BAB II.
LANDASAN TEORI

Bab II Landasan Teori menyajikan kerangka konseptual yang mencakup teori, konsep, serta hasil penelitian terdahulu yang relevan sebagai dasar dalam menganalisis hubungan antara data sentimen pasar dan data historis saham pada bank digital. Pembahasan dalam bab ini disusun ke dalam tiga bagian utama, yaitu tinjauan pustaka, keaslian penelitian, dan dasar teori yang mendukung pendekatan penelitian yang digunakan.

Penjelasan dimulai dengan tinjauan pustaka terhadap penelitian-penelitian relevan yang dikelompokkan ke dalam beberapa tema utama sesuai dengan fokus kajian. Selanjutnya, dilakukan analisis keaslian penelitian melalui perbandingan terhadap studi-studi sebelumnya untuk menunjukkan kontribusi dan kebaruan penelitian ini. Bagian terakhir menyajikan dasar-dasar teori yang menjadi landasan konseptual dalam merancang pendekatan integratif berbasis *machine learning* untuk prediksi harga saham bank digital.

2.1 Tinjauan Pustaka

Penelitian terkait prediksi harga saham berbasis *machine learning* telah mengalami perkembangan yang signifikan dalam beberapa tahun terakhir. Studi-studi tersebut menerapkan pendekatan algoritmik untuk menangkap pola kompleks yang terkandung dalam data historis saham maupun informasi eksternal seperti sentimen pasar. Tinjauan pustaka ini diklasifikasikan ke dalam empat kelompok utama yang relevan dengan fokus penelitian, yaitu: (1) prediksi harga saham menggunakan *machine learning*, (2) integrasi data historis saham dan analisis sentimen, (3) studi mengenai bank digital atau sektor keuangan digital, dan (4) perbandingan algoritma *machine learning* dalam prediksi harga saham.

Dalam rangka membangun landasan konseptual yang kokoh bagi penelitian ini, penting untuk mengkaji secara komprehensif berbagai studi terdahulu yang relevan. Tinjauan pustaka berikut disusun untuk mengidentifikasi dan mengkritisi

literatur yang berkaitan dengan prediksi harga saham berbasis *machine learning*, integrasi data historis dan sentimen pasar, serta penerapan pendekatan ini dalam konteks bank digital. Dengan memahami perkembangan metodologis, pendekatan yang digunakan, serta kelemahan dan keterbatasan dalam studi sebelumnya, penelitian ini bertujuan untuk mengidentifikasi celah yang masih terbuka dan merumuskan kontribusi yang berarti. Kajian ini diawali dengan pembahasan mengenai penerapan *machine learning* dalam prediksi harga saham.

2.1.1. Prediksi Harga Saham Menggunakan Machine Learning

Berbagai pendekatan *machine learning* telah digunakan untuk memprediksi harga saham, baik dengan algoritma tradisional maupun model deep learning. Penelitian yang dilakukan oleh (Deevenapalli dkk., 2023) merumuskan strategi prediksi harga saham yang akurat dengan mempertimbangkan kompleksitas faktor-faktor yang memengaruhi pergerakan pasar. Studi ini mengimplementasikan empat algoritma *machine learning* yang umum digunakan, yaitu *K-Nearest Neighbors* (KNN), *Naive Bayes, Support Vector Machine* (SVM), dan *Random Forest*, dalam proses analisis dan peramalan harga saham di berbagai kondisi pasar. Evaluasi kinerja model dilakukan berdasarkan metrik akurasi, presisi, dan *recall* menggunakan dataset harga saham yang relevan. Hasil kajian ini menunjukkan bahwa penggunaan beragam algoritma *machine learning* dalam skenario yang bervariasi dapat memberikan pendekatan yang lebih tangguh dan andal dalam proses prediksi harga saham. Pendekatan ini menawarkan kontribusi penting dalam mengatasi tantangan prediksi pasar yang bersifat dinamis, serta memperkuat dasar metodologis bagi pengembangan model prediksi multimetode di masa mendatang.

Selanjutnya, penelitian yang dilakukan oleh (Hu dkk., 2023) menunjukkan bahwa model *machine learning* mampu memprediksi harga saham yang kompleks dengan rasio sinyal terhadap *noise* yang rendah melalui pemodelan data nonlinier yang bersifat *fuzzy*. Berbeda dari studi sebelumnya yang cenderung fokus pada peningkatan satu algoritma, penelitian ini membandingkan kinerja tiga model *LightGBM, Random Forest*, dan *Logistic Regression* dalam klasifikasi naik turun harga saham pada berbagai jendela waktu. Dengan menggunakan indikator teknikal

seperti *Relative Strength Index* (RSI) dan *Simple Moving Average (*SMA), hasilnya menunjukkan perbedaan signifikan antara prediksi jangka pendek dan jangka panjang. Temuan ini memberi panduan penting bagi investor dalam memilih algoritma yang sesuai untuk strategi investasi berdasarkan jangka waktu prediksi.

(Koratamaddi dkk., 2021) meneliti tantangan dalam pemodelan alokasi portofolio saham, yang bertujuan menemukan strategi investasi optimal guna memaksimalkan imbal hasil dan meminimalkan risiko. Pendekatan *deep reinforcement learning* digunakan untuk melatih agen cerdas berbasis data historis harga saham. Namun, pendekatan sebelumnya umumnya belum mempertimbangkan sentimen pasar, padahal sentimen yang terbentuk melalui media sosial dan berita daring terbukti berpengaruh terhadap keputusan investor. Menanggapi hal ini, penelitian tersebut mengusulkan model baru berbasis *deep reinforcement learning* yang tidak hanya memanfaatkan data historis, tetapi juga mengintegrasikan sentimen pasar dalam pengelolaan portofolio saham perusahaan anggota Dow Jones. Hasilnya menunjukkan bahwa pendekatan ini lebih unggul dibanding metode konvensional, ditinjau dari metrik seperti *Sharpe ratio* dan keuntungan tahunan investasi.

Secara umum, masih terdapat masalah yang belum terpecahkan dalam kelompok ini, yaitu belum adanya pendekatan yang secara efektif menggabungkan data frekuensi tinggi, indikator teknikal, dan sentimen pasar dalam satu model prediktif yang adaptif terhadap perubahan pola pasar.

2.1.2. Integrasi Data Historis Saham dan Analisis Sentimen

Integrasi antara data historis saham dan analisis sentimen telah menjadi pendekatan yang menjanjikan dalam meningkatkan akurasi prediksi harga saham. (Polireddi, 2024) mengembangkan model *deep reinforcement learning* yang menggabungkan data historis dengan sentimen dari media sosial dan berita keuangan. Pendekatan ini menunjukkan bahwa integrasi sentimen dapat meningkatkan akurasi prediksi portofolio saham. Namun, studi ini tidak difokuskan pada saham individual dan belum mengevaluasi efektivitas berbagai sumber sentimen secara komparatif.

Selain itu, dinamika temporal dari sentimen jangka pendek terhadap jangka panjang belum dianalisis secara mendalam.

Masalah yang belum terpecahkan dalam konteks ini adalah belum adanya studi yang menguji integrasi data sentimen *real-time* dengan data frekuensi tinggi untuk prediksi saham individual, serta belum adanya eksplorasi terhadap kualitas dan pengaruh relatif dari berbagai sumber sentimen seperti Twitter, Reddit, atau forum investor.

2.1.3. Studi pada Bank Digital dan Sektor Keuangan Digital

Studi tentang penerapan kecerdasan buatan (AI) dan *machine learning* dalam sektor keuangan digital menunjukkan potensi yang signifikan dalam meningkatkan efisiensi operasional dan akurasi pengambilan keputusan.. (De Oliveira Silva dkk., 2023) menyoroti peran penting kecerdasan buatan (AI) dan *machine learning* dalam sektor perbankan digital, termasuk dalam deteksi penipuan, pemrosesan pinjaman otomatis, dan personalisasi layanan. Studi ini menekankan bahwa kecerdasan buatan (AI) dapat meningkatkan efisiensi operasional dan keamanan sistem keuangan. Namun, penelitian ini bersifat konseptual dan tidak menyajikan model prediktif atau eksperimen empiris yang relevan dengan prediksi harga saham. Selain itu, belum ada eksplorasi terhadap bagaimana teknologi kecerdasan buatan (AI) yang digunakan di sektor perbankan dapat diadaptasi untuk memodelkan pergerakan harga saham, khususnya di sektor keuangan digital.

