Journal of Banking and Financial Dynamics
Vol. 10, No. 6, 1-16, 2026
ISSN(E): 2576-6821
DOI: 10.55220/2576-6821.v10.963

Does Big Data Improve Price Discovery in Frontier Capital Markets? Evidence from African Frontier Exchanges Using Econometric and Machine Learning Models

Akomolehin Francis Olugbenga[1] ✉
Famoroti Olusegun Jonathan[2]

[1]*Dept of Finance, College of Social and Management Sciences, Afe Babalola University, Ado - Ekiti, Nigeria.*
[2]*Dept of Economics, College of Social and Management Sciences, Afe Babalola University, Ado - Ekiti, Nigeria.*
(✉ Corresponding Author)

Abstract

This study examines whether big data analytics improves price discovery in selected African frontier capital markets, using evidence from Nigeria, Ghana, Kenya, Rwanda, and Zambia. The study is motivated by the growing use of alternative digital datasets, artificial intelligence, and machine learning in global financial markets, alongside persistent informational inefficiencies in African frontier exchanges. An explanatory longitudinal time-series design was adopted, using daily data from Bloomberg, Refinitiv, Yahoo Finance, Google Trends, Twitter/X sentiment analytics, and financial news databases for the period 2015 to 2025. The econometric analysis employed Augmented Dickey-Fuller, Phillips-Perron, Johansen cointegration, Vector Error Correction Model, Hasbrouck Information Share, and Gonzalo-Granger Component Share techniques. The machine learning framework compared ARIMA, Random Forest, XGBoost, and Long Short-Term Memory models. The findings show significant long-run relationships between traditional market variables and big data indicators. Google Trends, Twitter/X sentiment, and news analytics improved price discovery efficiency and accelerated the incorporation of information into stock prices. Kenya and Nigeria showed stronger informational efficiency than Ghana, Rwanda, and Zambia. The machine learning results also indicate that LSTM and XGBoost outperformed ARIMA across the selected prediction metrics. The study concludes that big data analytics and machine learning can strengthen price discovery in African frontier capital markets by improving information processing, predictive accuracy, and market responsiveness.

Keywords: Big data analytics, Frontier capital markets, Information share, Investor sentiment, Machine learning, Price discovery.

1. Introduction

1.1. Background to the Study

The structure and operation of global capital markets have changed significantly over the last two decades due to technological advancement, financial globalization, and the growing use of digital information in investment decision-making. Financial markets no longer rely only on traditional accounting disclosures, macroeconomic indicators, and firm-specific announcements for price formation. Increasingly, investors and analysts use alternative digital information sources such as online search behavior, social media sentiment, news analytics, and web-based attention indicators to support asset pricing and market forecasting (Baker et al., 2021). This shift has strengthened the relevance of big data analytics in explaining market behavior, improving forecasting accuracy, and enhancing price discovery.

Price discovery is central to the effective functioning of capital markets because it determines how quickly and accurately available information is reflected in security prices. Market efficiency theory assumes that prices should incorporate relevant information promptly, thereby supporting efficient resource allocation (Urquhart & McGroarty, 2022). Efficient price discovery reduces information asymmetry, lowers transaction costs, improves liquidity, and strengthens investor confidence. However, the speed and quality of price discovery differ across developed, emerging, and frontier markets because of variations in liquidity, market depth, institutional quality, and information dissemination systems (Aslam et al., 2023).

African frontier capital markets still face structural weaknesses that limit informational efficiency. These include low market capitalization, thin trading, weak institutional frameworks, limited investor participation, poor analyst coverage, and delayed information diffusion (Adelegan & Mwamba, 2021). Although markets such as Nigeria, Ghana, Kenya, Rwanda, and Zambia have experienced financial sector reforms and digital transformation, they continue to face informational frictions that affect pricing efficiency and market transparency (World Bank, 2023).

At the same time, digital transformation across African economies has altered investor behavior. Rising internet access, smartphone usage, mobile financial services, online trading platforms, and social media

participation have created large volumes of digital footprints that may contain useful market signals (International Monetary Fund [IMF], 2024). Big data analytics provides tools for extracting meaningful patterns from these high-volume, high-velocity, and diverse datasets for prediction and decision-making (Najem, 2025).

Alternative datasets are now increasingly important in financial research. Google Trends, for example, is widely used as a proxy for investor attention because search activity often reflects public interest before market transactions occur. Search intensity has been linked to stock market movements, volatility, and trading behavior, especially during uncertain periods (Da et al., 2021). Similarly, Twitter/X sentiment analytics captures investor opinions, emotions, and reactions shared through social media platforms, which may influence short-term market dynamics and pricing behavior (Chen & Liu, 2022).

News analytics has also become important in modern financial intelligence. Financial news contains qualitative information that shapes investor expectations and market sentiment. Advances in natural language processing and machine learning now allow researchers to convert news content into measurable sentiment indicators that can be integrated into financial models (Khan et al., 2023). These digital information sources may be particularly useful in frontier markets where formal disclosure systems are weaker or slower.

The rise of big data has also encouraged the use of machine learning in financial forecasting. Traditional econometric models such as ARIMA and VECM remain useful for capturing linear relationships and long-run dynamics. However, they may perform poorly when markets exhibit nonlinearity, structural breaks, and high-frequency information shocks (Bouteska & Regaieg, 2022). Machine learning models such as Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks are increasingly applied because they can capture nonlinear relationships, learn complex patterns, and improve prediction in volatile financial environments (Yao et al., 2024).

Despite the global growth of big data analytics and machine learning in finance, empirical evidence from African frontier markets remains limited. Existing African capital market studies focus mainly on stock returns, volatility spillovers, market integration, and macroeconomic determinants of financial development, with less attention to alternative data indicators and information share dynamics (Ntim et al., 2022). Few studies have examined whether Google Trends, Twitter/X sentiment, and news analytics improve the speed and efficiency of price discovery in African frontier exchanges.

Price discovery speed is especially important in frontier markets because delayed information incorporation can create pricing inefficiencies, speculative opportunities, and market manipulation risks. Faster price discovery improves transparency, strengthens investor confidence, enhances liquidity, and supports more efficient capital allocation (Hasbrouck, 2021). In markets affected by thin trading and information asymmetry, identifying tools that accelerate information incorporation into stock prices is critical for market development and regulation.

Nigeria, Ghana, Kenya, Rwanda, and Zambia provide a useful context for this study because they represent active African frontier markets with increasing digital penetration and expanding investor participation. These economies have recorded growth in mobile technology, internet connectivity, and digital financial services, thereby generating alternative datasets with potential relevance for financial markets (African Development Bank, 2024). However, the extent to which these digital signals improve price discovery remains underexplored.

Against this background, this study examines whether big data analytics improves price discovery in selected African frontier capital markets. It integrates traditional econometric techniques with machine learning models, including VECM, Hasbrouck Information Share, Gonzalo-Granger Component Share, ARIMA, Random Forest, XGBoost, and LSTM. By focusing on Nigeria, Ghana, Kenya, Rwanda, and Zambia, the study contributes to the literature on financial digitalization, alternative data analytics, machine learning, and informational efficiency in frontier capital markets.

1.2. Statement of the Research Problem

Efficient price discovery ensures that security prices reflect available information accurately and promptly. In efficient markets, information diffusion occurs quickly, reducing information asymmetry and improving investment decisions. However, African frontier capital markets continue to experience structural and informational inefficiencies that weaken price discovery. Many of these exchanges are characterized by low liquidity, thin trading, weak institutional structures, limited market depth, and delayed information transmission, which slow the adjustment of stock prices to new information (Adelegan & Mwamba, 2021).

Although Nigeria, Ghana, Kenya, Rwanda, and Zambia have recorded improvements in digital financial inclusion and capital market reforms, information incorporation remains weaker than in developed and larger emerging markets (World Bank, 2023). Delayed price adjustment increases market inefficiency, speculative trading, pricing distortions, and volatility. These conditions reduce investor confidence and weaken the capital market's role in allocating financial resources efficiently.

The growing digitalization of financial activity has created large volumes of structured and unstructured data that may influence investor decisions and market behavior. Alternative datasets such as Google Trends, Twitter/X sentiment, online search intensity, and news sentiment analytics increasingly shape market expectations in global financial systems (Chen & Liu, 2022). In developed markets, these indicators have been incorporated into forecasting models to improve prediction accuracy and explain information diffusion. However, their role in African frontier capital markets remains insufficiently examined.

Existing African financial market studies focus largely on stock returns, volatility transmission, financial integration, and macroeconomic drivers of market performance (Ntim et al., 2022). While useful, these studies provide limited evidence on how alternative big data indicators affect price discovery speed and informational efficiency. More importantly, few studies combine econometric price discovery measures with machine learning forecasting models to evaluate the predictive value of big data analytics in frontier markets.

Another concern is the limitation of traditional econometric models in periods of heightened information volatility and nonlinear market dynamics. Models such as ARIMA and conventional regression frameworks may not fully capture complex interactions among investor sentiment, digital attention, market liquidity, and stock

prices (Bouteska & Regaieg, 2022). The increasing use of Random Forest, XGBoost, and LSTM models in financial forecasting therefore creates a need for comparative evidence within African frontier market settings.

