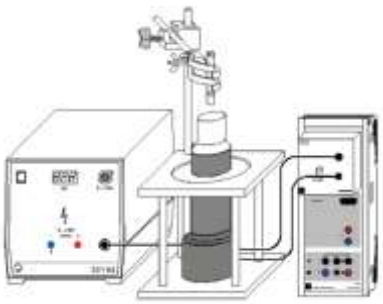
Berbicara tentang *machine learning*, *k-Nearest Neighbors* (kNN) adalah algoritma *machine learning classifier* populer yang paling sederhana. kNN pertama kali diperkenalkan oleh T. Cover dan P. Hart pada tahun 1967 dimana algoritma ini mengklasifikasikan kelas sampel berdasarkan kelas tetangga terdekatnya [1]. kNN sering disebut juga sebagai *lazy learner*, karena kNN mempelajari dan mengklasifikasi data tanpa membangun sebuah model. Tidak seperti algoritma klasifikasi berbasis model, *classifier* kNN hanya perlu mengingat semua data *training* dalam memori [2]. Seiring dengan kepopulerannya, kNN banyak digunakan untuk melakukan klasifikasi data dalam bidang sains dan teknik maupun ekonomi dan bisnis. Beberapa publikasi terkini yang menerapkan kNN meliputi: prediksi penyakit jantung [3], diagnosis penyakit diabetes [4], prediksi permintaan pasar [5], klasifikasi protein [6] dan desain radar detektor [7]. Selain itu, kNN juga banyak digunakan sebagai basis pengembangan algoritma *machine learning* yang lebih canggih [8-12].

Data mining adalah proses ekstraksi pengetahuan dari sejumlah data [13]. Dalam bidang fisika, penerapan metode *data mining* sedang sangat berkembang pesat dalam dua dekade terakhir ini [14]. Hal ini dipicu oleh semakin banyaknya jumlah data eksperimen yang dihasilkan dan metode *data mining* yang cenderung lebih efisien dibandingkan metode analisis tradisional yang telah lama diterapkan. Salah satu penerapan *data mining* dalam bidang fisika ialah dalam mengolah data spektroskopi, seperti pada spektroskopi gamma.

Data spektroskopi gamma biasa digunakan sebagai media untuk mempelajari sifat dari sumber radiasi gamma seperti identifikasi isotop dan estimasi aktivitas radioaktif. Untuk spektroskopi gamma beresolusi tinggi, data dapat dianalisis menggunakan metode tradisional seperti *fitting* dan identifikasi puncak. Pustaka khusus yang didedikasikan untuk mempelajari data gamma spektroskopi beresolusi tinggi telah tersedia [16]. Namun, metode ini tidak cukup baik ketika digunakan pada data gamma spektroskopi beresolusi rendah. Beberapa studi menunjukkan bahwa algoritma *machine learning* seperti *neural network* (NN) [15-16] dan *support vector machine* (SVM) [17] sangat efektif untuk mempelajari data spektroskopi gamma beresolusi rendah. Berdasarkan hal tersebut, pada penelitian ini kami mencoba mengevaluasi penggunaan kNN untuk mengidentifikasi sebuah radioisotop berdasarkan data spektroskopi gamma beresolusi rendah. Penelitian ini bertujuan untuk mengetahui apakah algoritma sederhana seperti kNN dapat diandalkan untuk proses identifikasi radioisotop.

METODE

Data spektroskopi gamma diambil dengan meletakkan radioisotop di depan detektor sintilator NaI(Tl). Pengambilan data dilakukan dalam beberapa variasi waktu cacahan, yaitu 60, 120 dan 180 detik. Radioisotop yang digunakan adalah Co-60, Na-22, Am-241, Sr-90 dan campuran Sr-90, Am-241 dan Cs-137. Setiap radioisotop diambil data cacahan radiasi gamma sebanyak 200 kali dan total data yang diambil adalah 1000 data untuk 5 unsur. Data tersebut akan digunakan sebagai data *training* pada algoritma kNN. Sementara itu, untuk data tes, diambil data cacahan dengan waktu 45, 90, 150 dan 240 detik dengan total data sebanyak 100 data untuk 5 unsur. Alat-alat yang digunakan untuk pengambilan data dapat dilihat pada Gambar 1.