Dengan demikian, masih terdapat celah riset dalam menjembatani teknologi kecerdasan buatan (AI) di sektor perbankan dengan aplikasinya dalam prediksi saham sektor keuangan digital, terutama dalam konteks integrasi data pasar dan pendekatan prediktif berbasis machine learning.

2.1.4. Perbandingan Algoritma Machine Learning untuk Prediksi Saham

Perbandingan algoritma *machine learning* dalam konteks prediksi harga saham telah dilakukan oleh beberapa peneliti. (Bhandari dkk., 2022) membandingkan performa *Logistic Regression, Random Forest*, dan *LightGBM* dalam memprediksi arah pergerakan harga saham berdasarkan indikator teknikal. Hasilnya menunjukkan bahwa *LightGBM* unggul dalam prediksi jangka panjang, sementara *Random Forest* lebih baik untuk jangka pendek. Namun, studi ini hanya

menggunakan data teknikal dan tidak mempertimbangkan data sentimen maupun makroekonomi, serta tidak menguji performa model dalam kondisi pasar ekstrem.

(Lanbouri dan Achchab, 2020) juga menunjukkan bahwa model tradisional dapat mengungguli model *deep learning* apabila dilatih secara spesifik untuk masing-masing saham. Namun, pendekatan semacam ini tidak bersifat diterapkan secara luas dan belum mampu mengakomodasi dinamika perubahan pola data secara *real-time*. Di samping itu, hingga kini belum terdapat pendekatan komparatif yang secara holistik mengintegrasikan berbagai jenis data seperti data historis, teknikal, dan sentimen serta menguji ketahanan model terhadap perubahan konsep data seiring waktu dan dinamika pasar yang berlangsung cepat.

Dengan demikian, meskipun telah terjadi berbagai kemajuan dalam pengembangan model prediksi harga saham, masih terdapat sejumlah permasalahan yang belum terselesaikan. Beberapa di antaranya meliputi keterbatasan dalam mengintegrasikan berbagai jenis data seperti data frekuensi tinggi, sentimen pasar, dan indikator teknikal kurangnya kemampuan adaptasi terhadap perubahan pasar secara *real-time*, serta minimnya studi yang secara khusus menguji efektivitas model pada saham individual dalam kondisi volatilitas tinggi. Oleh karena itu, masih terbuka peluang riset yang signifikan untuk merancang model prediksi harga saham yang lebih adaptif, komprehensif, dan responsif terhadap dinamika pasar yang kompleks dan cepat berubah

Berdasarkan tinjauan terhadap berbagai studi sebelumnya, dapat disimpulkan bahwa meskipun telah banyak dilakukan penelitian mengenai prediksi harga saham dengan pendekatan *machine learning*, integrasi data historis dan sentimen, serta penerapan teknologi kecerdasan buatan (AI) dalam sektor keuangan digital, masih terdapat sejumlah keterbatasan dan permasalahan yang belum sepenuhnya teratasi. Beberapa studi memang menunjukkan hasil yang menjanjikan, namun umumnya masih terbatas pada jenis data tertentu, rentang waktu yang sempit, atau belum mampu mengakomodasi dinamika pasar yang berubah secara cepat. Di samping itu, masih sedikit penelitian yang secara komprehensif menggabungkan data frekuensi tinggi, indikator teknikal, dan sentimen pasar dalam satu kerangka prediktif yang adaptif dan responsif. Oleh karena itu, diperlukan

pendekatan baru yang tidak hanya mampu menjawab tantangan-tantangan tersebut, tetapi juga mengisi celah riset yang ada, guna mengembangkan model prediksi harga saham yang lebih akurat, fleksibel, dan aplikatif dalam konteks pasar keuangan modern yang semakin kompleks dan dinamis

Sebagai langkah selanjutnya, akan disajikan tabel keaslian penelitian untuk menegaskan orisinalitas pendekatan yang ditawarkan dalam studi ini, serta memperjelas kontribusinya dalam membedakan diri dari penelitian-penelitian sebelumnya.

2.2 Keaslian Penelitian

Keaslian (originalitas) merupakan unsur esensial dalam penelitian doktoral karena mencerminkan kontribusi nyata terhadap pengembangan ilmu pengetahuan, khususnya di bidang sistem informasi. Keaslian tidak hanya ditentukan oleh kebaruan data atau metode, tetapi juga oleh pendekatan konseptual, integrasi teori, konteks penelitian, serta pengembangan model yang memberikan nilai tambah secara akademik dan praktis.

Penelitian ini bertujuan untuk menganalisis korelasi antara data historis saham dan sentimen media sosial pada perusahaan bank digital, dengan *machine learning* digunakan sebagai alat analisis pola pergerakan harga saham. Pendekatan ini memiliki unsur kebaruan karena penelitian terdahulu umumnya masih menggunakan satu jenis data secara terpisah serta belum secara khusus mengkaji saham bank digital di pasar negara berkembang seperti Indonesia.

Untuk menegaskan orisinalitas penelitian ini, dilakukan analisis komparatif terhadap 26 studi terdahulu yang relevan, terdiri atas artikel jurnal, makalah konferensi/prosiding, dan *book chapter* yang telah terindeks dalam basis data Scopus. Analisis dilakukan berdasarkan aspek judul, peneliti, tahun publikasi, jenis artikel, peringkat kuartil jurnal, variabel yang digunakan, metode yang diterapkan, serta hasil utama yang diperoleh. Penyajian ini bertujuan untuk memetakan posisi penelitian dalam lanskap keilmuan yang ada, sekaligus menyoroti kontribusi khas yang ditawarkan melalui pendekatan yang diusulkan dalam disertasi ini.

Tabel 2. 1 Analisis Penelitian Terdahulu untuk Menunjukkan Keaslian Gagasan.

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
1.	Measuring Investor Sentiment of China's Growth Enterprises Market with ERNIE	Junxiao Gui, Jichun Pu, Nathee Naktnasukanjn, Xi Yu, Lei Mu, Heping Pan	2022	Conference Paper (The 2022 International Conference on Business and Information)	-	Sentimen investor, indeks GEM	ERNIE, semi-covariance, regresi	Sentimen investor berpengaruh signifikan terhadap return indeks GEM
2.	S_I_LSTM: Stock Price Prediction Based on Multiple Data Sources and Sentiment Analysis	Shengting Wu, Yuling Liu, Ziran Zou, Tien-Hsiung Weng	2022	Jurnal (Connection Science)	Q1	Data historis, indikator teknikal, berita keuangan, forum saham	CNN untuk sentimen, LSTM dengan attention	Data multi-sumber meningkatkan akurasi prediksi; MAE lebih rendah dibanding metode lain
3.	Stock Price Forecasting with Deep Learning: A Comparative Study	Tej Bahadur Shahi, Ashish Shrestha, Arjun Neupane, William Guo	2020	Jurnal (Mathematics)	Q1	Data historis, sentimen berita keuangan	LSTM, GRU, analisis sentimen (VADER)	Sentimen berita meningkatkan akurasi prediksi; LSTM dan GRU sama-sama efektif secara dinamis

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
4.	Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor	Junaid Maqbool, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, Ishfaq Ali Ganaie	2023	Conference Paper (Procedia Computer Science)	-	Data harga saham historis, sentimen berita keuangan (VADER, TextBlob, FLAIR), label relevansi	MLP Regressor + kombinasi skor sentimen	Akurasi prediksi tren hingga 90% untuk 10 hari; FLAIR paling efektif; Tata Motors paling sulit diprediksi
5.	Artificial Intelligence in Innovation Research: A Systematic Review, Conceptual Framework, and Future Research Directions	Marcello M. Mariani, Isa Machado, Vittoria Magrelli, Yogesh K. Dwivedi	2023	Jurnal (Technovation)	Q1	Adopsi AI, inovasi, big data, IoT, digital platforms, sustainability, efisiensi, kapabilitas organisasi	Systematic Literature Review + Bibliometric Analysis	AI mendorong inovasi melalui efisiensi, keunggulan kompetitif, dan pengembangan teknologi baru
6.	Integrating Social Media Data and Historical Stock Prices for Predictive Analysis: A Reinforcement	Mei Li, Ye Zhang	2023	Jurnal (International Journal of Advanced Computer Science and Applications)	-	Sentimen media sosial, berita keuangan, harga saham historis	Sentiment Analysis (BERT + RL + DE) + Attention-based LSTM	Model mengurangi error prediksi hingga 14%; RL meningkatkan akurasi klasifikasi minor class secara signifikan

