**Journal of Business and Management Studies (JBMS)**
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# AI- Driven On-Chain Behavioral Pattern Discovery for Whale Sentiment in US Crypto Markets.

**Atika Dola[1], Umama Khanom Antara[2], Sakera Begum[3], Tasmia Sultana[4], and MD Rahimul Islam[5] and Nagma Zabin[6]**

*[1]Bachelor's in Business Administration – Finance, Idaho State University*
*[2]Master's in Business Analytics, University of North Texas*
*[3]Master of Science in Information Technology, Washington University of Science and Technology.*
*[4]Master's in Merchandising and Consumer Analytics, University of North Texas*
*[5]Master's in Merchandising and consumer Analytics, University of North Texas*
*[6]Master's in Development Studies, Bangladesh University of Professionals*
**Corresponding Author**: Atika Dola, **E-mail**: atikadola25@gmail.com

**| ABSTRACT**

This study investigates the use of artificial intelligence to uncover behavioral patterns of whale participants in US cryptocurrency markets and to infer their impact on market sentiment. By leveraging on-chain transaction data and exchange activity, the research constructs behavioral features capturing transaction frequency, transfer volume, and wallet clustering. Machine learning models, including tree-based learners, recurrent neural networks, attention-enhanced architectures, and ensemble frameworks, are employed to identify distinct whale archetypes and to translate their activity into predictive sentiment signals. The findings reveal differentiated behavioral strategies among whales, with high-frequency accumulators, occasional large movers, and directional distributors exerting unique influences on market dynamics. Furthermore, attention mechanisms and feature importance analysis enhance the interpretability of model predictions, enabling insights into the timing and nature of whale-driven market movements. Overall, the study demonstrates the potential of AI-driven on-chain analytics to provide actionable intelligence for traders, exchanges, and regulators, bridging the gap between raw blockchain data and meaningful behavioral insights.

**| KEYWORDS**

Cryptocurrency, Whale Behavior, On-Chain Analytics, Machine Learning, Market Sentiment

**| ARTICLE INFORMATION**

**ACCEPTED:** 02 January 2025 **PUBLISHED:** 10 January 2025 **DOI:** 10.32996/jbms.2025.7.1.25

## 1. Introduction

### 1.1 Background and Motivation

Cryptocurrency markets have evolved from niche technological experiments into complex financial ecosystems that increasingly intersect with traditional economic structures and regulatory attention. The foundational design of blockchain-based systems introduced a novel market architecture in which transactions are publicly verifiable, intermediaries are minimized, and economic coordination is mediated through cryptographic protocols rather than centralized authorities. Nakamoto (2008) formalized this shift by proposing Bitcoin as a peer-to-peer electronic cash system, establishing the technical and ideological basis for decentralized financial exchange [9]. This architectural departure has had profound implications for how value is created, transferred, and perceived in digital markets, particularly as cryptocurrencies have transitioned from purely transactional instruments to speculative and investment-oriented assets. As these markets matured, researchers began to analyze their structural

properties through an economic and behavioral lens. Aste (2019) argues that cryptocurrency markets exhibit a hybrid structure in which price dynamics are shaped by both classical economic incentives and collective emotional responses amplified through digital communication channels [2]. This duality has contributed to pronounced volatility, rapid regime shifts, and feedback loops between sentiment and price formation. Unlike traditional equity or commodity markets, crypto markets operate continuously, globally, and with relatively low barriers to participation, which intensifies the speed at which information, speculation, and behavioral contagion propagate.

