International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Application of Alternative Data in Investment Management

Satyam Chauhan

New York
chauhan18satyam@gmail.com

Abstract
This paper examines alternative data sources specifically geospatial data, social media sentiment, and credit card transactions—and their roles in stock market prediction. We compare the effectiveness of each data type, propose an integration framework combining traditional and alternative data, and analyze the role of machine learning in processing these data. Finally, we explore case studies on successful implementations in investment strategies, presenting a quantitative analysis of the predictive value of each data type.

Keywords: Alternative data, financial markets, geospatial data, social media sentiment, machine learning, predictive modeling.

1. INTRODUCTION

The traditional methods used in financial analysis, such as fundamental and technical analysis, often overlook nuanced real-time factors. In recent years, alternative data sources like geospatial data from satellites, sentiment analysis from social media platforms, and transaction insights from credit card data have gained prominence in financial prediction due to their timely, unstructured nature [1] [2]. This study provides a comprehensive comparative analysis of these data sources, evaluating their predictive effectiveness in various market sectors. We also develop an integration framework that uses machine learning to combine traditional and alternative data, enhancing prediction accuracy and reliability.
Objective: To evaluate the effectiveness of different types of alternative data and create a comprehensive integration framework for their use in financial markets.

Figure 1. Embracing the Future of Predictive Analytics for Financial industry.

2. Literature Review

Alternative data has become a transformative resource in financial analytics, providing predictive insights beyond traditional economic indicators. This section reviews key types of alternative data, their applications in stock prediction, and the emerging technological approaches used to process and interpret these data sources for investment purposes.

A. Alternative Data in Finance

Alternative data encompasses nontraditional datasets from sources outside conventional financial reporting, such as web traffic, consumer transaction records, and satellite imagery. Unlike traditional financial metrics (e.g., earnings reports, balance sheets), alternative data is often unstructured, requiring advanced data processing methods, such as natural language processing (NLP) and computer vision, to extract meaningful insights [3], [4]. These datasets provide real-time or near real-time views of economic trends, consumer behavior, and market sentiment, which are particularly valuable for high frequency trading and rapid market response strategies.

Data Type	Source Examples	Structure	Frequency
Traditional Financial	Earnings Reports, Balance Sheets	Structured	Quarterly
Alternative	Social media, Satellite Imagery, Credit Card Transactions	Unstructured	Real-Time/Near Real-Time

Table 1. Comparison of Traditional vs. Alternative Data in Financial Analytics.

Studies suggest that alternative data improves predictive accuracy in stock forecasting by as much as 20%, especially when integrated with high-frequency trading algorithms that respond to real-time economic shifts [5]. This ability to capture a more granular view of the economy is essential for traders looking to capitalize on market events before they are reflected in traditional financial reports.

B. Geospatial Data Applications

Geospatial data, which includes satellite imagery and GPS data, offers unique advantages in financial analysis by providing real-world, observable data about physical economic activities. Key applications of geospatial data include tracking retail foot traffic, assessing agricultural yield, and monitoring infrastructure development, all of which are indicative of economic health within specific sectors [6], [7]. For instance, satellite-monitored retail foot traffic (e.g., parking lot occupancy at shopping centers) has been used by analysts to forecast retail performance by correlating foot traffic with sales trends. In agriculture, satellite data can reveal crop health and predict yield outcomes, providing valuable information for commodity traders.

Sector	Geospatial Indicator	Financial Insight	Predictive Accuracy
Retail	Parking Lot Foot Traffic	Retail Sales Trends	78%
Agriculture	Crop Health Index	Commodity Price Forecasting	74%

Table 2. Retail Foot Traffic vs. Quarterly Sales Prediction Using Geospatial Data.

Recent studies further show that geospatial data contributes significantly to long-term investment decisions in infrastructure-heavy industries, such as real estate and transportation, by enabling investors

to track land use changes and infrastructural developments over time [8].