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
	Learning Approach							
7.	The Impact of Sentiment and Attention Measures on Stock Market Volatility	Francesco Audrino, Fabio Sigrist, Daniele Ballinari	2020	Jurnal (International Journal of Forecasting)	Q1	Sentimen media sosial, volume pencarian Google, volume pesan StockTwits, VIX, turnover ratio	Adaptive Lasso Regression + HAR Model + Kalman Filter + Deep-MLSA Sentiment Analysis	Variabel atensi (Google Trends, StockTwits) signifikan meningkatkan akurasi prediksi volatilitas jangka pendek
8.	Optimization of Investment Strategies through Machine Learning	Jiaqi Li, Xiaoyan Wang, Saleem Ahmad, Xiaobing Huang, Yousaf Ali Khan	2023	Jurnal (Heliyon)	Q2	EVA, PCA, MACD, KDJ, LSTM, ANN, SVR, indikator keuangan dan teknikal	EVA + PCA untuk seleksi saham, LSTM/ANN/SVR untuk prediksi harga, MACD/KDJ untuk sinyal jual	LSTM dengan EVA menghasilkan return tahunan tertinggi (27.165%), mengungguli benchmark 9.67%
9.	Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives	Sara Oleiro Araújo, Ricardo Silva Peres, José Cochicho Ramalho, Fernando	2023	Jurnal (Agronomy)	Q1	ML, RF, SVM, CNN, LSTM, crop/water/soil/animal management, PRISMA	Systematic Literature Review (PRISMA), analisis tren dan tantangan ML di sektor pertanian	RF dan SVM paling banyak digunakan; ML meningkatkan efisiensi pertanian, namun tantangan data dan adopsi masih signifikan

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
		Lidon, José Barata						
10.	Artificial Intelligence Techniques in Financial Trading: A Systematic Literature Review	Fatima Dakalbab, Manar Abu Talib, Qassim Nasir, Tracy Saroufil	2024	Jurnal (Journal of King Saud University - Computer and Information Sciences)	Q1	AI, ML, DL, RL, DRL, RSI, MACD, S&P 500, EUR/USD, BTC, Sharpe Ratio, RMSE, Yahoo Finance	Systematic Literature Review (143 artikel, 2015–2023), analisis pasar, teknik AI, metrik evaluasi	Deep learning paling dominan (30%), hanya 16% model yang otomatis, RSI indikator teknikal paling umum, RMSE dan Sharpe Ratio metrik evaluasi utama
11.	The Use of Predictive Analytics in Finance	Daniel Broby	2022	Jurnal (The Journal of Finance and Data Science)	Q1	Time series, regression, classification, clustering, ANN, SVM, GARCH, credit scoring, fraud detection	Systematic Literature Review (SPAR-4-SLR), eksplorasi metode statistik dan komputasional	Prediksi berbasis data internal dan eksternal digunakan untuk forecasting ekonomi, harga saham, risiko, dan perilaku pelanggan
12.	A Hybrid Model Combined Deep Learning Approaches in Stock Price Prediction	Zidong Huang, Yiming Lin, Haoran Xue	2022	Konferensi (IEEE International Conference on Electronic Technology, Communication	-	LSTM, SVM, EMD, CNN, investor sentiment, financial news, SSE Index, Citi Group stock	Review dan eksperimen model hybrid: LSTM, SVM, EMD, CNN, HTM, sentiment analysis	Model hybrid LSTM+CNN+EMD dan LSTM+sentiment analysis menunjukkan akurasi tinggi; tantangan tetap pada

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
				and Information)				generalisasi lintas negara dan integrasi data sosial
13.	Prediction of Stock Market Using Sentiment Analysis and Ensemble Learning	Archana Y. Chaudhari, Smita Mahajan	2025	Jurnal (MethodsX)	Q1	DRL (A2C, PPO2, SAC), PACTRO, RSI, MACD, BB, ROC, sentiment analysis, Yahoo Finance data	Deep Reinforcement Learning + Sentiment Analysis + Technical Indicators (PACTRO algorithm)	Model PACTRO meningkatkan akurasi prediksi dengan menggabungkan indikator teknikal dan sentimen; validasi dilakukan pada saham AAPL
14.	Artificial Intelligence and Numerical Weather Prediction Models: A Technical Survey	Muhammad Waqas, Usa W. Humphries, Bunthid Chueasa, Angkool Wangwongchai	2024	Jurnal (Natural Hazards Research)	Q1	AI, ML, DL, NWP, WRF, CNN, LSTM, GRU, GAN, uncertainty quantification, extreme weather	Systematic Literature Review (2000–2024), integrasi AI dengan NWP di berbagai tahap	AI meningkatkan akurasi prediksi cuaca, terutama dalam post-processing dan prediksi ekstrem, namun belum menggantikan NWP sepenuhnya; tantangan utama: interpretabilitas, ketergantungan data, dan prediksi cuaca ekstrem

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
15.	Using Financial News Sentiment for Stock Price Direction Prediction	Bledar Fazlija, Pedro Harder	2022	Jurnal (Mathematics, MDPI)	Q1	FinBERT, sentiment score (title & content), S&P 500, random forest, Brier score, MCC	FinBERT fine-tuned untuk klasifikasi sentimen, digunakan dalam strategi sederhana dan random forest untuk prediksi arah harga	Sentimen dari konten berita lebih efektif daripada judul; model random forest berbasis konten mengungguli random walk dan strategi sederhana
16.	Integrating AI Techniques for Enhanced Financial Forecasting and Budgeting Strategies	Vineet Jain, Parth A. Kulkarni	2023	Jurnal (SSRG International Journal of Economics and Management Studies)	-	AI, ML, LSTM, Random Forest, AutoML, NLP, BNN, RL, FP&A, budgeting, forecasting	Studi literatur dan studi kasus; eksplorasi model AI seperti LSTM, ensemble, NLP, AutoML, BNN, RL	AI meningkatkan akurasi prediksi, efisiensi alokasi anggaran, dan kualitas analisis varians; studi kasus dari Amex, HSBC, JP Morgan menunjukkan dampak nyata
17.	Economic Forecasting with Big Data: A Literature Review	Wencan Lin, Yunjie Wei	2024	Jurnal (Journal of Management Science and Engineering)	Q1	Big data, economic forecasting, bibliometric analysis, structural variation analysis, SCM, ML, NLP	Bibliometric analysis, co-citation, keyword evolution, structural variation analysis (SVA)	Big data meningkatkan akurasi dan ketepatan waktu prediksi ekonomi; SCM dan machine learning menjadi

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
								tema dominan; identifikasi artikel dengan potensi transformasional tinggi
18.	Stock Price Prediction using Technical Indicators: A Predictive Model using Optimal Deep Learning	Manish Agrawal, Asif Ullah Khan, Piyush Kumar Shukla	2019	Jurnal (International Journal of Recent Technology and Engineering)	-	LSTM, STIs (RSI, MACD, MFI, CCI, CMF, ADX), OHLC, correlation tensor	Optimal LSTM dengan tensor korelasi adaptif STIs	Akurasi prediksi tertinggi 65.64%, rata-rata 59.25%; lebih baik dari SVM, LR, dan ELSTM
19.	Predictive Analytics in Customer Behavior: Anticipating Trends and Preferences	Hamed GhorbanTanhaei et al.	2024	Jurnal (Results in Control and Optimization)	Q1	Customer behavior, RF, LR, SVM, Gradient Boosting, DT, precision, recall, F1, ROC-AUC	Evaluasi komparatif model ML untuk prediksi perilaku pelanggan	RF dan LR menunjukkan performa terbaik; LR unggul dalam recall (1.0), RF seimbang dalam semua metrik
20.	Machine Learning Sentiment Analysis, COVID-19 News and Stock	Michele Costola, Oliver Hinz, Michael Nofer, Loriana Pelizzon	2023	Jurnal (Research in International Business and Finance)	Q1	COVID-19 news, FinBERT, sentiment score, S&P 500, volatility, trading volume	FinBERT untuk analisis sentimen berita COVID-19 dari MarketWatch, NYTimes, Reuters	Sentimen positif berkorelasi dengan return S&P 500; berita bisnis NYTimes paling