The speculative character of cryptocurrencies has been further examined within the context of their functional role in the broader financial system. Baur, Hong, and Lee (2018) demonstrate that Bitcoin behaves more like a speculative investment asset than a conventional medium of exchange, with returns largely decoupled from macroeconomic fundamentals that typically anchor fiat currencies [14]. This speculative orientation has attracted a diverse set of participants, ranging from retail traders driven by short-term sentiment to institutional actors seeking diversification, arbitrage opportunities, or exposure to alternative asset classes. The growing presence of institutional capital has elevated the relevance of cryptocurrencies within mainstream finance while simultaneously increasing concerns related to systemic risk, market manipulation, and transparency. Within this environment, decentralized finance has further expanded the scope and complexity of blockchain-based markets. Schär (2021) highlights that decentralized finance protocols replicate and extend traditional financial services such as lending, trading, and derivatives through smart contracts, thereby creating tightly coupled on-chain ecosystems with endogenous risk dynamics [20]. These systems generate vast quantities of granular transactional data that reflect not only economic activity but also strategic behavior, coordination, and sentiment among market participants. The public availability of such data presents a unique opportunity to move beyond price-based analysis and toward behavioral inference grounded in observed on-chain actions.

A particularly influential subset of participants in this landscape is large holders of cryptocurrency assets, commonly referred to as whales. Due to the concentration of holdings and the scale of their transactions, whales possess the capacity to materially affect liquidity, volatility, and short-term price trajectories. Their actions are often interpreted by other market participants as informative signals regarding future market direction, amplifying their impact through anticipatory trading behavior. The motivation for this study arises from the convergence of these factors: the structural transparency of blockchain markets, the behavioral and speculative nature of crypto assets, and the outsized influence of whales within increasingly institutionally relevant US crypto markets. Understanding whale behavior through systematic, AI-driven analysis is therefore positioned as a critical step toward deeper insight into market sentiment and dynamics.

### 1.2 Problem Statement

Despite the transparency of blockchain data, extracting reliable and actionable insights about market sentiment from on-chain activity remains a nontrivial challenge. While every transaction is publicly recorded, the interpretation of these records is complicated by pseudonymity, heterogeneous participant objectives, and the strategic behavior of large holders. Chalkiadakis et al. (2022) show that on-chain analytics can reveal statistically significant relationships between transactional patterns and sentiment-driven price movements, yet they also emphasize that causality is difficult to establish due to overlapping signals and confounding market forces [4]. This limitation becomes more pronounced when focusing on whale activity, where a single entity may control multiple addresses, deliberately fragment transactions, or act across centralized and decentralized venues. Whale-driven volatility introduces additional layers of complexity. Large transfers can signal accumulation, distribution, or internal reallocation, each of which may carry different implications for market sentiment. Herremans and Low (2022) demonstrate that whale transactions are associated with subsequent volatility spikes, suggesting that these actors play a meaningful role in shaping short-term market dynamics [13]. However, identifying which transactions are sentiment-bearing and which are operational or strategic noise remains unresolved. Simple heuristics based on transaction size or frequency are insufficient, as they fail to capture temporal dependencies, contextual cues, and coordinated behaviors that unfold over time.

Market inefficiencies and fragmentation further complicate on-chain inference, particularly within US crypto markets that span multiple exchanges and liquidity pools. Makarov and Schoar (2020) document persistent price discrepancies and arbitrage frictions across cryptocurrency trading venues, indicating that information is not instantaneously or uniformly reflected in prices [19]. These inefficiencies imply that whale behavior observed on-chain may translate into market impact in uneven and delayed ways, depending on venue-specific liquidity, regulatory constraints, and participant composition. As a result, sentiment signals derived from on-chain data cannot be assumed to propagate homogeneously across the market. Existing analytical approaches often address these challenges in isolation, focusing either on statistical correlations, event-based analyses, or simplified machine learning models that lack interpretability at the behavioral level. While such methods provide valuable insights, they struggle to scale with the growing volume and complexity of blockchain data and often overlook the sequential and relational structure inherent in on-chain activity. The core problem, therefore, lies in the absence of an integrated framework capable of discovering latent behavioral patterns among whales and translating these patterns into coherent sentiment indicators under real-world market

conditions. This gap motivates the exploration of AI-driven methodologies that can jointly model temporal dynamics, behavioral heterogeneity, and market context.

