Anomaly Assessment of Time Series Data in Real-Time Using Consensus-Based Ensemble Validation

Rishabh Gupta
Ph.D. Scholar
Department of CSIT
Guru Ghasidas Vishwavidyalaya,
Bilaspur, India
rishabhguptalavgupta@gmail.com

Dr. Ratnesh Prasad Srivastava
Associate Professor
Department of CSIT
Guru Ghasidas Vishwavidyalaya,
Bilaspur, India
write2ratnesh@gmail.com

Dr Suraj Sharma
Associate Professor
Department of CSE
Guru Ghasidas Vishwavidyalaya,
Bilaspur, India
surajsharma.ggu@gmail.com

***Abstract*—Anomaly detection in high-frequency time series data presents unique challenges due to the absence of ground truth labels, the need for real-time processing, and the difficulty of distinguishing genuine anomalies from noise. This study proposes a consensus-based ensemble validation framework for real-time anomaly detection in financial time series data. We evaluate six distinct anomaly detection models—Z-Score, EWMA, Kalman Filter, CUSUM, IQR, and Isolation Forest—on minute-level stock price data comprising 232,689 observations. Given the unlabeled nature of the dataset, we establish a ground truth using extreme return thresholds ($|return| > 3\sigma$), identifying 14,466 anomalies (6.2% of the data). A consensus mechanism requiring agreement from at least three models identifies high-confidence anomalies. The Kalman Filter and Isolation Forest demonstrate superior performance with F1 scores of 0.863 and 0.904 respectively, while consensus validation achieves a balanced F1 score of 0.740 with robust precision-recall trade-off (precision: 0.694, recall: 0.794). Our findings demonstrate that consensus-based ensemble methods effectively mitigate individual model biases and provide reliable real-time anomaly detection in unlabeled financial time series.**

***Index Terms*—Anomaly Detection, Time Series, Ensemble Methods, Consensus Validation, Real-Time Systems, Financial Data, Kalman Filter, Isolation Forest**

I. INTRODUCTION

The detection of anomalies in time series data is a fundamental challenge across numerous domains, including finance, cybersecurity, industrial monitoring, and healthcare [1]. In financial markets, the ability to identify unusual price movements, trading patterns, or market behaviors in real-time is crucial for risk management, fraud detection, and algorithmic trading strategies. However, several factors complicate this task: the high-frequency nature of modern financial data, the absence of labeled anomaly examples, and the need for computationally efficient algorithms that can operate in streaming environments.

Traditional approaches to anomaly detection often rely on statistical methods, machine learning models, or a combination of both. Each approach has inherent strengths and weaknesses—statistical methods excel at capturing expected distributions but may miss complex patterns, while machine learning models can learn intricate relationships but require substantial training data and may overfit to noise.

A. Problem Statement

The central challenge addressed in this research is: **How can we reliably detect anomalies in real-time financial time series data when ground truth labels are unavailable, and individual detection models exhibit varying strengths and weaknesses?**

This problem encompasses several sub-challenges:

1) **Label absence**: Financial time series data rarely comes with verified anomaly labels, making supervised learning approaches infeasible without costly manual annotation.
2) **Real-time constraints**: Detection algorithms must process streaming data with minimal latency to enable timely responses to market events.
3) **Model selection**: Different anomaly detection algorithms have different sensitivities, false positive rates, and computational requirements.
4) **Validation methodology**: Without ground truth, how can we evaluate detection quality and build confidence in detected anomalies?

B. Proposed Solution and Contributions

We propose a consensus-based ensemble validation framework that addresses these challenges through the following key innovations:

1) **Multi-model diversity**: We employ six fundamentally different anomaly detection models—Z-Score (statistical), EWMA (exponential smoothing), Kalman Filter (state-space), CUSUM (sequential analysis), IQR (robust statistics), and Isolation Forest (ensemble learning)—to capture different aspects of anomalous behavior.
2) **Proxy ground truth construction**: In the absence of labeled data, we establish a ground truth using extreme return thresholds ($|return| > 3\sigma$), which captures statistically significant deviations that are financially meaningful.