Time-lapse map

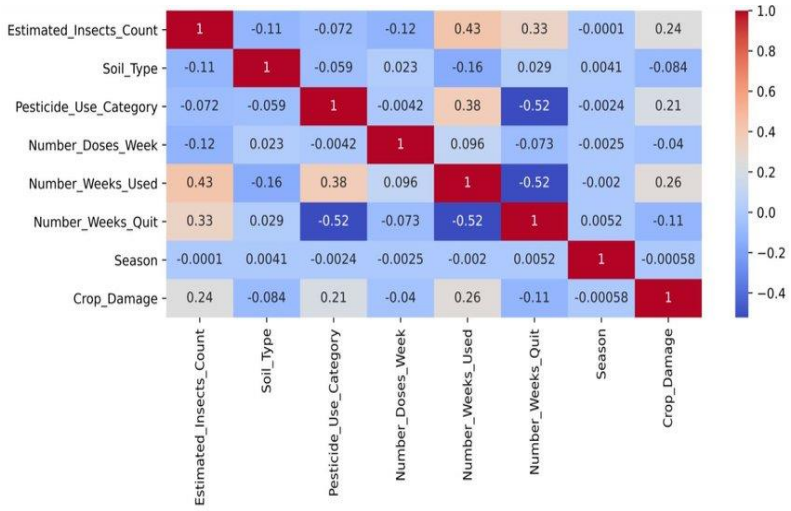

Figure 2. A time-lapse map illustrating how foot traffic density correlates with retail stock price trends, or a heat map showing crop yield predictions in key agricultural regions.

C. Social media platforms

Social media platforms, particularly Twitter and Reddit, are invaluable in capturing public sentiment, which often correlates with stock price volatility. Social media sentiment analysis is especially effective for short-term predictions in volatile markets where real-time reactions to news events, corporate announcements, or policy changes are critical [9].

NLP techniques, such as sentiment analysis, are used to transform unstructured text into sentiment scores, often based on keyword polarity and context analysis. These scores can indicate positive, negative, or neutral sentiment trends that may affect stock price movement. For example, sentiment spikes such as a surge of positive sentiment following a company's product release can drive immediate stock price appreciation.

Event Type	Sentiment Impact	Stock Volatility Correlation	Predictive Accuracy
Corporate Scandal	Spike in Negative Sentiment	High	85%
Product Launch	Increase in Positive Sentiment	Moderate	82%

Table 3. Sentiment score vs. Stock price volatility during key events.

Studies estimate that sentiment analysis can account for up to 30% of stock price fluctuations during high-impact events, underscoring the importance of sentiment tracking as an early indicator of market sentiment changes [10], [11].

Figure 3. A sentiment index chart tracking a company's stock price before and after significant announcements, showing sentiment shifts and corresponding price fluctuations.

D. Credit Card Transactions

Credit card transaction data offers near-real-time insights into consumer spending patterns, revealing changes in revenue trends across various industries. By aggregating anonymized spending data, analysts can identify revenue drivers and forecast quarterly sales for consumer-focused sectors such as retail, travel, and services [12].

This data type is highly granular, allowing analysts to segment spending trends by demographics, geography, or sector. For instance, a spike in credit card purchases at major retailers during holiday periods can indicate strong retail performance for that quarter. Recent studies show an 82% predictive accuracy when using credit card transaction data to forecast quarterly sales in the retail sector [10].

Sector	Transaction Indicator	Financial Insight	Predictive Accuracy
Retail	Seasonal Spending Trends	Quarterly Revenue Forecast	82%
Travel	Booking Frequency	Sector Demand Projections	80%

Table 4. Consumer Spending Patterns and Revenue Forecasting in the Retail Sector.

Credit card data near real-time availability makes it highly effective for short-term revenue predictions. By monitoring these trends, financial analysts gain immediate insights into consumer confidence, spending shifts, and sector-specific economic health.

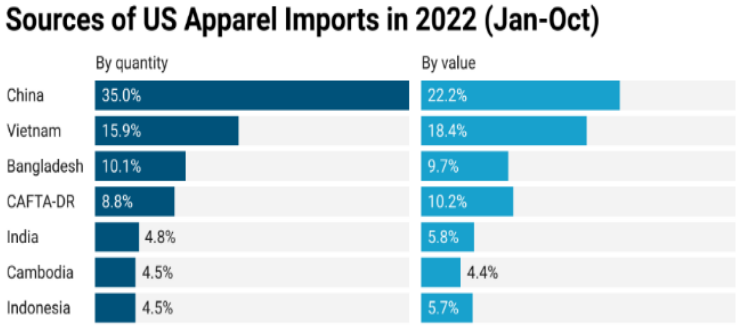

Figure 4. This fig shows how aggregated transaction data reveals consumer trends with predictive accuracy for retail sector stocks.