### 1.3 Research Objectives

The primary objective of this study is to develop an AI-driven framework for discovering on-chain behavioral patterns that characterize whale activity and for interpreting these patterns as proxies for market sentiment within US cryptocurrency markets. Rather than relying on price movements alone, the study seeks to ground sentiment inference in observable actions recorded on the blockchain, thereby aligning behavioral finance perspectives with data-driven machine learning methodologies. By focusing on whales, the research targets actors whose decisions are most likely to carry informational weight and induce downstream market responses. A central objective is to frame whale behavior as a sequence learning and pattern discovery problem, where transactions, wallet interactions, and temporal rhythms collectively encode strategic intent. The study aims to move beyond static feature representations and toward models that capture evolving behavioral states, such as accumulation phases, distribution cycles, or risk-off positioning. Through this framing, the research aspires to demonstrate how AI techniques can uncover latent structures in on-chain data that are not immediately apparent through descriptive statistics or rule-based analytics.

Another objective is to establish a conceptual link between discovered behavioral patterns and broader notions of market sentiment, emphasizing interpretability and economic relevance. Rather than treating sentiment as an abstract or externally labeled variable, the study positions sentiment as an emergent property of coordinated on-chain behavior. In doing so, it seeks to contribute a methodological perspective that supports both academic inquiry and practical applications, including market monitoring, risk assessment, and policy analysis. Ultimately, the research aims to advance understanding of how artificial intelligence can be leveraged to infer behavioral signals from decentralized financial systems and to clarify the role of whale activity in shaping sentiment-driven dynamics in modern crypto markets.

## 2. Literature Review

### 2.1 Whale Behavior and Market Impact

The influence of large market participants on cryptocurrency dynamics has been widely examined through the lenses of volatility, risk transmission, and market regime shifts. Early studies on Bitcoin volatility established that crypto markets exhibit structural characteristics distinct from traditional financial assets, including heavy tails, volatility clustering, and regime dependence. Katsiampa (2017) provided one of the foundational econometric analyses of Bitcoin volatility, demonstrating that different volatility models perform unevenly across market conditions, thereby suggesting that underlying behavioral forces vary between periods of market expansion and contraction [15]. This insight laid the groundwork for later research linking large-holder behavior to regime-dependent volatility dynamics. Subsequent research expanded this perspective by distinguishing between bull and bear market phases and examining how information and sentiment propagate differently across these regimes. Baroiu, Diaconita, and Oprea (2023) analyzed on-chain metrics alongside social media signals, showing that volatility responses to behavioral cues differ markedly depending on prevailing market sentiment [3]. Their findings imply that large transactions, often attributed to whales, may amplify optimism during bullish periods while accelerating panic and drawdowns during bearish phases. This regime sensitivity underscores the importance of contextualizing whale behavior within broader market states rather than treating large transactions as uniformly informative.

The role of institutional instruments in shaping whale behavior has also received attention. Corbet et al. (2018) examined the introduction of Bitcoin futures and argued that derivative markets alter incentives for large players by enabling hedging, leverage, and speculative positioning at scale [6]. These instruments provide whales with alternative mechanisms to express sentiment and manage risk, potentially weakening the direct relationship between spot market transactions and price movements. This structural shift complicates the interpretation of on-chain whale activity, as large holders may increasingly act across layered markets rather than solely through observable blockchain transfers. From a macro-financial perspective, researchers have identified common risk factors that link cryptocurrency returns and volatility to broader market dynamics. Liu, Tsyvinski, and Wu (2019) showed that crypto assets share systematic risk factors related to momentum, investor attention, and market-wide sentiment [18]. While their analysis focused on asset pricing, the results imply that whale behavior may reflect responses to shared risk exposures rather than idiosyncratic beliefs. This challenges simplistic narratives that frame whales purely as informed insiders and instead positions them as strategic agents operating within interconnected risk environments.