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
	Market Reactions							berpengaruh terhadap return pasar
21.	Social Media and Stock Market Prediction: A Big Data Approach	Mazhar Javed Awan et al.	2021	Jurnal (Computers, Materials & Continua)	Q2	Historical stock data, social media sentiment, LR, GLR, RF, DT, NB, Logistic Regression	Spark MLlib (PySpark), Databricks, sentiment analysis, structured & unstructured data	GLR dan LR akurat untuk data historis (97% dan 95%); NB dan Logistic Regression akurat untuk data sentimen (80% dan 77%)
22.	A Comparative Study of the Stock Market using Machine Learning Algorithms	Karthik Deevenapalli et al.	2023	Prosiding (International Conference on Electronics and Renewable Systems)	-	KNN, SVM, Naive Bayes, Random Forest, OHLC, volume, adj close, Kaggle dataset	Evaluasi akurasi, precision, recall pada data historis saham (2013–2019)	Random Forest dan SVM unggul dalam akurasi; pendekatan multi-model meningkatkan prediksi saham
23.	Comparison of Stock Price Prediction Based on Different Machine Learning Approaches	Qianqiao Hu, Songshan Qin, Shuai Zhang	2023	Book Chapter (Springer)	-	RSI, SMA, MACD, OBV, CR, VR, DMI, Logistic Regression, Random Forest, LightGBM	Rolling window, evaluasi akurasi, precision, recall, F1 pada 3 saham Shanghai Stock Exchange	LightGBM unggul untuk prediksi jangka panjang; Random Forest lebih baik untuk jangka pendek; Logistic Regression kurang akurat
24.	An Effective Role of Artificial Intelligence and Machine	Naga Simhadri Apparao Polireddi	2024	Jurnal (Measurement: Sensors)	Q2	AI, ML, fraud detection, credit scoring, customer service, risk prediction	Studi literatur, analisis korelasi dan regresi, use case AI/ML di sektor keuangan	AI/ML meningkatkan efisiensi, deteksi penipuan, layanan pelanggan, dan pengambilan

No	Judul	Peneliti	Tahun	Jenis Artikel	Quartile (Q)	Variabel	Metode	Hasil
	Learning in Banking Sector							keputusan di sektor perbankan
25.	Stock Market Prediction on High Frequency Data using Long-Short Term Memory	Zineb Lanbouri, Said Achchab	2020	Prosiding (Procedia Computer Science)	Q3	S&P500 intraday data, LSTM, EMA, MACD, Bollinger Bands, OHLC, volume	LSTM untuk prediksi 1, 5, 10 menit ke depan; evaluasi RMSE dengan/ tanpa indikator teknikal	LSTM efektif untuk prediksi jangka pendek; tanpa indikator teknikal hasilnya lebih baik dalam kasus Amazon
26.	Stock Price Movement Prediction based on Optimized Traditional Machine Learning Models	José Junior de Oliveira Silva, Roberto Souto Maior de Barros, Silas Garrido Teixeira de Carvalho Santos	2023	Prosiding (IEEE Symposium Series on Computational Intelligence)	-	OHLC, volume, moving averages (5, 10, 15 hari), % perubahan harga	Model per saham, optimasi hyperparameter, LR, SVM, NB, KNN, evaluasi MCC, AUC, F1	Logistic Regression mengungguli model deep learning (StockNet, Adv-ALSTM) dalam akurasi dan MCC

Secara *state of the art*, penelitian ini menempati posisi yang unik dan menonjol dalam ranah literatur sistem informasi dan ilmu data, dengan karakteristik akademik yang khas serta menampilkan kebaruan ilmiah yang signifikan. Pertama, penelitian ini mengusung pendekatan integratif dengan menggabungkan data numerik berupa data historis harga saham dan data tekstual berupa analisis sentimen dari media sosial ke dalam satu kerangka kerja prediktif yang komprehensif dan adaptif. Pendekatan tersebut memungkinkan pemodelan yang lebih menyeluruh terhadap dinamika pasar, mencakup aspek teknis maupun aspek emosional yang memengaruhi pergerakan harga saham.

Kedua, penelitian ini dilakukan dalam konteks sektoral yang spesifik, yakni pada institusi perbankan digital di Indonesia. Fokus ini merepresentasikan kontribusi orisinal mengingat masih terbatasnya kajian ilmiah yang secara eksplisit mengeksplorasi prediksi harga saham berbasis pembelajaran mesin pada sektor bank digital di Indonesia. Dengan demikian, penelitian ini mengisi kekosongan literatur yang relevan dengan kebutuhan kontekstual serta perkembangan ekonomi digital nasional.

Ketiga, penelitian ini berlandaskan paradigma sistem informasi sebagai sistem pendukung keputusan, yang memandang sistem informasi tidak semata sebagai entitas teknologi, melainkan sebagai kesatuan yang mengintegrasikan unsur manusia, teknologi, data, dan proses secara sinergis. Pendekatan ini diarahkan untuk menghasilkan pengetahuan prediktif yang dapat dimanfaatkan secara strategis oleh para pemangku kepentingan dalam proses pengambilan keputusan yang berbasis data dan teknologi informasi.

Keempat, dari segi kontribusi, penelitian ini memberikan sumbangan teoretis dan praktis secara simultan. Kontribusi teoretis terlihat melalui pengembangan model prediksi berbasis algoritma *machine learning* dalam kerangka sistem informasi, sementara kontribusi praktis diwujudkan dalam bentuk rekomendasi implementatif yang dapat diaplikasikan di sektor keuangan digital, khususnya untuk meningkatkan kualitas pengambilan keputusan investasi di era transformasi digital.

Dengan demikian, penelitian ini menunjukkan orisinalitas yang kuat baik dari sisi pendekatan metodologis, integrasi variabel, konteks penerapan, maupun kontribusi terhadap penguatan landasan keilmuan sistem informasi. Hal tersebut menjadikan penelitian ini relevan dan signifikan dalam menjawab tantangan akademik maupun praktis di era ekonomi digital yang terus berkembang.

Sebagai landasan yang kuat dalam pengembangan model prediksi harga saham, perlu dipahami terlebih dahulu konsep-konsep dasar yang mendasari teknologi dan teori yang digunakan. Oleh karena itu, pada sub bab berikutnya akan dibahas secara mendetail tentang Dasar Teori yang meliputi konsep *machine learning*, algoritma-algoritma populer, analisis sentimen, teori pasar efisien, serta evaluasi kinerja model prediksi. Pemahaman mendalam terhadap teori-teori ini menjadi fondasi utama dalam merancang dan mengimplementasikan model prediksi harga saham yang akurat dan efektif.

2.3 Dasar Teori

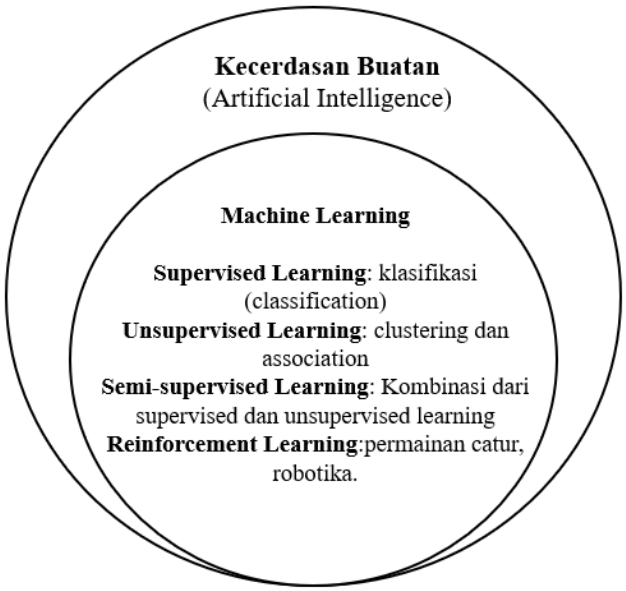
Pada subbab dasar teori, akan dijelaskan konsep-konsep dasar yang menjadi dasar teori dalam penelitian ini. Penjelasan dimulai dengan pengertian dan komponen utama dari *machine learning* sebagai teknologi inti dalam pengembangan model prediksi harga saham. Selanjutnya, akan dibahas berbagai algoritma *machine learning* populer yang relevan dengan prediksi harga saham, termasuk regresi linier, *random forest*, dan *neural networks*. Selain itu, dijelaskan pula konsep analisis sentimen sebagai pendekatan untuk menggali informasi subjektif dari data teks yang berpengaruh pada pasar saham. Kemudian, akan dibahas teori pasar efisien yang menjadi kerangka pemahaman mengenai perilaku harga saham dalam pasar modal. Terakhir, bab ini menguraikan pentingnya evaluasi kinerja model prediksi sebagai tahap krusial dalam memastikan keandalan dan akurasi hasil penelitian.