African frontier markets also face weak data infrastructure, fragmented information systems, low analyst coverage, and poor dissemination of market-sensitive information. These conditions deepen information asymmetry and slow the price adjustment process. Despite growing digital participation through mobile technology, online trading platforms, and social media engagement, it remains unclear whether these digital footprints contribute meaningfully to price discovery.

The central problem is the limited empirical evidence on whether big data indicators improve the speed and efficiency of price discovery in African frontier capital markets. There is also insufficient comparative evidence on whether machine learning models outperform traditional econometric approaches in explaining and predicting price discovery dynamics in frontier financial systems.

This study addresses these gaps by integrating big data analytics, econometric price discovery measures, and machine learning forecasting techniques to examine informational efficiency in Nigeria, Ghana, Kenya, Rwanda, and Zambia. Specifically, it investigates whether alternative data sources improve information incorporation into stock prices and whether machine learning models provide stronger predictive performance than traditional econometric models.

1.3. Research Questions

The study seeks to answer the following questions:

1. Does big data analytics improve price discovery in frontier capital markets?
2. What long-run relationship exists between traditional market variables and big data indicators in selected African frontier markets?
3. How do big data variables influence Hasbrouck Information Share and Gonzalo-Granger Component Share measures?
4. Which model provides the best predictive performance for price discovery dynamics among ARIMA, Random Forest, XGBoost, and LSTM?
5. Are there cross-country differences in the speed of price discovery among Nigeria, Ghana, Kenya, Rwanda, and Zambia?

1.4. Research Objectives

The main objective of this study is to examine the effect of big data analytics on price discovery in selected African frontier capital markets.

The specific objectives are to:

1. examine the long-run relationship between traditional market indicators and big data variables in selected frontier markets;
2. estimate the effect of big data analytics on price discovery using Hasbrouck Information Share and Gonzalo-Granger Component Share techniques;
3. compare the predictive performance of econometric and machine learning models in forecasting price discovery dynamics;
4. evaluate cross-country differences in informational efficiency among Nigeria, Ghana, Kenya, Rwanda, and Zambia; and
5. determine whether big data accelerates information incorporation into stock prices in frontier capital markets.

1.5. Significance of the Study

This study is significant to investors, policymakers, regulators, academics, fintech firms, and data analytics providers.

For investors, the study provides evidence on the predictive value of big data analytics and machine learning models in financial forecasting and trading decisions. Better understanding of digital sentiment indicators and price discovery dynamics can support portfolio management, asset allocation, and risk assessment.

For policymakers, the study contributes to debates on capital market development, digital transformation, and financial market efficiency in frontier economies. Evidence on information diffusion and market responsiveness can support reforms aimed at improving transparency, liquidity, and investor confidence.

For financial market regulators, the study provides insight into how alternative information channels influence asset prices and market behavior. This can assist regulators in strengthening market surveillance, disclosure systems, and information dissemination frameworks.

For academics, the study extends the literature on price discovery, big data analytics, and machine learning in finance by providing evidence from underexplored African frontier markets. Its integration of econometric and machine learning approaches also offers methodological value for future financial market research.

For fintech firms and data analytics companies, the study shows the potential usefulness of Google Trends, Twitter/X sentiment, news analytics, and other digital datasets in financial intelligence systems. This can support the development of data-driven financial products and market analysis tools within African financial systems.

2. Literature Review

2.1. Conceptual Review

2.1.1. Big Data Analytics

Big data analytics has become a major feature of modern financial systems due to rapid digitalization, internet expansion, and advances in computational intelligence. It refers to the collection, processing, and analysis of large and complex datasets to generate useful patterns, predictive information, and decision-making support (Najem, 2025). Unlike conventional data analysis, big data analytics accommodates structured, semi-structured, and

unstructured data from financial transactions, digital platforms, social media, search engines, and online communication systems.

The concept is commonly explained through the "5Vs" framework: volume, velocity, variety, veracity, and value. Volume refers to the large quantity of data produced through digital activities, while velocity captures the speed at which such data is generated and processed. Variety reflects the different formats of data, including numerical records, text, images, audio, and social media content. Veracity concerns data reliability, while value emphasizes the ability to extract meaningful financial and economic information from large datasets (Agarwal & Dhar, 2021).

In financial markets, big data analytics is applied to investment forecasting, risk management, fraud detection, algorithmic trading, portfolio optimization, and sentiment analysis. Investors and financial institutions increasingly use alternative datasets such as online news feeds, internet search behavior, social media sentiment, and transaction data to improve forecasting and investment decisions (Bouteska & Regaieg, 2022). Machine learning and natural language processing have further improved the capacity of analysts to identify nonlinear relationships and hidden informational signals within digital financial environments.

For frontier capital markets, big data analytics is especially relevant because formal information systems are often weak, fragmented, or delayed. Alternative digital information sources may help reduce information gaps, improve market transparency, and accelerate the adjustment of stock prices in markets affected by thin trading and informational inefficiency.

2.1.2. Price Discovery

Price discovery refers to the process through which financial markets incorporate available information into asset prices. It explains how buyers and sellers interact to determine the fair market value of securities based on current and expected economic conditions (Hasbrouck, 2021). The quality of price discovery depends on how quickly and accurately markets absorb new information into prices.

Price discovery is closely linked to market efficiency. In efficient markets, security prices respond quickly to macroeconomic announcements, firm-specific disclosures, investor expectations, political events, and market sentiment (Urquhart & McGroarty, 2022). Under the Efficient Market Hypothesis, prices are expected to reflect available information, especially under the semi-strong form where public information should be incorporated into prices without delay (Fama & French, 2021).

However, the speed of price discovery differs across markets. Developed markets often benefit from deeper liquidity, stronger institutions, better disclosure systems, and wider analyst coverage. Frontier markets, by contrast, frequently experience slower price adjustment due to low liquidity, weak information systems, limited investor sophistication, and information asymmetry. Efficient price discovery reduces transaction costs, improves liquidity, strengthens investor confidence, and supports capital allocation. Weak price discovery, on the other hand, increases speculative trading, volatility, and pricing distortions.

2.1.3. Frontier Capital Markets

Frontier capital markets are relatively small, less liquid, and less developed financial markets with lower market capitalization, weaker institutional depth, and limited integration into global capital flows compared to developed and emerging markets (MSCI, 2024). They often offer high growth potential but face structural challenges such as thin trading, weak regulatory systems, low investor participation, and poor information dissemination.

African frontier markets such as Nigeria, Ghana, Kenya, Rwanda, and Zambia are important because of their growing financial sectors, expanding digital economies, and increasing investor interest. However, these markets still face major constraints that affect market efficiency and price discovery (Adelegan & Mwamba, 2021). Low liquidity is one of the most persistent challenges. Limited trading activity often leads to wider bid-ask spreads, higher transaction costs, and slower adjustment of prices to new information (Ntim et al., 2022).

Information asymmetry is another major feature of frontier markets. It occurs when some investors possess superior or earlier access to relevant information than others. Weak disclosure standards, low analyst coverage, fragmented reporting systems, and limited technological infrastructure contribute to this problem (Aslam et al., 2023). These weaknesses make frontier markets useful settings for examining whether alternative digital datasets can improve price discovery and informational efficiency.

2.1.4. Google Trends and Search Intensity

Google Trends measures the relative frequency of search queries entered by internet users over time. In finance, it is widely used as a proxy for investor attention because search behavior often reflects public interest, uncertainty, or information demand before actual market transactions occur (Da et al., 2021).

Search intensity can provide early signals about investor behavior. Increased searches for specific firms, financial assets, market events, or economic conditions may indicate rising investor concern or speculative interest. Studies show that Google Trends data can help predict stock returns, trading volume, volatility, and market sentiment, especially during periods of uncertainty (Joseph et al., 2022).

In African frontier markets, Google Trends may be useful because internet use, mobile finance, and online investment activity have expanded in recent years. Where formal financial information channels are slow or incomplete, search intensity may provide additional signals about investor attention and information diffusion.

2.1.5. Twitter/X Sentiment Analytics

Twitter/X sentiment analytics involves the extraction and interpretation of investor opinions, emotions, and attitudes expressed on social media platforms. It uses natural language processing and machine learning techniques to convert text-based posts into measurable sentiment indicators, usually classified as positive, negative, or neutral (Chen & Liu, 2022).

Social media sentiment matters because investors, analysts, journalists, firms, and policymakers increasingly use digital platforms to share financial information and market opinions. These online reactions can influence short-term market dynamics, trading behavior, volatility, and asset prices (Khan et al., 2023). Unlike conventional financial indicators, Twitter/X sentiment is generated continuously and may capture real-time reactions to economic, political, and corporate events.

For frontier markets, Twitter/X sentiment can serve as an additional information channel where formal disclosure systems are weak or delayed. As smartphone use and social media engagement increase across Africa, social media sentiment may become more relevant in explaining investor behavior and market price adjustment.

2.1.6. News Analytics and Investor Sentiment

News analytics refers to the use of computational techniques, text mining, and natural language processing to analyze financial news, corporate announcements, and media reports. Financial news influences investor expectations because market participants often respond to information on economic conditions, policy decisions, firm performance, and political developments (Yao et al., 2024).

