Journal of Quality Measurement and Analysis *JQMA* 21(4) 2025, 329-353
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2104.2025.17

HYBRID INDICATOR SATURATION MACHINE LEARNING FRAMEWORK FOR OUTLIER DETECTION AND VOLATILITY FORECASTING

(*Kerangka hybrid ketepuan Petunjuk - Pembelajaran Mesin bagi Pengesanan Pencilan dan Peramalan Volatiliti*)

WONG HUI SHEIN*, FARID ZAMANI CHE ROSE, JAYANTHI ARASAN & SIM CHONG YANG

ABSTRACT

Unaccounted for outlier can severely affect the stability of models and degrade the forecasting accuracy. In financial time series, an outlier in the prior unconditional mean can introduce systematic biases in forecasts and result in model misspecification, potentially distort the parameter estimates and inference. In order to address these issues, this study proposes a hybrid approach designed to detect and model outlier effectively. The methods build upon the principles of impulse indicator saturation (IIS) by incorporating unsupervised machine learning technique density-based spatial clustering of applications with noise (DBSCAN) to identify the clusters and refine the detection of outliers. The log returns series is served as the primary input for outlier detection and volatility modeling. The proposed hybrid method IIS-DBSCAN-GARCH is assessed through a Monte Carlo simulation study and subsequently applied to five Asia-Pacific stock market daily return series spanning from 20 June 1994 to 23 December 2024. The results consistently demonstrate that IIS-DBSCAN-GARCH outperforms the classical GARCH model that does not explicitly account for outlier in both simulation and empirical study. Besides effectively identify and account for extreme fluctuations, the proposed method enhances volatility estimation and improves out-of-sample forecasting accuracy. In empirical study, the proposed hybrid methods have been compared to the classical indicator saturation method and DBSCAN approach. The findings indicate that the proposed hybrid method is more accurate with respect to identifying the outliers and shows superior in forecasting the financial volatility.

Keywords: outliers, indicator saturation, DBSCAN, GARCH, volatility forecasting

ABSTRAK

Kehadiran pencilan yang tidak diambil kira boleh memberi kesan ketara terhadap kestabilan model dan mengurangkan ketepatan ramalan. Dalam siri masa kewangan, pencilan dalam purata tak bersyarat sebelumnya boleh memperkenalkan bias sistematik dalam ramalan, mengakibatkan spesifikasi model kurang tepat serta menjejaskan anggaran parameter dan inferens. Bagi menangani isu ini, kajian ini mencadangkan satu pendekatan hibrid yang berkesan untuk mengesan dan memodelkan pencilan. Kaedah ini dibina berasasan prinsip Ketepuan Petunjuk Impuls (IIS) dengan menggabungkan teknik pembelajaran mesin tak diselia, iaitu pengelompokan spatial berasaskan ketumpatan dengan hingar (DBSCAN) untuk mengenal pasti kelompok dan memperhalusi pengesanan pencilan. Data pulangan log digunakan sebagai input utama dalam proses pengesanan pencilan dan pemodelan volatiliti. Kaedah hibrid yang dicadangkan, IIS-DBSCAN-GARCH, dinilai melalui kajian simulasi Monte Carlo dan seterusnya diaplikasikan kepada lima siri data pulangan harian pasaran saham Asia-Pasifik dari 20 Jun 1994 hingga 23 Disember 2024. Hasil kajian menunjukkan bahawa model IIS-DBSCAN-GARCH mengatasi prestasi model GARCH klasik yang tidak mengambil kira pencilan, baik dalam kajian simulasi mahupun kajian empirikal. Selain daripada mengenal pasti pencilan secara berkesan, kaedah ini juga meningkatkan anggaran volatiliti serta menambah baik

ketepatan ramalan luar sampel. Dalam kajian empirikal, kaedah ini turut dibandingkan dengan IIS dan DBSCAN secara berasingan, dan menunjukkan prestasi yang lebih baik dari segi pengesanan pencilan dan ramalan volatiliti kewangan.

Kata kunci: pencilan, ketepuan petunjuk, DBSCAN, GARCH, ramalan volatiliti

1. Introduction

The last two decades have seen a growing trend toward outlier detection, leading a proliferation of studies among many researchers and practitioners in recent years. Identifying outliers remains a significant challenge in time series. This task could be defined as identifying observations with atypical features, or trends with unexpected behaviors in data set. Normally, outlier detection approaches are investigated to lessen or remove negative effects on ordinary observations in dataset. However, data mining researchers recognize that, in addition to data cleaning, outlier detection can uncover hidden and significant information across diverse datasets and applications. These include healthcare fraud (Massi *et al.* 2020), network intrusion detection (Catillo *et al.* 2023), fraudulent credit card activities (Islam *et al.* 2023), temporal irregularities in traffic data (Kalair & Connaughton 2021), and disease diagnosis through medical image analysis (Li *et al.* 2023).

The examination of outlier remains an enduring concern in the study of economic and financial time series. In the regional stock indices return distributions show heavy tails which indicate extreme movements happen at increased rate compared to standard models suggest. Previous researches often link this situation to the presence of abnormal data in the dataset. Irregular data in the financial and economic time series can affect the application of modelling approaches, with implications for various areas such as risk management, policy evaluation and forecasting. In today' global economy, outlier analysis has become a research priority in financial economics areas (Shukla & Sengupta 2020; Venkateswarlu *et al.* 2022; Savić *et al.* 2022; Hajek *et al.* 2023). In relation to stock markts, identifying and interpreting outliers in stock market trends is important for attaining accurate volatility analysis and forecasting (Ranjan *et al.* 2021), risk assessment (Savić *et al.* 2022), and asset valuation (Yaqoob & Maqsood 2024). These irregularities can alter model outcomes and give a direct impact in financial decisions making.

The Indicator Saturation (IS) approach was first introduced by Hendry (1999) which represents a promising recent technique for identifying outliers. It follows a general-to-specific methodology, starting with impulse indicator saturation (IIS), which was developed to identify anomalies arising at uncertain points in time and varying length. In this approach, a pulse dummy variable is introduced for every observation in the data set as a potential intervention. When a dummy variable shows statistically significant at a given time point, it signals the presence of an additive outlier. However, in term of computational perspective, IIS procedure can cause high generation of regressors compared to data points. This issue can be addressed by spliting the dummy set into smaller blocks and apply sequential identification (Hendry & Krolzig 2005). Several studies have explored this approach in various economic contexts (Castle *et al.* 2012; Santos *et al.* 2008; Johansen & Nielsen 2009; Ericsson & Reisman 2012; Rose *et al.* 2021; Muhammadullah *et al.* 2022; Mohamed *et al.* 2024; Khan *et al.* 2024).

The IIS approach has proven effective in detecting outliers by testing multiple potential indicators, either simultaneously or sequentially, under strict significance levels (Johansen & Nielsen 2009; Rose *et al.* 2021; Muhammadullah *et al.* 2022; Mohamed *et al.* 2024). This process, often executed using the general-to-specific (GETS) modeling approach, adjusts

rejection thresholds for numerous comparisons to minimize false positives. However, in small samples, the reliance on numerous indicator variables can still lead to an increased rate of false positives, making IIS a conservative method for outlier detection. Firstly, while this conservatism helps control Type I errors, it may also fail to distinguish between genuine outliers and random noise, particularly in datasets with complex dynamics and varying densities. Secondly, impulse indicator saturation method is primarily linear and may struggle with datasets containing non-linear patterns or varying densities which are common in financial time series. Additionally, most existing methods focus exclusively on either statistical modeling or clustering techniques. To the best of our knowledge, there are scarce studies have explored hybrid approaches that combine both methodologies for enhanced detection. High-frequency financial data often exhibit extreme volatility, noise, and frequent structural changes, which make outlier detection more challenging. The challenges serve as a motivation for this study to propose a hybrid approach that leverages different attributes of anomalies to provide a versatile solution applicable to different scenarios.

The primary objectives of this study are (1) to address the gaps in existing indicator saturation methods for detecting outliers, (2) to improve the accuracy of outlier detection, and (3) to enhance the precision of volatility forecasting in high-frequency financial time series. The present work explores the volatility modeling using GARCH(1,1) model, assuming that the return series has a zero conditional mean and therefore does not require a separate mean model. This assumption is common in financial volatility forecasting, particularly when analysing high-frequency financial data.

To achieve these objectives, this research yields several key contributions. First, a hybrid approach for outlier detection is proposed in addressing the conservativeness of impulse indicator saturation method, which can result in false positives or failures to identify the distiction between noise and genuine outliers, especially in datasets with complex dynamics or varying densities. It also highlights the challenge of applying these approaches to non-linear datasets, such as financial time series, which often exhibit non-linear patterns, high volatility, and frequent structural changes. Second, to achieve highly accurate results with both efficiency and stability, three statages are included in the outlier detection process, namely filtering, refinement, and feature engineering. In this proposed hybrid approach, density-based spatial clustering of applications with noise (DBSCAN) is incorporated as a second stage refinement into indicatior saturation method which act as a refinement to identify the presence of outliers.

Thirdly, proposed approach is applied to GARCH(1,1) model and compared with one-period-ahead GARCH forecasts against realized returns. Evidence from Monte Carlo simulations confirms the robustness of the technique in detecting the outliers effectively. Following the approach of Franses and Ghijsels (1999), the out-of-sample predictive performance of proposed outlier-corrected model is benchmarked against the standard GARCH model, which does not account for outliers. By integrating these methodologies, the study offers a more versatile solution capable of addressing the limitations of traditional indicator saturation methods in dealing with non-linear patterns and varying densities, such as high-frequency financial data.

The remaining structure of the paper is outlined as follows. The related works about classical outlier detection techniques are discussed in Section 2. Section 3 presents the algorithms of indicator saturation, DBSCAN and the proposed hybrid method. Section 4 describes about the design and performance of Monte Carlo simulation. Section 5 covers empirical study with real stock market data. Section 6 discusses the findings on the detection efficiency of the proposed hybrid approach and comparisons with different techniques in terms of volatility forecasting and followed by Section 7 conclusion.

2. Literature Review

Detecting, analysing, and addressing outliers has become a prominent focus across various disciplines and applications. Financial and economic time series, particularly stock market, are critical areas of study due to their significance in modern economies. Understanding the historical occurrence and impact of outliers is essential for optimizing resource allocation across diverse investments and facilitating the trading of publicly listed companies' securities. A substantial body of research has examined these anomalies and their effects. This literature explores various techniques for identifying outlying data in the stock market and related assets. From a classical perspective, Hawkins (1980) has provided a commonly accepted clarification of the term "outlier". Outlier is an observation which deviates markedly from the rest of the data, to the extend that it trigger concern it was caused by a distinct procedure. Thus, outliers can be regarded as observations that diverge from the expected pattern.

