

Prediksi Return Saham Berbasis *Clustering* Menggunakan *K-Means Clustering* dan *Convolutional Neural Network*

Rizaldi Maulid Fathurachman¹, Deni Saepudin²

^{1,2}Fakultas Informatika, Universitas Telkom, Bandung

¹rizaldimaulidf@students.telkomuniversity.ac.id, ²denisaepudin@telkomuniversity.ac.id

Abstrak

Prediksi harga saham di masa depan merupakan sebuah masalah yang menantang. Dalam penelitian ini telah diterapkan pendekatan *clustering* untuk mengelompokkan saham-saham yang menunjukkan kesamaan dalam prediksi harga. Proses ini membantu dalam mempermudah prediksi harga, karena saham yang ada dalam 1 cluster yang sama diharapkan mempunyai pola yang lebih mirip. Melalui *clustering*, saham-saham yang memiliki prediksi harga yang sama dikelompokkan berdasarkan cluster yang telah terbentuk. Kemudian, dilakukan training model menggunakan data rata-rata dari harga saham dalam setiap cluster untuk menghasilkan prediksi harga yang lebih tepat. Dalam penelitian ini metode *K-Means* digunakan untuk mengelompokkan saham berdasarkan rata-rata return dan standar deviasi, sedangkan *CNN* digunakan untuk memprediksi harga saham harian. Selanjutnya prediksi return dapat dihitung berdasarkan prediksi harga yang telah diperoleh. Dataset harga saham LQ45 dibagi menjadi dua bagian, yaitu training data dan test data. Data training digunakan untuk melatih model *CNN* untuk menganalisis data harga saham historis yang dapat membantu memprediksi return saham di masa depan, sedangkan data test digunakan untuk mengevaluasi kinerja model yang telah dilatih. Hasil penelitian ini menunjukkan bahwa, dengan melihat kriteria elbow maka *clustering* dengan metode *K-Means* akan diambil dengan jumlah clusternya adalah sebanyak 4 cluster. Dan setelah itu dilakukan prediksi melalui *CNN* dan hasilnya mampu menghasilkan prediksi harga saham harian dengan baik, dibuktikan dengan nilai *RMSE* prediksi untuk 1 hari ke depan pada *cluster 1* adalah 0,09, Untuk *cluster 2*, *RMSE* yang diperoleh adalah 0,05, Sementara itu, *cluster 3* memiliki *RMSE* sebesar 0,07, dan untuk *cluster 4* memiliki *RMSE* yang diperoleh yaitu 0,57. Setelah itu, dihasilkan dari prediksi saham-saham di setiap *cluster* yang menghasilkan prediksi *return* tertinggi di setiap harinya. Kemudian dibuat prediksi kinerja portofolio menggunakan *equal weight* dengan hasil rata-rata *return* sebesar 0,0009 dan standar deviasi sebesar 0,0209 yang lebih baik dibandingkan dengan indeks LQ45.

Kata kunci: Prediksi *return* saham, *clustering*, *K-means*, *Convolutional Neural Network (CNN)*, portofolio investasi.

Abstract

Predicting future stock prices is a challenging problem. In this research, a clustering approach will be applied to group stocks that show similarities in price movements. This process helps in making price prediction easier, as stocks in the same cluster are expected to have more similar patterns. Through clustering, stocks that have similar price movements are grouped based on the clusters that have been formed. Then, a training model is performed using average data from stock prices in each cluster to produce more precise price predictions. In this research, the K-Means method is used to group stocks based on average return and standard deviation, while CNN is used to predict daily stock prices. Furthermore, return predictions can be calculated based on the price predictions that have been obtained. The LQ45 stock price dataset is divided into two parts, namely training data and test data. Training data is used to train CNN models to analyze historical stock price data that can help predict future stock returns, while test data is used to evaluate the performance of the model that has been trained. The results of this study indicate that, by looking at the elbow criteria, clustering with the K-Means method will be taken with the number of clusters is 4 clusters. And after that, predictions are made through CNN and the results are able to produce daily stock price predictions well, as evidenced by the RMSE value of predictions for the next 1 day in cluster 1 is 0.09, For cluster 2, the RMSE obtained is 0.05, Meanwhile, cluster 3 has an RMSE of 0.07, and for cluster 4 has an RMSE obtained which is 0.57. After that, the predictions of stocks in each cluster are generated which produce the highest return predictions on each day. Then a portfolio performance prediction is made using equal weight with an average return of 0.0009 and a standard deviation of 0.0209 which is better than the LQ45 index.

Keywords: Stock return prediction, clustering, K-means, Convolutional Neural Network, investment portfolio.

1. Pendahuluan

1.1 Latar Belakang

Dalam penelitian ini dibahas mengenai tantangan dalam memprediksi harga saham di masa depan sebagai masalah utama dalam prediksi return saham. Masalah ini menarik karena dapat membantu investor memahami

karakteristik saham tertentu untuk keputusan investasi yang lebih baik. Untuk mengatasi tantangan tersebut, dalam penelitian ini diterapkan pendekatan *clustering* untuk mengelompokkan saham-saham yang menunjukkan kesamaan dalam prediksi harga. Proses ini membantu dalam mempermudah prediksi harga, karena saham yang ada dalam 1 cluster yang sama diharapkan mempunyai pola yang lebih mirip. Melalui *clustering*, saham-saham yang memiliki prediksi harga yang sama dikelompokkan berdasarkan *cluster* yang telah terbentuk. Kemudian, dilakukan *training model* menggunakan data rata-rata dari harga saham dalam setiap *cluster* untuk menghasilkan prediksi *return* yang lebih tepat.

Banyak penelitian telah dilakukan untuk mengatasi masalah ini. Penelitian Hongxiang He dan tim pada tahun 2018, Dimana dalam penelitian ini menemukan bahwa CNN mampu menghasilkan prediksi yang lebih akurat dibandingkan metode lain seperti regresi linier dan model ARIMA. Dalam eksperimen, CNN menunjukkan akurasi prediksi yang signifikan dengan akurasi 73,2% hingga 75,2% setelah 500.000 iterasi pada berbagai set data. Selain itu, F1-score tertinggi mencapai 73,0%, yang menegaskan kemampuan CNN untuk menangkap pola kompleks dalam data harga saham, memberikan keunggulan dalam memproyeksikan prediksi harga saham dibandingkan dengan model tradisional lainnya[4].