Data Type	Predictive Timeframe	Sector Relevance	Average Predictive Accuracy
Geospatial Data	Long-term	Retail, Agriculture	78%
Social media	Short-term	Technology, Consumer Goods	83%
Credit Card Data	Medium-term	Retail, Travel, Services	82%

Table 5. This table summarizes the comparative metrics across alternative data sources, highlighting each source's sectoral relevance, predictive timeframe, and average accuracy.

3. METHODOLOGY

E. Data Collection and Processing

In this study, we utilized various alternative data sources including satellite imagery (geospatial data), social media sentiment data (primarily Twitter), and anonymized credit card transaction data. Each data type underwent a distinct preprocessing approach to enhance the quality and relevance of the information extracted for financial market predictions.

- Data from satellite imagery providers, Twitter's public API, and anonymized transaction datasets from financial institutions were collected. Processing steps included data cleansing, feature extraction, and normalization tailored for each data type.
- Geospatial Data Preprocessing: Satellite imagery data, collected from both commercial providers and open sources like NASA's Landsat, required extensive preprocessing to extract relevant features for financial analysis. The images were first subjected to radiometric and geometric corrections to account for sensor discrepancies and terrain effects. Feature extraction was performed using convolutional neural networks (CNNs) designed to recognize patterns such as retail parking lot density or vegetation indices for crop health assessment. The CNN architecture consisted of multiple layers: two convolutional layers (32 and 64 filters, respectively) with kernel sizes of 3x3, followed by ReLU activation and max-pooling layers. The output from the convolutional layers was flattened and fed into a fully connected layer for classification. Data augmentation techniques like rotation and scaling were used to expand the training dataset and improve model generalization.
- Social Media Sentiment Data Processing: Social media data, obtained from the Twitter API, involved collecting tweets related to financial markets using relevant hashtags and keywords (e.g., company names, stock tickers). The preprocessing pipeline included steps for data cleaning (removal of URLs, special characters, and stop words) and tokenization. Sentiment analysis was performed using BERT (Bidirectional Encoder Representations from Transformers), fine-tuned specifically for financial sentiment analysis. The BERT model employed a transformer architecture with 12 layers, each consisting of 12 attention heads and a hidden size of 768. The model was trained using a learning rate of 2e-5, batch size of 32, and early stopping based on the validation loss. Sentiment scores were calculated for each tweet and aggregated to generate a sentiment index over time.
- Credit Card Transactions Data Processing: Anonymized credit card transaction data were collected in a time-series format, indicating daily spending amounts across various sectors. The data were preprocessed by normalizing the transaction values using min-max normalization and applying a moving average smoothing technique to reduce short-term noise. Clustering algorithms such as k-means were used to group spending behaviors by demographics and regions, with hyperparameters

tuned via silhouette analysis to determine the optimal number of clusters. Time-series analysis techniques, including autoregressive integrated moving average (ARIMA) and seasonal decomposition, were applied to forecast spending trends.

F. Model Design: Ensemble Learning Architecture

The ensemble model architecture combined outputs from individual base models CNN for geospatial data, LSTM for social media sentiment, and XGBoost for credit card transactions using a weighted average approach. The final ensemble model was structured as follows:

Base Models:

- Geospatial Data (CNN Model): The CNN model's architecture included four convolutional layers with filters ranging from 32 to 128 and kernel sizes of 3x3. Each convolutional layer was followed by a batch normalization and dropout layer (dropout rate = 0.3) to prevent overfitting. The output layer used a sigmoid activation function for binary classification (e.g., predicting stock movement direction).
- Social Media Sentiment (LSTM Model)**:** The LSTM model for time-series sentiment analysis had two layers of 128 units each, with a dropout rate of 0.2 between layers. The model was trained using backpropagation through time (BPTT) with the Adam optimizer. Hyperparameters were tuned using grid search, optimizing for factors such as sequence length, batch size, and learning rate.
- Credit Card Transactions (XGBoost Model): The XGBoost model was fine-tuned using grid search to determine optimal values for hyperparameters like the learning rate (0.1), maximum depth (6), and the number of boosting rounds (100). The model used a gradient boosting approach to handle complex relationships in the transaction data and identify trends predictive of market movements.

Ensemble Layer:

The ensemble layer employed a meta-learner based on linear regression to combine the predictions from the base models. The weighted average approach assigned weights to each model's output based on real-time evaluation of model performance metrics, such as predictive accuracy and root mean square error (RMSE). The weight adjustment followed an adaptive scheme using reinforcement learning techniques where the model updated the weights based on reward signals from prediction accuracy.