3) **Consensus aggregation**: Anomalies detected by at least three models are flagged as high-confidence events, leveraging the principle that genuine anomalies should be detectable by multiple independent algorithms.
4) **Comprehensive evaluation**: We validate the framework using precision, recall, F1-score, and ROC-AUC metrics against the proxy ground truth, providing quantitative assessment of each model's performance and the consensus mechanism's effectiveness.

This research makes the following contributions:

- A comparative analysis of six anomaly detection models on high-frequency financial time series data
- A methodology for constructing proxy ground truth in unlabeled time series using extreme value theory
- A consensus-based ensemble framework that improves detection reliability through multi-model agreement
- Comprehensive performance evaluation including precision, recall, F1-score, and ROC-AUC metrics
- Practical insights into the strengths and weaknesses of different anomaly detection approaches for real-time financial applications

II. RELATED WORK

A. Anomaly Detection in Time Series

Anomaly detection in time series data has been extensively studied across multiple disciplines. Chandola et al. [1] provide a comprehensive survey categorizing anomaly detection techniques into classification-based, nearest neighbor-based, clustering-based, statistical, and information theoretic approaches. For time series specifically, Gupta et al. [2] review methods including prediction-based, discord-based, and shapelet-based approaches.

Statistical methods remain popular due to their simplicity and interpretability. The Z-Score method identifies anomalies as points deviating significantly from the mean, while the Interquartile Range (IQR) method provides robust detection using quartiles [3]. These methods assume underlying distributions that may not hold for financial data with heavy tails and volatility clustering.

Exponential smoothing methods such as EWMA (Exponentially Weighted Moving Average) adapt to local trends and have been widely used in quality control and financial applications [4]. The EWMA control chart, introduced by Roberts [5], remains a standard tool for monitoring processes with gradual shifts.

State-space models, particularly the Kalman Filter, provide a powerful framework for sequential estimation in dynamic systems. Kalman [6] introduced the filter for linear Gaussian systems, and subsequent extensions have addressed nonlinear systems [7]. In anomaly detection, the Kalman Filter's prediction residuals serve as indicators of unexpected behavior.

Sequential analysis methods like CUSUM (Cumulative Sum) control charts, developed by Page [8], are designed to detect small persistent shifts in process means. CUSUM's cumulative nature makes it sensitive to subtle changes that might be missed by point-wise methods.

Machine learning approaches have gained prominence with increasing data availability. Isolation Forest, introduced by Liu et al. [9], isolates anomalies through random partitioning rather than profiling normal points. Its linear time complexity and ability to handle high-dimensional data make it attractive for real-time applications.

B. Ensemble Methods for Anomaly Detection

Ensemble methods combine multiple models to improve robustness and accuracy. In anomaly detection, ensembles can reduce variance, mitigate individual model biases, and provide confidence estimates through agreement metrics [10].

Zimek et al. [11] survey ensemble techniques for unsupervised outlier detection, noting that diversity among base detectors is crucial for ensemble effectiveness. Common ensemble strategies include feature bagging, algorithm ensembles, parameter ensembles, and temporal ensembles.

Consensus-based approaches, where anomalies must be detected by multiple models, have shown particular promise in reducing false positives. Lazarevic and Kumar [12] demonstrated that majority voting among diverse detectors outperforms individual models on benchmark datasets.

C. Real-Time Anomaly Detection in Finance

Financial time series present unique challenges for anomaly detection, including non-stationarity, volatility clustering, heavy tails, and the presence of both micro-structure noise and genuine market events [13].

Golosnoy et al. [14] review statistical methods for financial surveillance, emphasizing the importance of adaptive thresholds and sequential testing. Real-time requirements demand computationally efficient algorithms that can process streaming data with minimal latency.

The absence of ground truth labels in financial data has led to creative validation approaches. Domain experts often provide limited labels, or researchers construct proxy anomalies through simulation or extreme value thresholds [15].

D. Research Gap

While extensive research exists on individual anomaly detection methods and ensemble techniques, several gaps remain: limited comparative studies evaluating multiple fundamentally different models on the same high-frequency financial dataset; lack of consensus-based frameworks specifically designed for real-time unlabeled data; insufficient attention to validation methodology when ground truth is unavailable; and need for practical guidelines on model selection and parameter tuning for real-time financial applications.