Penelitian yang dilakukan oleh Zheng Fang dkk. 2021, menggunakan algoritma K-means yang dioptimalkan dengan algoritma Artificial Fish Swarm (AFSA) untuk mengelompokkan saham berdasarkan karakteristik tertentu. Hasilnya menunjukkan bahwa metode KAFSA efektif dalam mengelompokkan saham menjadi dua kategori: saham berkinerja tinggi dan rendah, dengan koefisien siluet sebesar 0,99. Saham dalam kelompok B (berkinerja tinggi) memiliki harga penutupan, rasio harga-pendapatan, laba per saham, dan pengembalian aset bersih yang lebih tinggi dibandingkan dengan saham dalam kelompok A (berkinerja rendah), menawarkan panduan yang baik bagi investor dalam membuat keputusan investasi[2].

Adapun paper yang menjadi referensi acuan adalah Sarfaraz Hashemkhani Zolfani dkk pada tahun 2023, menggunakan metode PSO-CNN+MVF dan *K-means clustering* untuk mengelompokkan saham-saham berdasarkan mean dan variance dari fitur input dan return aset selama periode pelatihan. Hasil menunjukkan bahwa model ini unggul dalam return dan pengelolaan risiko. Model LSTM memberikan return harian 0,0071, sementara CNN+MVF mencatat 0,0046 setelah biaya transaksi. CNN+1/N mencatat risiko terendah, dan secara tahunan, metode yang diusulkan mencapai pengembalian tertinggi 4,9007 dengan rasio Sortino 9,0543, mengungguli model lain dalam efisiensi dan risiko yang lebih rendah, menunjukkan potensi besar dalam strategi investasi berbasis data[1].

Dalam penelitian ini, dilakukan *clustering* saham dengan menggunakan metode *k-means*. Fitur-fitur yang digunakan dalam clustering ini adalah rata-rata *return* dan standar deviasi. Setelah *return* saham terbentuk, semua saham telah diclusterkan berdasarkan karakteristik tersebut. Selanjutnya, untuk memprediksi tren pasar, digunakan algoritma *Convolutional Neural Network* (CNN) untuk memprediksi *return* saham harian. Hasil dari prediksi ini kemudian digunakan sebagai dasar dalam memilih saham untuk portofolio investasi. Saham-saham dengan *return* tertinggi kemudian dimasukkan dalam portofolio dan dengan menggunakan metode *equal weight*, kemudian portofolionya diuji dibandingkan dengan kinerja dari indeks LQ45.

Dengan demikian, penelitian ini bertujuan untuk menghadirkan solusi baru dalam analisis pasar saham, membantu para investor membuat keputusan investasi yang lebih informasional dan terinformasi.

1.2 Topik dan Batasannya

Topik yang dibahas dalam penelitian ini adalah bagaimana penerapan clustering dan penggunaan CNN dapat mempengaruhi prediksi return saham untuk seleksi awal dalam optimasi portofolio.

Adapun batasan dari penelitian ini adalah dataset indeks LQ45 yang diperoleh dari situs yahoo finance. Dataset yang digunakan adalah data harga harian selama 10 tahun mulai dari 1 Januari 2013 sampai 1 Januari 2023. Dataset ini terdiri dari 39 saham yang tergantung dalam indeks LQ45.

1.3 Tujuan

Tujuan dari penelitian ini adalah menerapkan metode *K-means clustering* untuk mengelompokkan saham berdasarkan rata-rata *return* dan standar deviasi, membangun model prediksi menggunakan *Convolutional Neural Network* (CNN), serta mengembangkan metode pengukuran kinerja portofolio yang terbentuk dari saham-saham yang dipilih berdasarkan prediksi return. Setelah itu menerapkan metode prediksi untuk menyeleksi saham dalam portofolio, Tujuannya adalah untuk membantu investor membuat keputusan investasi yang lebih baik dengan menemukan saham dengan potensi pertumbuhan terbaik.

2. Studi Terkait

2.1 Penelitian Terkait

Pada tahun 2018, Chen dan He menggunakan metode CNN untuk memprediksi harga saham di pasar saham China, dengan menggunakan data seperti harga pembukaan, tertinggi, terendah, penutupan, dan *volume* saham. Hasil penelitian menunjukkan bahwa metode ini efektif untuk memprediksi harga saham di China[4].

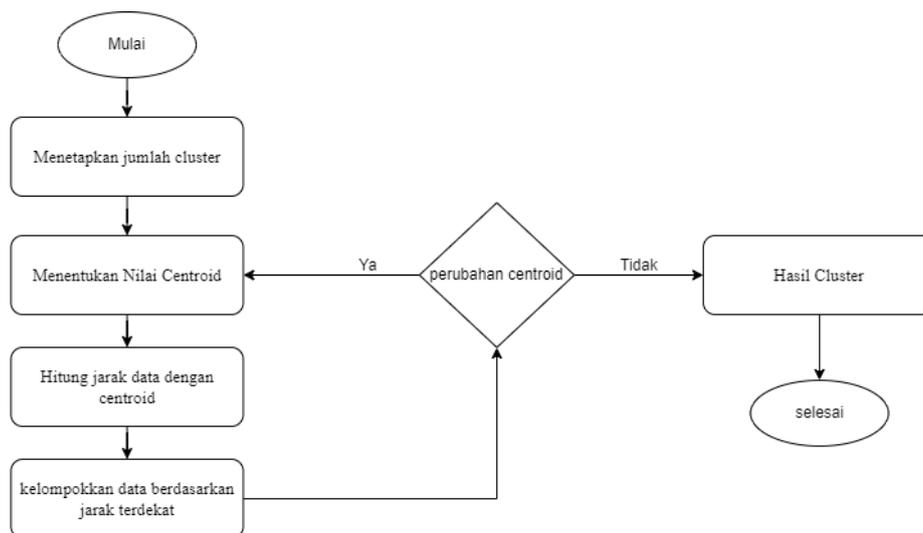
Pada tahun 2021, penelitian lain menggabungkan model pembelajaran mesin dan pembelajaran mendalam seperti *Random Forest* (RF), *Support Vector Regression* (SVR), LSTM, *Deep Multilayer Perceptron* (DMLP), dan CNN untuk pra-pemilihan saham. Penelitian ini bertujuan untuk mengintegrasikan hasil prediksi dari model-model tersebut dalam optimasi portofolio menggunakan model *mean-variance* (MV) dan omega[7].

Pada tahun 2021, Shawel, et al. menggunakan k-means clustering untuk mengklasifikasikan pemanfaatan spektrum dari 639 stasiun pangkalan dengan tujuan meningkatkan akurasi prediksi dalam pengambilan keputusan. Untuk prediksi masa depan, mereka menggunakan RNN (khususnya LSTM) dan CNN untuk memprediksi tingkat pemanfaatan pada level cluster[5].