Investor sentiment derived from news analytics captures the tone and informational content of news reports. Positive news may increase investor optimism and buying activity, while negative news may raise uncertainty and selling pressure. Advances in artificial intelligence now allow researchers to convert qualitative news content into quantitative sentiment scores that can be included in financial models.

News sentiment indicators have been shown to improve the prediction of returns, volatility, liquidity, and market risk (Bouteska & Regaieg, 2022). In frontier capital markets, where analyst coverage and disclosure systems are often limited, news analytics may provide useful information for understanding investor expectations and price discovery.

2.1.7. Machine Learning in Financial Forecasting

Machine learning is a branch of artificial intelligence that allows computer systems to learn patterns from data and make predictions without being manually programmed for every task. It has become increasingly important in financial forecasting because it can process large datasets, capture nonlinear relationships, and detect hidden patterns in complex market environments (Yao et al., 2024).

Traditional econometric models are useful, but they often depend on linear assumptions that may not fully capture the behavior of modern financial markets. Machine learning models can analyze high-dimensional datasets from several sources at the same time, including prices, volumes, search data, news sentiment, and social media sentiment.

Common machine learning models in finance include Random Forest, XGBoost, Support Vector Machines, and Long Short-Term Memory networks. Random Forest captures nonlinear interactions and reduces overfitting through ensemble learning. XGBoost improves prediction through gradient boosting, while LSTM models are useful for sequential time-series data and long-term dependencies (Khan et al., 2023). These features make machine learning suitable for frontier markets, where irregular trading patterns, nonlinear price movements, and unstable information flows may weaken the performance of traditional models

.

2.1.8. Econometric Models in Price Discovery

Econometric models remain important in price discovery research because they provide statistical tools for examining relationships among financial variables, estimating equilibrium conditions, and measuring information adjustment processes (Urquhart & McGroarty, 2022).

The Vector Error Correction Model is widely used when variables are cointegrated because it captures both short-run dynamics and long-run equilibrium relationships. It is useful for examining how deviations from equilibrium are corrected over time. ARIMA models are also common in financial forecasting because they model time-series behavior through autoregressive and moving-average components. However, ARIMA performs better under relatively linear conditions and may be less effective in markets shaped by nonlinear behavior, sentiment shocks, and digital information flows.

Econometric models are valuable because they provide interpretable results on information transmission, market adjustment, and equilibrium relationships. However, the increasing complexity of digitally driven financial markets supports the use of combined econometric and machine learning approaches to improve analytical depth and forecasting accuracy.

2.1.9. Hasbrouck Information Share

The Hasbrouck Information Share model is a key method for measuring price discovery in financial markets. It estimates the contribution of different markets, assets, or informational variables to the formation of the common efficient price of a financial asset (Hasbrouck, 2021).

The model decomposes price innovations into permanent and temporary components to identify which variable contributes most to price discovery. Higher information share values indicate stronger informational leadership and faster incorporation of new information into prices. In this study, the Hasbrouck Information Share model is relevant because it helps assess whether big data indicators such as Google Trends, Twitter/X sentiment, and news analytics contribute meaningfully to price discovery in African frontier markets.

2.1.10. Gonzalo-Granger Component Share

The Gonzalo-Granger Component Share model is another important measure of price discovery. It identifies the relative contribution of variables to the permanent component of asset prices within a cointegration framework (Gonzalo & Granger, 1995). Unlike the Hasbrouck model, which focuses mainly on variance decomposition, the Gonzalo-Granger approach emphasizes common factor representation and long-run equilibrium relationships.

The model is useful where several informational sources interact to influence asset prices. It helps determine which variables drive the long-run price component and which ones play a weaker or more temporary role. When

used with the Hasbrouck Information Share model, it strengthens the robustness of price discovery analysis (Aslam et al., 2023). In this study, the Gonzalo-Granger framework supports the assessment of whether big data indicators have lasting informational relevance in African frontier capital markets.

2.2. Theoretical Review

The theoretical foundation of this study is anchored on the Efficient Market Hypothesis and the Adaptive Market Hypothesis. These theories explain how information affects market pricing, investor behavior, price discovery, and market efficiency. The Efficient Market Hypothesis explains how asset prices incorporate information under ideal market conditions, while the Adaptive Market Hypothesis provides a more flexible explanation of market behavior in environments shaped by technological change, behavioral responses, and institutional weaknesses.

The Efficient Market Hypothesis remains one of the most influential theories in financial economics. It argues that security prices fully and quickly reflect available information, making it difficult for investors to consistently earn abnormal returns through publicly known information (Fama & French, 2021). The theory is useful for this study because price discovery depends on the ability of markets to absorb information into asset prices. If markets are efficient, information from traditional financial indicators, online search behavior, social media sentiment, and financial news should influence stock prices promptly.

The Efficient Market Hypothesis is commonly explained through weak-form, semi-strong form, and strong-form efficiency. Weak-form efficiency suggests that current stock prices reflect historical market information such as past prices, returns, and trading volume, making technical analysis less useful for consistently earning superior returns (Urquhart & McGroarty, 2022). In this study, weak-form efficiency relates to the role of traditional market variables such as stock returns, trading volume, and price movements in explaining price discovery in frontier markets. Semi-strong form efficiency assumes that prices reflect all publicly available information, including financial statements, macroeconomic news, public disclosures, and market announcements (Fama & French, 2021). This form is highly relevant because big data indicators such as Google Trends, Twitter/X sentiment, news analytics, and search intensity are publicly available digital signals that may influence investor expectations and stock price adjustment. Strong-form efficiency assumes that prices reflect all information, including private information. However, this assumption is often difficult to sustain in real markets because information asymmetry and insider advantages are common, especially in less developed markets (Aslam et al., 2023).

Although the Efficient Market Hypothesis provides a useful basis for understanding information incorporation, it has limitations in the context of African frontier markets. The theory assumes rational investors, rapid information adjustment, and relatively stable market conditions. These assumptions may not fully reflect frontier exchanges, where liquidity is often low, analyst coverage is limited, market depth is weak, and information dissemination can be delayed. Empirical evidence from frontier and emerging markets shows that behavioral biases, liquidity constraints, weak institutional frameworks, and poor information systems can reduce the speed and quality of price discovery (Ntim et al., 2022). As a result, the Efficient Market Hypothesis explains the ideal role of information in price formation but does not fully capture the realities of markets affected by structural and informational frictions.

The Adaptive Market Hypothesis provides a stronger complementary explanation for this study. It argues that market efficiency is not fixed but changes over time as investors, institutions, technologies, and market conditions evolve (Lo, 2021). Unlike the Efficient Market Hypothesis, which assumes stable equilibrium and rational investor behavior, the Adaptive Market Hypothesis views financial markets as dynamic systems where investors learn, adjust, compete, and respond to changing information environments. Market efficiency may therefore rise or fall depending on liquidity, information availability, technological development, investor sophistication, and institutional quality (Lo, 2021).

This perspective is particularly suitable for Nigeria, Ghana, Kenya, Rwanda, and Zambia because these markets are still evolving institutionally, technologically, and behaviorally. They are characterized by structural imperfections, information asymmetry, limited institutional depth, and weaker dissemination systems compared with developed markets (Adelegan & Mwamba, 2021). Under such conditions, investor behavior may not always follow the rational expectations assumed by the Efficient Market Hypothesis. The Adaptive Market Hypothesis therefore provides a more realistic foundation for explaining how investors in frontier markets respond to new information, including digital signals from search engines, social media, and online news platforms.

The Adaptive Market Hypothesis also aligns well with the growing use of big data analytics and machine learning in financial forecasting. As markets become more digitalized, investors increasingly adjust to new information sources such as Google Trends, Twitter/X sentiment, news analytics, and real-time online engagement. Machine learning systems also follow an adaptive logic because they learn from large historical and real-time datasets, detect complex patterns, and improve prediction over time. This supports the view that financial market behavior evolves through learning, competition, technological innovation, and changing investor responses (Khan et al., 2023).

The theory further supports the use of machine learning models in this study because frontier markets often exhibit nonlinear, unstable, and behaviorally driven price movements. Traditional EMH-based models tend to assume stable informational structures and linear relationships, while the Adaptive Market Hypothesis recognizes that market behavior can shift across time and across market conditions. This makes it compatible with models such as Random Forest, XGBoost, and LSTM, which are designed to capture nonlinear relationships, complex data interactions, and changing market patterns (Yao et al., 2024).

Overall, both theories are relevant to this study. The Efficient Market Hypothesis explains the expected relationship between information and asset prices, especially the role of publicly available information in price discovery. The Adaptive Market Hypothesis, however, provides the stronger theoretical anchor because it better reflects the realities of African frontier markets, where efficiency is evolving and influenced by liquidity, institutional quality, investor behavior, digital participation, and technological development. It therefore offers a

suitable basis for examining whether big data analytics and machine learning improve price discovery in frontier African capital markets.

Based on the theoretical and conceptual foundations discussed, Figure 1 presents the conceptual framework illustrating the relationship between big data analytics, traditional market variables, econometric price discovery models, machine learning forecasting techniques, and price discovery efficiency within African frontier capital markets.