Researchers have extensively applied GARCH-type models to explore the effects of extreme values on volatility dynamics. These models had been applied to examine how the extreme values affect the prolonged volatility in varying fields and the finding showed accounting for outliers improves the volatility modelling. Charles (2008) applied GARCH models to detect and adjust for outliers in returns from 17 French stocks and the CAC40 index. The research discovered that parameter estimates in the volatility equation are vulnerable to the distortion from outliers, and correcting for outliers leads to better volatility forecasts.

Grané and Veiga (2010) introduced a wavelet-based method for outlier detection and correction, applicable to various volatility models, proving its effectiveness on three major stock indices which are S&P500, FTSE100 and Dow Jones. Additionally, Ané *et al.* (2008) proposed an AR(1)-GARCH(1,1) approach for identifying anomalies in weekly return data from five Asian stock markets. The findings showed that this technique improved forecasting accuracy compared to traditional GARCH models. Charles and Darné (2012) investigated outliers in ten international stock market indices to study the influence of the September 11, 2001 terrorist attacks in USA, finding that these events caused both temporary and permanent significant shocks in global stock markets.

A wavelet-based detection and correction method was proposed by Grané and Veiga (2010) that is able to apply to a broad spectrum of volatility models. This approach effectively identifies both individual outliers and clusters of outliers while significantly reducing false positivess. The findings also revealed that addressing outliers in the data reduces skewness and excess kurtosis in return distributions, resulting in more accurate return predictions than when outliers are left uncorrected.

The issue of detecting outliers and time-varying jumps was addressed by Dutta and Bouri (2022) in cryptocurrency markets. The results indicated that the outliers were present only in Bitcoin returns, emphasizing the significance of the outliers and the presence of time-varying jumps in Bitcoin returns once outlying observations were adjusted for. Shehadeh *et al.* (2022) employed a boxplot approach to check and analyze extreme returns in daily basis across fourteen stock market indices internationally. The findings indicated that the proportion of the outlying returns ranged from 4% to 10%, with an average of 6%. In more conservative terms, approximately 1.4% of return series observations were classified as extreme outlying data. Negative outliers were observed to be severe, impactful, transmissible and frequent.

A technique was proposed by Akpan *et al.* (2023) who intended to mitigate the impact of outliers from existing heteroscedastic models to more effectively capture excess kurtosis in return series. Both existing and outlier-corrected models were assessed, including autoregressive conditional heteroscedasticity (ARCH), generalized autoregressive conditional heteroscedasticity (GARCH), exponential GARCH (EGARCH), and Glosten, Jagannathan, and

Runkle GARCH (GJR-GARCH) under both normal and Student-t distributions. These models were assessed based on ability to capture excess kurtosis compared to the theoretical kurtosis value. Liu *et al.* (2024) examined the effectiveness of generalized autoregressive conditional heteroskedasticity mixed data sampling (GARCH-MIDAS) and additional outliers corrected GARCH-MIDAS model (AO-GARCH-MIDAS) in forecasting stock market volatility by incorporating the volatility impacts of macroeconomic variables. The importance of including realized volatility alongside macroeconomic factors are highlighted in the study to better capture these volatility effects. Cai *et al.* (2021) introduced the Realized GARCH model with additive and innovative outliers (Realized GARCH-AI model) for volatility forecasting. Via this technique, the atypical returns and realized volatility are identified and rectified through model prameters estimation and outlier tests.

In summary, existing studies have provided noteworthy contributions to understanding of outlier detection and volatility forecasting. However, several gaps remaining, particularly concerning the drawbacks of impulse indicator saturation, the integration of advanced indicator saturation method and their application to volatility forecast in high-frequency financial time series. These gaps bring attention to the necessity for a more robust approach, which this study aims to address. The following section outlines the methodology employed to overcome the challenges and proposes a novel solution.

3. Methodology

This section details the methodologies used in this study, covering both theoretical foundations and implementation procedures. First, the core techniques are introduced, namely Impulse Indicator Saturation (IIS) for detecting potential outliers, DBSCAN for refining outlier classification, and GARCH for volatility modeling. Next, the proposed hybrid IIS-DBSCAN-GARCH method is presented, outlining its step-by-step implementation and integration into the forecasting framework. The detailed explanation of each method ensures clarity in understanding how they contribute to improve outlier detection and volatility forecasting.

3.1. *Impulse indicator saturation*

Indicator saturation refers to a general-to-specific framework which an indicator of a particular type is added to the set of candidate regressors for each observation. This indicates that T indicator variables are introduced for T observations. IIS was the first method introduced by Hendry (1999) to detect the outliers occurring at different positions with uncertain magnitudes. While IIS is capable of generating a complete set of indicator variables, not all indicators are integrated as regressors to prevent overfitting in general-to-specific modelling process. Including all indicators can result in more regressors than observations in the model, and leading to degree of freedom deficiency (Castle *et al.* 2021). If $I_j(t)$ denotes an indicator variable, then in the IIS case $I_j(t)$ is a pulse dummy that takes the value 1 for $j = t$ and equal to 0 otherwise for $j = 1, \dots, T$.

The distributional characteristics of IIS has been studied (Santos *et al.* 2008; Johansen & Nielsen 2009) when the data is simulated in accordance with the model $y_t = \mu + \varepsilon_t$, $t = 1, \dots, T$, where ε_t is normally and independently distributed with mean zero and variance σ_ε^2. In pursuit of this purpose, IIS is integrated into the model for y_t using the split half approach. To be specific, for the initial half of the observations, $T/2$ impulse indicators are incorporated in the model and thus leads to

$$y_t = \mu_t + \sum_{j=1}^{\frac{T}{2}} \varphi_j I_j(t) + \varepsilon_t, \quad \text{for } t = 1, \dots, \frac{T}{2} \tag{1}$$

where $I_j(t)$ represents impulse indicator vector and ε_t is the error term which $\varepsilon_t \sim NID\ (0, \sigma_\varepsilon^2)$. Simultaneously, the another half of the sample results to

$$y_t = \mu_t + \sum_{j=1}^{T-\frac{T}{2}} \varphi_j I_j(t) + \varepsilon_t, \quad \text{for t} = \frac{T}{2}, \dots, T. \tag{2}$$

At the specified significance level α, significant indicators are identified based on the t-statistic value in the initial half of the observations. The position of the significant indicators will be captured and recorded. Next, $\frac{T}{2}$ impulse indicator is added to the second half of the observations, $T - \frac{T}{2}$ following the repeated execution of the identification procedure to identify the significant indicators under null hypothesis, assuming no outliers. At last, two sets of significant dummy variables are combined to contruct a terminal model. The identification of significant indicators follows sequential selection process where non-significant indicators are removed one at a time at chosen significance level. As an alternative, non-sequential selection eliminates non-sigificanct indicators simultaneously at the chosen α for every segment. The retained indicators are considered as significant indicators. This method is always feasible if the number of regressors, N equals to the number of observations, T. However, if the total number of regressors, N, exceeds the number of observations, T, a cross block algorithm proposed by Hendry and Krolzig (2005) is considered. This algorithm divides all indicators into m blocks, and the selection algorithm is applied repeatedly. Therefore, this yields a total of $m - (m-1)/2$ runs of the identification process. This approach has been applied by researchers in the context of basic structural models (BSM) (Marczak & Proietti 2016) and local level models (Nasi *et al.* 2022).

However, one of the major drawbacks of impulse indicator saturation is that these indicators are tested against strict significance levels, often using a general-to-specific (GETS) modelling approach. While this controls for false positives or Type I errors, it also makes the approach conservative, potentially failing to distinguish between genuine outliers and noise, especially in datasets with complex dynamics or varying density. To address this, the present stuidy proposes extending IIS by incorporating an unsupervised machine learning method DBSCAN.

3.2. DBSCAN

Clustering is one of the well-known issue in detection of outlier, and there are myriad methods of clustering algorithms in the past literature. The methods include hierarchical approaches, density-based approaches, partitioning approaches and grid-based techniques. Density-based spatial clustering of applications with noise (DBSCAN) was developed by Eskin *et al.* (2002). This technique is widely applied in data mining and machine learning to group observations according to their spatial proximity. DBSCAN is effective in anomaly detection as it can differentiate dense clusters from isolated noise points. Many researches showed that DBSCAN outperforms other algorithms in detecting anomalies, especially in large and dense datasets (Dokuz *et al.* 2020; Saeedi Emadi & Mazinani 2018; Hahsler *et al.* 2019).

In this study, DBSCAN approach is chosen to identify the cluster of the outliers in financial time series data. Its advantage includes ability to identify clusters of different shapes while handling noisy data in spatial and non-spatial high dimensional datasets. This versatility is beneficial when handling volatile financial time series. In addition, density-based clustering method do not assume parametric distributions or use variance. Hence, DBSCAN posses the

ability of searching arbitrarily shaped clusters, to deal with high numbers of noise, and do not necessitate prior knowledge about how to configure the cluster number. To understand the concept, it is crucial to first review the key definitions used in DBSCAN and related algorithms. The definitions and pseudo code provided in this section are derived from Eskin *et al.* (2002) but have been adjusted for consistency with the other algorithms examined in this study.

Clustering begins with a dataset D which is comprised of a set of points $p \in D$. A fundamental principle of DBSCAN is that, for a point to be part of a cluster, its neighborhood defined by a specified radius (ε) must contain at least a minimum number of point (minPts). In other words, the neighborhood' cardinality must surpass a specific threshold. DBSCAN calculates the density around each point using the concept of ε-neighborhood.

Definition 1. ε-neighborhood. The ε-neighborhood of an arbitrary point $p \epsilon D$ is denoted as

$$N_\varepsilon(p) = \{q \epsilon D \mid dist(p, q) < \varepsilon\}$$

where D is the database of objects and dist is the distance measure and $\varepsilon \in \mathbb{R}^+$. Since $p \in D$, this indicates that point p is always part of its own ε -neighborhood, which $p \in N_\varepsilon(p)$consistently applies.