Pada tahun 2023, Ashrafzadeh, et al. mengembangkan sebuah model hibrida yang mengintegrasikan *Convolutional Neural Network* (CNN) dengan *Particle Swarm Optimization* (PSO) untuk pra-seleksi saham dan model *mean-variance with forecasting* (MVF) untuk optimalisasi portofolio. Tujuan utama dari penelitian ini adalah untuk meningkatkan kinerja portofolio dengan mengelola aset keuangan secara efektif, meminimalkan risiko, dan menghasilkan keuntungan positif bagi investor[1].

2.2 Algoritma K-Means

Algoritma *K-Means* adalah metode pengelompokan data, yang bertujuan untuk membagi sekumpulan data menjadi K kelompok (*cluster*) yang berbeda dengan menggunakan pendekatan iteratif. *K-means clustering* adalah teknik pengelompokan data non-hierarkis yang efektif. Metode ini membagi kumpulan data menjadi beberapa *cluster*, di mana setiap *cluster* terdiri dari data yang memiliki kesamaan. Dengan menggunakan pendekatan ini, data dapat dikategorikan ke dalam kelompok-kelompok yang memudahkan analisis dan interpretasi. Gambar 1 adalah flowchart dari algoritma *K-Means* [3].



Gambar 1 : Flowchart algoritma k-means

Berikut langkah-langkah dalam algoritma *K-means Clustering*:

1. Menetapkan jumlah *cluster*
2. Menentukan nilai *centroid*
3. Menghitung jarak antara titik *centroid* dan setiap titik objek
4. Pengelompokan objek dalam menentukan anggota *cluster* dilakukan dengan mempertimbangkan jarak minimum antar objek
5. Kembali ke tahap 2, proses ini diulang hingga nilai *centroid* stabil dan tidak ada anggota cluster yang berpindah ke cluster lain.

2.3 Metode Elbow

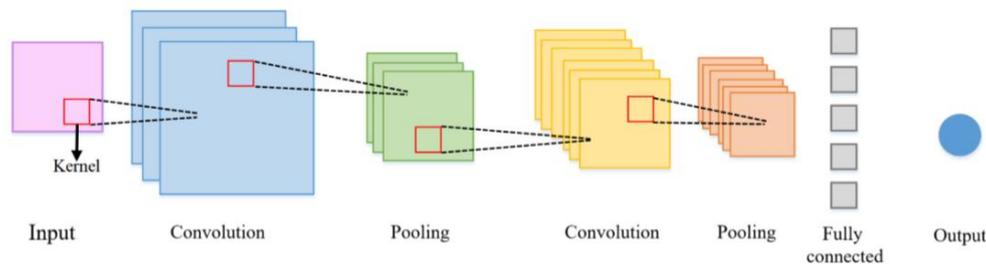
Metode *Elbow* dalam konteks saham digunakan untuk menentukan jumlah optimal *cluster* (kelompok) saat melakukan analisis *cluster* pada data saham. Analisis *cluster* bertujuan untuk mengelompokkan saham-saham dengan karakteristik serupa berdasarkan berbagai fitur seperti volatilitas, pengembalian, atau indikator teknis.

Metode *Elbow* membantu menentukan berapa banyak *cluster* yang paling sesuai untuk menangkap struktur data yang mendasarinya [6].

Metode *Elbow* adalah teknik heuristik yang digunakan untuk menentukan jumlah optimal *cluster* (nilai K) dalam algoritma pengelompokan seperti *K-Means*. Metode ini mengevaluasi variasi total dalam *cluster* (*Within-Cluster Sum of Squares*/WCSS atau *inersia*) untuk berbagai nilai K . WCSS adalah ukuran seberapa baik objek dikelompokkan bersama, dengan nilai yang lebih rendah menunjukkan pengelompokan yang lebih baik. Dalam metode *Elbow*, nilai K yang berbeda diplot terhadap WCSS yang sesuai. Titik di mana grafik membentuk "siku" atau penurunan paling tajam dalam WCSS menunjukkan jumlah *cluster* yang optimal. Pada titik ini, menambahkan lebih banyak *cluster* tidak akan secara signifikan meningkatkan kualitas pengelompokan.

2.4 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) untuk saham adalah model pembelajaran mendalam (*deep learning*) yang terinspirasi oleh cara kerja visual korteks pada otak manusia. CNN dirancang untuk memproses data yang memiliki struktur grid seperti gambar dan telah berhasil diterapkan dalam berbagai tugas pengenalan pola, termasuk analisis saham. Dalam konteks saham, CNN digunakan untuk menganalisis data historis harga saham, volume perdagangan, indikator teknis, dan informasi relevan lainnya yang disajikan dalam bentuk grafik atau matriks. CNN dapat mempelajari pola-pola kompleks dalam data ini, seperti tren, volatilitas, dan korelasi antar variabel, yang dapat digunakan untuk memprediksi harga saham di masa depan (Gambar 2).



Gambar 2 : arsitektur *Convolutional Neural Network* (CNN) 1D

Arsitektur dasar dari *Convolutional Neural Network* (CNN) adalah suatu struktur *neural network* yang sangat efektif untuk analisis dan pengolahan data visual seperti gambar dan video. CNN biasanya terdiri dari beberapa lapisan utama yang berfungsi saling terkait untuk mengidentifikasi fitur penting dari data visual, mempelajari pola, dan membuat prediksi yang akurat. Gambar tersebut menjelaskan arsitektur dasar dari *Convolutional Neural Network* (CNN) 1D yang sangat efektif untuk aplikasi seperti prediksi saham. Dalam CNN, data masuk melalui lapisan input dan diolah menggunakan operasi *convolution*, di mana filter atau kernel mengekstrak fitur penting dari data. Langkah selanjutnya adalah *pooling*, yang mengurangi dimensi peta fitur sambil mempertahankan fitur penting, membantu mengurangi komputasi dan membuat model lebih stabil terhadap perubahan posisi dalam data. Setelah itu, lapisan *fully connected* mengintegrasikan semua fitur yang telah diekstrak untuk menghasilkan output akhir, yaitu prediksi, berdasarkan pola yang dipelajari dari data. Proses ini sangat meningkatkan akurasi prediksi dan efisiensi komputasi dalam menganalisis data keuangan[1].