Dengan memahami dasar teori ini, diharapkan pembaca memperoleh gambaran yang jelas mengenai konsep, metode, dan pendekatan yang digunakan dalam penelitian prediksi harga saham berbasis *machine learning* dan analisis sentimen pasar.

Untuk memulai pembahasan dasar teori, terlebih dahulu akan dijelaskan mengenai *machine learning* sebagai fondasi utama dalam pengembangan model prediksi harga saham pada penelitian ini. Pembahasan mencakup definisi, jenis-jenis, serta peran *machine learning* dalam menganalisis data pasar saham secara efektif dan efisien.

2.3.1. Machine Learning

Machine learning adalah cabang dari *Artificial Intelligence* (*kecerdasan buatan*) yang berfokus pada pengembangan algoritma dan teknik yang memungkinkan komputer untuk "belajar" dari data serta membuat prediksi atau keputusan berdasarkan data tersebut (Lac dkk., 2024).



Gambar 2. 1 Kedudukan Machine Learning.

Machine learning memiliki beberapa konsep dasar:

1. Definisi dan tujuan

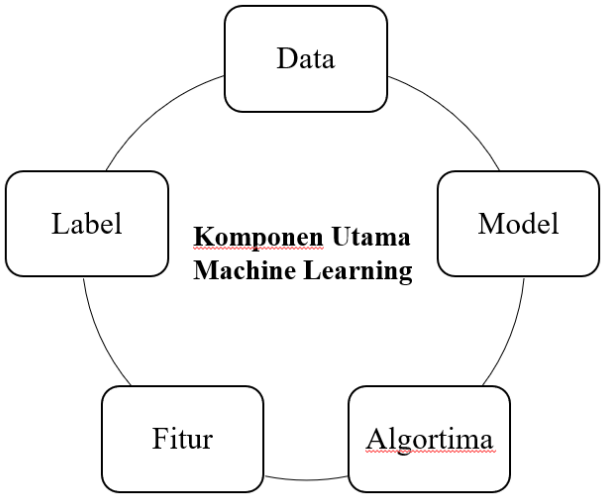
Machine learning memiliki definisi yaitu proses di mana komputer menggunakan data untuk meningkatkan kinerja pada tugas tertentu tanpa diprogram secara eksplisit untuk tugas tersebut. Tujuan *Machine learning* adalah untuk meningkatkan akurasi dan kemampuan prediktif dari model melalui pengalaman atau data.

2. Jenis-jenis *Machine Learning*

a. *Supervised Learning*: Algoritma belajar dari data yang sudah diberi label. Contoh: klasifikasi dan regresi.
b. *Unsupervised Learning*: Algoritma belajar dari data tanpa label dan menemukan pola atau struktur dalam data. Contoh: *clustering* dan *association*.
c. *Semi-supervised Learning*: Kombinasi dari *supervised* dan *unsupervised learning*, menggunakan sedikit data berlabel dan banyak data tidak berlabel.
d. *Reinforcement Learning*: Algoritma belajar melalui interaksi dengan lingkungan dan menerima *feedback* dalam bentuk *reward* atau *punishment*. Contoh: permainan catur, robotika.

3. Komponen Utama dalam *Machine Learning*

a. Data: Kumpulan informasi yang digunakan untuk melatih model.
b. Model: Struktur matematis atau algoritma yang digunakan untuk membuat prediksi atau keputusan.
c. Algoritma: Metode atau prosedur yang digunakan untuk mengoptimalkan model berdasarkan data.
d. Fitur: Variabel input yang digunakan untuk membuat prediksi.
e. *Label*: Hasil atau output yang diinginkan dalam *supervised learning.*

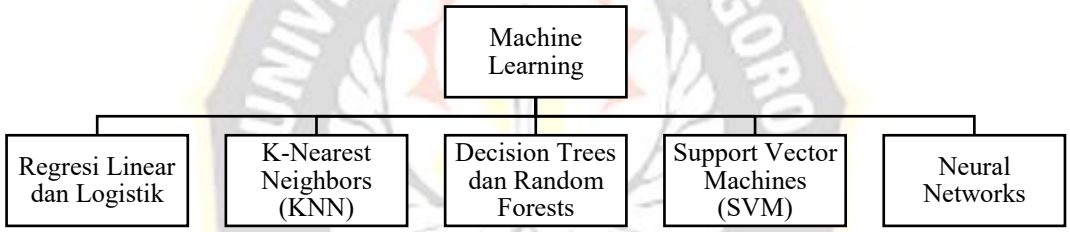


Gambar 2. 2 Komponen utama Machine Learning.

4. Algoritma Populer dalam *Machine Learning*

a. Regresi Linear dan Logistik: Untuk prediksi kontinuitas dan klasifikasi biner.
b. *K-Nearest Neighbors* (KNN): Algoritma sederhana untuk klasifikasi dan regresi.
c. *Decision Trees* dan *Random Forests*: Algoritma berbasis pohon untuk klasifikasi dan regresi.
d. *Support Vector Machines* (SVM): Algoritma untuk klasifikasi dengan margin maksimum.
e. *Neural Networks*: Algoritma untuk berbagai tugas, terutama deep learning.

Gambar 2.3 berikut ini disajikan ilustrasi dari beberapa algoritma populer dalam *machine learning* yang telah dijelaskan sebelumnya.



Gambar 2. 3 Algortima Populer Machine Learning.

2.3.2. Machine Learning dalam Prediksi Harga Saham

Menurut (Murphy, 1990) dalam "*Machine Learning: A Probabilistic Perspective*" pendekatan probabilistik dalam *machine learning* memungkinkan penanganan ketidakpastian dan variabilitas dalam data pasar saham. *Machine learning* menyediakan berbagai algoritma yang mampu mempelajari pola dari data historis dan melakukan prediksi berdasarkan pola tersebut. Algoritma populer yang sering digunakan dalam prediksi harga saham meliputi Regresi linear, Regresi logistik, *Support Vector Machines* (SVM), *Random Forest*, dan *deep learning*.

1. Linear Regression

Regresi linier adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (fitur) dan satu variabel dependen (target). Dalam konteks prediksi saham, variabel independen bisa berupa harga

saham sebelumnya, volume perdagangan, atau skor sentimen, sedangkan variabel dependen adalah harga saham yang ingin diprediksi.

Langkah-langkah proses dimulai dengan mengumpulkan data historis saham sebagai dasar analisis. Setelah itu, dilakukan pemilihan fitur-fitur relevan yang akan digunakan dalam model, seperti harga saham pada hari sebelumnya, volume perdagangan, serta indikator teknikal seperti *moving average*. Selanjutnya, model dilatih untuk mencari nilai optimal dari parameter β (*beta*) dengan menggunakan metode *Least Squares*, yang bertujuan meminimalkan selisih antara prediksi dan nilai aktual. Setelah model selesai dilatih, performanya diuji menggunakan metrik evaluasi seperti *Root Mean Squared Error* (RMSE) atau *Mean Absolute Error* (MAE) untuk menilai akurasi prediksi yang dihasilkan.

Metode ini memiliki kelebihan berupa kesederhanaan, kemudahan interpretasi, dan efisiensi dalam menangani *dataset* besar. Proses pelatihannya yang cepat juga memungkinkan pemodelan hubungan linier antara fitur input dan variabel target secara jelas.

Namun demikian, metode ini juga memiliki kekurangan. Metode linear regression tidak mampu menangkap hubungan non-linier yang kompleks antar variabel, sehingga kadang kurang akurat pada data yang bersifat dinamis dan tidak linier. Selain itu, model ini sangat sensitif terhadap keberadaan outlier yang dapat memengaruhi hasil prediksi secara signifikan, serta rentan terhadap masalah multikolinearitas antar fitur yang dapat menyebabkan ketidakstabilan dalam estimasi parameter.

Rumus Dasar untuk model *linear regression* sederhana:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (2.1)$$

Di mana:

- $\hat{y}$: Prediksi harga saham
- x_i: Fitur-fitur input
- β_i: Koefisien regresi (yang diestimasi oleh model)
- ϵ: *Error* (residu)

2. Random Forest

Random Forest adalah algoritma *ensemble learning* yang menggunakan gabungan dari beberapa *decision tree* untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting*. Pada kasus regresi, *output* dari semua pohon tersebut dirata-ratakan untuk menghasilkan prediksi akhir.