In accordance with this definition, the magnitude of the neighborhood $|N_\varepsilon(p)|$ is a straightforward unnormalized kernel density estimate around point p, utilizing a uniform kernel with a bandwidth of ε. In DBSCAN approach, $N_\varepsilon(p)$ and a threshold referred to as minPts are used to identify dense areas and categorize the points as core, border, or outlier in the dataset.

Definition 2. Point classes. Given a dataset D, a point $p \in D$ is classified as follows based on the cardinality of its ε-neighborhod $N_\varepsilon(p)$.

$$Point\ class\ (p) = \begin{cases} \text{core point}, & \text{if } |N_\varepsilon(p)| \geq \text{minPts} \\ \text{border point}, & \text{if } \exists q \in D \text{ such that p} \in N_\varepsilon(p) \text{ and } q \text{ is a core point} \\ \text{outlier}, & \text{otherwise} \end{cases}$$

where minPts $\in Z^+$ is a user-specified density threshold.

The magnitude of the neighborhood for some data points is illustrated as a circle, and an annotation denotes as their class as displayed in Figure 1. DBSCAN approach establishes the concepts of reachability and connectedness to group individual points into contiguous dense regions.

Definition 3. Directly density-reachable. A point $q \in D$ is directly density-reachable from a point $p \in D$ with respect to ε and minPts if, and only if, $|N_\varepsilon(p)| \geq$ minPts and $q \in N_\varepsilon(p)$, where p is a core point and q is in its ε-neighborhood.

Definition 4. Density-reachable. If the presence of an arranged collection of points $\{p_1, p_2, \dots, p_n\}$ with $q = p_1$ and $p = p_n$, then a point p is considered as density-reachable. In other words, p_{i+1} has direct reachability from $p_i \ \forall\ i \in \{1, 2, \dots, n-1\}$.

Definition 5. Density connected. If a single point $o \in D$, then the point $p \in D$ is a density-connected to a point $q \in D$. In other words, p and q are density-reachable from o. As shown in Figure 2, the notion of density-connection can be applied to generate clusters as contiguous dense regions.

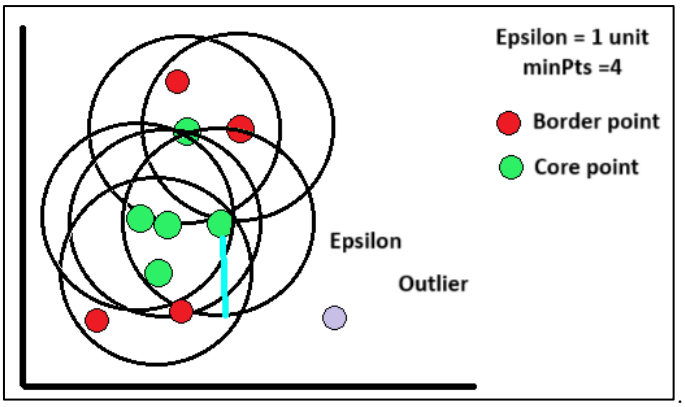


Figure 1: Illustration of DBSCAN concept with core points, border points and outlier.

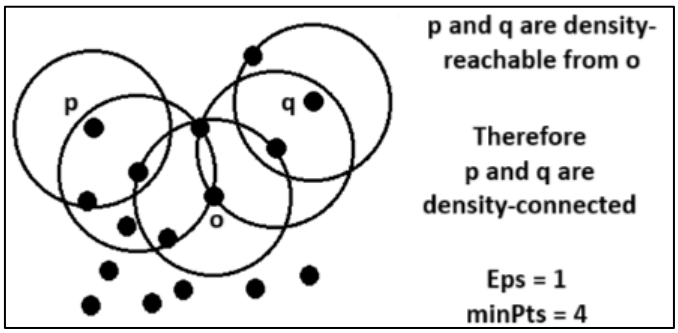


Figure 2: Illustration of density reachability and density-connectivilty.

Definition 6. Density-based cluster. A density-based cluster C is a non-empty subset of D fulfilling the criteria of maximality and connectivity. For maximality, $q \in C$ if $p \in C$ and q is density-reachable from p while connectivity implies that $\forall\, p, q \in C, p$ is density-connected to q.

Clusters are detected by DBSCAN algorithm by locating core points and expanding them to include entire points that are density-reachable. The algorithm starts with a randomly selected point p and its ε-neighborhood is retrieved. A new cluster will be initiated if it is a core point, followed by expansion to encompass entire points within its neighborhood in the cluster. If there is no additional core points in the expanded neighborhood, then the cluster is considered as complete. The remaining unclustered points are subsequently analyzed to check if there are any other core points presents to initiate a new cluster. The points that are not assigned to any cluster after all the points have been analysed will be categorized as outliers (Monko & Kimura 2023).

One of the most popular clustering algorithms, DBSCAN, is frequently cited in scientific research (Birant & Kut 2007). DBSCAN algorithm identifies clusters as dense regions of data points separated by low-density areas. Many researchers have utilized the DBSCAN algorithm in their studies due to its benefits. For instance, Saeedi Emadi and Mazinani (2018) proposed a novel DBSCAN-based algorithm for anomaly detection. With the growing interest in blockchain technology, DBSCAN algorithm has also been applied to identify anomalies in Bitcoin price data (Dokuz *et al.* 2020). Additionally, a modified version of DBSCAN was introduced by Jain *et al.* (2022) to detect anomalies in seasonal time-series datasets. For example, by examining the dataset as a whole, DBSCAN can effectively detect global

anomalies in monthly temperature data, but it might not be able to identify local anomalies within individual months.

DBSCAN is a density-based spatial clustering algorithm that capable of identifying anomalies within series of data. Data points are grouped into clusters using this density-based clustering technique according to their density and proximity within the data space. DBSCAN is widely applied owing to its capability to capture clusters of arbitrary shapes and its robustness in identifying outliers or noise. Neighborhood distance epsilon (ε) and minimum number of points (minPts) are two user-defined parameters that DBSCAN relies on (Çelik *et al.* 2011). For a given point, the points in the ε distance are considered as its neighbors. Those points form a cluster when the number of neighboring points of a point exceeds minPts. The data points are classified as core points, border points and outlier points by DBSCAN. Core points are those with at least minPts points within the ε distance. Border points are not core points but are neighbors of core points. Outlying points, on the other hand, fall into neither core or border points categories.

3.3. GARCH

In this study, standard GARCH(1,1) model is assumed as the primary volatility modelling framework. This model is widely regarded as a benchmark for capturing volatility clustering in financial markets. Volatility clustering describes a pattern in which the variance of a time series depends on its past variances. This means that periods of large stock price fluctuations tend to be followed by further large fluctuations, regardless of direction, while periods of small price movements are followed by continued low volatility. This phenomenon results in a strong autocorrelation in squared returns. It is formally known as Autoregressive Conditional Heteroskedasticity (ARCH) or simply the ARCH effect.

In this study, the lag order is fixed to GARCH(1,1) following the principles of parsimony, as this model is widely used and captures volatility clustering effectively. While higher-order models AR(p)-GARCH(m,s) could be explored, Bollerslev (1986) and Charles (2008) suggested that GARCH(1,1) often provides comparable performance with lower computational costs.

The idea of time series models allowing for heteroscedasticity was initiated by introducing autoregressive conditional heteroscedastic (ARCH) model (Engle 1982). This idea was then extended to GARCH model by Bollerslev (1986) which provide more parsimonious outcome than ARCH models. The GARCH(1,1) model specifies that the conditional variance depends on previous squared returns and past conditional variances.

Let R_t represents the market return of an asset at time t and is derived from $R_t = (logP_t - logP_{t-1}) \times 100$, where P_t denote the asset price at time *t*. The standard GARCH(1,1) model, which is the simplest and most widely used form of the GARCH family, is represented by

$$R_t = \sigma_t \varepsilon_t, \varepsilon_t \sim N(0,1), \tag{3}$$

$$\sigma_t^2 = \omega + \alpha R_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{4}$$

where $\omega > 0, \alpha \geq 0$ and $\beta \geq 0$ for the positivity of variance and $0 \leq \alpha + \beta \leq 1$ for the stationarity of variances, while R_t corresponds to demeaned return series (mean-adjusted), implying that the conditional mean $E[R_t]$ is assumed to be zero. This assumption is common

in volatility modelling when the primary focus is on the conditional variance dynamics rather than the mean process.

Hansen and Lunde (2005) uncovered no indication that the GARCH(1,1) model is deficient to other ARCH-type models from the conclusion of the basis of empirical study. GARCH(1,1) model is the simplest specification in the GARCH family, assumes that an asset return R_t can be decomposed into a forecastable part, namely the conditional expectation $E(R_t|F_{t-1})$ and a random noise ε_t. F_{t-1} represents the data filtering at time $t-1$. The error term, ε_t, is assumed to be an independent and identically distributed (i.i.d.) process with zero mean and unit variance. Under classical assumption, this implies a conditional Gaussian distribution which is expressed as

$$\varepsilon_t|F_{t-1} \sim \text{i.i.d. } N(0,1). \tag{5}$$

While various extensions of the GARCH model exist, including Exponential GARCH (EGARCH) and Glosten-Jagannathan-Runkle GARCH (GJR-GARCH), the standard GARCH(1,1) model is used in this study as a baseline due to its well-established effectiveness in volatility forecasting. The key distinction is that an AR(1)-GARCH(1,1) model specifies equations for both the conditional mean and the conditional variance, whereas the standard GARCH(1,1) model specifies only the conditional variance and assumes a constant (or zero) mean.

A notable advantage of GARCH models is their ability to generate multi-step-ahead forecasts with relative ease (Ané *et al*. 2008). With the model defined in Eq. (3) and Eq. (4), the one-period-ahead return forecasts is expressed as

$$R_{t,t+1} = E(R_{t+1}|F_t) = a_0 + a_1 R_t. \tag{6}$$

The variance associated with this forecast is derived in closed-from as follows

$$\sigma^2_{t,t+1} = var(R_{t+1}|F_t) = \alpha_0 + (\alpha_1 + \beta)\sigma_t^2. \tag{7}$$

Given the conditional distribution of standardized errors in Eq. (7), the derivation of the one-period-ahead return interval forecast is relatively direct. At $(1-\alpha)\%$ confidence level, the conditional return at time $t+1$, given the available information set at time t , fall within the interval

$$R_{t+1} \in \left[R_{t,t+1} \pm F\left(1-\frac{\alpha}{2}\right)\sigma_{t,t+1}\right], \tag{8}$$

where $F\left(1-\frac{\alpha}{2}\right) = P\left(Z_t \leq 1-\frac{\alpha}{2}\right)$ corresponds to a specific quantile of the assumed conditional distribution.