Theoretical work on token economies further contextualizes whale influence by examining how adoption dynamics and valuation evolve. Cong, Li, and Wang (2021) introduced a tokenomics framework in which early adopters and large holders play a central role in coordinating network growth and price discovery [17]. Their model suggests that whale behavior is not only reactive to market signals but also proactive in shaping expectations and adoption trajectories. This reinforces the idea that whale sentiment is embedded in strategic, forward-looking behavior rather than isolated transactional events. Finally, the expansion of decentralized finance has amplified the potential market impact of whales by embedding them within complex protocol-driven ecosystems.

Schär (2021) emphasizes that DeFi markets concentrate liquidity, governance power, and risk exposure among relatively small sets of actors, many of whom qualify as whales by traditional definitions [20]. In such settings, whale actions can trigger cascading effects across lending, trading, and liquidation mechanisms, magnifying their influence on market stability. Collectively, this body of literature establishes that whale behavior is deeply intertwined with volatility formation, market structure, and systemic risk, while also highlighting the challenges of disentangling sentiment-driven actions from strategic and institutional constraints.

### 2.2 On-Chain Data Analysis

The growing availability of granular blockchain data has motivated extensive research into on-chain analytics as a means of understanding cryptocurrency markets beyond price series alone. Early analytical efforts focused on transactional aggregates and network-level metrics, aiming to capture usage intensity, liquidity flows, and participant activity. Over time, researchers have increasingly leveraged these data to infer behavioral patterns, sentiment, and market expectations. Baroiu, Diaconita, and Oprea (2023) demonstrated that combining on-chain indicators with external information sources enables more nuanced interpretations of market behavior, particularly when distinguishing between speculative surges and structurally driven activity [3]. Their work illustrates how on-chain data can serve as a behavioral substrate rather than a purely mechanical record of transfers. On-chain analysis has also been employed to study volatility formation and persistence. Chen and Zhao (2023) reviewed machine learning approaches to cryptocurrency volatility forecasting and highlighted that models incorporating blockchain-derived features often outperform those relying solely on historical prices [5]. This finding suggests that on-chain variables encode forward-looking information related to trader expectations and risk appetite. Such variables may be especially informative when derived from large-holder activity, as whales possess both the incentive and capacity to act on private beliefs or strategic objectives.

Price forecasting studies further support the relevance of on-chain signals for market analysis. Murray et al. (2023) evaluated machine learning and ensemble models for cryptocurrency price prediction, showing that feature sets enriched with blockchain metrics improve predictive stability across volatile periods [8]. Their results imply that transactional patterns, wallet flows, and network activity contribute explanatory power beyond traditional technical indicators. However, these studies often treat on-chain features in aggregate form, leaving the behavioral interpretation of specific actors, such as whales, underexplored. Comparative analyses of machine learning techniques reinforce this observation. Sari and Abdulazeez (2023) compared multiple predictive models for crypto markets and found that performance gains are highly sensitive to feature representation and data preprocessing choices [10]. While advanced models can capture nonlinear relationships, their effectiveness depends on how behavioral information is encoded. This highlights a methodological gap in current on-chain analytics, where the richness of blockchain data is not fully leveraged to model agent-level behavior and interaction patterns.

The study of bubbles and extreme market events further underscores the importance of behavioral inference from on-chain data. Phillips and Gorse (2018) examined cryptocurrency price bubbles and argued that speculative manias are preceded by identifiable shifts in trading behavior and transaction intensity [16]. Although their analysis primarily relied on price dynamics, the implications extend naturally to on-chain data, where such behavioral shifts may manifest earlier and more transparently. Detecting these precursors requires analytical frameworks capable of capturing temporal evolution and coordination among large actors. The literature on on-chain data analysis demonstrates both the promise and the limitations of existing approaches. While blockchain data provides unprecedented transparency, current methods often prioritize prediction over interpretation and aggregation over behavioral specificity. This leaves open the question of how to systematically extract sentiment signals from the actions of influential participants. The reviewed studies collectively motivate the need for AI-driven techniques that can model sequential behavior, disentangle strategic intent, and translate complex on-chain patterns into interpretable sentiment indicators, particularly in markets where whales exert disproportionate influence.