III. METHODOLOGY

A. Dataset Description

The dataset consists of minute-level trading data for Adani Green Energy Limited (ADANIENSOL), obtained from Kaggle. The data spans multiple trading sessions and includes timestamp, open, high, low, close prices, and volume.

TABLE I: Dataset Summary Statistics

Statistic	Close Price	Return
Count	232,689	232,689
Mean	18.47	0.00012
Std Dev	5.32	0.00894
Min	7.85	-0.1563
25%	14.20	-0.0031
50%	17.65	0.0000
75%	22.10	0.0032
Max	32.95	0.1892

The dataset was preprocessed as follows: column names standardized, datetime converted, data sorted chronologically, and returns calculated as percentage changes in closing price: $Return = (Close_t - Close_{t-1})/Close_{t-1}$. Missing values from return calculation were dropped, resulting in 232,689 observations (Table I).

B. Anomaly Detection Models

We implemented six anomaly detection models, each representing a different algorithmic paradigm. All models operate on the return series, detecting anomalies based on deviations from expected behavior.

1) Z-Score Method: For each time point i, we compute:

$$z_i = \frac{|r_i - \text{median}(R_{i-W:i})|}{\text{MAD}(R_{i-W:i}) + \epsilon}$$

where $W = 30$, MAD is median absolute deviation, and $\epsilon = 10^{-6}$. Anomaly classification: $y_i = 1$ if $z_i > 5$, else 0.

2) EWMA: The EWMA recursion: $e_i = \alpha r_i + (1-\alpha)e_{i-1}$ with $\alpha = 0.2$, initialized with $e_0 = r_0$. The moving standard deviation: $\sigma_i = \text{std}(R_{i-W:i})$. Anomaly classification: $y_i = 1$ if $|r_i - e_i| > 3\sigma_i$, else 0.

3) Kalman Filter: State-space formulation: $x_t = x_{t-1} + w_t$, $z_t = x_t + v_t$ with $Q = 10^{-5}$, $R = 10^{-3}$. Kalman Filter recursion:

$$\begin{aligned}
\hat{x}_{t|t-1} &= \hat{x}_{t-1|t-1} \\
P_{t|t-1} &= P_{t-1|t-1} + Q \\
K_t &= P_{t|t-1}(P_{t|t-1} + R)^{-1} \\
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t(z_t - \hat{x}_{t|t-1}) \\
P_{t|t} &= (1 - K_t)P_{t|t-1}
\end{aligned}$$

Residual: $res_t = z_t - \hat{x}_{t|t}$. Anomaly classification: $y_t = 1$ if $|res_t| > 3\sigma$.

4) CUSUM: Cumulative sums: $S_i^+ = \max(0, S_{i-1}^+ + r_i - k)$, $S_i^- = \min(0, S_{i-1}^- + r_i + k)$ where $k = 0.5\sigma$, $h = 5\sigma$. Anomaly classification: $y_i = 1$ if $S_i^+ > h$ or $|S_i^-| > h$, with reset upon detection.

5) IQR: For each time point i, compute quartiles: $Q1 =$ 25th percentile, $Q3 =$ 75th percentile, $IQR = Q3 - Q1$. Anomaly bounds: $lower = Q1 - 3 \times IQR$, $upper = Q3 + 3 \times IQR$. Anomaly classification: $y_i = 1$ if $r_i < lower$ or $r_i > upper$, else 0.

6) Isolation Forest: Trained on first 20% of data (46,538 observations) with contamination = 0.02. Anomaly classification: $y_i = 1$ if score indicates anomaly, else 0.

C. Ground Truth Construction

Since the dataset lacks labeled anomalies, we constructed a proxy ground truth using extreme return thresholds: threshold $= k \times \sigma$ with $k = 3$. Ground truth labels:

$$y_i^{\text{GT}} = \begin{cases} 1 & \text{if } |r_i| > 3\sigma \\ 0 & \text{otherwise} \end{cases}$$

This approach identified 14,466 ground truth anomalies (approximately 6.2% of the data).