3.4. *Proposed hybrid IIS-DBSCAN-GARCH method*

The proposed multi-stage method for detecting and handling outliers, incorporating IIS and DBSCAN for improved accuracy is presented in this section. The hybrid outliers detection procedure are comprised of three main phases. An initial filtering stage is started with IIS method, then followed by a refinement stage where DBSCAN is applied to validate potential outliers, and a final feature engineering stage to analyze the characteristics of the confirmed outliers. The outlier corrected model is then applied to forecast volatility using GARCH model,

and its effectiveness is evaluated through a comparative analysis of corrected and uncorrected GARCH variance. The proposed procedures are illustrated in Figure 3 and explained as follows.

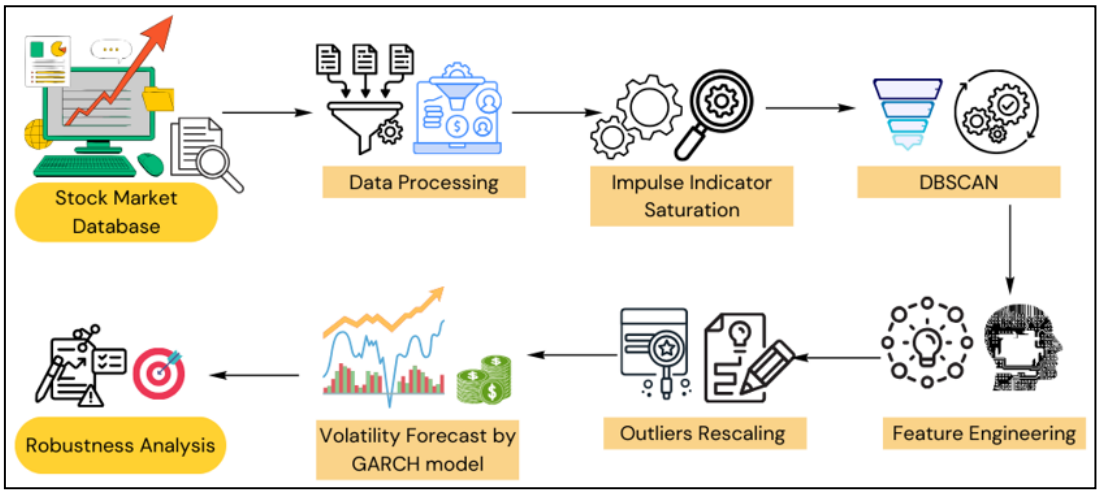


Figure 3: Procedure of proposed hybrid IIS-DBSCAN-GARCH method

Step 1. Data Preprocessing
The first step involves calculating log returns for the stock market's closing prices. Log returns are used as they standardize price changes and are appropriate for modeling the proportional changes in financial data. The log returns are computed for the entire time series and serve as the primary input for outlier detection and volatility modeling.

Step 2. Proposed hybrid detection procedure
The hybrid outlier detection is conducted in three stages, beginning with initial detection using Impulse Indicator Saturation, followed by refinement using DBSCAN, and lastly with feature extraction for detailed analysis.

- Stage 1: Impulse indicator Saturation Method
  The IIS method is applied to identify potential outliers in the log return series. The method detects abrupt deviations from expected patterns by introducing indicator variables into the regression model. Each indicator variable corresponds to a specific time point and isolates significant anomalies. The procedure is carried out using a general-to-specific (GETS) modeling framework. Significant outliers are detected according to pre-specified thresholds or statistical conditions. The output of IIS includes a set of candidate outliers, which are passed to the next stage for validation and refinement.

- Stage 2: DBSCAN Refinement
  DBSCAN is employed to refine the outlier detection process by validating and clustering the anomalies identified by IIS. In order to determine the parameter neighborhood distance epsilon (eps), *k*-nearest neighbour (kNN) plot is used. The distance to the *k*-th nearest neighbor for each candidate outlier is calculated, after which the distances are sorted in descending order and plot to identify the elbow point which indicates the optimal. The minPts parameter in DBSCAN is often set to a default value of 4 which was proposed by Schubert *et al.* (2017), aimed at refining the density estimation. For large datasets, the minPts is commonly set as 4 for two dimensional

data (Ester *et al.* 1996). However, Sander *et al.* (1998) recommended setting it to twice the dimensionality of dataset. For noisy, high-dimensional and extensive datasets, or those containing many duplicates, increasing the minPts value can often yield improved outcomes. DBSCAN is applied to the potential outliers identified by IIS using the optimal parameters derived from the *K*-distance plot. The purpose is to validate whether these points form clusters, indicating dense regions of anomalies, or remain isolated signifying true outliers. A true outlier is defined as an observation that significantly deviates from the expected pattern of a dataset due to genuine underlying causes rather than random variation or noise. This step enhances precision by reducing false positives and refining the detection of outliers.

- Stage 3: Feature Engineering for Finalized Outliers
  In this stage, feature engineering is applied to extract meaningful characteristics from the detected outliers. First, the magnitude of each outlier is quantified as its absolute deviation from the expected value, allowing for a structured comparison of extreme fluctuations. Second, the time gaps between consecutive outliers are measured to detect potential clustering patterns, which may indicate periods of heightened market instability. These engineered features provide deeper insights into the behavior of outliers, aiding in refining volatility modeling and enhancing predictive accuracy.

Step 3. Outlier Replacement
Detected outliers are replaced to create a cleaned dataset for subsequent analysis. The replacement strategy ensures that the adjusted data retains realistic values. As suggested by Ané *et al.* (2008), extreme positive (negative) outlier is substituted with the corresponding upper (lower) bound of the confidence interval as defined in Eq. (8).

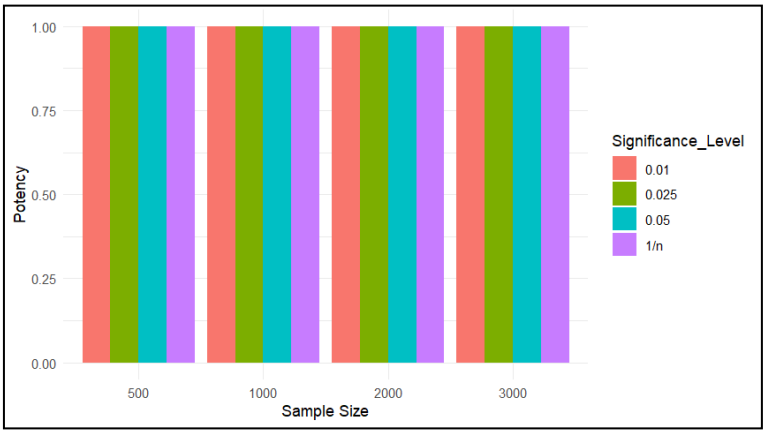
Step 4. Volatility Forecasting
The cleaned dataset is used to compute the adjusted GARCH variance for volatility forecasting. This corrected variance is then compared against the GARCH variance derived from the original log return data containing outliers as well as the realized variance.

4. Monte Carlo Simulation

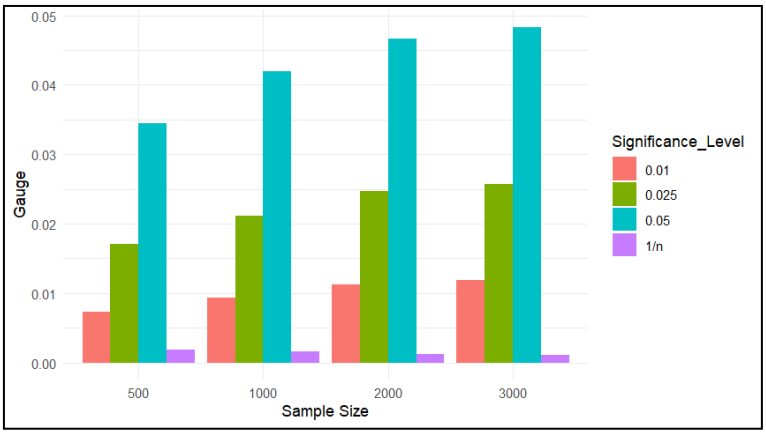
Monte Carlo simulation is carried out to assess the efficacy of the proposed hybrid outlier detection method in identifying outliers in univariate GARCH(1,1) model. By utilizing R language and *gets* package, the procedure begins with generating observations for GARCH(1,1) under Gaussian distribution with the setting of parameters $\alpha = 0.1, \beta = 0.8$ and $\omega = 1 - \alpha - \beta$. The estimation procedure applies certain constraints, with the selected parameters aligning with the approaches used in the previous studies (Carnero *et al*. 2012; Ismail & Nasir 2020; Marczak & Proietti 2016). The simulation is repeated for 1000 times with significance level, $0.01, 0.025, 0.05,$ and $1/n$. The sample size of each simulation are set as $n = \{500, 1000, 2000, 3000\}$ representing typical sample sizes for financial time series. The sample size of 3000 denotes over 10 years of daily data which is generally considered sufficient to comprehend the variability within dataset. The sequential procedure is applied in the stage 1 IIS method as the sequential method is proved outperform to the non-sequential method. Multiple additive outliers (AOs) with varying magnitudes (3σ, 5σ, 10σ, and 15σ) which

encompass both positive and negative, were introduced into the simulated return series, with standard deviation, σ, of the stock return series.

The performance of the IIS-DBSCAN-GARCH method in identifying the outliers is assessed through the concept of potency and gauge as proposed (Marczak & Proietti 2016). Potency is defined as the mean of retention proportion of significant indicator that are retained to the model. Gauge refers to the mean retention frequency of retaining indicator that is statistically significant but irrelevant to the model. If the potency is 70%, this indicates that the method retains 7 out of 10 relevant variables in average. Conversely, when α=0.02, this implies that two irrelevant indicators out of every hundred are expected to be included in the model on average.



(a)



(b)

Figure 4: (a) Potency rate (b) Gauge rate of GARCH(1,1) simulations.