Gambar 1 Susunan alat untuk pengambilan data spektroskopi gamma [18]

Algoritma kNN bekerja dengan cara menghitung jarak tiap titik pada data tes dengan data latihan tiap kelas. Lalu, diurutkan dari jarak terdekat ke jarak terjauh dan akan dipilih jarak terdekat antara data tes dengan data latihan sejumlah *k*. Kelas yang memiliki jarak terdekat dengan data tes akan menjadi kelas data tes tersebut. Jika diambil lebih dari 1 tetangga, maka kelas data tes akan ditentukan dengan mayoritas kelas terdekat [19].

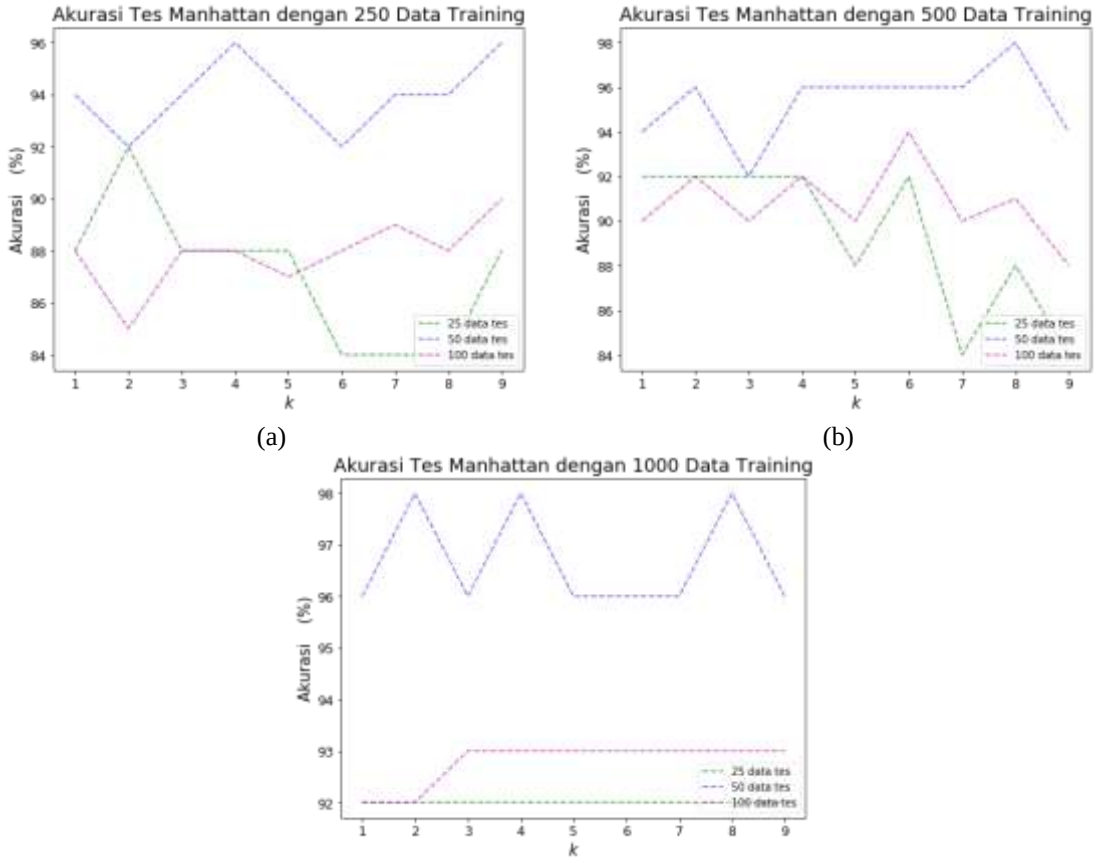
Algoritma kNN memiliki 3 *metric* atau cara menghitung jarak antar data, yaitu *Manhattan distance*, *Euclidean distance* dan *Minkowski distance*. Secara *default*, algoritma kNN menggunakan *metric Minkowski distance*. *Minkowski distance* secara matematis dapat ditulis seperti persamaan berikut:

$$d(\mathbf{x}.\mathbf{y}) = |\sum_{i=1}^{n}(x_i - y_i)^p|^{\frac{1}{p}} \quad (1)$$

Dengan *d* ialah jarak antara titik **x** dengan titik **y** di ruang fitur. *n* adalah jumlah fitur. Ketika $p = 1$, maka persamaan tersebut akan berubah menjadi *Manhattan distance*. Sementara jika $p = 2$, maka akan berubah menjadi *Euclidean distance*.