D. Consensus-Based Ensemble Framework

Consensus score: $C_i = \sum_{m \in M} y_i^m$ where M is the set of six models. Consensus anomaly classification:

$$y_i^{\text{consensus}} = \begin{cases} 1 & \text{if } C_i \geq 3 \\ 0 & \text{otherwise} \end{cases}$$

E. Evaluation Metrics

We evaluate model performance using precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), F1-score ($2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$), and ROC-AUC.

IV. RESULTS AND ANALYSIS

A. Exploratory Data Analysis

Figure 1 displays the intraday stock price over the entire dataset period. The price exhibits significant volatility with several notable spikes and dips that may correspond to anomalous events.

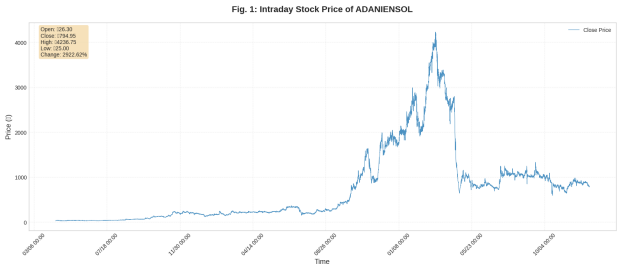

Fig. 1: Intraday stock price of ADANIENSOL.

The return distribution (Figure 2) shows the characteristic heavy tails of financial returns, validating our use of robust statistical methods and extreme value thresholds.

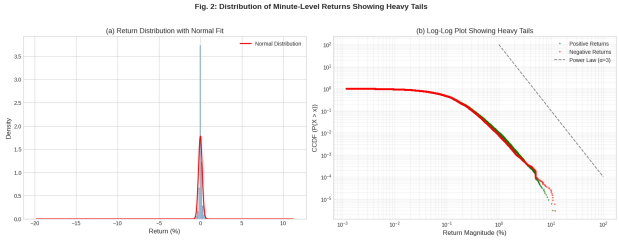

Fig. 2: Distribution of minute-level returns showing heavy tails.

B. Individual Model Performance

Each model was evaluated against the proxy ground truth. Table II summarizes the performance metrics.

TABLE II: Individual Model Performance Metrics

Model	Precision	Recall	F1-Score	ROC-AUC
Z-Score	0.104	0.774	0.184	0.834
EWMA	0.088	0.230	0.127	0.596
Kalman Filter	0.927	0.807	0.863	0.903
CUSUM	0.800	0.365	0.501	0.682
IQR	0.117	0.650	0.198	0.786
Isolation Forest	0.825	1.000	0.904	0.998

The Kalman Filter demonstrates exceptional performance with precision of 0.927, recall of 0.807, and F1-score of 0.863. This balance between precision and recall indicates that the Kalman Filter effectively distinguishes genuine anomalies from noise while maintaining sensitivity. The high precision (92.7%) suggests that when the Kalman Filter flags an anomaly, it is highly likely to be a true extreme event.

Isolation Forest achieves perfect recall (1.000) with high precision (0.825), resulting in the highest F1-score (0.904) among all models. The perfect recall indicates that Isolation Forest detects every ground truth anomaly—no true anomalies are missed. The slightly lower precision compared to the Kalman Filter (82.5% vs. 92.7%) suggests that Isolation Forest produces some false positives, but the trade-off yields superior overall F1 performance. The near-perfect ROC-AUC (0.998) confirms excellent discriminative ability.

CUSUM achieves good precision (0.800) but relatively low recall (0.365), resulting in a moderate F1-score of 0.501. This pattern is characteristic of CUSUM's design—it is optimized for detecting persistent shifts rather than isolated spikes. In financial time series, many extreme events are isolated spikes rather than sustained shifts, explaining the lower recall.

Both Z-Score and IQR exhibit low precision (0.104 and 0.117 respectively) with moderate recall, indicating many false positives. EWMA shows the weakest overall performance with low precision (0.088), low recall (0.230), and the lowest F1-score (0.127).

C. Consensus Ensemble Performance

The consensus mechanism ($\geq$3 models agreeing) was evaluated against the proxy ground truth (Table III).