After performed several runs, in stage 2 DBSCAN, the minPts value is fixed from 10 to 60 and then determining the eps value. The simulation results exhibit similar pattern. Thus, for the remaining of the paper, a minPts value of 10 was chosen for the dataset to ensure completeness. Figure 4(a) shows the summary of potency rate of the hybrid method for different sample sizes. It can be observed that as the number of sample size rise, the potency rate for α of 0.01, 0.025, 0.05 and $1/n$ increase substantially to 100%. This indicates that the hybrid method becomes more effective in detecting outliers as more data points are available, reinforcing its robustness for larger datasets. Figure 4(b) illustrates the gauge values using sequential method across

various significance level. The gauge values tend to align closely with the significance level. This indicates that the proposed hybrid method effectively controls false positives. Among the tested significance levels, $\alpha = 1/n$ appears to provide the best balance between potency and gauge control. Therefore, sequential method with setting $\alpha = 1/n$ is the considered as the best practice to ensures a reliable trade-off between detecting the true outliers and minimizing false positives.

5. Empirical Study

For empirical evaluation, the effectiveness of the proposed hybrid method is tested using daily Morgan Stanley Capital International (MSCI) stock index series from five Asia-Pacific countries, namely Hong Kong, Japan, Korea, Singapore and Taiwan. The data series span from 20 June 1994 to 23 December 2024 and were obtained from Datastream. This period includes several episodes of non-normal market conditions, such as Asian Financial Crisis in 1997, Dot-com Bubble burst in 2000, Global Financial Crisis in 2008, COVID-19 pandemic-induced market turbulence in 2020, and periods of heightened volatility in 2022–2023 linked to geopolitical tensions and inflationary pressures. These events provide a diverse range of market environments for testing the robustness of the proposed method.

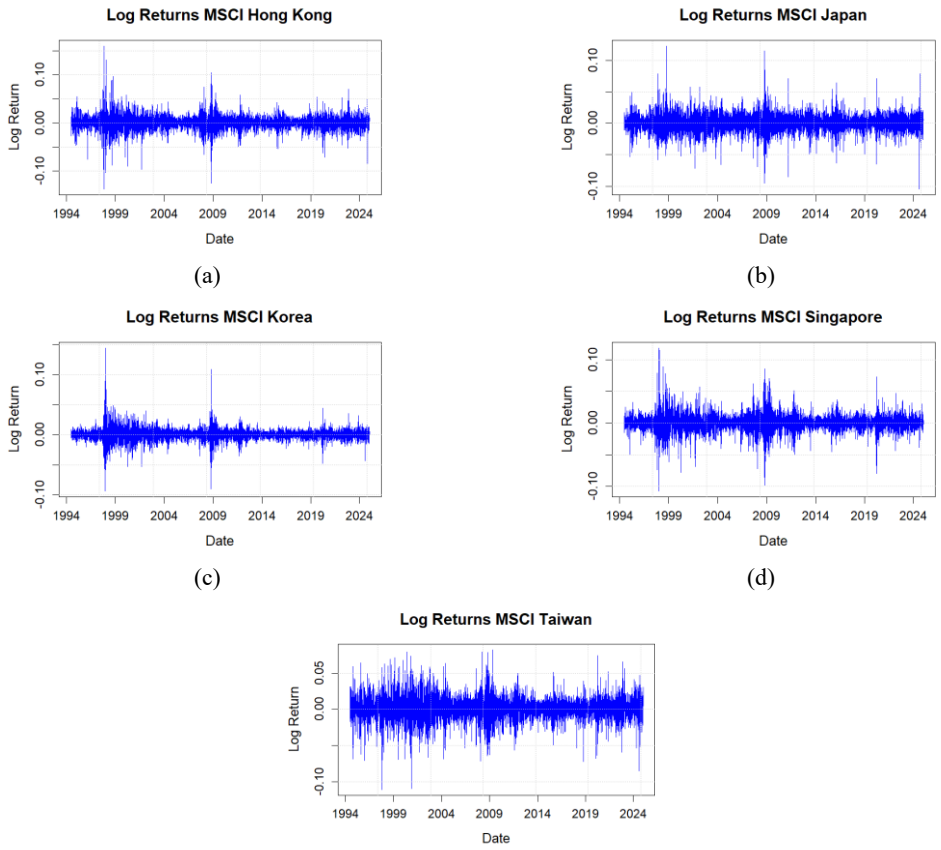


Figure 5: Daily log returns of Asia-Pacific stock index: (a) Hong Kong, (b) Japan, (c) Korea, (d) Singapore, (e) Taiwan

The closing price of the MSCI stock indices are transformed to log return which are computed as the first differenced value of the natural logarithm. Transformation of MSCI stock market indices into log returns is to stabilize variance and ensure stationarity, which is a prerequisite for reliable volatility modeling. The daily log returns of five countries are shown in Figure 5 which show the stock market data volatility.

Table 1 summarizes the descriptive statistics for the MSCI returns of all five stock indices. All the MSCI indices display positive expected returns, with values close to zero mean with negative skewness, excluding Korea. The standard deviations are remain low and nearly uniform across the nations, implying that the variability between return series of Asia-Pacific countries is low. The result reflects the presence of fat-tailed distribution as the kurtosis values of the return series are above 3. Moreover, due the outcome shows significant in Jarque–Bera (JB) statistics, all the MSCI indices indicates not standard normally distributed. The stationarity and residual diagnostics in Table 2 are derived from the GARCH(1,1) model, assessing the adequacy of the volatility specification. According to the result of residual diagnostics test, there appears to have an autoregressive conditional heteroskedasticity (ARCH) impact within the data, implying time-varying volatility in the data series. The Ljung-Box Q-statistics reveals the presence of serial correlation up to 20 lags. However, Augmented Dickey-Fuller (ADF) and KPSS tests indicate that the indices are stationary. In the stock returns, to further to examine long memory behavior, both semiparametric GPH test (Geweke & Porter-Hudak 1983) and Gaussian semiparametric (GSP) test (Robinson & Henry 1999) were performed. The results show that strong evidence of long memory characteristics across all stock index returns. It suggests that the market incorporates new information gradually rather than instantaneously, a behaviour that can be captured using long memory models.

Table 1: Descriptive statistics for MSCI stock index returns

	Descriptive statistics for MSCI stock index returns				
	Hong Kong	Japan	Korea	Singapore	Taiwan
Mean	0.00007	0.00001	0.00011	0.00006	0.00015
S.D.	0.01425	0.01343	0.02103	0.01295	0.01517
Minimum	-0.13772	-0.10491	-0.21644	-0.10760	-0.11130
Maximum	0.16005	0.12272	0.33210	0.11846	0.08235
Skewness	-0.01004	-0.02578	0.40164	-0.03632	-0.11938
Kurtosis	9.61147	4.97366	19.95351	7.36805	3.59754

6. Results and Discussion

This section presents the discussion of the empirical findings. The performance of IIS, DBSCAN and proposed hybrid methods are compared with respect to accuracy in detecting the outliers in the MSCI data series of five Asia-Pacific countries. Overall, based on Table 2 the outliers detected by IIS, DBSCAN (denoted by DB) and hybrid method range between 24 to 88 outliers, which the outlying observations is ranging between 0.0033% to 1.78% over approximately 30 years data series. Similarly, the three methods rank Korea first, recording the most outliers detected(IIS: 67, DBSCAN: 88, Hybrid: 48), then followed by Singapore (IIS: 59, DBSCAN: 67, Hybrid: 43). The hybrid method records Taiwan and Japan with the fewest outliers, at 24 abnormal values, whereas IIS and DBSCAN identified Japan and Taiwan as lowest outliers counts respectively. Number of outliers and magnitude of the extreme values

detected by hybrid method are highly consistent with the unconditional level of kurtosis based on the descriptive statistics for each nation in Table 1.

Table 2: Stationary and residual diagnostic tests for MSCI stok index returns

	Stationary and Residual Diagnostic Tests for MSCI stock index returns				
	Hong Kong	Japan	Korea	Singapore	Taiwan
Jarque-Bera	30561.180**	8189.130**	131876.700**	17954.880**	4298.780**
Q(20)	38.728**	56.875**	138.110**	116.750**	55.8540**
$Q^2(20)$	4011.900**	3016.400**	5718.900**	6204.900**	1974.700**
ARCH test	1445.200**	988.760**	1773.500**	1655.500**	732.560**
ADF	-19.293**	-20.234**	-18.925**	-19.057**	-18.972**
KPSS	0.05309	0.12351	0.06567	0.05092	0.14807
GPH	-0.08073**	0.04363**	0.01076**	-0.02276**	0.01670**
GSP	0.07411**	0.08707**	0.08125**	0.07728**	0.06750**

** signifies significance at the 5% level.

The preliminary analysis of the series in Table 1 indicates that any detected significant asymmetry is persistently negative. This is driven by the higher quantity of negative outliers. Overall, the quantity of negative outliers surpasses the positive outliers except for Japan, Singapore and Taiwan which is closely aligned with the level of development of the countries. The most extreme outlier in absolute value is typically negative, followed by a small magnitude positive outlier which is in line with with Longin (1996) findings.

In addition to investigating the quantity, sign and magnitude of outliers, it is essential for financial analysts to consider the timing and patterns of extraordinary returns. The analysis considers the presence and patterns of consecutive positive or negative outliers, where outliers can occur in continuous sequences of days. Besides, lasting outliers can involve occasional outlying observations within a particular period independent of consecutive occurence. The five return series of Hong Kong, Japan, Korea, Singapore and Taiwan show consecutive, lasting and temporary outlying observations occur. Based on the result in Table 3, each series shows over half of the outliers are consecutive and lasting. If no specific trend can be discerned for the continuous extreme observations, the lasting outlying observation includes negative outliers followed by another negative value. In addition, temporary outliers are also found in each series with varying counts of detections.

By leveraging the strength of DBSCAN, the existence of clusters of outlying observation can be investigated for Asia-Pacific stock market. According to Figure 6, the yellow dots signify potential outlier detected by IIS method while the red dots exemplify finalized outliers after the process of validation and refinement by DBSCAN. The finalized outliers are then plotted into one visual graph to analyse the time and pattern of occurrence of the outliers.

Based on hybrid approach results, a total of five clusters are determined and plotted as displayed in Figure 7. The time frame of the detected clusters of outliers aligns with five major historical economic crises. The first cluster, occurred between 1997 and 1998, corresponded to the Asian financial crisis. This crisis began in Thailand following a sudden devaluation of the Thai baht in mid-1997 (Berg 1999). The devaluation triggered a domino effect that caused currency exchange rates across the Asian region to collapse in a narrow time span (Wu 1998).