Pada penelitian ini, ketiga *metric* digunakan untuk mengetahui *metric* mana yang memiliki akurasi tertinggi pada algoritma kNN yang digunakan. Selain menggunakan ketiga *metric* tersebut, dilakukan juga beberapa variasi lain yaitu variasi jumlah data *training* (250, 500 dan 1000 data *training*), variasi jumlah data tes (25, 50 dan 100 data tes) serta variasi nilai *k* (dari 1 sampai 9). Hal ini dimaksudkan untuk mengetahui bagaimana performa algoritma kNN dalam mengidentifikasi sebuah radioisotop berdasarkan data spektroskopi gamma yang dimiliki.

HASIL DAN PEMBAHASAN

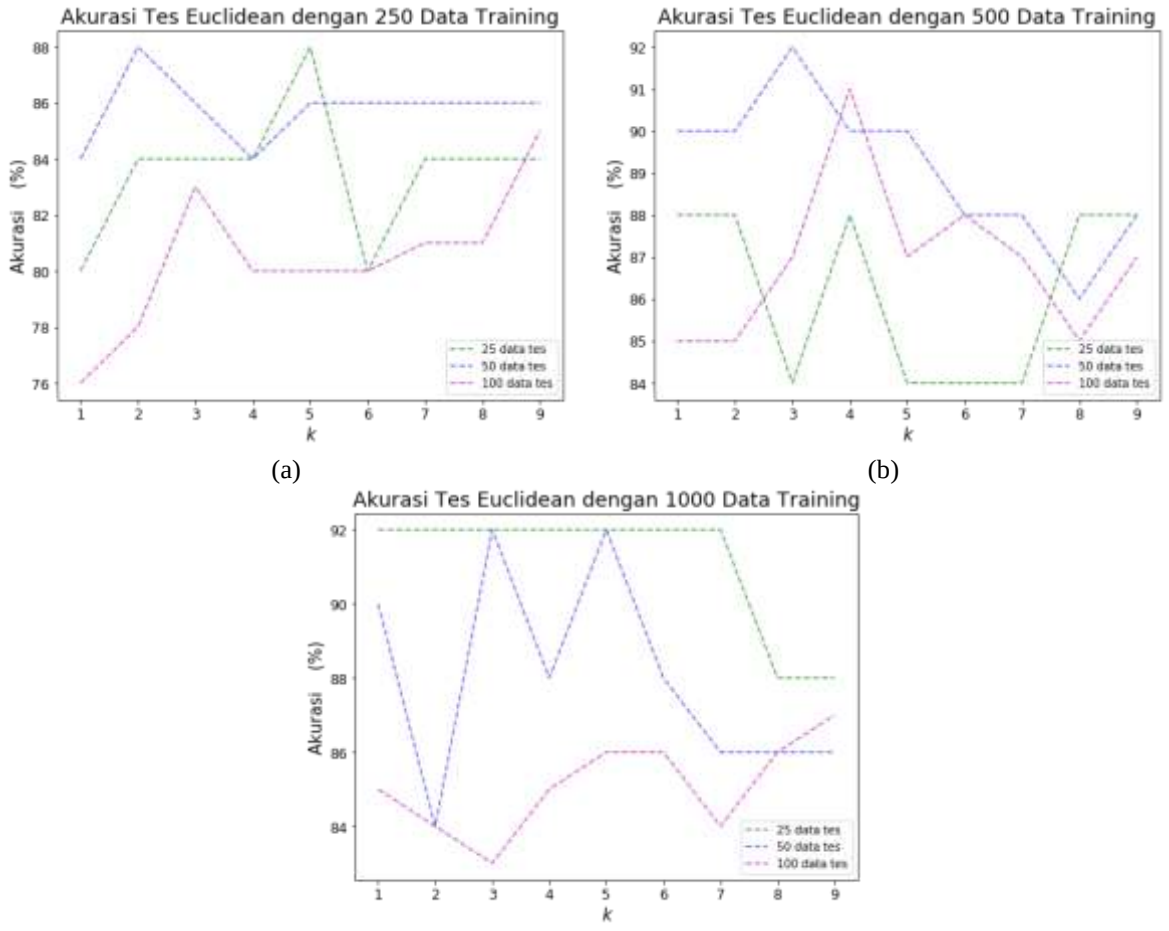


Gambar 2 Grafik Akurasi Tes *Metric Manhattan* (a) dengan 250 Data *Training*, (b) dengan 500 Data Training dan (c) dengan 1000 Data *Training*