Prinsip kerja algoritma ini dimulai dengan membagi *dataset* pelatihan menjadi beberapa subset menggunakan teknik *bootstrapping*, yaitu pengambilan sampel secara acak dengan penggantian dari data asli. Untuk setiap subset yang dihasilkan, dibuat sebuah *decision tree* yang dibangun berdasarkan subset acak dari fitur-fitur yang tersedia, proses ini dikenal dengan istilah *feature sampling*. Setelah semua *decision tree* selesai dibuat, hasil prediksi dari seluruh pohon tersebut dirata-ratakan untuk menghasilkan prediksi akhir yang lebih stabil dan akurat.

Rumus Umum Prediksi *Random Forest*: Jika terdapat N pohon keputusan, dan masing-masing menghasilkan prediksi $\hat{y}_i$, maka prediksi akhir $\hat{y}$ dihitung sebagai:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \qquad (2.2)$$

Rumus ini menjelaskan bahwa untuk menghasilkan prediksi akhir pada algoritma *Random Forest*, kita mengambil rata-rata dari semua prediksi individu yang diberikan oleh masing-masing pohon keputusan.

Di mana:

- N adalah jumlah total pohon keputusan yang dibangun dalam model *Random Forest*.
- $\hat{y}_i$ adalah prediksi yang dihasilkan oleh pohon keputusan ke-i.
- $\hat{y}$ adalah prediksi akhir yang diperoleh dari gabungan semua pohon.

Dengan cara ini, *Random Forest* menggabungkan kekuatan beberapa pohon keputusan yang berbeda, sehingga prediksi akhir menjadi lebih stabil dan akurat dibandingkan prediksi dari satu pohon saja. Rata-rata ini juga membantu

mengurangi risiko overfitting dan menangkap pola data yang lebih kompleks secara keseluruhan.

3. Neural Networks:

Neural Network adalah model komputasi yang terinspirasi dari cara kerja otak manusia. Model ini sangat fleksibel dan kuat untuk menangani pola *non-linier* yang kompleks, terutama dalam kasus *time series* seperti prediksi harga saham.

Struktur Umum Jaringan terdiri dari:

a. *Input layer*: Menerima data fitur (misalnya harga, volume, sentimen).
b. *Hidden layers:* Lapisan yang melakukan transformasi non-linier.
c. *Output layer*: Menghasilkan prediksi akhir (harga saham).

Rumus dasar *neural network*:

$$\hat{y} = f(Wx + b) \qquad (2.3)$$

Keterangan:

- $\hat{y}$: output prediksi
- x: vektor input (misal: fitur harga historis, sentimen, dll.)
- W: matriks bobot dari layer (bobot antar neuron)
- b: vektor bias
- f: fungsi aktivasi (non-linearitas), seperti ReLU, sigmoid, tanh

Algoritma seperti *neural network* memiliki sejumlah keunggulan yang menjadikannya sangat fleksibel dalam menangani data dengan pola yang kompleks dan dinamis. Kemampuannya untuk melakukan pemodelan non-linier memungkinkan algoritma ini mencapai akurasi tinggi, terutama ketika diterapkan pada data yang tidak mengikuti hubungan linier sederhana. *Neural network* juga sangat cocok digunakan pada *big data* serta integrasi berbagai jenis fitur multivariat, termasuk data numerik seperti harga saham dan volume perdagangan, maupun data teks seperti sentimen dari media sosial.

Namun, di balik keunggulannya, algoritma ini memiliki beberapa kekurangan. Untuk mencapai hasil yang optimal, *neural network* memerlukan data

dalam jumlah besar. Proses pelatihannya pun memakan waktu lebih lama dan membutuhkan sumber daya komputasi yang tinggi. Selain itu, model ini sulit diinterpretasikan karena tergolong sebagai *black-box model*, yang berarti pengguna tidak dapat dengan mudah memahami bagaimana keputusan atau prediksi dihasilkan.

Sebagai contoh aplikasinya, fitur input dalam *neural network* untuk prediksi harga saham dapat mencakup harga saham lima hari terakhir, volume harian, skor sentimen dari Twitter, serta indikator teknikal seperti RSI (*Relative Strength Index*) atau MACD (*Moving Average Convergence Divergence*). Model ini akan belajar mengidentifikasi korelasi antar fitur tersebut serta mengenali pola-pola temporal yang relevan untuk memprediksi harga saham pada hari berikutnya.

2.3.3. Analisis Sentimen

Analisis sentimen adalah teknik dalam *Natural Language Processing* (NLP) yang digunakan untuk mengidentifikasi dan mengekstraksi informasi subjektif dari teks. Analisis ini sering digunakan untuk memahami opini publik atau sentimen pasar yang dapat mempengaruhi harga saham. Pendekatan umum dalam analisis sentimen melibatkan beberapa langkah berikut:

1. *Preprocessing*: Membersihkan teks dari elemen yang tidak relevan seperti tanda baca, stop words, dan lain-lain.
2. *Feature Extraction*: Menggunakan metode seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) atau *embeddings* untuk merepresentasikan teks.
3. *Sentiment Classification*: Menggunakan algoritma *machine learning* seperti *Naive Bayes*, *Support Vector Machines* (SVM), atau *deep learning* untuk mengklasifikasikan teks sebagai positif, negatif, atau netral.

2.3.4. Teori Pasar Efisien

Teori Pasar Efisien atau *Efficient Market Hypothesis* (EMH) merupakan teori yang dikembangkan oleh Eugene Fama pada tahun 1970, yang menyatakan bahwa harga pasar mencerminkan seluruh informasi yang tersedia pada waktu tertentu (Han, 2025). Dengan kata lain, tidak ada investor yang dapat secara konsisten

mengalahkan pasar karena semua informasi relevan telah tercermin dalam harga saham.

Dalam konteks ini, pasar modal dikatakan efisien apabila harga saham bergerak secara acak dan tidak dapat diprediksi hanya dengan menggunakan informasi publik yang tersedia. Berdasarkan teori *Efficient Market Hypothesis* (EMH), analisis teknikal maupun fundamental yang mendalam dianggap tidak memberikan keuntungan signifikan, karena seluruh informasi yang relevan telah tercermin dalam harga saham saat ini. Menurut Eugene Fama, efisiensi pasar dibagi menjadi tiga bentuk utama, yaitu: efisiensi lemah, di mana harga saham mencerminkan seluruh data historis; efisiensi setengah kuat, yang mencerminkan seluruh informasi publik; dan efisiensi kuat, yang mencerminkan seluruh informasi publik maupun privat.

1. Efisiensi Lemah

   Harga saham saat ini mencerminkan seluruh informasi historis harga. Dalam bentuk ini, analisis teknikal dianggap tidak efektif karena pola masa lalu tidak bisa digunakan untuk memprediksi harga masa depan.
2. Efisiensi Setengah Kuat

   Harga saham mencerminkan seluruh informasi publik, termasuk laporan keuangan, berita ekonomi, dan kebijakan perusahaan. Dalam kondisi ini, baik analisis teknikal maupun fundamental tidak akan memberikan keunggulan signifikan.
3. Efisiensi Kuat

   Harga saham mencerminkan seluruh informasi, termasuk informasi rahasia atau insider information. Dengan kata lain, bahkan investor dalam perusahaan pun tidak dapat memperoleh keuntungan abnormal secara konsisten.

Jika pasar benar-benar efisien, maka model prediksi harga saham menggunakan *machine learning* hanya akan menangkap *noise* dari data, bukan pola yang dapat digunakan untuk memperoleh keuntungan. Namun, berbagai penelitian menunjukkan bahwa pada kondisi tertentu terutama di pasar negara berkembang atau saham dengan kapitalisasi kecil pasar belum sepenuhnya efisien. Hal ini

membuka ruang bagi pemanfaatan model prediktif berbasis data historis dan sentimen pasar.