The second cluster, detected between 2000 and 2001, coincided with the Dot-Com Bubble. Sparked by Netscape Communications Incorporated's initial public offering in 1995, investors began pouring money into internet-based companies regardless of their profitability and

business models (DeLong & Magin 2006). During this period, the Nasdaq Composite Index which was dominated by tech stocks, soared due to speculative investments. However, by early 2000, scepticism about the profitability of dot-com companies grew. Disappointing earnings reports and concerns over inflated valuations triggered a massive selloff, leading to a stock market crash in the U.S (Wheale & Amin 2003). Many Asian countries, particularly Korea, Japan, and Taiwan which were heavily reliant on technology and electronic exports (notably semiconductors), were significantly impacted as global demand for ICT products collapsed (Liu 2011).

Table 3: Outlier detection outcome of return series

	Outlier detection outcome of the return series														
	Hong Kong			Japan			Korea			Singapore			Taiwan		
	IIS	DB	Hybrid	IIS	DB	Hybrid	IIS	DB	Hybrid	IIS	DB	Hybrid	IIS	DB	Hybrid
Outliers summary															
Total outliers	46	62	29	41	66	24	67	88	48	59	67	43	46	61	24
% of outliers	0.57	0.78	0.37	0.52	0.83	0.30	0.84	1.78	0.60	0.74	0.84	0.54	0.006	0.77	0.003
+ outliers	25	40	8	17	23	17	33	43	14	27	34	27	24	35	12
- outliers	21	22	21	24	43	7	34	45	34	32	33	16	22	26	12
Outliers severity															
Magnitude of 1σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Magnitude of 2σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Magnitude of 3σ	6	22	3	8	33	0	7	28	4	10	18	4	10	25	0
Magnitude of 4σ	15	15	5	19	19	9	32	32	16	26	26	16	29	29	15
Magnitude of 5σ	11	11	7	8	8	6	13	13	13	12	12	12	5	5	7
Magnitude of 6σ	8	8	8	2	2	5	8	8	8	7	7	7	0	0	0
Magnitude of 7σ	2	1	2	2	2	1	2	2	2	1	1	1	2	2	2
Magnitude of 8σ	1	1	1	1	1	2	1	1	1	2	2	2	0	0	0
Magnitude of 9σ	2	2	2	1	1	1	2	2	2	1	1	1	0	0	0
Magnitude of 10σ	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
Magnitude of 11σ	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0
Magnitude of 12σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Magnitude of 13σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Outliers classification															
Consecutive outliers	21	26	10	18	32	10	30	51	27	26	34	21	18	26	11
+/-	11	14	5	11	15	5	15	22	8	12	14	8	13	17	7
-/+	12	14	5	12	16	5	15	21	8	13	16	9	13	17	6
-/-	6	5	4	5	10	2	8	10	6	7	8	5	4	3	3
+/+	9	9	3	3	10	3	7	7	4	6	12	6	6	12	1
Lasting outliers	16	26	16	18	21	7	17	17	16	22	25	18	15	17	8
+/-	6	6	5	6	8	3	4	6	4	6	4	1	1	7	6
-/+	7	7	2	2	4	4	2	4	1	2	5	1	2	6	3
-/-	5	8	7	7	6	5	8	8	8	8	9	7	6	7	7
+/+	2	3	2	0	1	1	1	1	1	1	2	2	1	1	2
Temporary outliers	9	10	3	5	13	5	6	20	5	8	11	4	9	12	5

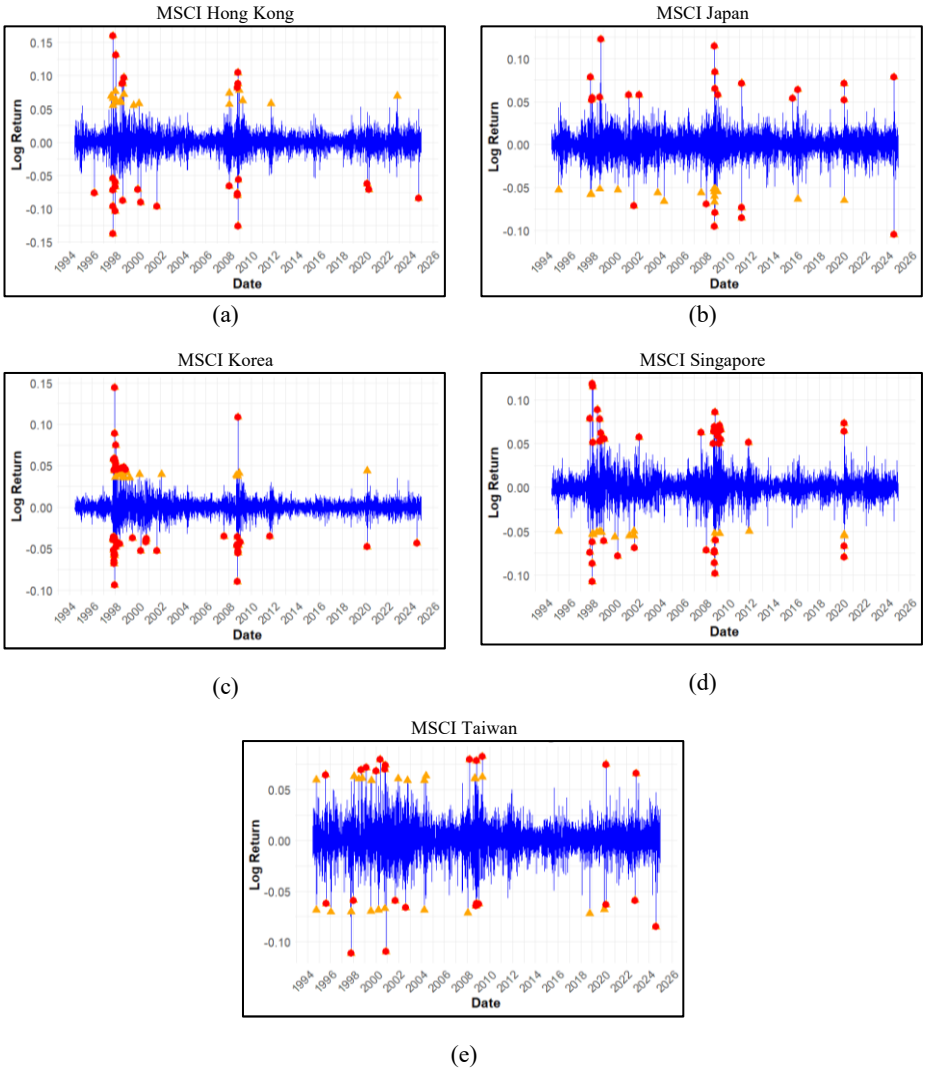


Figure 6: Outlier detection using the proposed hybrid method and the Impulse Indicator Saturation (IIS) method for Asia-Pacific stock indices: (a) Hong Kong, (b) Japan, (c) Korea, (d) Singapore, and (e) Taiwan.

The third cluster, detected between 2008 and 2009, coincided with the global financial crisis. Originating in the U.S. housing market, banks issued risky subprime mortgages to unqualified borrowers (Kramer 2022). The combination of rising property prices, low interest rates, and lax lending policies encouraged speculative borrowing. These mortgages were subsequently bundled into mortgage-backed securities and sold globally, creating a widespread network of financial vulnerability (Baily *et al.* 2008). The fourth cluster coincided with the European sovereign debt crisis in 2011. Triggered by high sovereign debt levels in several European countries, namely Greece, Portugal, Ireland, Spain and Italy, this crisis exposed fiscal vulnerabilities exacerbated by the skyrocketing government debt levels as these countries borrowed heavily to stabilise their economies following financial crisis happen globally in 2008 (Coillignon 2012).

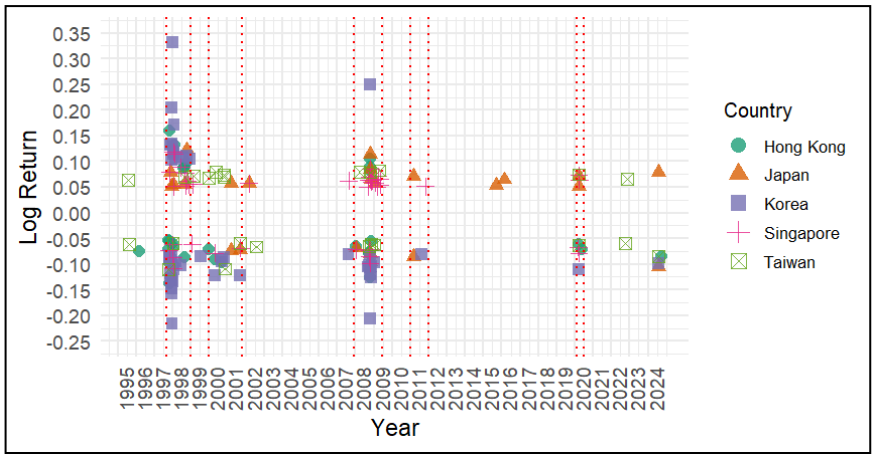


Figure 7: Outliers detection by five Asia-Pacific countries from 1994 till 2024

The final cluster occurred in 2020 during the COVID-19 pandemic, which originated in China in late 2019. The pandemic quickly spread worldwide, causing widespread health, economic, and social disruptions. In response to the pandemic, governments globally enforced stringent measures, such as quarantines, lockdowns and travel bans. These interventions, while effective in containing the virus, resulting in a sharp economic slowdown driven by declining consumer demand (Naseer *et al*. 2023).

Before proceeding to volatility forecast, this study conducted a statistical treatment of outliers with the purpose of substituting of the outlying data, thereby restoring a filtered series that is suitable for subsequent forecasting and myriad applications. This method was also employed by Ané *et al*. (2008) and their findings indicated an improvement in forecasting future values. Thus, this study applied the similar method in replacing the finalized outliers with the interval forecasts as specified in Eq. (8).

The descriptive statistics is carried out after the replacement of outliers detected by hybrid method as presented in Table 4. While the values show no significant changes, a decrease is observed in the standard deviation of the unconditional distribution of the stock return series, with a reduction of up to approximately 19% for the MSCI Korea series. Additionally, according to Table 5, the residual diagnostics are based on the fitted GARCH(1,1) model. The normality test is no longer significant at the 5% level, indicating that normality is achieved in the series after the adjustment. This suggests that the abnormalities often seen at the daily frequency are primarily driven by a few observations with extremely large absolute values, which are adequate to reject the standard Gaussian model.