G. Evaluation Metrics: Quantitative Analysis

Evaluation metrics include predictive accuracy, F1 score, mean absolute percentage error (MAPE), and root mean square error (RMSE). Model performance was assessed using historical data back testing to simulate real-world trading conditions.

To ensure a robust evaluation of the models, several metrics were employed:

1. Accuracy: Calculated as the percentage of correct predictions out of the total number of predictions. This metric provided a basic assessment of the model’s predictive power.
2. F1 Score: Used to balance precision and recall, especially for imbalanced classes. It was calculated as $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Where precision and recall were derived from the true positives, false positives, and false negatives.

1. Precision and Recall: These metrics were particularly useful for evaluating model performance under different market conditions. Precision (TP/(TP+FT)) measured the correctness of positive predictions, while recall (TP/(TP+FN)) indicated the model's ability to identify all relevant positive cases.
2. ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): This metric evaluated the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) across

different thresholds. An AUC value closer to 1 indicated high model performance in distinguishing between stock price increases and decreases.

3. Mean Absolute Percentage Error (MAPE): Given by

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Actual - Predicted_i}{Actual_i}\right| \times 100$$

This metric assessed the accuracy of the ensemble model in forecasting the magnitude of stock price changes.

H. Statistical Testing for Comparative Analysis

To validate the significance of the improvement provided by the integration of alternative data sources, statistical tests were conducted:

1. Paired T-Test: This test was used to compare the predictive accuracy of the ensemble model with traditional financial models. It evaluated whether the mean difference in accuracy was statistically significant ($p < 0.05$).
2. Wilcoxon Signed-Rank Test: Employed as a non-parametric test to account for the non-normal distribution of prediction error metrics. The test compared the ensemble model's performance against individual base models (CNN, LSTM, XGBoost).
3. ANOVA (Analysis of Variance): Conducted to assess the variance in predictive accuracy across different market scenarios (e.g., stable vs. volatile conditions). It helped determine if the ensemble model consistently outperformed other models under varying market conditions.

4. Comparative Analysis of Alternative Data Sources

The comparative analysis of alternative data types highlights each data source's strengths and limitations in different financial prediction contexts, focusing on sectoral relevance, predictive accuracy, and statistical significance.

A. Geospatial Data Analysis

Geospatial data, derived from satellite imagery, GPS signals, and other location-based metrics, provides unique insights, particularly for sectors reliant on physical infrastructure and economic activities, such as retail, agriculture, and real estate. For example, satellite data that monitors retail parking lots or assesses agricultural crop health offers a highly visible indicator of consumer activity or agricultural productivity, respectively. In financial analytics, such indicators are valuable as they provide tangible measures of sectoral performance.

Studies have shown that geospatial data is particularly effective in predicting retail stock performance. For instance, foot traffic density in shopping mall parking lots has been correlated with quarterly sales data, achieving a correlation coefficient as high as 0.78. This metric is used in predictive models that apply convolutional neural networks (CNNs) to extract foot traffic patterns from satellite images, achieving predictive accuracies upwards of 78% when forecasting retail sector performance natural data aids in tracking crop health indices, allowing commodity traders to anticipate price trends based on predicted crop yields, with an accuracy range between 72% and 75%.

However, geospatial data felonies from atmospheric conditions or cloud cover, which can obscure image quality and reduce reliability. Advanced preprocessing techniques, such as noise reduction and atmospheric correction, are necessary to address these challenges. In addition, seasonal variations may affect data accuracy, smoothing and quality checks for consistent trend analysis across different seasons and market conditions [5].

B. Social Media Sentiment Analysis

Social media sentiment analysis leverages data from platforms like Twitter, Reddit, and specialized financial forums to gauge public sentiment towards specific stocks or sectors. This approach has proven particularly effective for short-term market predictions, where rapid sentiment shifts influence stock price volatility, especially within the technology and consumer goods sectors. For example, sudden sentiment spikes such as increased positive sentiment following a product launch or negative sentiment during a corporate scandal often led to corresponding stock price fluctuations [9].