Secara matematis, teori ini dapat direpresentasikan dalam bentuk berikut:

$$P_t = E(P_{t+1} \mid \Phi_t) \tag{2.4}$$

Keterangan:

- P_t: Harga saham saat ini
- $E(P_{t+1} \mid \Phi_t)$: Ekspektasi harga saham di masa depan berdasarkan seluruh informasi yang tersedia saat ini

Atau dalam bentuk perubahan harga saham:

$$r_t = P_t - P_{t-1} = \epsilon_t \tag{2.5}$$

Keterangan:

- r_t: Return saham pada waktu tt
- ϵ_t: Komponen *error* acak (*white noise*) yang tidak dapat diprediksi

Persamaan di atas menegaskan bahwa perubahan harga saham (r_t) bersifat acak dan mengikuti distribusi probabilistik normal, sehingga tidak mungkin memperoleh keuntungan abnormal secara konsisten hanya dengan mengandalkan informasi yang sudah diketahui publik.

Walaupun EMH merupakan teori dominan dalam ekonomi keuangan, teori ini juga mendapat kritik, terutama dari kalangan *behavioral finance*. Kritik utama adalah bahwa pasar sering kali tidak rasional akibat faktor psikologis investor, seperti *herding behavior, overreaction*, dan *underreaction*, yang menyebabkan deviasi harga dari nilai fundamental. Oleh karena itu, analisis berbasis *machine learning* yang menggabungkan sentimen pasar dapat memberikan insight tambahan dalam kondisi pasar yang tidak sepenuhnya efisien.

2.3.5. Evaluasi Kinerja Model Prediksi

Evaluasi model merupakan tahap penting dalam proses pembangunan model prediksi, khususnya dalam konteks *machine learning*. Tujuannya adalah untuk

menilai seberapa baik model dalam melakukan generalisasi terhadap data yang belum pernah dikenal sebelumnya. Evaluasi dilakukan dengan menggunakan berbagai metrik yang dapat mengukur ketepatan, akurasi, dan kesesuaian hasil prediksi model terhadap data aktual. Dalam konteks prediksi harga saham, evaluasi kinerja model sangat krusial karena kesalahan kecil dalam prediksi dapat berdampak signifikan terhadap keputusan investasi.

Dalam proses evaluasi model prediksi, diperlukan sejumlah metrik evaluasi yang dapat memberikan gambaran objektif mengenai tingkat akurasi dan efisiensi prediksi model terhadap data yang sebenarnya. Metrik-metrik ini berfungsi untuk mengukur sejauh mana hasil prediksi model mendekati nilai aktual, serta seberapa besar kesalahan yang terjadi selama proses prediksi. Beberapa metrik evaluasi yang umum digunakan dalam penelitian prediksi harga saham antara lain adalah *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan *R-Squared* atau R^2. Berikut ini adalah uraian masing-masing metrik evaluasi tersebut.

1. *Mean Absolute Error* (MAE)

MAE adalah metrik evaluasi yang mengukur rata-rata dari selisih absolut antara nilai aktual dan nilai yang diprediksi. Nilai MAE memberikan informasi langsung mengenai seberapa besar kesalahan prediksi secara rata-rata. Selain itu MAE memiliki keunggulan dalam interpretasi yang mudah karena menggunakan satuan yang sama dengan target prediksi.

Rumus MAE:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (2.6)$$

Komponen:

- n: jumlah data
- y_i: nilai aktual
- $\hat{y}_i$: nilai prediksi
- $| y_i - \hat{y}_i |$: nilai absolut dari kesalahan prediksi

2. *Root Mean Square Error* (RMSE)

RMSE adalah metrik yang menghitung akar dari rata-rata kuadrat kesalahan antara nilai aktual dan prediksi. RMSE lebih sensitif terhadap kesalahan besar karena efek kuadrat, menjadikannya penting untuk model yang harus sangat akurat. RMSE sering digunakan ketika kesalahan besar lebih tidak diinginkan dibandingkan kesalahan kecil.

Rumus RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.7}$$

Komponen:

- n: jumlah data
- y_i: nilai aktual
- $\hat{y}_i$: nilai prediksi
- $(y_i - \hat{y}_i)^2$: kuadrat dari selisih nilai aktual dan prediksi

3. *R-squared* (R^2)

R^2 atau koefisien determinasi digunakan untuk menilai seberapa baik model dapat menjelaskan variasi data. Nilai R^2 berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa model menjelaskan sebagian besar variabilitas target. Semakin tinggi nilai R^2, semakin baik model dalam menjelaskan data.

Rumus R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{2.8}$$

Komponen:

- y_i: nilai aktual
- $\hat{y}_i$: nilai prediksi
- $\bar{y}$: rata-rata nilai aktual
- $\sum(y_i - \hat{y}_i)^2$: *residual sum of squares* (RSS)
- $\sum\left(y_i - \bar{y}\right)^2$: total *sum of squares* (TSS)

2.3.6. Strategi Pembagian Data dan Validasi Model

Dalam pengembangan model *machine learning*, strategi pembagian data dan validasi model merupakan aspek penting untuk memastikan bahwa model yang dibangun tidak hanya akurat pada data pelatihan, tetapi juga mampu melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Salah satu pendekatan umum yang digunakan adalah membagi dataset menjadi dua bagian utama: data pelatihan dan data pengujian, dengan proporsi yang bervariasi tergantung pada ukuran dan karakteristik data.

Dalam penelitian ini, digunakan strategi pembagian data sebesar 70:30, di mana 70% data digunakan untuk melatih model dan 30% sisanya digunakan untuk menguji performa model. Pendekatan ini memberikan keseimbangan antara jumlah data yang cukup untuk pelatihan dan data yang representatif untuk evaluasi. Strategi ini juga sejalan dengan praktik umum dalam literatur *machine learning*, terutama ketika ukuran *dataset* tidak terlalu besar namun cukup beragam.

Dalam studi oleh (Mundra dkk., 2025), strategi pembagian data dilakukan dengan pendekatan yang disesuaikan untuk skenario *few-shot learning*. *Dataset* HBN-*tweet* yang dikembangkan dalam penelitian tersebut dibagi menjadi *support set* dan *query set*, yang masing-masing berfungsi sebagai data pelatihan dan pengujian dalam kerangka episodic training. Meskipun pendekatannya berbeda secara teknis, prinsip dasarnya tetap sama: memisahkan data untuk pelatihan dan evaluasi guna menghindari data *leakage* dan memastikan validitas hasil.

Validasi model dalam konteks ini dilakukan dengan mengukur akurasi klasifikasi pada *query set* setelah model dilatih menggunakan *support set*. Dalam pendekatan *few-shot learning*, validasi dilakukan secara berulang dalam bentuk *episodic training*, di mana model dilatih dan diuji dalam banyak episode dengan kombinasi data yang berbeda. Hal ini bertujuan untuk meningkatkan kemampuan generalisasi model terhadap data baru dengan jumlah label yang terbatas.

Dalam konteks penelitian ini, pembagian 70:30 digunakan untuk membangun model prediksi harga saham berbasis *machine learning*, di mana data historis saham dan data sentimen media sosial digabungkan. Strategi ini memungkinkan evaluasi yang objektif terhadap kemampuan model dalam memprediksi harga saham berdasarkan data yang belum pernah dilihat sebelumnya.

Dengan demikian, strategi pembagian data dan validasi model yang digunakan dalam penelitian ini tidak hanya mengikuti praktik terbaik dalam *machine learning*, tetapi juga memperhatikan kebutuhan spesifik dari pendekatan yang digunakan.

2.3.7. Penyetelan Hyperparameter dalam Model Machine Learning

Dalam pengembangan model *machine learning*, penyetelan *hyperparameter* merupakan tahap krusial yang secara langsung memengaruhi performa model. *Hyperparameter* adalah parameter yang nilainya ditentukan sebelum proses pelatihan dimulai, seperti *learning rate*, jumlah *neuron*, jumlah lapisan tersembunyi, fungsi aktivasi, dan ukuran *batch* (Manivannan dan Senthilkumar, 2025). Tidak seperti parameter model yang dipelajari selama pelatihan, *hyperparameter* harus ditentukan melalui proses eksploratif dan eksperimental.

Penyetelan *hyperparameter* yang tidak optimal dapat menyebabkan model mengalami *overfitting* atau *underfitting*, serta memperlambat konvergensi selama pelatihan. Oleh karena itu, berbagai pendekatan telah dikembangkan untuk mengotomatisasi dan mengoptimalkan proses ini, mulai dari *grid search, random search*, hingga algoritma optimasi berbasis *metaheuristic*.