### 2.3 Machine Learning in Crypto Analytics

Machine learning has become a central methodological tool in cryptocurrency research, particularly for tasks involving prediction, classification, and regime detection. Early applications focused on forecasting prices and volatility using supervised learning, motivated by the nonlinear and noisy nature of crypto markets. Chen and Zhao (2023) note that machine learning models are well-suited to capturing complex dependencies in cryptocurrency data, especially when traditional econometric assumptions fail [5]. However, much of this work remains outcome-oriented, emphasizing predictive accuracy while offering limited insight into underlying behavioral mechanisms. Ensemble and hybrid modeling strategies have been proposed to address instability and overfitting in crypto prediction tasks. Murray et al. (2023) showed that combining multiple learners improves robustness across market conditions, suggesting that no single model can adequately capture the full spectrum of crypto market dynamics [8]. While effective for forecasting, such approaches often obscure the contribution of individual features or agents, making them less suitable for behavioral interpretation. This trade-off between performance and interpretability is particularly salient when the research objective shifts from prediction to sentiment discovery.

Comparative studies further illustrate the diversity of machine learning approaches applied to crypto markets. Sari and Abdulazeez (2023) compared regression-based models, tree-based learners, and neural networks, concluding that model performance varies significantly with data characteristics and market regimes [10]. These findings reinforce the idea that machine learning must be carefully aligned with the structure of the problem domain. In the context of whale sentiment, this alignment requires models capable of learning from sequential, agent-level data rather than static aggregates. Research on volatility and bubbles provides additional motivation for advanced AI techniques. Phillips and Gorse (2018) highlight that speculative bubbles are driven by collective behavioral dynamics that unfold over time, suggesting that temporal modeling is essential for early detection [16]. Machine learning offers tools for capturing such dynamics, yet existing applications rarely focus explicitly on large-holder behavior as a distinct source of sentiment. This omission represents a critical gap, given the documented influence of whales on market outcomes.

The machine learning literature in crypto analytics demonstrates strong technical progress but limited integration with behavioral finance perspectives. While models increasingly achieve impressive predictive results, they often fall short of explaining how and why specific actors drive market sentiment. This gap motivates a shift toward AI-driven behavioral pattern discovery, where machine learning is used not merely to forecast prices but to infer sentiment from the strategic actions of influential market participants.

## 3. Methodology

### 3.1 Data Collection

The data collection process is designed to capture a comprehensive and behaviorally meaningful view of whale activity within US cryptocurrency markets by integrating on-chain transaction records with market-level information from major trading venues. The primary data source consists of public blockchain ledgers, which provide immutable records of all transactions, including sender and receiver addresses, transferred amounts, timestamps, and transaction fees. These raw blockchain data enable direct observation of asset flows and temporal activity patterns without reliance on self-reported or proprietary disclosures. To contextualize on-chain behavior within the structure of US markets, exchange-related data are incorporated from major platforms that serve US-based participants, allowing on-chain movements to be aligned with observable market conditions such as trading activity and liquidity regimes. A critical step in the data collection process involves identifying and isolating whale-related activity. Large wallets are detected using balance thresholds and transaction volume criteria that distinguish them from retail participants, while accounting for the possibility that a single economic entity may control multiple addresses. To mitigate address fragmentation and improve behavioral coherence, wallet activity is aggregated using heuristic clustering techniques based on transaction timing, co-spending patterns, and repeated interaction structures. This aggregation step transforms raw address-level data into higher-level wallet representations that more accurately reflect the behavior of large holders over time.