Pada dasarnya, otomasi identifikasi radioisotop dapat dilakukan menggunakan algoritma kNN, karena kNN merupakan algoritma *classifier*. Untuk itu, algoritma kNN terlebih dahulu dilatih dengan total jumlah data *training* sebanyak 1000 data (waktu cacahan 60, 120 dan 180 detik). Sementara itu, total jumlah data tes yang digunakan adalah 100 data (waktu cacahan 45, 90, 150 dan 240 detik) dan dibagi ke dalam tiga variasi jumlah yaitu 25, 50 dan 100 data tes. Data *training* yang dimiliki dibagi ke dalam tiga variasi jumlah data yaitu 250, 500 dan 1000 data *training*. Semua variasi jumlah data *training* digunakan pada model kNN dengan *metric Manhattan, Euclidean* dan *Minkowski.* Setiap model dilatih dengan melakukan iterasi nilai *k* sebanyak sembilan kali (*k* = 1, 2, 3, …, 9) dan memprediksi tiap variasi jumlah data tes yang digunakan.

Algoritma kNN dengan tiga *metric* berbeda dan variasi jumlah data latihan menunjukkan hasil yang berbeda satu sama lain ketika memprediksi data tes. Bahkan, untuk *metric* yang sama, namun menggunakan jumlah data *training* yang berbeda, terdapat perbedaan nilai akurasi saat memprediksi data tes untuk tiap variasi jumlah data tes.

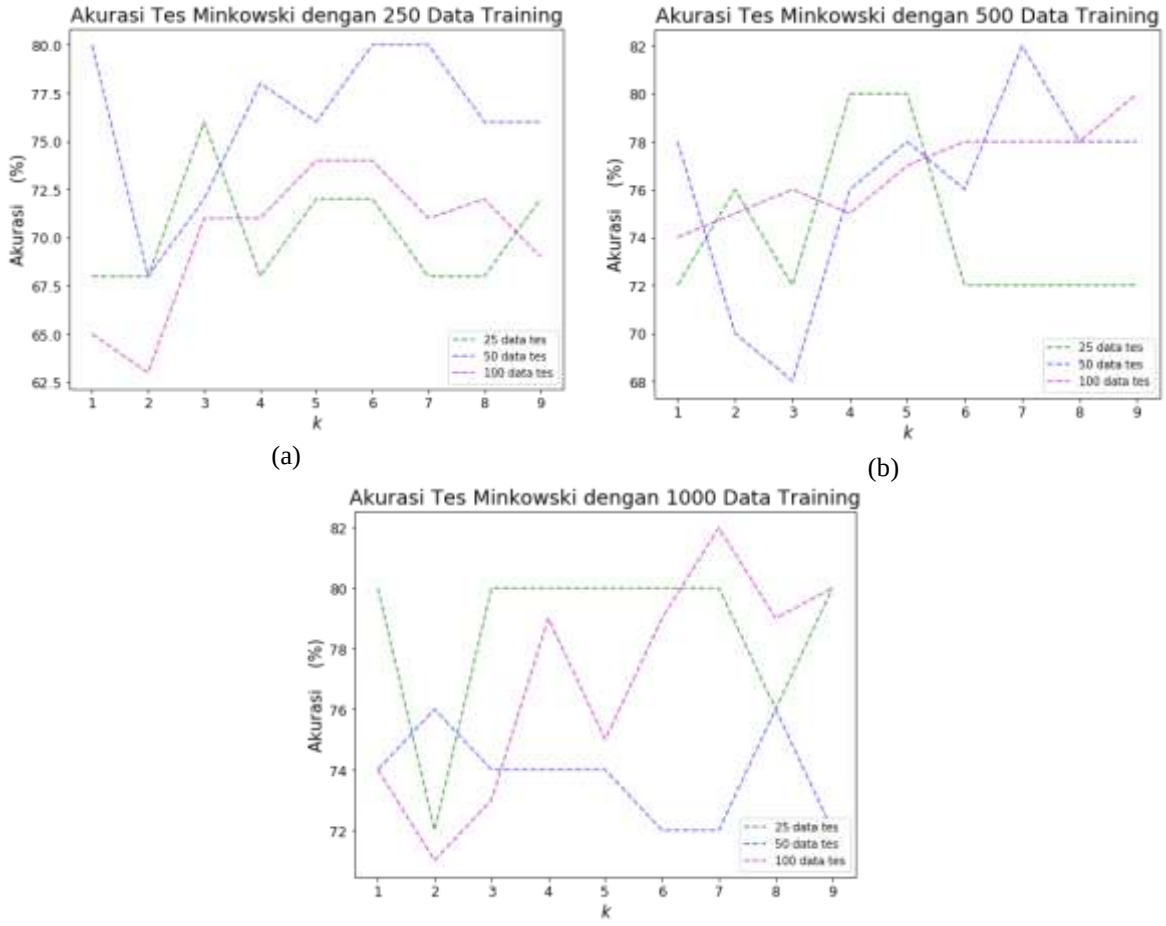
Pertama, model kNN dilatih dengan menggunakan *metric Manhattan* (*p* = 1). Pada gambar 2 dapat terlihat bahwa untuk algoritma kNN dengan *metric Manhattan* mendapatkan akurasi tes tertinggi sebesar 98% pada model dengan 500 dan 1000 data *training* ketika memprediksi 50 data tes. Hasil tersebut didapat ketika nilai *k* = 8 untuk model dengan 500 data *training* dan nilai *k* = 2, 4 dan 8 untuk model dengan 1000 data *training*.