Dalam penelitian ini, penyetelan *hyperparameter* dilakukan menggunakan metode *grid search*, yaitu salah satu pendekatan pencarian ekshaustif yang paling umum digunakan dalam *machine learning*. *Grid search* bekerja dengan cara mengevaluasi semua kombinasi yang mungkin dari nilai-nilai *hyperparameter* yang telah ditentukan sebelumnya dalam sebuah ruang pencarian. Meskipun metode ini bersifat *brute-force* dan dapat menjadi mahal secara komputasi, keunggulannya

terletak pada kesederhanaan dan kemampuannya untuk menemukan kombinasi parameter terbaik secara sistematis.

Sebagai contoh, jika model *machine learning* memiliki dua *hyperparameter* seperti *learning rate* dan jumlah *neuron*, maka *grid search* akan mencoba semua kombinasi nilai yang mungkin dari kedua parameter tersebut, seperti:

- *learning rate*: [0.001, 0.01, 0.1]
- *neuron*: [64, 128, 256]

Maka total kombinasi yang diuji adalah 3 × 3 = 9. Setiap kombinasi akan dievaluasi menggunakan teknik validasi silang untuk mengukur performa model, dan kombinasi dengan skor terbaik akan dipilih sebagai konfigurasi akhir.

Meskipun dalam studi oleh Manivannan dan Senthilkumar (2025) digunakan algoritma *Fox Optimization* untuk menyetel *hyperparameter* secara adaptif, pendekatan *grid search* tetap relevan dan efektif, terutama ketika ruang pencarian relatif kecil dan sumber daya komputasi mencukupi. Dalam konteks prediksi harga saham, *grid search* memberikan kontrol penuh terhadap eksplorasi parameter dan dapat menghasilkan model yang stabil serta dapat direproduksi.

Dalam studi oleh Manivannan dan Senthilkumar (2025), penyetelan *hyperparameter* dilakukan menggunakan algoritma optimasi berbasis perilaku hewan, yaitu *Fox Optimization Algorithm* (*FOX*). FOX meniru perilaku berburu rubah dalam mencari mangsa, dengan menyeimbangkan *eksplorasi* (*global search*) dan *eksploitasi* (*local refinement*). Pendekatan ini terbukti efektif dalam menemukan kombinasi *hyperparameter* terbaik untuk model *Adaptive Recurrent Neural Network* (ARNN), yang digunakan dalam sistem deteksi intrusi jaringan.

Pendekatan FOX ini dapat diadaptasi dalam konteks prediksi harga saham, terutama ketika menggunakan model *deep learning* seperti LSTM atau ARNN yang memiliki banyak *hyperparameter*. Dengan mengintegrasikan algoritma optimasi seperti FOX atau metode sistematis seperti *grid search*, proses *tuning* menjadi lebih efisien dan hasil prediksi menjadi lebih akurat.

Secara umum, penyetelan *hyperparameter* yang efektif tidak hanya meningkatkan akurasi model, tetapi juga mempercepat waktu pelatihan dan mengurangi kebutuhan komputasi. Oleh karena itu, dalam penelitian ini, pendekatan optimasi berbasis algoritma cerdas dipertimbangkan sebagai bagian integral dari *pipeline* prediksi harga saham berbasis *machine learning*.

2.3.8. Analisis Residual dan Visualisasi Prediktif

Dalam evaluasi performa model *machine learning* untuk tugas regresi seperti prediksi harga saham, analisis tidak hanya bergantung pada metrik numerik seperti *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan R^2. Pendekatan visual seperti grafik prediksi vs aktual, *scatter plot*, dan analisis *residual* juga sangat penting dalam menilai akurasi dan keandalan model.

Visualisasi ini memberikan pemahaman tambahan tentang sejauh mana model dapat menangkap pola dalam data, serta membantu mengidentifikasi bias, *outlier*, dan potensi *overfitting* atau *underfitting*.

Salah satu cara visual yang paling umum dan efektif untuk menilai performa model regresi adalah dengan membandingkan langsung nilai prediksi dengan nilai aktual dalam bentuk grafik deret waktu. Grafik ini memungkinkan peneliti untuk melihat sejauh mana model mengikuti pola asli data secara temporal.

1. Grafik Prediksi vs Aktual

Grafik ini menyajikan perbandingan visual antara nilai yang diprediksi oleh model dan nilai aktual dari data dalam bentuk deret waktu. Grafik semacam ini memungkinkan peneliti untuk menilai apakah model mengikuti pola data dengan baik. Dalam penelitian oleh (Hussain dkk., 2025), grafik prediksi dan aktual digunakan untuk menampilkan kinerja model *Extra Trees Regressor* (ETR) terhadap konsumsi energi kendaraan listrik.

Penulis menyatakan: "*Plot deret waktu untuk model Extra Trees Regressor (ETR) pada Gambar 4(a) menunjukkan bahwa nilai prediksi sangat sesuai dengan tren konsumsi energi aktual, dengan deviasi yang sangat kecil*."

Pernyataan tersebut menunjukkan bahwa grafik ini dapat menggambarkan sejauh mana model mendekati kenyataan, dan memberikan indikasi visual terhadap kesalahan prediksi.

2. *Scatter Plot* dan Analisis *Residual*

Residual adalah selisih antara nilai aktual dan prediksi yang dihasilkan oleh model. Untuk memahami distribusi dan pola kesalahan ini, *scatter plot* antara nilai aktual dan nilai prediksi digunakan. Jika titik-titik data menyebar dekat dengan garis diagonal (y = x), hal ini menunjukkan bahwa prediksi model mendekati nilai aktual.

(Hussain dkk., 2025) menjelaskan bahwa: "*Gambar 4(b) menunjukkan scatter plot yang mengindikasikan korelasi yang kuat antara nilai aktual dan prediksi. Titik-titik data tampak berkumpul rapat di sepanjang garis diagonal*."

"*Scatter plot menunjukkan hubungan linear yang kuat, meskipun dengan sebaran yang lebih besar dibandingkan model-model unggulan lainnya*."

Visualisasi ini membantu menilai apakah kesalahan tersebar secara merata, atau terdapat pola yang mengindikasikan kelemahan model, seperti *overfitting* pada *subset* data tertentu.

3. Distribusi *Residual* (Interpretasi Visual)

Distribusi *residual*, yaitu selisih antara nilai aktual dan prediksi, menjadi aspek penting dalam menilai performa model regresi. Dalam studi Hussain dkk. (2025), meskipun tidak menampilkan histogram *residual* secara eksplisit, analisis visual terhadap scatter plot prediksi vs aktual dan grafik deret waktu membantu mengamati pola kesalahan model. *Residual* yang tersebar acak di sekitar garis nol menunjukkan bahwa kesalahan prediksi bersifat *random* dan model tidak mengandung bias sistematis.

Sebaliknya, apabila *residual* menunjukkan pola tertentu, seperti penumpukan di area tertentu atau perubahan variansi residual, ini mengindikasikan adanya bias atau masalah seperti *overfitting* dan *underfitting*. Pola tersebut mengisyaratkan bahwa model kurang mampu menangkap hubungan kompleks dalam data atau tidak stabil di beberapa segmen. Selain itu, keberadaan *outlier residual* yang jauh dari nol juga perlu diperhatikan karena menandakan prediksi ekstrem yang kurang akurat.

Analisis visual distribusi *residual* ini sangat membantu dalam memvalidasi asumsi dasar model, seperti variansi *residual* yang konstan, serta memberikan wawasan untuk perbaikan model lebih lanjut. Dengan pendekatan ini, evaluasi model tidak hanya bergantung pada metrik numerik, tetapi juga pada pemahaman mendalam terhadap karakteristik kesalahan yang terjadi, sehingga meningkatkan keandalan prediksi.

Dengan dasar teori yang kuat dalam aspek-aspek tersebut, penelitian ini dapat memberikan dasar konseptual yang kuat untuk mendukung penelitian, sementara algoritma *Machine learning* yang disebutkan dapat digunakan untuk mengintegrasikan analisis sentimen pasar dengan fluktuasi harga saham perusahaan Bank digital pada industri *finTech*.

Setelah memahami dasar-dasar teori yang mendasari penelitian ini, langkah berikutnya adalah menjelaskan metode penelitian yang digunakan untuk mengembangkan dan menguji model prediksi harga saham. Bab metodologi penelitian memaparkan secara rinci waktu dan tempat pelaksanaan penelitian, serta prosedur dan jadwal kegiatan yang akan dijalankan selama masa penelitian. Penjelasan ini penting agar proses penelitian dapat dilakukan secara sistematis dan terukur sesuai dengan tujuan yang telah ditetapkan.