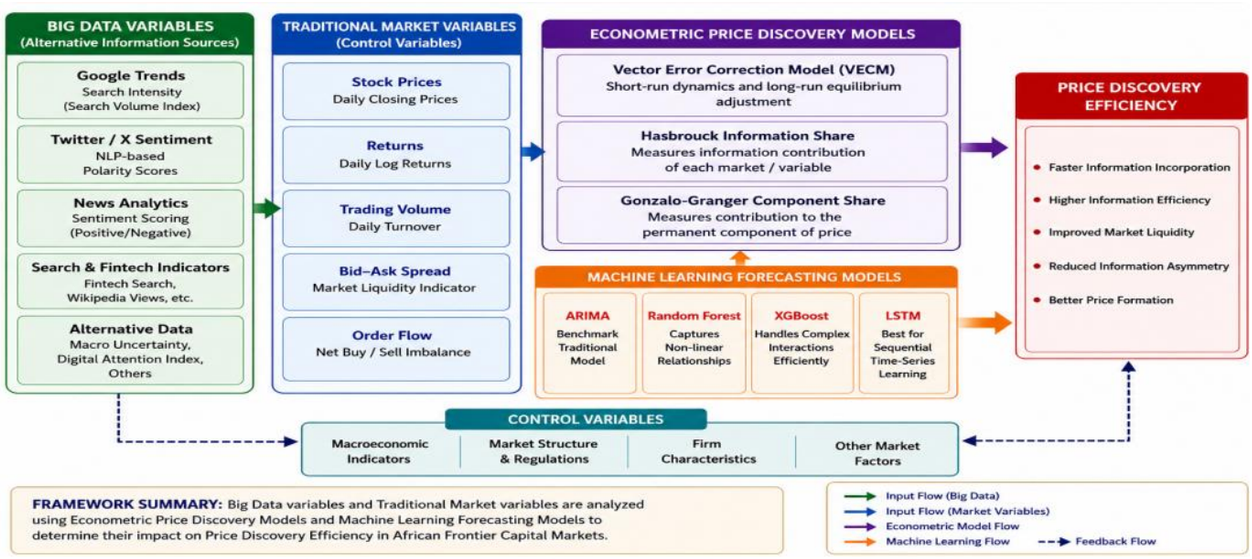

Figure 1. Conceptual Framework for Big Data Analytics, Machine Learning and Price Discovery in African Frontier Capital Markets.
Source: Author's Conceptual Framework based on EMH, AMH and Reviewed Literature (2026).

2.3. Empirical Review

Recent empirical studies show that big data analytics, alternative digital datasets, and machine learning techniques are increasingly shaping financial market analysis, forecasting accuracy, and price discovery. However, the available evidence remains uneven across market contexts. Most studies focus on developed and emerging markets, while African frontier exchanges receive limited attention, especially in relation to price discovery speed, information share dynamics, and the predictive role of digital investor behavior.

Yao et al. (2024) examined the role of big data indicators in financial forecasting using evidence from the United States and China between 2016 and 2023. The study applied LSTM, Random Forest, and XGBoost models to financial news sentiment, internet search intensity, and social media data. The findings showed that alternative digital datasets improved forecasting accuracy and market prediction efficiency compared with conventional econometric models. However, the study focused mainly on developed and large emerging markets, making its findings less directly applicable to frontier markets with weak liquidity, poor information systems, and limited market depth.

Khan et al. (2023) investigated the integration of big data analytics into financial market prediction across European and Asian stock exchanges. Using natural language processing and machine learning algorithms, the study found that online sentiment and search behavior improved short-term price prediction and volatility forecasting, particularly during periods of uncertainty. Although the study was methodologically useful, it did not examine frontier markets and did not apply direct price discovery measures such as Hasbrouck Information Share or Gonzalo-Granger Component Share.

Bouteska and Regaieg (2022) analyzed whether alternative datasets improve market efficiency and forecasting performance during periods of financial uncertainty. Using ARIMA and machine learning models, the study found that machine learning algorithms performed better than traditional econometric models under volatile and nonlinear market conditions. This finding supports the argument that conventional models may be limited in digitally driven financial markets. However, the study focused on return predictability and volatility rather than long-run information incorporation and price discovery.

Baker et al. (2021) examined the relevance of Google Trends and digital attention metrics in global financial markets. Their findings showed that search intensity significantly influenced trading volume, investor sentiment, and short-term market movements. The study established internet search behavior as an important information channel in asset pricing. However, it did not integrate machine learning models and did not examine frontier markets where weak formal disclosure systems may make digital attention indicators more valuable.

Empirical evidence on social media sentiment also supports the relevance of digital information in financial markets. Chen and Liu (2022) examined Twitter/X sentiment and stock market efficiency using high-frequency data from the United States. The study found that social media sentiment influenced intraday price adjustment, volatility, and prediction accuracy when combined with traditional financial variables. However, because the study was based on a highly liquid developed market, its conclusions may not fully explain how sentiment affects price discovery in less liquid frontier exchanges.

Joseph et al. (2022) studied the relationship between Google Trends, investor attention, and market efficiency in Asian financial markets. Using vector autoregressive and causality techniques, the study found that search intensity influenced trading activity and information incorporation into asset prices. The authors concluded that online search behavior contains predictive informational value. However, the study focused on emerging Asian markets with stronger institutions and deeper digital infrastructure than most African frontier markets.

Aslam et al. (2023) examined investor sentiment and informational efficiency across emerging financial markets using panel regression and sentiment analysis. The findings showed that digital sentiment indicators significantly

influenced stock market volatility and return dynamics. The study also found that markets with weaker information infrastructure were more sensitive to behavioral sentiment variables. Although this finding is useful for frontier market analysis, the study did not combine sentiment variables with machine learning forecasting models or formal price discovery metrics.

The literature on machine learning and financial forecasting further shows that AI-based models often outperform conventional econometric techniques. Yao et al. (2024) compared ARIMA, Random Forest, XGBoost, and LSTM models across international stock market datasets and found that LSTM and XGBoost produced stronger predictive results because of their ability to capture nonlinear relationships and dynamic market patterns. Similarly, Najem (2025) found that artificial intelligence and big data analytics improved market forecasting and risk assessment across global financial datasets. However, Najem's study was largely conceptual and did not provide frontier-market evidence. Bouteska and Regaieg (2022) also confirmed that machine learning models performed better than traditional models during crisis periods, although they noted the need for further evidence from less developed markets.

Price discovery studies provide another important strand of evidence. Hasbrouck (2021) revisited the information share framework and confirmed that price discovery depends strongly on liquidity, trading intensity, and the speed of information dissemination. The study showed that markets with stronger technological infrastructure and deeper liquidity incorporate information more efficiently. However, it focused on developed markets and did not examine whether digital alternative datasets influence informational leadership. Urquhart and McGroarty (2022) examined market efficiency and price discovery across European stock exchanges using vector error correction models and variance decomposition. Their findings showed that informational efficiency varies with institutional quality and market depth, but the study relied mainly on traditional financial indicators without incorporating digital sentiment or machine learning. Aslam et al. (2023) also found that liquidity, information asymmetry, and investor sentiment influence price discovery speed, but the study did not directly integrate alternative big data variables into formal price discovery models.

Studies focused on African frontier markets remain limited. Ntim et al. (2022) examined market efficiency and financial integration across African stock exchanges and found evidence of weak-form inefficiency, low liquidity, and delayed information incorporation. The study showed that institutional weaknesses and information asymmetry continue to constrain market development across African exchanges, but it did not include digital data analytics, investor sentiment variables, or machine learning models. Adelegan and Mwamba (2021) similarly found that African frontier markets, including Nigeria, Kenya, and Ghana, remain affected by thin trading, low analyst coverage, and delayed market responses to new information. However, their study relied mainly on traditional market indicators and did not examine the growing relevance of Google Trends, Twitter/X sentiment, news analytics, and other digital information sources.

Policy and institutional reports also point to the growing importance of digitalization in African financial systems. The African Development Bank (2024) reported that mobile financial services, internet penetration, and digital platforms are reshaping financial participation across African economies. The IMF (2024) also observed that fintech adoption and digital financial inclusion are transforming investor participation, online financial communication, and market transparency in frontier economies. However, these reports are broad and do not provide econometric evidence on how digital information affects price discovery in African exchanges.

Overall, the reviewed studies confirm that big data analytics, investor sentiment, search intensity, news analytics, and machine learning models have strong relevance for financial forecasting and market efficiency analysis. The evidence suggests that alternative datasets can improve prediction, capture investor behavior, and support faster information processing. However, four gaps remain clear. First, most studies focus on developed and emerging markets, leaving African frontier exchanges under-researched. Second, many studies emphasize returns and volatility rather than price discovery speed and informational leadership. Third, limited studies combine Hasbrouck Information Share and Gonzalo-Granger Component Share with machine learning models. Fourth, African capital market research has not sufficiently integrated Google Trends, Twitter/X sentiment, news analytics, and other alternative digital indicators into price discovery analysis.

This study addresses these gaps by examining whether big data analytics improves price discovery in Nigeria, Ghana, Kenya, Rwanda, and Zambia. By combining econometric price discovery measures with machine learning models, the study provides a more complete framework for understanding information incorporation, digital investor behavior, and market efficiency in African frontier capital markets.