2.5 Portofolio (*Equal Weight Portfolio*)

Equal-Weighted Portfolio adalah strategi alokasi aset di mana setiap aset dalam portofolio diberi bobot yang sama, tidak peduli besar kecilnya perusahaan atau variabel lainnya. Strategi ini sering digunakan karena kesederhanaannya dan karena memberikan eksposur yang lebih merata di seluruh aset yang berbeda, yang bisa mengurangi risiko khusus sektor atau perusahaan dan potensi bias dalam portofolio. Dengan pendekatan ini, investor tidak perlu melalui proses analisis yang rumit atau memilih aset dengan sangat selektif. Karena semua aset diberi bobot yang sama, proses pembuatan dan pemeliharaan portofolio menjadi lebih mudah dan langsung[8].

$$W_i = \frac{1}{n}, i = 1 \dots n \quad (1)$$

Keterangan:

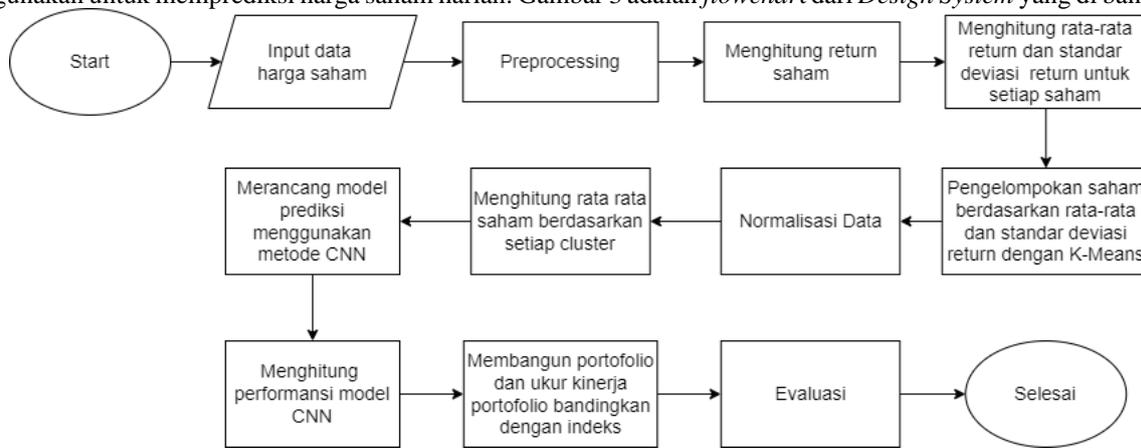
W_i : Bobot dari saham,

i dan n : Jumlah indeks saham yang akan digunakan.

3. Sistem yang Dibangun

3.1 Design System

Alur dari sistem pemodelan ini yang menerapkan metode *K-means clustering* untuk mengelompokkan saham berdasarkan rata-rata *return* dan standar deviasi, Selanjutnya *Convolutional Neural Network* (CNN) digunakan untuk memprediksi harga saham harian. Gambar 3 adalah *flowchart* dari *Design System* yang di bangun.



Gambar 3: *Flowchart Design System*

3.2. Data Harga Saham LQ45

Dataset harga saham LQ45 menyajikan harga saham perusahaan di Bursa Efek Indonesia yang termasuk dalam indeks LQ45, yang diperoleh dari situs yahoo finance. Dataset yang digunakan adalah data harga harian selama 10 tahun mulai dari 2 Januari 2013 sampai 30 Desember 2022. Dataset ini terdiri dari 39 saham yang tergantung dalam indeks LQ45.

Tabel 1 : Data harga saham LQ45

Date	ADRO	AMRT	ANTM	ASII	BBCA	BBNI	BBRI	BBTN	BFIN	BMRI
1/2/2013	1.740	525	1.117	7.500	1.820	1.863	1.410	1.480	203	2.063
1/3/2013	1.750	525	1.142	7.850	1.830	1.863	1.440	1.460	195	2.050
1/4/2013	1.750	500	1.134	7.850	1.810	1.900	1.470	1.500	191	2.063
1/7/2013	1.730	493	1.126	7.750	1.840	1.900	1.480	1.500	192	2.050
.....

3.3 Pre-Processing

Preprocessing data untuk saham adalah serangkaian langkah penting yang dilakukan sebelum data saham digunakan untuk analisis atau pemodelan. Tujuannya adalah untuk meningkatkan kualitas data dengan mengeliminasi noise, menangani masalah missing values, dan mengonversi data ke format yang sesuai dengan algoritma machine learning atau teknik analisis lainnya[10]. Selanjutnya, *split* data menggunakan data dari 2019 kebawah untuk data train dan 2020 keatas untuk data test.

3.4 Menghitung Return Harga saham

Setelah melakukan *input* data, langkah selanjutnya adalah menghitung *return* untuk setiap harinya. *Return* merupakan keuntungan yang didapat investor dari investasi sahamnya. Dapat dihitung dengan persamaan berikut ini:

$$R_i = \frac{S_i - S_{i-1}}{S_{i-1}} \quad (2)$$

Keterangan:

R_i : Return saham,

S_i : Harga saham pada periode ke-i,

S_{i-1} : Harga saham sebelum periode ke i.

3.5 Menghitung Rata-rata Return dan Standar Deviasi

Pada tahap ini, dilakukan penghitungan nilai rata-rata *return* serta penghitungan standar deviasi untuk masing-masing sahamnya. Hasil dari penghitungan tersebut digunakan untuk proses *clustering* menggunakan metode *elbow* dan *K-Means*.

Rata-rata *return* saham adalah ukuran statistik yang menunjukkan berapa rata-rata keuntungan atau kerugian yang dihasilkan oleh investasi saham dalam periode tertentu. Berikut dapat diperoleh untuk melakukan perhitungan tersebut :

$$m = \frac{1}{n} \sum_{i=1}^n R_i \quad (3)$$

Keterangan:

m : mean (rata rata) return,

n : banyaknya data,

R_i : nilai *return* dari masing-masing saham periode ke i .