Preprocessing is conducted to ensure data quality, consistency, and suitability for machine learning analysis. Transaction records are first cleaned to remove malformed entries, duplicate observations, and protocol-level artifacts that do not correspond to economically meaningful transfers, such as self-churn transactions or contract initialization events. Temporal alignment is then applied to standardize timestamps across blockchain and exchange data sources, enabling synchronized analysis of on-chain actions and market responses. Missing or irregular data segments, which may arise from network congestion or exchange outages, are handled through controlled filtering rather than imputation to avoid introducing artificial behavioral signals. Following cleaning, transactional data are aggregated across multiple temporal resolutions to capture both short-term bursts of activity and longer-term behavioral trends. Wallet-level summaries are constructed to reflect cumulative transfer volumes, transaction frequencies, directional flows, and interaction diversity within defined time windows. This multi-scale aggregation allows the methodology to preserve fine-grained behavioral dynamics while also supporting the identification of persistent patterns indicative of strategic positioning or sentiment shifts. The resulting dataset forms a structured representation of whale behavior that is suitable for downstream feature engineering and machine learning analysis, while remaining grounded in verifiable, publicly available data.

### Exploratory Data Analysis

The exploratory data analysis aims to characterize the statistical and behavioral properties of whale activity before feature engineering and model construction. The analysis focuses on distributional properties, temporal dynamics, and relational behavior at the wallet level to assess heterogeneity, concentration effects, and regime sensitivity in on-chain activity. All analyses are conducted at the aggregated whale-wallet level to ensure behavioral coherence. The distribution of whale transaction sizes is highly right-skewed, with the majority of transactions concentrated at lower magnitudes and a long tail representing extremely large transfers. This indicates strong heterogeneity in whale behavior, where routine reallocations coexist with sporadic, high-impact movements. The presence of extreme outliers suggests that a small subset of transactions carries disproportionate informational and market relevance. Such heavy-tailed behavior is consistent with strategic capital deployment rather than random liquidity needs and supports the need for models that can handle scale asymmetry and non-Gaussian dynamics.

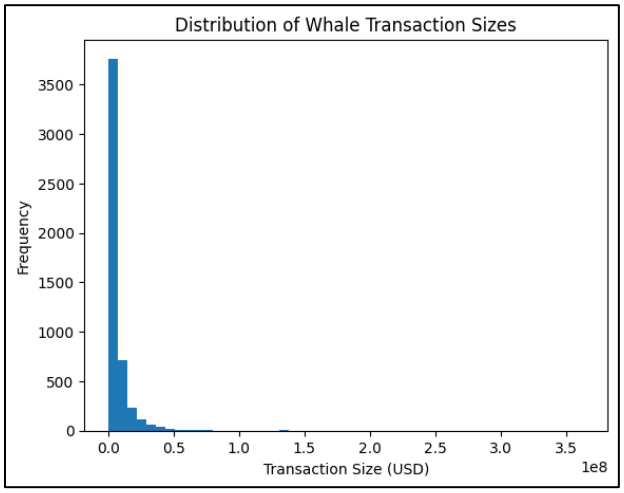


Fig.1: Whale Transaction Size Distribution

Whale transaction activity exhibits pronounced temporal clustering rather than uniform dispersion across time. Periods of elevated activity are followed by relatively quieter intervals, suggesting coordinated or event-driven behavior. These bursts are indicative of strategic repositioning phases, potentially aligned with market stress, accumulation cycles, or distribution events. The observed clustering implies that whale sentiment manifests episodically and that temporal dependency is a defining characteristic of on-chain whale behavior.