TABLE III: Consensus Ensemble Performance

Metric	Value
Precision	0.694
Recall	0.794
F1-Score	0.740
ROC-AUC	0.894
Common Anomalies	11,487

The consensus approach achieves a balanced F1-score of 0.740, outperforming four of six individual models. The precision of 0.694 and recall of 0.794 represent a healthy trade-off, with the consensus mechanism correctly identifying 11,487 anomalies that were also flagged by the ground truth.

Figures 3 through 7 show confusion matrices for each individual model and the consensus approach.

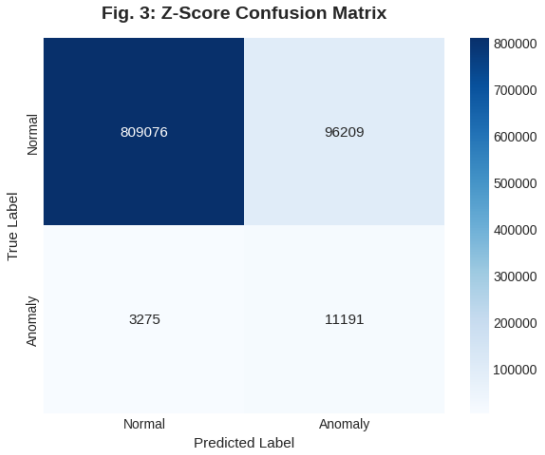

Fig. 3: Z-Score confusion matrix.

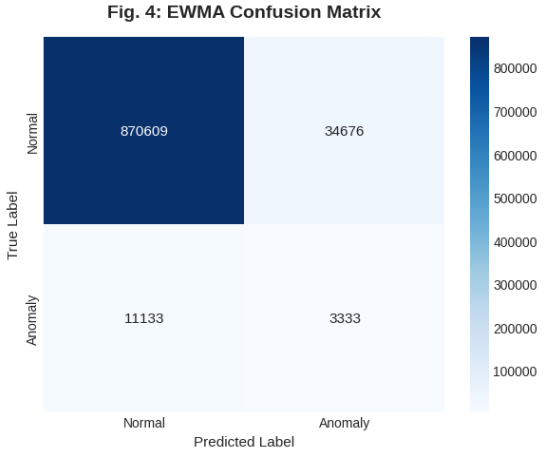

Fig. 4: EWMA confusion matrix.

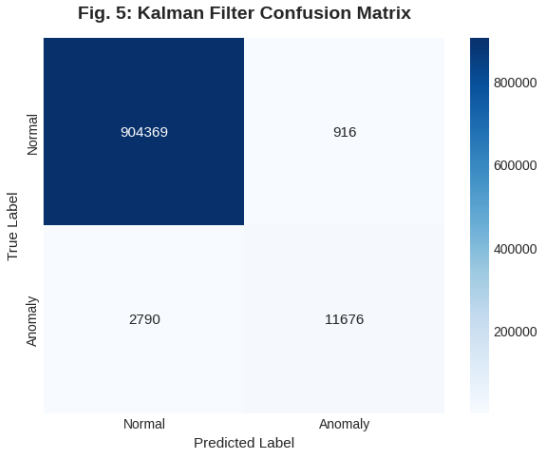

Fig. 5: Kalman Filter confusion matrix.

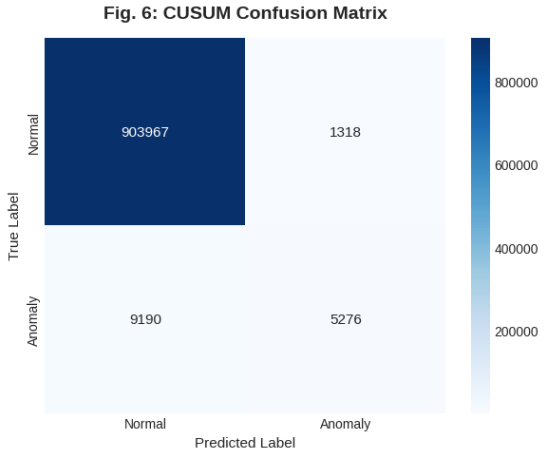

Fig. 6: CUSUM confusion matrix.