Gambar 3 Grafik Akurasi Tes *Metric Euclidean* (a) dengan 250 Data *Training*, (b) dengan 500 Data Training dan (c) dengan 1000 Data *Training*

Kedua, model kNN dilatih dengan menggunakan *metric Euclidean* (*p* = 2). Performa yang ditunjukkan oleh model dengan *metric Euclidean* berada di bawah performa model dengan *metric Manhattan.* Gambar 3 menunjukkan bahwa untuk algoritma kNN dengan *metric Euclidean*

mendapatkan akurasi tes tertinggi sebesar 92% pada model dengan 500 data *training* dengan nilai *k* = 3 dan model dengan 1000 data *training* dengan nilai *k* = 3 dan 5 ketika memprediksi 50 data tes serta model dengan 1000 data *training* dengan nilai *k* = 1 hingga 7 saat memprediksi 25 data tes.



Gambar 4 Grafik Akurasi Tes *Metric Minkowski* (a) dengan 250 Data *Training*, (b) dengan 500 Data Training dan (c) dengan 1000 Data *Training*

Terakhir, model kNN dilatih dengan menggunakan *metric Minkowski* (*p* = 3). Performa yang ditunjukkan oleh model dengan *metric Minkowski* adalah yang terburuk dari semua model dengan *metric* yang ada. Model dengan *metric Minkowski* tidak mampu mencapai akurasi di atas 90%. Akurasi terbaik pada model ini hanya sebesar 82%. Nilai akurasi tersebut didapat ketika model ini menggunakan 500 data *training* dengan nilai *k* = 7 memprediksi 50 data tes. Selain itu, angka tersebut juga diraih oleh model dengan *metric Minkowski* ketika menggunakan 1000 data *training* dengan nilai *k* = 7 memprediksi 100 data tes. Performa model dengan *metric Minkowski* dapat dilihat pada Gambar 4.

PENUTUP

Pada penelitian ini kami mengevaluasi kinerja algoritma *machine learning* kNN yang digunakan untuk keperluan identifikasi radioisotop berdasarkan data spektroskopi gamma beresolusi rendah. Berdasarkan beberapa variasi terhadap parameter-parameter karakteristik algoritma kNN yaitu nilai *p, k,* dan jumlah data *training* dan tes, teramati bahwa algoritma kNN tidak cukup presisi dalam memprediksi kelas dari radioisotop berdasarkan data spektroskopi gamma beresolusi rendah. Hal ini terlihat dari tidak adanya kecenderungan kenaikan nilai akurasi yang signifikan ketika data *training* bertambah. Meskipun teramati bahwa salah satu variasi