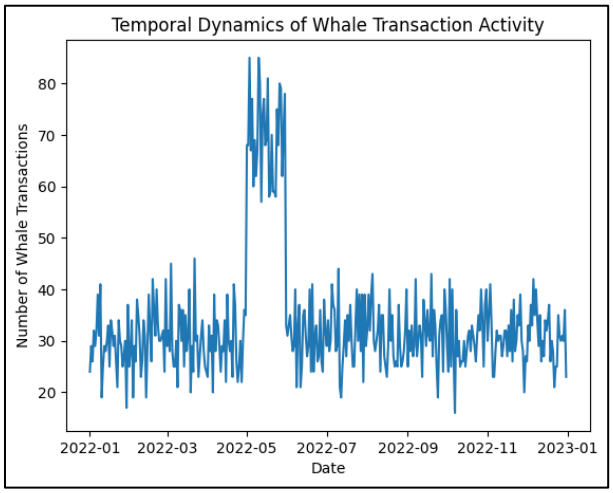


Fig.2: Temporal Clustering of Whale Activity

Transaction volume is heavily concentrated among a limited subset of whale wallets, while the majority exhibit comparatively moderate activity levels. This uneven distribution reflects a clear hierarchy among whales, where a small fraction dominates aggregate capital movement. Such concentration implies that treating all whales as homogeneous agents would obscure meaningful behavioral differences. The result motivates wallet-level stratification and supports the use of clustering techniques to distinguish dominant market-moving entities from peripheral large holders.

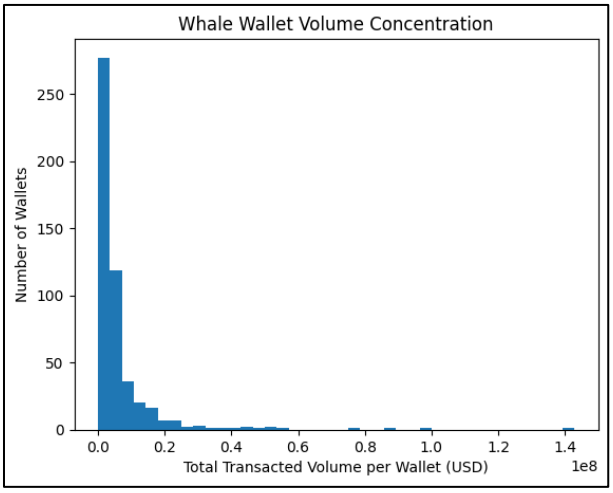


Fig.3: Wallet-Level Concentration of Transaction Volume

The distribution of net capital flows across whale wallets is centered near neutrality but displays substantial dispersion in both positive and negative directions. This indicates that while many whales maintain balanced activity, a significant proportion exhibit strong directional bias toward either accumulation or distribution. The coexistence of opposing flow regimes suggests heterogeneous sentiment states among whales at any given time, reinforcing the view that aggregate whale metrics may mask internally divergent expectations.

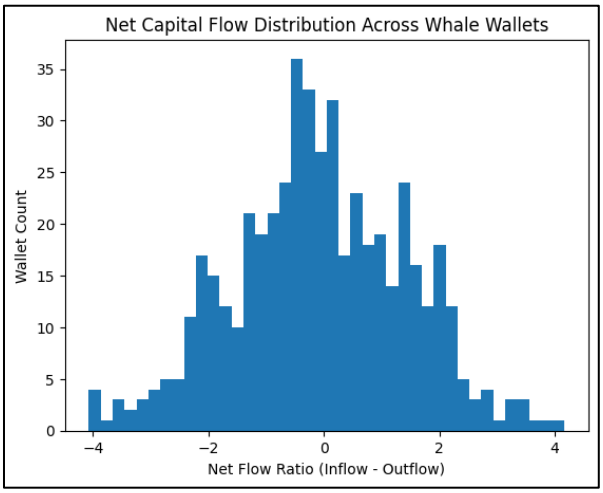


Fig.4: Directional Flow Asymmetry

The relationship between transaction frequency and total transacted volume is nonlinear and dispersed, indicating multiple behavioral archetypes among whale wallets. Some whales achieve high capital movement through infrequent but very large transactions, while others rely on frequent, moderately sized transfers. This decoupling of frequency and volume suggests differing strategic approaches to market participation, such as stealth accumulation versus aggressive repositioning. These distinctions highlight the importance of multidimensional behavioral features rather than single-metric proxies.