Standar deviasi sendiri adalah ukuran seberapa tersebar atau variatif *return* saham dari rata-ratanya. Standar deviasi *return* mengukur seberapa jauh *return* saham berbeda dari rata-rata *return*. Berikut dapat diperoleh untuk melakukan perhitungan tersebut :

$$sd = \sqrt{\frac{1}{n} \sum_{i=1}^n (R_i - m)^2} \quad (4)$$

Keterangan:

sd : Standar deviasi,

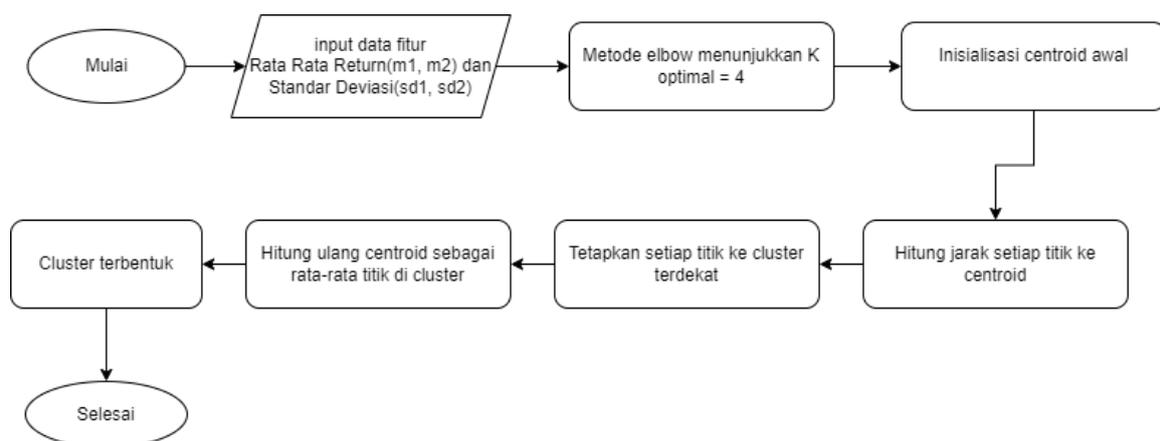
R_i : nilai *return* dari masing-masing saham periode ke i ,

m : nilai rata – rata saham,

N : banyaknya data.

3.6 Algoritma K-means dan Metode Elbow

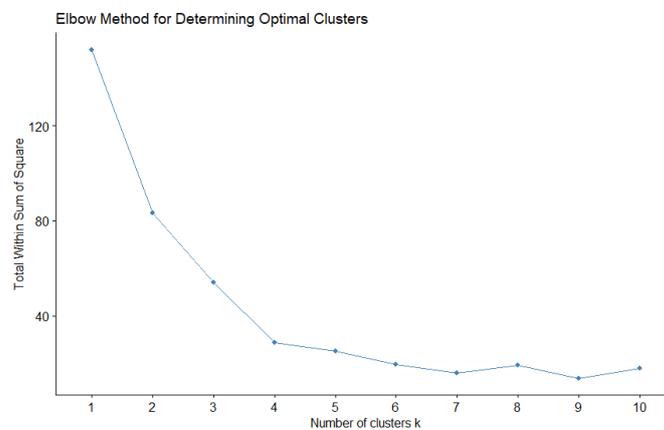
Setelah mendapatkan rata-rata *return* dan standar deviasi, hasil tersebut digunakan untuk menentukan anggota dari tiap *cluster* yaitu penentuan pada inersia atau titik siku yang mengalami penurunan untuk menentukan jumlah *cluster* optimal. Jumlah optimal *cluster* yang dihasilkan yaitu empat *cluster* (Gambar 4).



Gambar 4 : Flowchart K-means

Flowchart ini menjelaskan proses *clustering* menggunakan metode *K-Means*, khususnya dengan input data fitur Standar Deviasi dan Rata Rata *Return*. Berikut adalah penjelasan detail dari tiap langkah dalam flowchart tersebut:

- Langkah pertama ini menunjukkan bahwa proses clustering dimulai.
- Input data yang terdiri dari fitur Standar Deviasi dan Rata Rata *Return* dimasukkan ke dalam sistem untuk dianalisis. (m1 dan sd1 Diambil dari 5 Tahun awal, m2 dan sd2 Diambil dari 5 Tahun akhir.)
- Menggunakan metode elbow, jumlah *cluster* optimal yang dipilih adalah 4. Metode ini membantu menentukan jumlah *cluster* yang tepat dengan mencari titik di mana penurunan SSE (*Sum of Squared Errors*) mulai melambat.
- Inisialisasi centroid awal: Pada langkah ini, *centroid* awal dari keempat *cluster* dipilih secara acak.
- Setiap titik data kemudian dihitung jaraknya ke masing-masing *centroid* untuk menentukan kedekatannya.
- Berdasarkan jarak yang telah dihitung, setiap titik data ditetapkan ke *cluster* dengan centroid terdekat.
- Setelah semua titik ditetapkan ke *cluster*, *centroid* baru dihitung sebagai rata-rata posisi dari titik-titik dalam setiap *cluster*.
- Proses ini diulang beberapa kali hingga centroid tidak lagi berubah secara signifikan, menunjukkan bahwa *cluster* telah terbentuk dengan stabil.
- Proses *clustering* selesai dan hasil akhir adalah kumpulan *cluster* yang telah terbentuk berdasarkan data fitur yang dimasukkan.



Gambar 5 : Visualisasi metode elbow

Kemudian, hasil tersebut digunakan untuk menentukan anggota di setiap *cluster*-nya dengan cara melakukan perhitungan sesuai pendekatan dari hasil rata rata yang terdekat.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
BFIN EXCL TOWR	ADRO AMRT ANTM CPIN ERAA HRUM INCO ITMG JPFA MNCN PGAS PTBA TINS WIKA	ASII BBKA BBNI BBRI BBTN BMRI HMSP ICBP INDF INTP KLBF SMGR TBIG TLKM UNTR UNVR	BRPT EMTK INDY INKP MEDC TPIA

Tabel 2. Hasil *clustering*

3.7 Normalisasi Data

Setelah menetapkan anggota untuk setiap *cluster*, langkah berikutnya adalah melakukan normalisasi pada data harga saham dari tiap anggota *cluster* tersebut. Normalisasi ini bertujuan untuk menyeragamkan skala data sehingga perbandingan *return* saham antar *cluster* menjadi lebih valid.

Setelah data dinormalisasi, kita dapat menghitung rata-rata *return* harga saham harian untuk setiap *cluster*. Hasil perhitungan ini memberikan gambaran tentang kinerja rata-rata setiap *cluster*, sehingga dapat digunakan untuk membandingkan dan menganalisis karakteristik setiap kelompok saham.