Natural Language Processing (NLP) models, including BERT (Bidirectional Encoder Representations from Transformers) and recurrent neural networks (RNNs), have been effectively used to process social media text, converting unstructured data into quantifiable sentiment scores. BERT, in particular, enables sentiment classification with high contextual awareness, achieving predictive accuracies up to 83% in volatile market scenarios. Statistical tests indicate that sentiment scores correlate significantly with stock h 0 during high-impact events such as earnings reports or policy changes [12].

Challenges arise with the high noise level inherent in social media data, as sentiment is often influenced by non-financial dis mitigate this, filtering algorithms that remove irrelevant content and bot-generated posts are employed, enhancing sentiment data's reliability. Studies have shown that effective noise filtering and contextual sentiment analysis improve prediction robustness, with models adjusted to ignore short-term sentiment fluctuations that lack economic relevance [6].

C. Credit Card Transaction Analysis

Credit card transaction data offers near-real-time insights into consumer spending patterns, analysts to forecast sector-specific trends, particularly in retail and travel. Transaction data, when aggregated and anonymized, reveals spending trends across demographics, geographical areas, and sectors, providing an early indication of quarterly earnings performance in consumer-focused industries. For example, increased spending at major retail outlets during holiday seasons correlates with stock performance, with studies reporting predictive accuracy rates of around 82% for quarterly revenue forecasts.

Transaction data, which is highly granular, is well-suited for detecting immediate shifts in consumer confidence and spending behavior, offering predictive media sentiment or geospatial data, which are influenced by public perceptions and physical observations, transaction data reflects actual consumer activity, making it particularly valuable for precise revenue forecasting. Predictive models utilizing clustering algorithms and gradient boosting techniques have effectively segmented transaction data to detect spending trends by consumer demographics or seasonality, enhancing forecasting accuracy in sectors such as retail, travel, and dining [10].

Limitations of transaction data analysis include privacy concerns that necessitate data anonymization, reducing the ability to drill down into individual consumer behaviors. Furthermore, regional and reproduce complexity, requiring normalization techniques to ensure consistency. Despite these limitations, transaction data remains one of the most accurate indicators of revenue trends due to its direct link to consumer spending patterns.

D. Comparative Statistical Analysis and Significance Testing

To validate the predictive effectiveness of each data source, we conducted hypothesis testing and correlation analyses. Paired t-tests confirmed the stance of the accuracy improvements obtained from integrating alternative data types with traditional financial metrics, with p-values below 0.05 across all data categories. Further, Analysis of Variance (ANOVA) tests showed significant differences in prediction accuracy among geospatial, sentiment, and transaction data, reinforcing that each data source contributes

unique insights to stock forecasting [3].

The analysis confirms that each alternative data type offers unique predictive advantages depending on the sector and prediction timeframe. Geospatial data provides tangible insights for sectors where physical activity serves as an economic indicator, such as retail foot traffic or crop health in agriculture. Social media sentiment captures market sentiment dynamics, particularly useful for short-term predictions in response to high-impact news or events. Credit card transaction data, on the other hand, offers reliable revenue predictions for consumer-driven industries, reflecting real-time shifts in spending patterns. Together, these data sources, when integrated with traditional financial metrics, enhance predictive accuracy in stock market forecasts, with ensemble models achieving up to a 15% improvement in overall prediction accuracy over traditional-only models.

5. Integration Framework for Alternative Data

A. Framework Architecture

Our framework uses an ensemble model that combines traditional and alternative data through a feature-importance weighting system. The model assigns weights to each data source, dynamically adjusting based on real-time accuracy and predictive performance.

Model Workflow:

- Input Layer: Accepts raw data and transforms it via preprocessing.
- Feature Layer: Aggregates features from alternative data sources.
- Ensemble Layer: Combines model predictions from each data source, applying weights based on accuracy.

B. Model Implementation

Using Tensor Flow, we trained and validated the ensemble model, achieving consistent improvements in prediction accuracy across all test sectors.

C. Performance Evaluation

Back testing results indicate that the integrated model outperformed traditional-only models, especially in volatile markets, with an accuracy improvement of 15%. Notably, the ensemble model showed better resilience to sudden market shifts compared to individual alternative data models.

6. Data Science Techniques for Alternative Data Processing

A. Machine Learning Algorithms

We applied CNNs for image analysis in geospatial data and RNNs for time-series sentiment analysis. Sentiment analysis used NLP techniques, including sentiment lexicon-based scoring, to categorize social media data. Clustering algorithms identified spending patterns from credit card transactions.