Table 4: Descriptive statistics for refined MSCI stock index returns

	Descriptive statistics for Refined MSCI stock index returns				
	Hong Kong	Japan	Korea	Singapore	Taiwan
Mean	0.00009037	0.00001361	0.00010889	−0.00000012	0.00015548
S.D.	0.0129861	0.0126031	0.0177598	0.0117581	0.0144152
Minimum	-0.0791986	-0.0754087	-0.0801977	-0.0690889	-0.0866706
Maximum	0.0703438	0.0510002	0.0805632	0.0496667	0.0812454
Skewness	-0.0992455	-0.0928211	-0.0750833	-0.1797144	-0.0529831
Kurtosis	5.319802	4.408892	5.670733	5.276070	5.024139

Lastly, GARCH(1,1) is used to produce precice volatility forecasts for the evolution of key parameters. An out-of-sample dynamic analysis is presented to compare the performance of the classical GARCH, IIS-GARCH, DBSCAN GARCH and the proposed IIS-DBSCAN-GARCH models. The out-of-sample forecasting horizon cover from 17 November 2018 till 23 December 2024. This comparison employs a rolling estimation technique, simulating a real-world scenario, which daily updates to the database enable a model recalibration and subsequent volatility forecasts. These forecasts can be applied to various purposes, such as investment strategy risk assessment, and the valuation or risk management of financial instruments. This approach has been extensively used by many researchers to assess the predictive capability of various GARCH model specifications (McKenzie & Mitchell 2002). In term of evaluating the accuracy of one-step-ahead forecast performance, the results in Table 6 demonstrate that the hybrid IIS-DBSCAN-GARCH method consistently outperforms the classical GARCH, IIS-GARCH and DBSCAN-GARCH methods with lowest MSEV, MAED, MAE and RMSE. Table 7 shows the result of Diebold-Mariano (DM) test. The output shows that none of the competing models outperform the IIS-DBSCAN-GARCH model in forecasting volatility across the five Asia-Pacific markets.

Table 5: Stationary and residual diagnostic tests for refined MSCI stok index returns

	Stationary and Residual Diagnostic Tests for Refined MSCI stock index returns				
	Hong Kong	Japan	Korea	Singapore	Taiwan
Jarque-Bera	1792.730	668.182	2365.445	1755.292	1357.986
Q(20)	29.107**	48.273**	119.262**	105.291**	50.208**
$Q^2(20)$	3817.39**	2923.62**	5217.01**	5982.92**	1610.49**
ARCH test	1317.24**	917.18**	1193.36**	1267.43**	691.83**
ADF	-15.391**	-19.318**	-16.260**	-17.236**	-15.169**
KPSS	0.04746	0.11095	0.05187	0.05324	0.13979
GPH	-0.06475**	0.04417**	0.01942**	-0.03461**	0.02645**
GSP	0.06896**	0.05880**	0.08416**	0.06961**	0.07817**

** signifies significance at the 5% level.

Table 6: Out-of-sample forecasting assessment results.

Stock Market	Model	MSEV	MAED	MAE	RMSE
Hong Kong	GARCH	4.217108e-07	0.007901582	0.0002252911	0.0006493926
	IIS-GARCH	4.216401e-07	0.007868673	0.0002239923	0.0006493382
	DBSCAN-GARCH	4.209065e-07	0.007867228	0.0002239635	0.0006487731
	IIS-DBSCAN-GARCH	**4.184177e-07**	**0.007674568**	**0.0002134683**	**0.0006468522**
Japan	GARCH	2.027655e-07	0.007295987	0.0001904472	0.0004505672
	IIS-GARCH	2.024063e-07	0.007267026	0.0001892795	0.0004504633
	DBSCAN-GARCH	2.081212e-07	0.007260194	0.0001891387	0.0004562030
	IIS-DBSCAN-GARCH	**2.020540e-07**	**0.007088035**	**0.0001828047**	**0.0004501055**
Korea	GARCH	3.695310e-06	0.011026530	0.0004954992	0.0019223190
	IIS-GARCH	3.693729e-06	0.010984500	0.0004928380	0.0019219080
	DBSCAN-GARCH	3.693545e-06	0.010981860	0.0004926203	0.0019218600
	IIS-DBSCAN-GARCH	**1.417084e-07**	**0.004618023**	**8.661092e-05**	**0.0003764417**
Singapore	GARCH	2.200918e-07	0.006727666	0.0001773693	0.0004691395
	IIS-GARCH	2.200630e-07	0.006700246	0.0001763556	0.0004691088

Table 6 (Continued)

	DBSCAN-GARCH	2.248679e-07	0.006698601	0.0001763105	0.0004742024
	IIS-DBSCAN-GARCH	**2.196316e-07**	**0.006582136**	**0.0001700289**	**0.0004686487**
Taiwan	GARCH	2.717790e-07	0.008817934	0.0002559123	0.0005213243
	IIS-GARCH	2.715779e-07	0.008774250	0.0002545173	0.0005211314
	DBSCAN-GARCH	2.714865e-07	0.008772790	0.0002543790	0.0005210437
	IIS-DBSCAN-GARCH	**2.702549e-07**	**0.008571858**	**0.0002460195**	**0.0005198605**

Table 7: Diebold-Mariano test result

Compared methods	HK	Japan	Korea	Singapore	Taiwan
IIS-GARCH vs GARCH	-0.07271*	-0.31908*	-0.17782*	-0.056309*	-1.2301**
	(0.0942)	(0.06252)	(0.08589)	(0.09551)	(0.02187)
IIS-DB-GARCH vs GARCH	-1.9721**	-2.4741**	-0.19085 *	-1.8428*	-2.551*
	(0.04863)	(0.009933)	(0.08486)	(0.0654)	(0.01076)
IIS-DB-GARCH vs IIS-GARCH	-1.1251**	-1.2479*	-0.43674*	-1.1386**	-0.65149*
	(0.02606)	(0.08939)	(0.06623)	(0.02549)	(0.05148)

Note: The accompanying DM test p-values are provided within parentheses

** signifies significance at the 5% level, and

* signifies significance at the 10% level.

7. Conclusion

In this study, a new hybrid appraoch is proposed by incorporating DBSCAN into IIS method. The analysis provides evidence that the hybrid method achieves accurate identification of the outliers and eliminate false positives by distinguishing genuine outliers from noise. Besides, it adds value by increasing the detection accuracy, especially in financial time series with highly volatile or non-linear patterns. This addresses the weakness of conservative IIS method which is primarily linear and may struggle with datasets containing non-linear patterns or varying densities, a characteristics that is very common in financial time series.

Based on Monte Carlo simulation study result, the proposed approach performs efficiently by detecting the outliers at unknown location and magnitude with high potency and low gauge. Then, the research proceeds with testing on real data which is MSCI of five Asian Pacific countries. Based on the empirical study, the integration of DBSCAN introduces a non-parametric, density-based approach that is well-suited for identifying clusters of outliers in non-linear data structures. This makes the detection method adaptable to a broader range of financial datasets and market behaviors. The combination of statistical modeling and an unsupervised machine learning technique improves the performance of outlier detection. This also allows feature engineering to analyse the timing and patterns of exceptional returns to understand the financial and economic crises that threaten the global economy. By employing the hybrid method, the clusters of outliers detected correspond to five significant historical economic crises: the Asian financial crisis of 1997–1998, the Dot-Com Bubble of 2000–2001, the global financial crisis of 2008–2009, the European sovereign debt crisis in 2011, and the COVID-19 pandemic in 2020.

Furthermore, outlier treatment approach is applied that in accordance with the past research (Ané *et al*. 2008) which found that stock market volatility forecast can be improved with incorporating outlier correction procedure. The out-of-sample assessment results reveals that the proposed hybrid exhibits superior predictive power over the classical GARCH, IIS-GARCH

and DBSCAN-GARCH models. The predictive assessment and DM test yield consistent results, thereby validate the robustness of the findings.

In summary, the flexibility and robustness of IIS-DBSCAN-GARCH hybrid method improves reliability and generalizability in detecting irregularities in high frequency financial data which often exhibit extreme volatility, noise, and frequent structural changes that make outlier detection challenging. The findings in this study carry important implications for financial analysts, stock market participants and policymakers. Policymakers should accomodate potential market changes on the market, focusing on the vital of fostering stability and transparency in financial markets. These insights enable stock market participants to consolidate risk management strategies and optimise investment decisions making. This study specifically applies volatility models for outlier detection, as financial market shocks are often reflected in volatility instead of mean returns. Future research is recommended to explore outlier detection in mean models.

Acknowledgments

The authors would like to extend their sincere gratitude to University of Technology Sarawak for the financial support received to this work under UTS Research Grant (3/2024/08).

References

Akpan E.A., Lasisi K.E., Moffat I.U. & Abasiekwere U.A. 2023. Appraisal of excess kurtosis through outlier-modified GARCH-type models. *Communications in Statistics - Simulation and Computation* **52**(4): 1523–1537.

Ané T., Ureche-Rangau L., Gambet J.B. & Bouverot J. 2008. Robust outlier detection for Asia–Pacific stock index returns. *Journal of International Financial Markets, Institutions and Money* **18**(4): 326–343.

Baily M.N., Litan R.E. & Johnson M.S. 2008. The origins of the financial crisis: Fixing Finance Series Paper 3. Brookings Institution.

Berg M.A. 1999. The Asia crisis: Causes, policy responses, and outcomes. IMF Working Paper No. 99/138. International Monetary Fund.

Birant D. & Kut A. 2007. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering* **60**(1): 208–221.

Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**(3): 307–327.

Cai G., Wu Z. & Peng L. 2021. Forecasting volatility with outliers in Realized GARCH models. *Journal of Forecasting* **40**(4): 667–685.

Carnero M.A., Peña D. & Ruiz E. 2012. Estimating GARCH volatility in the presence of outliers. *Economics Letters* **114**(1): 86–90.

Castle J.L., Doornik J.A. & Hendry D.F. 2012. Model selection when there are multiple breaks. *Journal of Econometrics* **169**(2): 239–246.

Castle J.L., Doornik J.A. & Hendry D.F. 2021. The value of robust statistical forecasts in the COVID-19 pandemic. *National Institute Economic Review* **256**: 19–43.

Catillo M., Pecchia A. & Villano U. 2023. CPS-GUARD: Intrusion detection for cyber-physical systems and IoT devices using outlier-aware deep autoencoders. *Computers & Security* **129**: 103210.