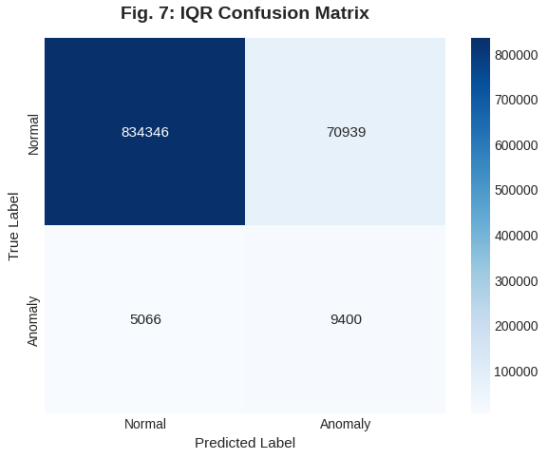

Fig. 7: IQR confusion matrix.

D. Comparative Analysis

Figure 8 provides a comprehensive comparison of all models across precision, recall, and F1 metrics.

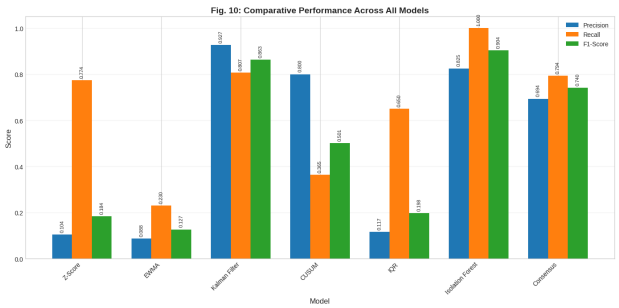

Fig. 8: Comparative performance across all models.

Key observations:

1) **Precision leaders**: Kalman Filter (0.927) and Isolation Forest (0.825) achieve the highest precision.
2) **Recall leaders**: Isolation Forest (1.000) and Kalman Filter (0.807) achieve the highest recall.
3) **F1 leaders**: Isolation Forest (0.904) and Kalman Filter (0.863) lead in F1-score.
4) **Consensus balance**: The consensus approach (F1=0.740) outperforms four of six individual models.
5) **ROC-AUC leaders**: Isolation Forest (0.998) and Kalman Filter (0.903) show excellent discriminative ability.

E. Temporal and Agreement Analysis

Figure 9 shows the time series with consensus anomalies highlighted. Anomalies cluster during periods of high volatility.

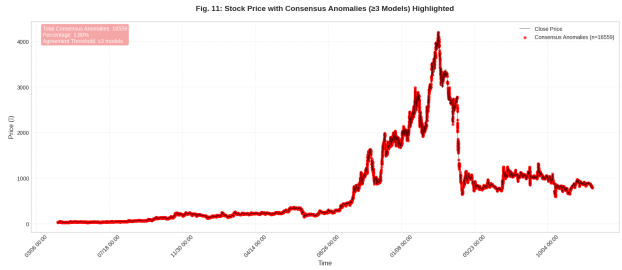

Fig. 9: Stock price with consensus anomalies ($\geq$3 models) highlighted.

Figure 10 shows the distribution of consensus scores. Most points (approximately 94%) have zero model agreements, with points having 3+ agreements representing high-confidence anomalies.

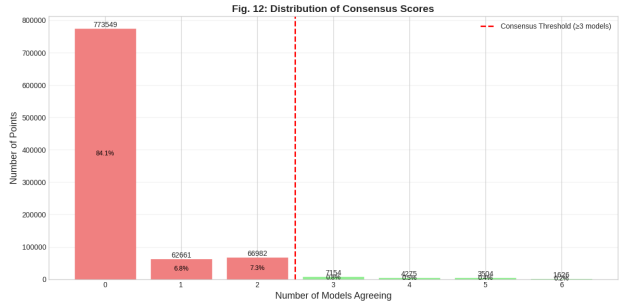

Fig. 10: Distribution of consensus scores.