3.8 Prediksi Return menggunakan Convolutional Neural Network (CNN)

Model yang digunakan untuk prediksi adalah sebagai berikut:

Tabel 3 : Model cnn

	Categories	Cluster 1	Cluster 2	Cluster 3	Cluster 4
conv1d (Conv1D)	Filters	64	64	64	64
	kernel_size	2	2	2	2
	activation	Relu	Relu	Relu	Relu
MaxPooling1D	pool_size	2	2	2	2
Fitting parameters	optimizer	adam	adam	adam	adam
	epochs	50	50	50	50
	batch_size	32	32	32	32
result	Train Loss	0.0028181	0.000552534	0.00100916	0.172316
	Test Loss	0.00446836	0.00152446	0.00432861	0.197435

Model yang digunakan adalah *Convolutional Neural Network* (CNN) dengan arsitektur *sequential* yang terdiri dari satu lapisan *convolutional* 1D, satu lapisan *max pooling*, satu lapisan *flatten*, dan dua lapisan *dense*. Model ini dilatih menggunakan *optimizer* 'adam' dan *loss function* 'mse'. Hasil pelatihan menunjukkan bahwa model mampu belajar dengan cepat dari data, dengan *train loss* dan *validation loss* yang menurun secara signifikan dalam beberapa *epoch* awal. Model mencapai *train loss* sebesar 0,0028 dan *validation loss* sebesar 0,0045 pada akhir pelatihan, menunjukkan kinerja yang baik dalam memprediksi data, baik pada data pelatihan maupun data baru. Meskipun demikian, perlu dilakukan evaluasi lebih lanjut dengan metrik yang sesuai dan analisis error untuk mendapatkan pemahaman yang lebih mendalam tentang kinerja model dan potensi.

3.9 RMSE (Root Mean Square Error)

Setelah menerapkan model prediksi, error prediksi dihitung menggunakan RMSE. Nilai RMSE yang rendah menunjukkan bahwa model prediksi memiliki tingkat akurasi yang tinggi dalam memprediksi harga saham. Ini adalah bagian penting dari analisis keuangan karena membantu investor dalam membuat keputusan investasi yang lebih baik berdasarkan prediksi yang akurat. Nilai RMSE dapat dihitung dengan persamaan sebagai berikut:

$$RMSE = \sqrt{\frac{\sum(\hat{Y}_i - Y_i)^2}{n}} \quad (5)$$

Keterangan:

- \hat{y}_i : Nilai prediksi saham
- y_i : Nilai aktual saham
- n : Banyaknya data

3.10 Menghitung return prediksi

Pada tahap ini dilakukan perhitungan *return* prediksi dari harga saham, dengan membandingkan harga yang diprediksi untuk setiap hari dengan harga aktual dari hari sebelumnya. Hal ini membantu dalam mengevaluasi

akurasi prediksi model keuangan, penting untuk pengambilan keputusan investasi dan strategi keuangan. Berikut dapat diperoleh untuk melakukan perhitungan tersebut:

$$R_p = \frac{PR_i - S_{i-1}}{S_{i-1}} \quad (6)$$

Keterangan:

PR_i : Nilai prediksi harga periode ke- i ,

S_{i-1} : Harga *actual* saham sebelum periode ke i .

3.11 Perhitungan Return Portofolio

Tahapan perhitungan *return* portofolio merupakan langkah krusial dalam mengevaluasi kinerja investasi.

Return portofolio dihitung dengan memperhitungkan bobot dari setiap saham dalam portofolio serta return dari saham-saham tersebut. Dapat dihitung dengan persamaan berikut ini:

$$rp = \sum_{i=1}^n W_i \cdot R_i \quad (7)$$

Keterangan:

rp : *Return* portofolio ,

W_i : Bobot saham i dalam portofolio,

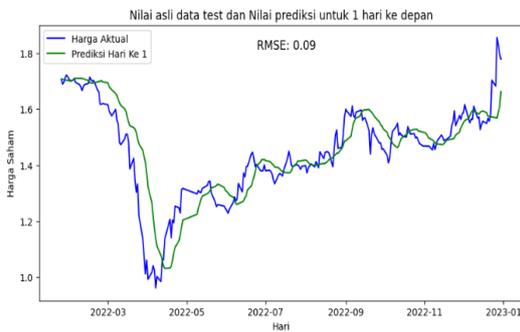
R_i : *Return actual* saham i pada portofolio,

n : Banyaknya saham pada portofolio.

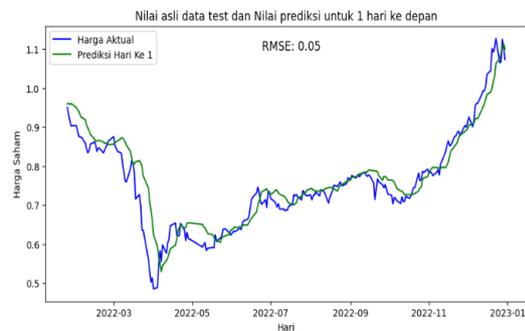
4. Evaluasi

4.1 Hasil Pengujian

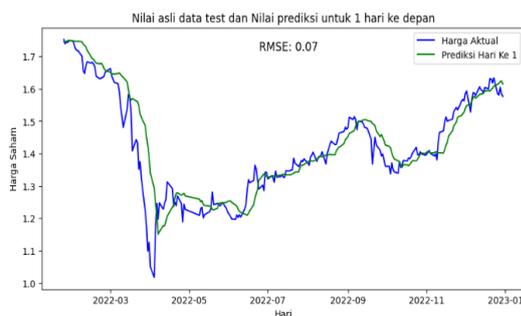
Pada pengujian ini terdapat hasil prediksi saham harian dari tiap cluster menggunakan model CNN dengan hasil seperti dibawah:



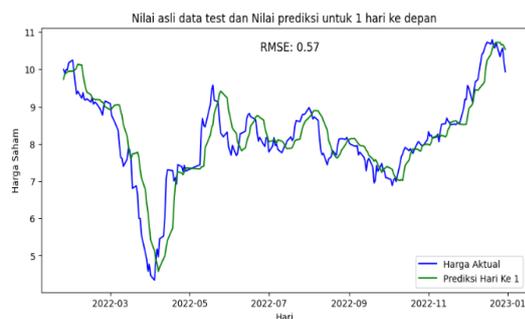
Gambar 6 : Prediksi 1 hari kedepan Cluster 1