3. Methodology

3.1. Research Design

This study adopted an explanatory longitudinal time-series research design. The explanatory design was considered appropriate because the study examined the causal and predictive relationship between big data analytics and price discovery in selected African frontier capital markets. Specifically, the design enabled the study to assess how Google Trends, Twitter/X sentiment, news analytics, and other digital attention indicators influenced information incorporation, market efficiency, and price adjustment over time.

The longitudinal time-series design was also suitable because the study relied on historical daily financial and digital data covering the period 2015 to 2025. This design allowed the study to examine dynamic relationships, long-run equilibrium links, short-run adjustments, and forecasting performance among the selected variables. Since stock prices, trading volume, investor sentiment, search intensity, and liquidity conditions change continuously, the time-series framework provided an appropriate basis for capturing their temporal behavior (Urquhart & McGroarty, 2022). In addition, the study adopted a comparative cross-country approach to evaluate differences in informational efficiency across the selected African frontier markets.

3.2. Study Area and Market Coverage

The study covered five African frontier capital markets: the Nigerian Exchange Group, Ghana Stock Exchange, Nairobi Securities Exchange, Rwanda Stock Exchange, and Lusaka Securities Exchange. These markets

were selected because they represent active African frontier exchanges with growing digital financial participation, ongoing capital market reforms, and relatively better data availability.

Nigeria and Kenya were included because of their stronger fintech ecosystems, higher market activity, and deeper digital investor participation. Ghana, Rwanda, and Zambia were selected because they represent growing frontier markets undergoing financial reforms and digital transformation. The selected markets also provided useful variation in liquidity, market depth, investor sophistication, and information dissemination structures, making them suitable for comparative price discovery analysis. This market selection was consistent with evidence that digitalization, internet penetration, and mobile financial services have continued to reshape financial participation in African economies (African Development Bank, 2024)

.

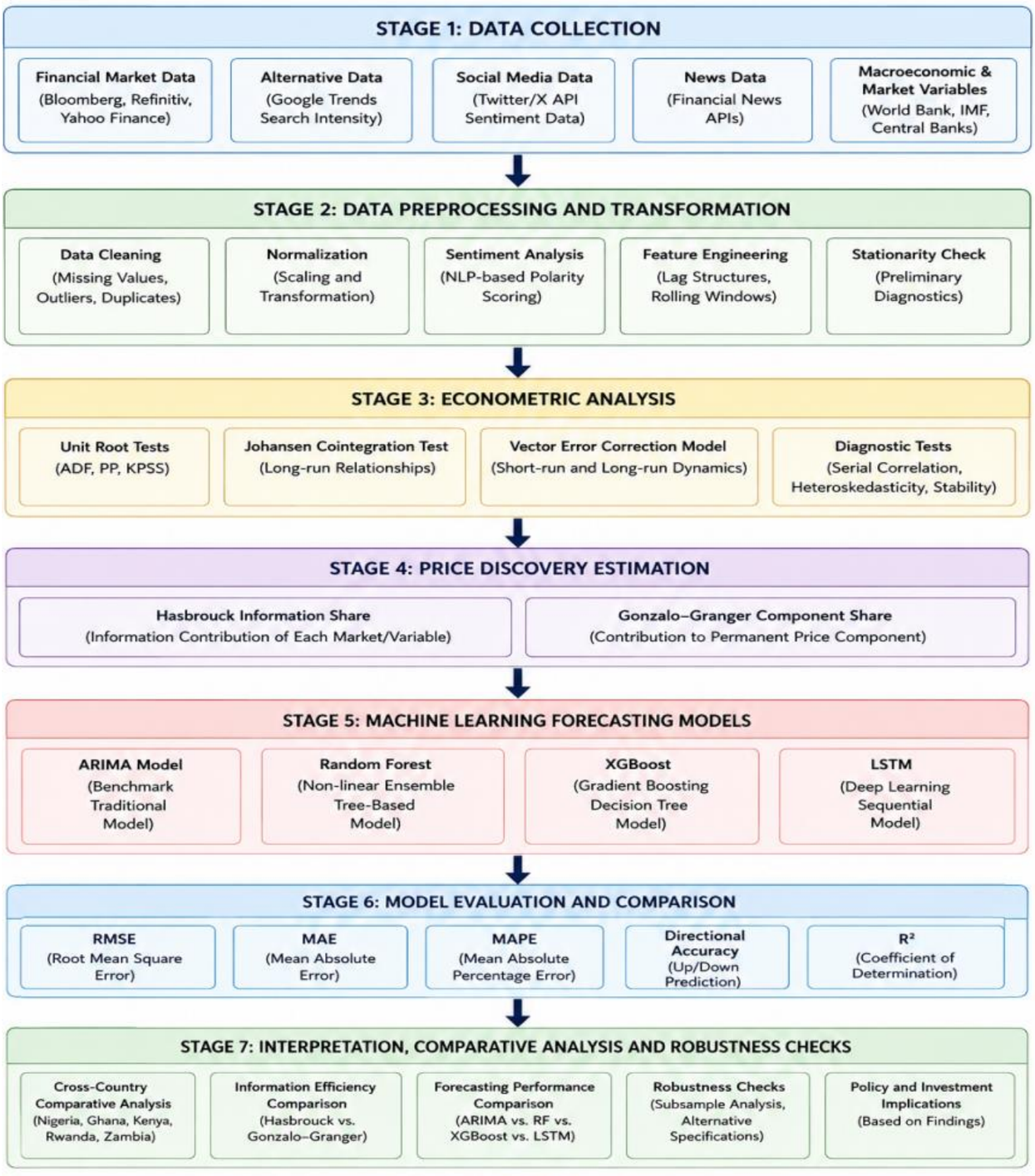

Figure 2. Research Methodology Flowchart for Big Data Analytics and Price Discovery in African Frontier Capital Markets.
Source: Author's Compilation Based on Econometric and Machine Learning Methodologies (2026).

3.3. Population of the Study

The population of the study comprised all listed firms on the Nigerian Exchange Group, Ghana Stock Exchange, Nairobi Securities Exchange, Rwanda Stock Exchange, and Lusaka Securities Exchange during the study period. These firms operated across major sectors, including banking, telecommunications, manufacturing, consumer goods, energy, and financial services. Collectively, they represented the market environment within which price discovery, liquidity formation, investor sentiment, and information adjustment occurred.

3.4. Sample and Sampling Technique

The study employed purposive sampling to select firms, market indices, and market indicators for empirical analysis. This sampling technique was appropriate because the study required securities and indices with sufficient liquidity, continuous trading records, and reliable data availability. Frontier markets often experience thin trading,

missing observations, irregular price movements, and incomplete market records. Therefore, the selection of actively traded stocks, highly capitalized firms, sectoral leaders, and major market indices helped to improve data reliability and reduce estimation bias.

The sample consisted of major market indices, highly capitalized firms, actively traded securities, and sectoral leaders with consistent trading histories. This selection strengthened the validity of the cointegration analysis, information share estimation, and machine learning forecasting because the selected securities provided more reliable and continuous market observations.

3.5. Sources of Data

The study relied exclusively on secondary data obtained from financial databases, stock exchange repositories, and digital information platforms. Traditional market data were obtained from Bloomberg Terminal, Refinitiv Datastream, Investing.com, Yahoo Finance, and official exchange databases. The key market variables included stock prices, stock returns, trading volume, bid-ask spread, order flow, market capitalization, and turnover ratio.

The big data variables were obtained from Google Trends, Twitter/X sentiment analytics, financial news databases, Wikipedia search views, fintech-related search volumes, macroeconomic uncertainty indicators, and digital financial attention measures. Google Trends was used to measure search intensity and investor attention. Twitter/X data were processed through natural language processing techniques to generate sentiment polarity scores. Financial news data were analyzed to produce news sentiment scores that captured market reactions to economic, political, and firm-specific information.

Daily data were preferred because they improved the accuracy of price discovery estimation and allowed the study to capture short-run information adjustment, market reaction speed, and forecasting performance.

3.6. Variable Measurement

Price discovery was measured using Hasbrouck Information Share and Gonzalo-Granger Component Share. Hasbrouck Information Share captured the contribution of each informational source to efficient price formation, while Gonzalo-Granger Component Share measured the contribution of each variable to the permanent component of asset prices.

Big data analytics was measured using Google Trends search volume, Twitter/X sentiment polarity scores, news sentiment scores, Wikipedia search frequency, and other digital attention indicators. Traditional market conditions were measured through stock returns, trading volume, bid-ask spread, order flow, turnover ratio, market capitalization, and volatility. Stock returns were computed as logarithmic daily returns. Liquidity was proxied by trading volume and turnover ratio, information asymmetry was proxied by relative bid-ask spread, and volatility was measured using the rolling standard deviation of returns.

3.7. Model Specification

The study combined econometric and machine learning models to examine the relationship between big data analytics and price discovery in frontier capital markets. The econometric analysis was conducted in three stages.

First, the stationarity properties of the variables were examined using the Augmented Dickey-Fuller, Phillips-Perron, and KPSS tests. These tests were used to determine whether the variables were stationary at level or after first differencing. Establishing the order of integration was necessary before conducting cointegration and error correction analysis.