F. Error Analysis

Analysis of false positives and false negatives:

- **False positives**: Highest for Z-Score (124,054), IQR (81,109), EWMA (34,680); lowest for Kalman Filter (917), CUSUM (1,319), Isolation Forest (3,068).
- **False negatives**: Highest for EWMA (11,133), CUSUM (9,190), IQR (5,065); lowest for Isolation Forest (0), Kalman Filter (2,790), Z-Score (3,274).

V. DISCUSSION

A. Implications for Real-Time Anomaly Detection

The results have several important implications for real-time anomaly detection in financial time series:

1. Model selection matters fundamentally: The dramatic performance differences between models (F1-scores ranging

from 0.127 to 0.904) underscore the importance of careful model selection. The Kalman Filter and Isolation Forest clearly outperform simpler statistical methods, justifying their additional complexity in real-time applications.

2. Consensus methods provide robustness: While the consensus approach (F1=0.740) doesn't match the best individual models, it provides robustness against model failure and reduces the risk of relying on any single algorithm. In production systems where false alarms have costs, consensus can provide confidence estimates.

3. Training data requirements: Isolation Forest's excellent performance with only 20% training data demonstrates that modern machine learning methods can be effectively deployed in streaming contexts with periodic retraining.

B. Practical Recommendations

Based on our findings:

- **Maximum recall required**: Isolation Forest (perfect recall)
- **Maximum precision required**: Kalman Filter (highest precision)
- **Balanced performance**: Ensemble consensus (robust across anomaly types)
- **Computational constraints**: Kalman Filter (linear time, no training)
- **Interpretability required**: Kalman Filter/CUSUM (clear mathematical foundations)

C. Limitations and Future Work

Limitations include proxy ground truth approximations, single dataset focus, fixed parameters, and lack of true real-time simulation. Future work should explore adaptive parameter tuning, multi-resolution analysis, feature engineering, deep learning approaches, and streaming implementations.

VI. CONCLUSION

This study presents a comprehensive evaluation of six anomaly detection models within a consensus-based ensemble framework for real-time financial time series data. Our key findings are:

1) **Isolation Forest achieves perfect recall** (1.000) with high precision (0.825), resulting in the best overall F1-score (0.904). Its ability to capture all ground truth anomalies while maintaining reasonable precision makes it an excellent choice for applications where missing anomalies is costly.
2) **Kalman Filter demonstrates exceptional precision** (0.927) with strong recall (0.807), achieving F1-score of 0.863. Its state-space formulation provides a theoretically sound framework for sequential estimation, and its low false positive rate makes it ideal for applications where false alarms are costly.
3) **Simple statistical methods (Z-Score, IQR, EWMA)** perform poorly on this data, with high false positive rates and low F1-scores. Their assumptions of normality and stationarity are violated by financial time series, limiting their effectiveness.
4) **Consensus ensemble ($\geq$3 models)** achieves balanced performance (F1=0.740) with reasonable precision (0.694) and recall (0.794). While not matching the best individual models, consensus provides robustness and confidence estimates valuable in production systems.

For practitioners deploying real-time anomaly detection systems, we recommend the Kalman Filter for applications prioritizing precision and Isolation Forest for applications prioritizing recall. For maximum robustness, an ensemble approach combining multiple models with consensus thresholds offers balanced performance and confidence estimation.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[2] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250-2267, 2014.

[3] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[4] J. S. Hunter, "The exponentially weighted moving average," *Journal of Quality Technology*, vol. 18, no. 4, pp. 203-210, 1986.

[5] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239-250, 1959.

[6] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.

[7] G. Welch and G. Bishop, "An introduction to the Kalman filter," University of North Carolina at Chapel Hill, Tech. Rep., 1995.

[8] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100-115, 1954.

[9] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413-422.

[10] C. C. Aggarwal, "Outlier analysis," in *Data mining*. Springer, 2015, pp. 237-263.

[11] A. Zimek, R. J. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: challenges and research questions," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 1, pp. 11-22, 2014.

[12] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 157-166.

[13] R. S. Tsay, *Analysis of financial time series*. John Wiley & Sons, 2005.

[14] V. Golosnoy, B. Gribisch, and R. Liesenfeld, "Statistical process control for financial surveillance," *Journal of Banking & Finance*, vol. 113, p. 105743, 2020.

[15] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016.