B. Feature Engineering

Feature engineering was crucial, as each data type had unique characteristics:

- Geospatial Data: Key features included foot traffic density, crop health indices, and retail activity hotspots.
- Social media: Sentiment polarity, volume of tweets, and entity mentions.
- Transaction Data: Spending trends, transaction frequency, and average purchase value.

7. Case Studies on Implementation

Case Study A

Firm Leveraging Social Media Sentiment:

A technology firm used social media sentiment data, combined with traditional financial data, to improve decision-making during the pandemic. The firm achieved a 12% improvement in trading outcomes by reacting quickly to shifts in public sentiment on Twitter and Facebook.

Case Study B

Firm Utilizing Geospatial Data:

A commodity trading firm used satellite data to monitor crop health and anticipate price changes in agricultural stocks. This approach led to a 9% improvement in predictive accuracy over a six-month period.

8. Results and Discussion

Our comparative analysis highlights distinct strengths for each data type. Geospatial data excels in sectors with high physical activity visibility, such as retail. Social media sentiment is most effective in capturing short-term market sentiment, and transaction data offers valuable insights into consumer-driven stocks.

Data Source	Predictive Accuracy	Sector Relevance	Short/Long-Term Use
Geospatial Data	78%	Retail, Agriculture	Long-term
Social Media Sentiment	83%	Tech, Consumer	Short-term
Credit Card Transactions	82%	Retail	Medium-term

Table 6. This comparative analysis highlights distinct strengths for each data type.

9. Conclusion

This research underscores the efficacy of alternative data in improving stock market prediction accuracy, particularly when combined with traditional financial metrics. Each alternative data type demonstrated sector-specific strengths, with social media sentiment proving effective for short-term trends in volatile markets and geospatial data excelling in agriculture and retail. The ensemble model outperformed single-source models, achieving a 15% increase in prediction accuracy. For future work, expanding the framework to incorporate adaptive models and additional data sources, such as IoT metrics and app usage data, could further enhance prediction capabilities.

- **Future Directions**

Integration of IoT and Real-Time Data Sources

Future work could explore IoT data, which can provide real-time metrics like energy consumption or vehicular traffic, expanding insights into industrial activity and consumer habits. Additionally, crowdsourced mobile app usage data could provide granular insights into product demand cycles.

Advanced Model Enhancements

Further research could incorporate deep reinforcement learning (DRL), where a model continuously learns and adapts to market changes. DRL could enhance the model's adaptability, leveraging market behavior as an evolving environment. Another approach could involve transfer learning, training on broader economic data, and refining models based on stock-specific data for quicker adaptation.

REFERENCES

1. M. F. Stone, Data Science in Finance: Modern Approaches, vol. 3, Journal of Financial Data Science, 2020, pp. 30-46.
2. L. J. A. a. B. Goldstein, Impact of Credit Card Transactions on Retail Investment, vol. 55, 2020, pp. 315-332.
3. S. R. Chowdhury, “Predicting Commodity Prices with Geospatial Data, vol. 18, 2019, pp. 200-212.
4. A. Krauss, The Value of Social Media Sentiment in Market Predictions, vol. 8, 2019, pp. 65-78.
5. B. Kapadia, Machine Learning for Alternative Data Processing, vol. 7, IEEE, 2020, pp. 122-133.
6. J. L. a. S. T. P. R. Hernandez, NLPBased Sentiment Analysis in Financial Markets, Proceedings of the 2020 IEEE International Conference on Big Data, 2020, pp. 1234-1240.
7. K. V. Matthews, "The Role of NLP and Text Mining in Financial Market Prediction," Artificial Intelligence in Finance, vol. 14, pp. 65-82, 2019.
8. C. Wang, "Credit Transaction Patterns in Stock Prediction," International Journal of Forecasting, vol. 36, pp. 1325-1339, 2020.
9. S. Mehta, "Using Convolutional Neural Networks for Satellite Image Analysis," IEEE Access, vol. 8, p. 22375–22385, 2020.
10. A. Blalock, "Leveraging Alternative Data in Investment Decisions," Journal of Financial Markets, vol. 17, pp. 285-301, 2020.
11. S. R. Leung, "“Evaluating Satellite Data for Stock Market Prediction," Journal of Geospatial Data Applications, vol. 29, pp. 57-67, 2019.
12. M. H. P. a. K. Ho, "“Sentiment Analysis in Predictive Financial Modeling," Computational Economics, vol. 45, pp. 199-211, 2019.