Gambar 7 : Prediksi 1 hari kedepan Cluster 2



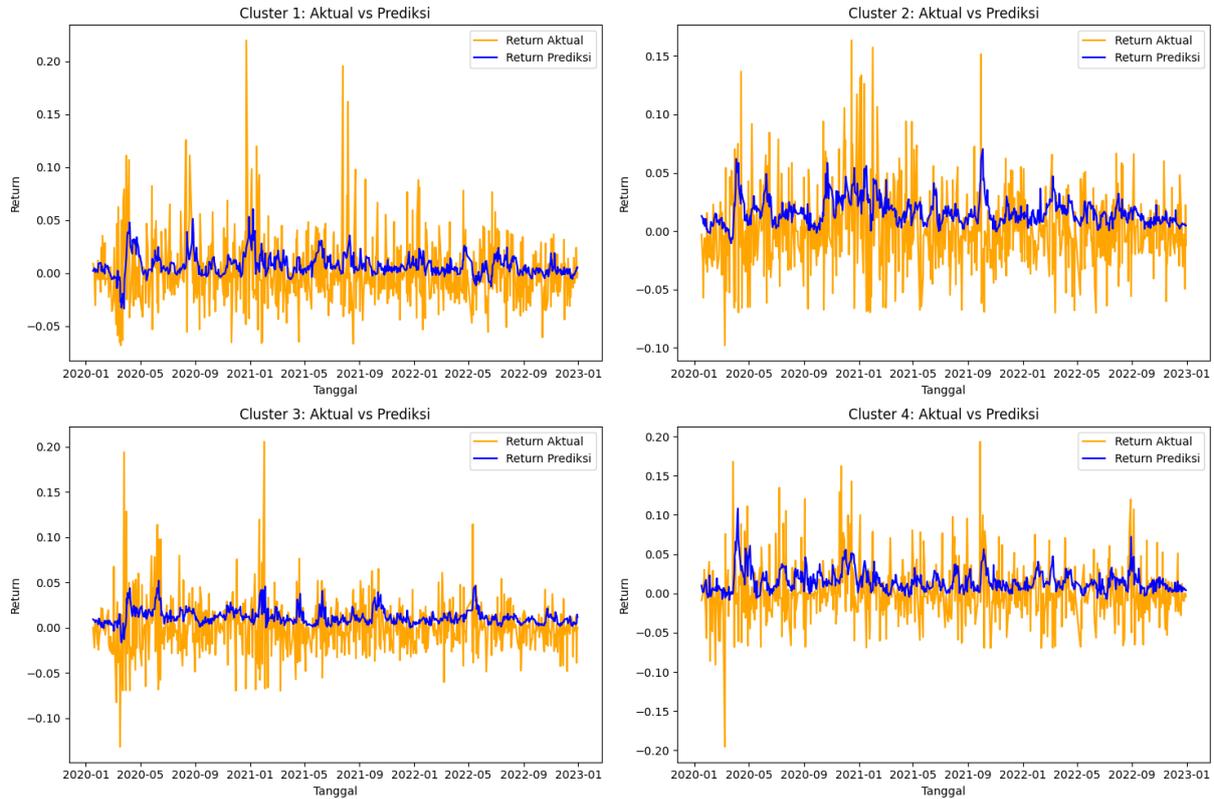
Gambar 8 : Prediksi 1 hari kedepan Cluster 3



Gambar 9 : Prediksi 1 hari kedepan Cluster 4

Dari hasil tersebut, kinerja model untuk memprediksi harga saham harian dinilai baik karena memiliki hasil RMSE yang rendah dan mendekati nilai aktual, yang menunjukkan efektivitas model dalam menangkap pola pasar. Berdasarkan analisis yang lebih detail, RMSE prediksi untuk 1 hari ke depan pada cluster 1 adalah 0,09, Untuk cluster 2, RMSE yang diperoleh adalah 0,05, Sementara itu, cluster 3 memiliki RMSE sebesar 0,07, dan untuk cluster 4 memiliki RMSE yang diperoleh yaitu 0,57. Kemudian dari hasil diatas, diambil nilai returnnya untuk menjadi prediktor memprediksi saham saham di setiap cluster.

Kemudian dihasilkan dari prediksi saham-saham di setiap cluster yang menghasilkan return tertinggi di setiap harinya. dapat dilihat hasil visualisasi sebagai berikut:

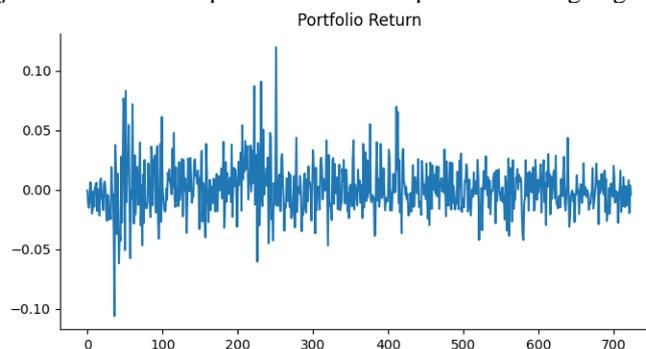


Gambar 10 : Visualisasi aktual dan prediksi *return* prediksi tertinggi per *cluster*

Visualisasi diatas menunjukkan *Cluster 3* memiliki kesesuaian terbaik antara *return* aktual dan prediksi, menunjukkan efektivitas model. Sementara itu, *Cluster 2* dan *4* memperlihatkan perbedaan signifikan, menunjukkan kebutuhan untuk penyesuaian model karena volatilitas tinggi. *Cluster 1* menunjukkan keseimbangan yang stabil antara return prediksi dan *actual*.

4.2 Analisis Hasil Pengujian

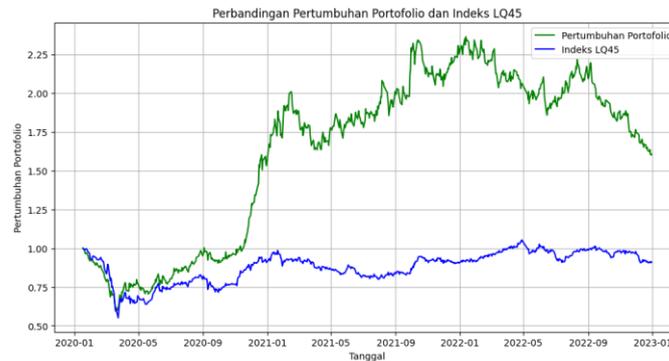
Dari hasil pengujian tersebut mendapatkan hasil return portofolio dengan gambar 11 dibawah ini :



Gambar 11 : Visualisasi *portfolio return*

Grafik di atas menampilkan *return* portofolio dari waktu ke waktu. Terlihat bahwa *return* portofolio berfluktuasi secara signifikan, dengan periode keuntungan dan kerugian yang bergantian. Terdapat beberapa lonjakan dan penurunan tajam, mengindikasikan volatilitas yang tinggi. Secara keseluruhan, *return* portofolio tampak beresilasi di sekitar nol, menunjukkan bahwa tidak ada tren kenaikan atau penurunan yang jelas dalam jangka panjang. Meskipun terdapat fluktuasi yang cukup besar, portofolio tampaknya mampu pulih dari periode kerugian dan menghasilkan keuntungan positif pada periode lainnya. Hal ini menunjukkan adanya diversifikasi yang baik dalam portofolio, dengan rata-rata *return* portofolio 0,0009 dan standar deviasi 0,0209. Standar deviasi yang tinggi ini mengindikasikan bahwa portofolio ini memiliki risiko yang relatif tinggi, sehingga investor perlu mempertimbangkan toleransi risiko mereka sebelum berinvestasi.