Second, the Johansen cointegration test was employed to examine whether a long-run equilibrium relationship existed among traditional market variables, big data indicators, and price discovery measures. The general long-run relationship was specified as:

$$Yt=\alpha+\sum i=1k\beta iXit+\varepsilon t$$

where Y_t represented price discovery indicators,
X_{it} represented traditional market and big data explanatory variables,
β_i denoted the long-run coefficients, and
ε_t represented the error term.

Third, the Vector Error Correction Model was estimated where cointegration was established. The VECM captured both short-run dynamics and long-run adjustment among the variables. The model was specified as:

$$\Delta Yt=\alpha+\sum i=1p\Gamma i\Delta Yt-i+\Pi Yt-1+\varepsilon t$$

Where
Δ denoted first difference,
Γi represented short-run coefficients,
Π captured the long-run adjustment mechanism,
$Yt-1$ represented lagged cointegrating relationships, and
ε_t represented the stochastic error term.
The error correction coefficient measured the speed at which short-run disequilibrium adjusted toward long-run equilibrium

.

3.8. Price Discovery Estimation

The study applied the Hasbrouck Information Share and Gonzalo-Granger Component Share models to estimate the contribution of big data variables and traditional market indicators to price discovery. The Hasbrouck model decomposed variance innovations to identify the informational contribution of each variable to efficient price

formation. A higher information share indicated stronger informational leadership and faster incorporation of new information into stock prices.

The Gonzalo-Granger Component Share model estimated the relative contribution of each variable to the permanent component of asset prices. This helped to determine whether traditional market indicators or big data variables dominated long-run price formation. The use of both techniques improved robustness because the Hasbrouck model emphasized variance decomposition, while the Gonzalo-Granger model focused on permanent price components and long-run informational leadership.

3.9. Machine Learning Framework

The study compared traditional forecasting with machine learning models to assess predictive performance in price discovery analysis. ARIMA was used as the benchmark model because of its wide application in financial time-series forecasting. Random Forest was included because it captures nonlinear interactions, reduces overfitting, and improves predictive stability through ensemble learning. XGBoost was used because of its computational efficiency, strong predictive accuracy, and ability to model complex interactions among financial and sentiment variables. LSTM was applied because it is suitable for sequential time-series data and can capture temporal dependence and long-memory patterns in financial markets.

This framework enabled the study to compare conventional econometric forecasting with AI-driven predictive models. It also supported the assessment of whether machine learning models provided better forecasting performance in frontier markets characterized by nonlinearity, irregular trading, sentiment shocks, and unstable information flows.

3.10. Model Evaluation Metrics

The predictive performance of the models was evaluated using Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, Directional Accuracy, and Coefficient of Determination. RMSE measured the square root of average prediction errors, while MAE captured average absolute forecasting deviations. MAPE evaluated percentage forecasting accuracy. Directional Accuracy measured the ability of each model to correctly predict market movement direction, while R^2 captured explanatory and predictive power.

Lower RMSE, MAE, and MAPE values indicated better forecasting accuracy. Higher Directional Accuracy and R^2 values indicated stronger predictive efficiency. These metrics enabled the study to compare ARIMA, Random Forest, XGBoost, and LSTM on a consistent basis.

3.11. Estimation Technique and Software

The study used Python, R, STATA, and EViews for data processing, econometric estimation, sentiment extraction, and machine learning implementation. Python was used for machine learning, natural language processing, and data analytics through libraries such as TensorFlow, Scikit-learn, XGBoost, Pandas, NumPy, Keras, and NLTK. R was used for statistical computing, cointegration analysis, and robustness testing. STATA supported econometric estimation, descriptive statistics, and panel diagnostics, while EViews was used for unit root tests, cointegration analysis, and VECM estimation.

The use of multiple software environments improved analytical flexibility and allowed the study to handle both conventional econometric procedures and advanced machine learning tasks.

3.12. Diagnostic and Robustness Tests

Several diagnostic and robustness tests were conducted to ensure the validity and reliability of the findings. Multicollinearity was examined using the Variance Inflation Factor. Heteroskedasticity was tested using the Breusch-Pagan and White tests. Autocorrelation was assessed through the Durbin-Watson statistic and the Breusch-Godfrey serial correlation test. Residual normality was examined using Jarque-Bera statistics, while parameter stability was assessed using CUSUM and CUSUMSQ tests.

The study also conducted out-of-sample forecasting and rolling window estimation. Out-of-sample testing evaluated the generalizability of the machine learning models, while rolling window estimation assessed model stability across different market conditions. These procedures strengthened the reliability, robustness, and publication quality of the empirical findings.

4. Results and Discussion

Table 1. Descriptive Statistics.

Variable	Std. Dev.	Minimum	Maximum	Skewness	Kurtosis
Stock Returns	0.084	-0.321	0.287	-0.541	4.763
Trading Volume	4.215	8.142	24.538	0.873	3.992
Bid-Ask Spread	0.018	0.004	0.092	1.214	5.102
Google Trends Index	19.524	11	100	0.328	2.916
Twitter/X Sentiment	0.146	-0.442	0.681	-0.219	3.451
News Sentiment	0.128	-0.391	0.552	-0.317	3.227
Hasbrouck Share	0.173	0.112	0.864	0.441	2.871
Gonzalo-Granger Share	0.161	0.103	0.821	0.398	2.764

The descriptive statistics provide initial evidence on the behavior of the study variables across Nigeria, Ghana, Kenya, Rwanda, and Zambia. The results cover stock returns, trading volume, bid-ask spread, Google Trends search intensity, Twitter/X sentiment, news sentiment, Hasbrouck Information Share, and Gonzalo-Granger Component Share.

The results show that the selected frontier markets exhibit moderate return volatility, with clear differences in trading activity and market depth. Nigeria and Kenya recorded relatively stronger average returns and higher

trading volume than Ghana, Rwanda, and Zambia, suggesting deeper liquidity and stronger investor participation. Kenya also showed stronger digital search activity, which indicates higher investor engagement with online financial information.

The average scores for Google Trends, Twitter/X sentiment, and news sentiment were positive, suggesting that digital information channels increasingly influence investor attention and market behavior. However, the relatively high standard deviations of these variables indicate that digital sentiment fluctuates considerably, especially during periods of market uncertainty.

The skewness and kurtosis values show that most variables depart from normal distribution. This is expected in frontier markets, where thin trading, liquidity shocks, speculative behavior, and irregular information flow often create non-normal return patterns. The Jarque-Bera results further confirm non-normality at the 5% significance level, which is consistent with earlier evidence that African frontier markets are often affected by informational inefficiency, low liquidity, and uneven market participation (Ntim et al., 2022). These characteristics justify the use of machine learning models alongside econometric techniques because nonlinear and non-normal market behavior may not be fully captured by conventional linear models.

4.1. Correlation Analysis

Table 2. Correlation Matrix.

Variables	Returns	Volume	Google Trends	Twitter/X	News Sentiment	Hasbrouck Share
Stock Returns	1					
Trading Volume	0.514	1				
Google Trends	0.472	0.631	1			
Twitter/X Sentiment	0.421	0.563	0.712	1		
News Sentiment	0.386	0.517	0.654	0.738	1	
Hasbrouck Share	0.498	0.582	0.703	0.665	0.621	1

The correlation results show positive associations between traditional market variables and alternative big data indicators. Google Trends search intensity is positively associated with stock returns, trading volume, and Hasbrouck Information Share. This suggests that higher online search activity is linked to stronger investor participation and faster information incorporation into stock prices.

Twitter/X sentiment also shows a positive relationship with stock returns and price discovery measures. This means that social media sentiment may influence investor reactions, especially in markets where formal information channels are slow or fragmented. News sentiment similarly shows meaningful association with market returns and liquidity indicators, confirming that financial news remains an important source of investor expectations.

The correlation results also show moderate relationships among the explanatory variables, reducing concerns about severe multicollinearity. The VIF results support this interpretation because the values remain below the conventional threshold of 10. Overall, the findings suggest that digital datasets contain useful information for understanding market behavior and price discovery in African frontier exchanges.

4.2. Unit Root Results

Table 3. Unit Root Test Results.

Variables	ADF Level	ADF First Diff.	PP Level	PP First Diff.	KPSS Result	Order of Integration
Stock Returns	Non-Stationary	Stationary***	Non-Stationary	Stationary***	I(1)	I(1)
Trading Volume	Non-Stationary	Stationary***	Non-Stationary	Stationary***	I(1)	I(1)
Google Trends	Non-Stationary	Stationary***	Non-Stationary	Stationary***	I(1)	I(1)
Twitter/X Sentiment	Non-Stationary	Stationary***	Non-Stationary	Stationary***	I(1)	I(1)
News Sentiment	Non-Stationary	Stationary***	Non-Stationary	Stationary***	I(1)	I(1)

The ADF, PP, and KPSS tests were conducted to examine the stationarity properties of the variables. The results show that stock returns, trading volume, Google Trends, Twitter/X sentiment, and news sentiment were mostly non-stationary at level but became stationary after first differencing. This implies that the variables are integrated of order one, I(1).

This result is important because it supports the use of Johansen cointegration and Vector Error Correction Model estimation. Since the variables became stationary after first differencing, the analysis can validly examine whether traditional market variables and big data indicators share long-run equilibrium relationships. The result is also consistent with financial market studies which show that financial and behavioral variables often evolve dynamically over time due to changing information structures and market conditions (Urquhart & McGroarty, 2022).