Çelik M., Dadaşer-Çelik F. & Dokuz A.Ş. 2011. Anomaly detection in temperature data using DBSCAN algorithm. *Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 91–95.

Charles A. & Darné O. 2012. Trends and random walks in macroeconomic time series: A reappraisal. *Journal of Macroeconomics* **34**(1): 167–180.

Charles A. 2008. Forecasting volatility with outliers in GARCH models. *Journal of Forecasting* **27**(7): 551–565.

Collignon S. 2012. Europe's debt crisis, coordination failure, and international effects. *ADBI Working Paper* No. 370. Asian Development Bank Institute.

DeLong J.B. & Magin K. 2006. A short note on the size of the dot-com bubble. NBER Working Paper No. 12011. National Bureau of Economic Research.

Dokuz A.Ş., Çelik M. & Ecemiş A. 2020. Anomaly detection in Bitcoin prices using DBSCAN algorithm. *European Journal of Science and Technology* **2020**: 436–443.

Dutta A. & Bouri E. 2022. Outliers and time-varying jumps in the cryptocurrency markets. *Journal of Risk and Financial Management* **15**(3): 128.

Engle R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4): 987–1007.

Ericsson N.R. & Reisman E.L. 2012. Evaluating a global vector autoregression for forecasting. *International Advances in Economic Research* **18**(3): 247–258.

Eskin E., Arnold A., Prerau M., Portnoy L. & Stolfo S. 2002. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Barbará D. & Jajodia S. (eds.). *Applications of Data Mining in Computer Security*: 77–101. New York: Springer.

Ester M., Kriegel H.P., Sander J. & Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231.

Franses P.H. & Ghijsels H. 1999. Additive outliers, GARCH and forecasting volatility. *International Journal of Forecasting* **15**(1): 1–9.

Geweke J. & Porter-Hudak S. 1983. The estimation and application of long memory time series models. *Journal of Time Series Analysis* **4**(4): 221–238.

Grané A. & Veiga H. 2010. Wavelet-based detection of outliers in financial time series. *Computational Statistics & Data Analysis* **54**(11): 2580–2593.

Hahsler M., Piekenbrock M. & Doran D. 2019. dbscan: Fast density-based clustering with R. *Journal of Statistical Software* **91**(1): 1–30.

Hajek P., Abedin M.Z. & Sivarajah U. 2023. Fraud detection in mobile payment systems using an XGBoost-based framework. *Information Systems Frontiers* **25**(5): 1985–2003.

Hansen P.R. & Lunde A. 2005. A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? *Journal of Applied Econometrics* **20**(7): 873–889.

Hawkins D.M. 1980. A single outlier in normal samples. In: *Identification of Outliers*: 27–41. Dordrecht, NL: Springer.

Hendry D.F. & Krolzig H.M. 2005. The properties of automatic GETS modelling. *The Economic Journal* **115**(502): C32–C61.

Hendry D.F. 1999. An econometric analysis of US food expenditure, 1931–1989. In Magnus J.R. & Morgan M.S. (eds.). *Methodology and Tacit Knowledge: Two Experiments in Econometrics*: 341–361. England: John Wiley and Sons.

Islam M.A., Uddin M.A., Aryal S. & Stea G. 2023. An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. *Journal of Information Security and Applications* **78**: 103618.

Ismail M.T. & Nasir I.N.M. 2020. Outliers and structural breaks detection in volatility data: A simulation study using Step Indicator Saturation. *Menemui Matematik (Discovering Mathematics)* **42**(2): 76–85.

Jain P.K., Bajpai M.S. & Pamula R. 2022. A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality. *International Arab Journal of Information Technology* **19**(1): 23–28.

Johansen S. & Nielsen B. 2009. An analysis of the indicator saturation estimator as a robust regression estimator. In Castle J.L. & Shephard N. (eds.). *The Methodology and Practice of Econometrics*: 1–36. Oxford: Oxford University Press.

Kalair K. & Connaughton C. 2021. Anomaly detection and classification in traffic flow data from fluctuations in the flow–density relationship. *Transportation Research Part C: Emerging Technologies* **127**: 103178.

Khan F., Muhammadullah S., Sharif A. & Lee C.C. 2024. The role of green energy stock market in forecasting China's crude oil market: An application of IIS approach and sparse regression models. *Energy Economics* **130**: 107269.

Kramer T.T. 2022. How the subprime mortgage crisis sparked new legislation and changed the way millennials purchase real estate. *Journal of Business & Entrepreneurship Law* **14**(1): 1–34.

Li J., Li J., Wang C., Verbeek F.J., Schultz T. & Liu H. 2023. Outlier detection using iterative adaptive mini-minimum spanning tree generation with applications on medical data. *Frontiers in Physiology* **14**: 1233341.

Liu B.J. 2011. Why world exports are so susceptible to the economic crisis: The prevailing "export overshooting" phenomenon. NBER Working Paper No. 16837. National Bureau of Economic Research.

Liu T., Choo W., Tunde M.B. & Xinping H. 2024. Application of AO-GARCH-MIDAS model based on volatility effect in stock market volatility forecasting. *International Journal of Academic Research in Business and Social Sciences* **14**(12): 3509–3521.

Longin F.M. 1996. The asymmetric distribution of extreme stock market returns. *Journal of Business* **69**(3): 383–408.

Marczak M. & Proietti T. 2016. Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting* **32**(1): 180–202.

Massi M.C., Ieva F. & Lettieri E. 2020. Data mining application to healthcare fraud detection: A two-step unsupervised clustering method for outlier detection with administrative databases. *BMC Medical Informatics and Decision Making* **20**(1): 160.

McKenzie M. & Mitchell H. 2002. Generalized asymmetric power ARCH modelling of exchange rate volatility. *Applied Financial Economics* **12**(8): 555–564.
Mohamed S.D., Ismail M.T. & Ali M.K.B.M. 2024. Cryptocurrency returns over a decade: Breaks, trend breaks and outliers. *Scientific Annals of Economics and Business* **71**(1): 1–20.
Monko G. & Kimura M. 2023. Optimized DBSCAN parameter selection: Stratified sampling for epsilon and grid search for minimum samples. *Computer Science and Information Technology (CS & IT)* **13**(20): 43–61.
Muhammadullah S., Urooj A., Mengal M.H., Khan S.A. & Khalaj F. 2022. Cross-sectional analysis of impulse indicator saturation method for outlier detection estimated via regularization techniques with application of COVID-19 data. *Computational and Mathematical Methods in Medicine* **2022**: 2588534.
Naseer S., Khalid S., Parveen S., Abbass K., Song H. & Achim M.V. 2023. COVID-19 outbreak: Impact on global economy. *Frontiers in Public Health* **10**: 1009393.
Nasi I.N.M., Ismail M.T. & Karim S.A.A. 2022. Detecting structural breaks and outliers for volatility data via impulse indicator saturation. In Abdul Karim S.A. (ed.). *Intelligent Systems Modeling and Simulation II: Machine Learning, Neural Networks, Efficient Numerical Algorithm and Statistical Methods*: 679–687. Cham: Springer.
Ranjan K.G., Tripathy D.S., Prusty B.R. & Jena D. 2021. An improved sliding window prediction-based outlier detection and correction for volatile time-series. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields* **34**(1): e2816.
Robinson P.M. & Henry M. 1999. Long and short memory conditional heteroskedasticity in estimating the memory parameter of levels. *Econometric Theory* **15**(3): 299–336.
Rose F.Z.C., Ismail M.T. & Tumin M.H. 2021. Outliers detection in state-space model using indicator saturation approach. *Indonesian Journal of Electrical Engineering and Computer Science* **22**(3): 1688–1696.
Sander J., Ester M., Kriegel H.P. & Xu X. 1998. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery* **2**(2): 169–194.
Saeedi Emadi H.S. & Mazinani S.M. 2018. A novel anomaly detection algorithm using DBSCAN and SVM in wireless sensor networks. *Wireless Personal Communications* **98**(2): 2025–2035.
Santos C., Hendry D.F. & Johansen S. 2008. Automatic selection of indicators in a fully saturated regression. *Computational Statistics* **23**(2): 317–335.
Savić M., Atanasijević J., Jakovetić D. & Krejić N. 2022. Tax evasion risk management using a hybrid unsupervised outlier detection method. *Expert Systems with Applications* **193**: 116409.
Schubert E., Sander J., Ester M., Kriegel H.P. & Xu X. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* **42**(3): 1–21.
Shehadeh A.A., Alwadi S.M. & Almaharmeh M.I. 2022. Detecting and analysing possible outliers in global stock market returns. *Cogent Economics & Finance* **10**(1): 2066762.
Shukla R.M. & Sengupta S. 2020. Scalable and robust outlier detector using hierarchical clustering and long short-term memory (LSTM) neural network for the Internet of Things. *Internet of Things* **9**: 100167.
Venkateswarlu Y., Baskar K., Wongchai A., Gauri Shankar V., Paolo Martel Carranza C., Gonzáles J.L. & Murali Dharan A.R. 2022. An efficient outlier detection with deep learning-based financial crisis prediction model in big data environment. *Computational Intelligence and Neuroscience* **2022**: 4948947.
Wheale P.R. & Amin L.H. 2003. Bursting the dot.com "Bubble": A case study in investor behaviour. *Technology Analysis & Strategic Management* **15**(1): 117–136.
Wu R.-I. 1998. Taiwan's role in the Asian financial crisis. *Review of Pacific Basin Financial Markets and Policies* **1**(4): 529–544.
Yaqoob T. & Maqsood A. 2024. The potency of time series outliers in volatile models: An empirical analysis of fintech, and mineral resources. *Resources Policy* **89**: 104666.

Institute for Mathematical Research (INSPEM)
Universiti Putra Malaysia (UPM)
43400 UPM Serdang
Selangor, MALAYSIA
*E-mail: wong.hui.shein@uts.edu.my**

Department of Mathematics and Statistics
Faculty of Science
Universiti Putra Malaysia (UPM)
43400 UPM Serdang
Selangor, MALAYSIA
E-mail: faridzamani@upm.edu.my, jayanthi@upm.edu.my

Centre of Technological Readiness and Innovation in Business and Technopreneurship
School of Business and Management
University of Technology Sarawak (UTS)
96000 Sibu
Sarawak, MALAYSIA
E-mail: wong.hui.shein@uts.edu.my[*]*, simcy@uts.edu.my*

Received: 9 June 2025
Accepted: 22 August 2025

[*]Corresponding author