algoritma kNN dengan *metric Manhattan* mendapatkan akurasi tes yang tinggi, yaitu sebesar 98% dengan jumlah data *training* dan nilai *k* yang berbeda, semua hasil variasi terlihat sangat fluktuatif. Oleh karena itu, cukup sulit bagi kami untuk mempercayai bahwa hasil yang diberikan oleh algoritma ini ialah hasil yang konsisten. Sehingga dapat ditarik kesimpulan bahwa algoritma ini tidak cocok untuk keperluan mempelajari data gamma spektroskopi beresolusi rendah.

REFERENSI

[1] T.Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[2] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, pp. 234–242, 2020.
[3] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013.
[4] R. Zubaedah, F. Xaverius, H. Jayawardana, and S. H. Hidayat, "Comparing euclidean distance and nearest neighbor algorithm in an expert system for diagnosis of diabetes mellitus," *Enfermería Clínica*, vol. 30, pp. 374–377, 2020.
[5] M. Kück and M. Freitag, "Forecasting of customer demands for production planning by local k-nearest neighbor models," *Int. J. Prod. Econ.*, p. 107837, 2020.
[6] R. Arian, A. Hariri, A. Mehridehnavi, A. Fassihi, and F. Ghasemi, "Protein kinase inhibitors' classification using K-Nearest neighbor algorithm," *Comput. Biol. Chem.*, vol. 86, p. 107269, 2020.
[7] A. Coluccia, A. Fascista, and G. Ricci, "A k-nearest neighbors approach to the design of radar detectors," *Signal Processing*, vol. 174, p. 107609, 2020.
[8] J. Jiang, Y. Chen, X. Meng, L. Wang, and K. Li, "A novel density peaks clustering algorithm based on k nearest neighbors for improving assignment process," *Phys. A Stat. Mech. its Appl.*, vol. 523, no. 20180101044, pp. 702–713, 2019.
[9] Ö. F. Ertuğrul and M. E. Tağluk, "A novel version of k nearest neighbor: Dependent nearest neighbor," *Appl. Soft Comput.*, vol. 55, pp. 480–490, 2017.
[10] B. Wang and Z. Mao, "A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule," *Inf. Fusion*, vol. 63, pp. 30–40, 2020.
[11] T. M. Tran, X.-M. T. Le, H. T. Nguyen, and V.-N. Huynh, "A novel non-parametric method for time series classification based on k-Nearest Neighbors and Dynamic Time Warping Barycenter Averaging," *Eng. Appl. Artif. Intell.*, vol. 78, pp. 173–185, 2019.
[12] Y. Pan, Z. Pan, Y. Wang, and W. Wang, "A new fast search algorithm for exact k-nearest neighbors based on optimal triangle-inequality-based check strategy," *Knowledge-Based Syst.*, vol. 189, p. 105088, 2020.
[13] J. Han, M. Kamber, and J. Pei, "1 - Introduction," in *The Morgan Kaufmann Series in Data Management Systems*, J. Han, M. Kamber, and J. B. T.-D. M. (Third E. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 1–38.
[14] N. M. Ball and R. J. Brunner, "Data Mining and Machine Learning in Astronomy," *Int. J. Mod. Phys. D*, vol. 19, no. 07, pp. 1049–1106, Jul. 2010
[15] M. Kamuda and C. J. Sullivan, "An automated isotope identification and quantification algorithm for isotope mixtures in low-resolution gamma-ray spectra," *Radiat. Phys. Chem.*, vol. 155, no. June 2018, pp. 281–286, 2019.
[16] M. Kamuda, J. Zhao, and K. Huff, "A comparison of machine learning methods for automated gamma-ray spectroscopy," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 954, no. October, 2020.
[17] H. Hata, K. Yokoyama, Y. Ishimori, Y. Ohara, Y. Tanaka, and N. Sugitsue, "Application of support vector machine to rapid classification of uranium waste drums using low-resolution γ-ray spectra," *Appl. Radiat. Isot.*, vol. 104, pp. 143–146, 2015.
[18] L. Didactic, Detecting γ radiation with a scintillation counter, Leybold Didactic, 2012.
[19] S. Raschka, *Python Machine Learning*. Packt Publishing, 2015.