4.3 Perbandingan Pertumbuhan Portofolio dengan Harga Saham Historis



Gambar 12 : Visualisasi perbandingan pertumbuhan portofolio dan indeks LQ45

Dalam grafik diatas, pertumbuhan portofolio (garis hijau) dibandingkan dengan pertumbuhan Indeks LQ45 (garis biru) dari awal tahun 2020 hingga awal 2023. Terlihat bahwa portofolio jauh lebih besar daripada Indeks LQ45 selama sebagian besar waktu tersebut. Portofolio menunjukkan pertumbuhan yang kuat sejak awal dengan rata-rata *return* 0,0009 dan standar deviasi 0,0209, melampaui Indeks LQ45 secara konsisten hingga pertengahan tahun 2021. Meskipun keduanya mengalami penurunan pada akhir 2021 dan sepanjang 2022, portofolio tetap di atas Indeks LQ45 dengan rata-rata *return* -0,0000 dan standar deviasi 0,0158. Pada awal 2023, terlihat sedikit pemulihan pada kedua investasi, tetapi portofolio tetap unggul.

Secara keseluruhan, grafik menunjukkan bahwa strategi investasi portofolio ini lebih efektif daripada hanya mengikuti Indeks LQ45. Ini menunjukkan potensi keuntungan dari pemilihan saham yang aktif dan manajemen portofolio yang baik.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Penelitian ini menunjukkan bahwa dengan mengklasterkan dapat menunjukkan hasil yang masih sangat bagus, dengan memprediksikan *return* saham harian dapat menghasilkan RMSE Cluster 1 dan 3 yang menunjukkan akurasi cukup baik dengan RMSE masing-masing 0,09 dan 0,07, menandakan model mampu menangkap pola data dengan baik. Cluster 2 memiliki akurasi terbaik RMSE 0,05, karena pola data yang konsisten dan model yang optimal. Sebaliknya, cluster 4 memiliki RMSE tertinggi 0,57, menunjukkan kesulitan model dalam memprediksi data yang disebabkan oleh ketidakcocokan model, data yang bervariasi, atau adanya outlier. Model ini berhasil mengidentifikasi saham-saham potensial dengan *return* tinggi, berkontribusi pada pembentukan portofolio yang menguntungkan.

Portofolio yang dibentuk berdasarkan prediksi model ini berhasil mengungguli Indeks LQ45 secara signifikan selama periode pengujian. Meskipun memiliki volatilitas yang tinggi, portofolio ini mencatatkan kinerja yang lebih baik dengan rata-rata *return* 0,0009 dan standar deviasi 0,0209, dibandingkan Indeks LQ45 yang memiliki rata-rata *return* -0,0000 dan standar deviasi 0,0158. Keunggulan ini terlihat jelas pada awal tahun 2021 dan bertahan hingga awal 2023, meskipun terjadi penurunan pada akhir 2021 dan sepanjang 2022. Secara keseluruhan, penelitian ini memberikan bukti kuat bahwa pendekatan clustering dan CNN dapat menjadi sistem yang efektif dalam prediksi *return* saham dan pembentukan portofolio investasi.

5.2 Saran

Hasil penelitian ini menghasilkan beberapa saran untuk pengembangan lebih lanjut. Pertama, perluasan dataset dengan memasukkan data saham dari berbagai indeks dan periode waktu yang lebih panjang, meningkatkan generalisasi model dan memberikan hasil yang lebih komprehensif. Kedua, eksplorasi metode *clustering* alternatif selain *K-Means*, seperti *hierarchical clustering* atau *density-based clustering*, dapat memberikan wawasan baru mengenai pengelompokan saham berdasarkan karakteristiknya.

Daftar Pustaka

- [1] Ashrafzadeh, M., Taheri, H. M., Gharehgozlou, M., & Zolfani, S. H. (2023). Clustering-based return prediction model for stock pre-selection in portfolio optimization using PSO-CNN+ MVF. *Journal of King Saud University-Computer and Information Sciences*, 35(9), 101737.
- [2] Fang, Z., & Chiao, C. (2021). Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization. *Journal of Computational Methods in Sciences and Engineering*, 21(5), 1081-1089.
- [3] Chusyairi, A., & Saputra, P. R. N. (2019). Pengelompokan Data Puskesmas Banyuwangi Dalam Pemberian Imunisasi Menggunakan Metode K-Means Clustering. *Telematika*, 12(2), 139-148.
- [4] Chen, S., & He, H. (2018, November). Stock prediction using convolutional neural networks. In *IOP Conference series: materials science and engineering* (Vol. 435, p. 012026). IOP Publishing.
- [5] Shawel, B. S., Bantigegn, F., Debella, T. T., Pollin, S., & Woldegebreal, D. H. (2022). K-Means Clustering Assisted Spectrum Utilization Prediction with Deep Learning Models. *Engineering Proceedings*, 18(1), 2.
- [6] Kassambara, A. (2023). Practical guide to cluster analysis in R: Unsupervised machine learning. STHDA.
- [7] Wu, J. M. T., Li, Z., Herencsar, N., Vo, B., & Lin, J. C. W. (2021). A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. *Multimedia Systems*, 1-20.
- [8] Yuliya Plyakha, Raman Uppal, Grigory Vilkov (2020). Why Does an Equal-Weighted Portfolio Outperform Value- and Price-Weighted Portfolios? EDHEC Risk Climate Impact Institute.
- [9] Tsantekidis, A., Passalis, N., Tefas, A., Kannianen, J., Gabbouj, M., & Iosifidis, A. (2020). Forecasting stock prices from the limit order book using convolutional neural networks. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 7928-7935). IEEE
- [10] Zhang, Z., Zohren, S., & Roberts, S. (2022). Deep learning for financial time series forecasting: A survey. *Journal of Financial Data Science*, 4(1), 70-91.