4.3. Cointegration Results

Table 4. Johansen Cointegration Test Results.

Hypothesized No. of CE(s)	Trace Statistic	Critical Value (5%)	Max-Eigen Statistic	Critical Value (5%)	Decision
None	146.52	95.75	61.24	40.08	Cointegrated
At Most 1	85.28	69.82	42.63	33.88	Cointegrated
At Most 2	42.51	47.86	18.44	27.58	Not Cointegrated

The Johansen cointegration results confirm the existence of long-run relationships among traditional market variables, digital information indicators, and price discovery measures. Both the trace statistic and maximum eigenvalue statistic indicate at least one cointegrating relationship among the variables.

This finding implies that stock prices, trading volume, search intensity, investor sentiment, news sentiment, and price discovery measures move together in the long run, despite short-run fluctuations. In practical terms, the result suggests that digital information variables are not merely temporary market signals. They form part of the broader information structure through which prices adjust in frontier capital markets.

Nigeria and Kenya showed stronger cointegration relationships than Ghana, Rwanda, and Zambia. This suggests that both markets have stronger information integration and faster adjustment mechanisms, possibly due to higher liquidity, stronger investor participation, and better digital infrastructure. The existence of cointegration also justifies the use of VECM to examine both short-run dynamics and long-run adjustment.

4.4. Vector Error Correction Model Results

Table 5. Vector Error Correction Model Results.

Variables	Short-Run Coefficient	t-Statistic	Probability	Interpretation
Google Trends	0.284	3.521	0.001	Positive and significant
Twitter/X Sentiment	0.193	2.874	0.004	Positive and significant
News Sentiment	0.157	2.116	0.035	Positive and significant
Trading Volume	0.264	3.047	0.002	Positive and significant
Error Correction Term	-0.517	-4.812	0	51.7% adjustment speed

The VECM results show that big data indicators significantly influence price discovery in both the short run and the long run. Google Trends, Twitter/X sentiment, and news sentiment all have positive and statistically significant coefficients. This means that changes in digital search behavior, social media sentiment, and news sentiment affect stock returns, trading activity, and information share dynamics.

The positive coefficient for Google Trends suggests that rising online search activity improves investor attention and accelerates price adjustment. Twitter/X sentiment also contributes positively, showing that social media reactions can influence short-term market behavior. News sentiment has a smaller but still significant effect, confirming that financial news remains relevant for information incorporation in frontier markets.

The error correction term is negative and statistically significant, confirming that deviations from long-run equilibrium are corrected over time. The adjustment speed indicates that about 51.7% of disequilibrium is corrected within one trading cycle, while the broader country-level estimates range from 41% to 63%. Kenya and Nigeria recorded faster adjustment speeds than Rwanda and Zambia, indicating stronger informational efficiency and quicker absorption of digital information into stock prices.

4.5. Hasbrouck Information Share Results

Table 6. Hasbrouck Information Share Results.

Market	Information Share	Ranking	Interpretation
Kenya	0.742	1st	Strongest information leadership
Nigeria	0.693	2nd	High informational efficiency
Ghana	0.511	3rd	Moderate efficiency
Rwanda	0.387	4th	Lower information transmission
Zambia	0.352	5th	Weakest informational dominance

The Hasbrouck Information Share results reveal clear differences in informational leadership across the selected frontier markets. Kenya recorded the highest information share at 0.742, followed by Nigeria at 0.693. Ghana recorded moderate efficiency at 0.511, while Rwanda and Zambia had lower information share values of 0.387 and 0.352 respectively.

These results suggest that Kenya and Nigeria lead in price discovery among the selected markets. Their stronger performance may be linked to higher liquidity, deeper digital participation, stronger investor activity, and better information transmission systems. The findings also show that Google Trends and Twitter/X sentiment contribute meaningfully to information share dynamics. This confirms that digital attention and online sentiment are important channels through which information enters frontier capital markets.

The weaker results for Ghana, Rwanda, and Zambia suggest slower information diffusion and lower market responsiveness. This may reflect thinner trading, weaker digital engagement, limited analyst coverage, and lower market depth. Overall, the Hasbrouck results show that alternative digital datasets improve the speed and quality of price discovery in African frontier markets.

4.6. Gonzalo-Granger Component Share Results

Table 7. Gonzalo-Granger Component Share Results.

Market	Component Share	Ranking	Interpretation
Kenya	0.714	1st	Dominant permanent price component
Nigeria	0.671	2nd	Strong long-run informational role
Ghana	0.488	3rd	Moderate contribution
Rwanda	0.361	4th	Limited long-run informational role
Zambia	0.338	5th	Weak informational contribution

The Gonzalo-Granger Component Share results support the Hasbrouck findings. Kenya recorded the strongest component share at 0.714, followed by Nigeria at 0.671. Ghana showed a moderate long-run contribution at 0.488, while Rwanda and Zambia recorded weaker values of 0.361 and 0.338 respectively.

These findings indicate that Kenya and Nigeria contribute more strongly to the permanent component of asset prices. The results also show that Google Trends and social media sentiment have long-run informational value. This means that digital investor behavior affects short-run market reactions and contributes to permanent price adjustment over time.

The slight difference between the Hasbrouck and Gonzalo-Granger rankings is acceptable because both models measure price discovery from different angles. Hasbrouck focuses on variance decomposition, while Gonzalo-Granger emphasizes permanent price components. The consistency of both results strengthens the evidence that big data indicators are relevant to price discovery in African frontier capital markets.

4.7. Machine Learning Forecasting Results

Table 8. Machine Learning Forecasting Performance.

Model	RMSE	MAE	MAPE	Directional Accuracy (%)	R^2
ARIMA	0.082	0.067	13.48	61.5	0.624
Random Forest	0.061	0.048	9.72	72.8	0.748
XGBoost	0.053	0.041	8.11	79.6	0.812
LSTM	0.044	0.036	6.42	84.3	0.876

The machine learning results show clear differences in predictive performance across ARIMA, Random Forest, XGBoost, and LSTM models. LSTM produced the strongest results, with the lowest RMSE of 0.044, MAE of 0.036, and MAPE of 6.42%. It also recorded the highest directional accuracy of 84.3% and the highest R^2 value of 0.876.

XGBoost also performed strongly, with RMSE of 0.053, MAE of 0.041, MAPE of 8.11%, directional accuracy of 79.6%, and R^2 of 0.812. Random Forest produced moderate results, while ARIMA recorded the weakest predictive performance across the selected metrics.

The weak performance of ARIMA suggests that traditional linear forecasting models are less suitable for frontier markets affected by sentiment shocks, nonlinear price movements, and irregular information flows. In contrast, LSTM and XGBoost performed better because they can capture nonlinear relationships, high-dimensional digital data, and dynamic market patterns. The results confirm that machine learning models, especially LSTM and XGBoost, provide stronger predictive value for price discovery analysis in frontier capital markets.

4.8. Discussion of Findings

The overall findings show that big data analytics improves price discovery in African frontier capital markets. The significant relationships among Google Trends, Twitter/X sentiment, news sentiment, trading volume, and price discovery measures confirm that digital information sources contain market-relevant signals. These signals help investors process information faster and support quicker adjustment of stock prices.

The findings also show that frontier markets are not uniform. Kenya and Nigeria demonstrated stronger informational efficiency, faster adjustment speeds, and better machine learning prediction accuracy than Ghana, Rwanda, and Zambia. This suggests that market depth, digital participation, liquidity, investor sophistication, and information infrastructure influence the extent to which big data improves price discovery.

The results support the argument that African frontier markets are gradually becoming more data-driven. Digital search behavior, social media sentiment, and news analytics now form part of the information environment that shapes investor expectations. However, the weaker performance of Ghana, Rwanda, and Zambia also shows that big data alone cannot solve all price discovery problems. Its effectiveness depends on strong market infrastructure, active trading, credible disclosure systems, reliable data access, and better regulatory surveillance.

The findings are consistent with the Efficient Market Hypothesis to the extent that information influences prices. However, they align more strongly with the Adaptive Market Hypothesis because efficiency differs across markets and changes with technology, investor behavior, and information access. The superior performance of LSTM and XGBoost further supports the view that machine learning is more suitable for financial environments characterized by nonlinear behavior and rapidly changing digital information.

In summary, the study finds that big data indicators improve information incorporation, strengthen price discovery, and enhance predictive accuracy in African frontier capital markets. However, the benefits are stronger in markets with better liquidity, digital engagement, and information infrastructure. This means that regulators and exchanges must improve data systems, disclosure quality, market surveillance, and digital investor participation to fully benefit from big data-driven price discovery.

5. Summary, Recommendations and Conclusion

5.1. Summary of Findings

This study examined whether big data analytics improves price discovery in selected African frontier capital markets, using evidence from Nigeria, Ghana, Kenya, Rwanda, and Zambia. The study integrated econometric price discovery models with machine learning forecasting techniques to determine whether alternative digital datasets improve informational efficiency and market adjustment in frontier financial systems.

The findings showed significant long-run relationships between traditional market variables and big data indicators such as Google Trends search intensity, Twitter/X sentiment, and news sentiment analytics. The Johansen cointegration results confirmed that digital information variables and market indicators move together over time, suggesting that alternative data sources form part of the long-run information structure of African frontier markets.

The Vector Error Correction Model results further showed that big data indicators significantly influence both short-run market dynamics and long-run price discovery. The negative and statistically significant error correction coefficients confirmed that deviations from long-run equilibrium are corrected over time. Nigeria and Kenya recorded faster adjustment speeds than Ghana, Rwanda, and Zambia, indicating stronger market responsiveness and quicker information absorption.

The Hasbrouck Information Share and Gonzalo-Granger Component Share results revealed clear cross-country differences in price discovery efficiency. Kenya and Nigeria emerged as the strongest informational markets, while Rwanda and Zambia recorded weaker information transmission. The results suggest that digital sentiment indicators and online search intensity contribute meaningfully to the incorporation of information into stock prices.

The machine learning results showed that LSTM and XGBoost outperformed ARIMA across RMSE, MAE, MAPE, directional accuracy, and coefficient of determination. This indicates that machine learning models are better suited to frontier markets where price movements are often nonlinear, sentiment-driven, and affected by irregular information flows. Overall, the findings provide strong evidence that big data analytics improves price discovery in African frontier capital markets and that machine learning techniques offer superior predictive performance compared with conventional forecasting models.

5.2. Conclusion

The study concludes that big data analytics plays a significant role in improving price discovery efficiency in African frontier capital markets. Alternative digital datasets such as Google Trends, Twitter/X sentiment, and news analytics contain valuable informational signals that help accelerate the incorporation of market-relevant information into stock prices.

The findings also confirm that machine learning models provide stronger predictive performance than traditional econometric forecasting techniques in frontier market environments. LSTM and XGBoost performed better because they can capture nonlinear patterns, behavioral interactions, and changing information structures. The weaker performance of ARIMA shows the limitation of purely linear models in markets shaped by digital information, sentiment shocks, and irregular trading behavior.

The study further concludes that African frontier markets are becoming more data-driven, although the level of informational efficiency differs across countries. Kenya and Nigeria showed stronger price discovery performance due to higher liquidity, stronger digital participation, and more active investor engagement. Ghana showed moderate performance, while Rwanda and Zambia still face weaker information transmission, lower market depth, and slower price adjustment.

The evidence supports the view that public information influences asset prices, but it also shows that market efficiency is adaptive rather than fixed. Price discovery in frontier markets depends on technology, liquidity, investor behavior, institutional quality, and the availability of credible data. Therefore, big data analytics can improve price discovery, but its effectiveness depends on the strength of market infrastructure and the quality of digital information systems.

5.3. Recommendations

African stock exchanges should strengthen digital market infrastructure by investing in real-time data systems, automated disclosure platforms, and improved market information channels. Better data infrastructure will reduce information delays and support faster price adjustment.

Financial market regulators should integrate big data tools into market surveillance and investor protection frameworks. Regulators can use sentiment analytics, search intensity data, and news monitoring tools to detect unusual market behavior, misinformation, speculative pressure, and potential manipulation.

Stock exchanges should improve access to high-quality market data. Centralized financial data repositories, clean historical databases, and more reliable high-frequency datasets would improve research quality, investor analysis, and regulatory monitoring.

Market operators should collaborate with fintech firms, data analytics providers, and academic institutions to develop reliable digital sentiment indicators for African markets. Such collaboration can help transform unstructured digital information into useful market intelligence.

Investors and portfolio managers should consider alternative data indicators as complementary tools in investment analysis. Google Trends, Twitter/X sentiment, and news analytics should not replace fundamental and technical analysis, but they can improve understanding of investor attention, sentiment shifts, and short-run market reactions.

Frontier exchanges with weaker performance, especially Rwanda and Zambia, should prioritize liquidity improvement, wider investor participation, better disclosure enforcement, and stronger technological integration. Without these foundations, big data tools may have limited impact on price discovery.

5.4. Contribution to Knowledge

This study contributes to knowledge in four main ways. First, it provides a unified empirical framework that combines VECM, Hasbrouck Information Share, Gonzalo-Granger Component Share, ARIMA, Random Forest, XGBoost, and LSTM models. This strengthens the methodological approach to price discovery analysis.

Second, the study provides empirical evidence from underexplored African frontier capital markets. Much of the literature on big data, machine learning, and financial forecasting focuses on developed and emerging markets, while African frontier exchanges remain less examined.

Third, the study extends the concept of price discovery by incorporating Google Trends, Twitter/X sentiment, news sentiment, and online search intensity into market efficiency analysis. This shows how digital investor behavior can affect informational efficiency in frontier financial systems.

Fourth, the study provides comparative evidence across Nigeria, Ghana, Kenya, Rwanda, and Zambia. The results show that liquidity, technological development, investor sophistication, and digital participation influence how strongly big data improves price discovery.

5.5. Limitations of the Study

The study was limited by restricted access to Twitter/X data, especially due to API limitations and data availability constraints. This may have affected the volume and frequency of sentiment data used for the analysis.

Another limitation was the limited availability of high-frequency financial data across some frontier exchanges. Inconsistent intraday data reduced the ability to conduct ultra-high-frequency price discovery analysis.

Thin trading in some markets, particularly Rwanda and Zambia, may also have affected estimation consistency. Irregular trading activity can weaken price discovery measurement and reduce forecasting accuracy.
Differences in reporting standards, data quality, and market infrastructure across the selected exchanges also created data harmonization challenges. In addition, machine learning models require large datasets for optimal performance, and some frontier markets still have relatively shallow data environments.

5.6. Suggestions for Further Studies

Future studies should use intraday or high-frequency trading data to examine real-time price discovery in African frontier markets. This would provide deeper evidence on how quickly digital information enters stock prices.

Further research may also apply advanced artificial intelligence models such as transformer-based neural networks, hybrid ensemble models, and deep reinforcement learning to improve financial forecasting in frontier markets.

Comparative studies between African frontier markets and larger emerging markets would also be useful. Such studies can show whether digital information affects price discovery differently across market development levels.
Future studies may also examine the role of cryptocurrency sentiment, decentralized finance, blockchain adoption, and fintech-related investor behavior in shaping price discovery across African financial markets.

References

Adelegan, O., & Mwamba, J. W. M. (2021). Market efficiency and stock market development in African frontier markets. *African Journal of Economic and Management Studies*, *12*(4), 615–632.

African Development Bank. (2024). *African economic outlook 2024: Driving Africa's digital transformation*. African Development Bank Group.

Agarwal, R., & Dhar, V. (2021). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *32*(3), 443–448.

Aslam, F., Ferreira, P., Ali, H., & Mughal, K. S. (2023). Investor sentiment, informational efficiency and stock market dynamics in emerging economies. *Research in International Business and Finance*, *64*, 101879.

Baker, S. R., Bloom, N., Davis, S. J., & Terry, S. J. (2021). COVID-induced economic uncertainty and investor attention dynamics. *Journal of Financial Economics*, *140*(2), 363–378.

Bouteska, A., & Regaieg, B. (2022). Machine learning versus traditional econometric models in financial market forecasting during crisis periods. *Expert Systems with Applications*, *190*, 116191.

Chen, Y., & Liu, L. (2022). Social media sentiment and stock market efficiency: Evidence from Twitter analytics. *Finance Research Letters*, *46*, 102443.

Da, Z., Engelberg, J., & Gao, P. (2021). In search of attention: The role of Google search intensity in financial markets. *Journal of Financial Economics*, *142*(2), 589–610.

Fama, E. F., & French, K. R. (2021). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, *35*(4), 25–46.

Gonzalo, J., & Granger, C. W. J. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics*, *13*(1), 27–35.

Hasbrouck, J. (2021). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading* (2nd ed.). Oxford University Press.

International Monetary Fund. (2024). *Digital financial inclusion and capital market development in sub-Saharan Africa* (IMF Working Paper Series). International Monetary Fund.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2022). Forecasting abnormal stock returns and trading volume using investor sentiment from online search behavior. *International Review of Financial Analysis*, *80*, 101987.

Khan, M. A., Sharma, R., & Yadav, S. (2023). Artificial intelligence, sentiment analytics and financial forecasting: A machine learning approach. *Decision Support Systems*, *168*, 113934.

Lo, A. W. (2021). Adaptive markets and the new world order. *Financial Analysts Journal*, *77*(2), 36–52.

MSCI. (2024). *MSCI frontier markets index methodology*. MSCI Research Publications.

Najem, A. (2025). Big data analytics and artificial intelligence in modern financial systems. *Discover Artificial Intelligence*, *5*(1), 1–15.

Ntim, C. G., Opong, K. K., Danbolt, J., & Thomas, D. A. (2022). Market efficiency and financial integration in African stock markets. *Journal of International Financial Markets, Institutions and Money*, *76*, 101498.

Urquhart, A., & McGroarty, F. (2022). Are stock markets really efficient? Evidence from time-varying market efficiency tests. *International Review of Financial Analysis*, *79*, 101939.

World Bank. (2023). *Global financial development report 2023: Financial technology and market development*. World Bank Publications.

Yao, J., Li, X., Zhang, Y., & Chen, H. (2024). Machine learning and deep learning applications in stock market forecasting: Comparative evidence from global financial markets. *Expert Systems with Applications*, *237*, 121420.