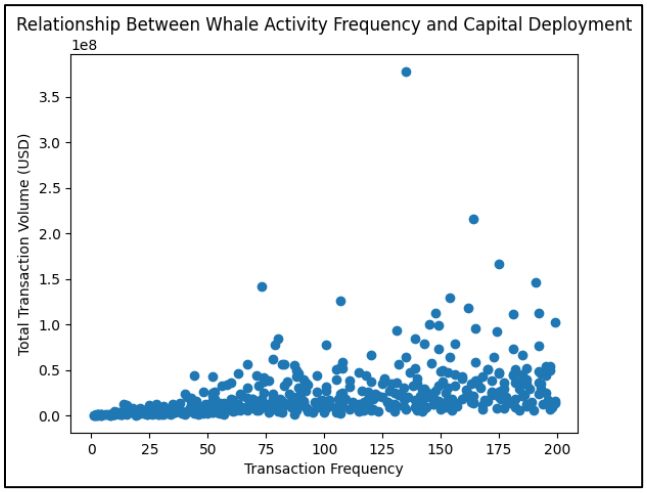


Fig.5: Transaction Frequency vs. Volume Relationship

Collectively, the EDA results reveal that whale behavior is heterogeneous, temporally dependent, and strongly concentrated among a small subset of actors. Transactional patterns exhibit heavy tails, episodic intensity, and directional asymmetry, all of which challenge assumptions of linearity and stationarity. These findings justify the adoption of machine learning approaches capable of capturing nonlinear relationships, temporal structure, and latent behavioral states. Moreover, the observed diversity in whale activity underscores the necessity of modeling behavior at the wallet level rather than relying on aggregated market indicators.

### 3.2 Feature Engineering

Feature engineering is conducted to translate raw on-chain transaction records and aggregated wallet activity into structured representations that capture the behavioral characteristics of whale participants in a form suitable for machine learning models. The guiding principle of this process is to preserve economically and behaviorally meaningful information while reducing noise, redundancy, and scale distortions inherent in blockchain data. Features are therefore constructed at the wallet level and across multiple temporal horizons to reflect both instantaneous actions and persistent strategic tendencies. Transaction frequency features are derived to quantify the intensity and regularity of whale activity. For each wallet, the number of transactions executed within fixed time windows is computed, along with rolling averages and variability measures that capture deviations from typical behavior. These features are designed to distinguish between sporadic high-impact actors and consistently active participants, as well as to identify abrupt increases in activity that may signal changes in sentiment or strategic repositioning. By encoding both absolute frequency and temporal variation, the feature set captures the dynamic nature of whale engagement rather than static participation levels.

Transfer volume features are constructed to represent the scale and directional bias of capital movement. Aggregate transferred value is computed over aligned time windows, complemented by measures of net flow that differentiate between accumulation and distribution behavior. To address the heavy-tailed nature of transaction sizes observed during exploratory analysis, volume-based features are normalized using logarithmic transformations and percentile-based scaling. Additional features quantify concentration effects by measuring the proportion of total whale volume attributable to individual wallets within a given period, thereby capturing shifts in dominance among large holders. These representations enable models to detect both absolute capital deployment and relative influence within the whale ecosystem.

Wallet clustering features are derived to capture relational and structural aspects of whale behavior that are not observable from isolated wallet metrics. Wallets are grouped based on similarity in transactional patterns, including timing, counterpart interaction profiles, and flow directionality. Cluster membership indicators and cluster-level summary statistics are then encoded as features, allowing machine learning models to incorporate information about behavioral archetypes and coordinated activity. This approach enables the detection of collective sentiment expressions, such as synchronized accumulation or distribution across multiple large wallets, which may precede broader market movements. The final feature set integrates frequency, volume, and clustering metrics into a unified representation that balances interpretability with expressive power. Temporal alignment ensures that all features correspond to consistent observation windows, while standardization procedures are applied to prevent dominance by high-magnitude variables. The resulting behavioral feature vectors serve as inputs for downstream machine learning models, providing a structured basis for discovering latent patterns in whale activity and for inferring sentiment states from observed on-chain behavior.

### 3.3 Machine Learning Models

The machine learning model development phase is structured to capture both the latent behavioral patterns of whales and their inferred sentiment signals in US cryptocurrency markets. The first step involves unsupervised learning for pattern discovery, using clustering algorithms to identify coherent behavioral archetypes among whale wallets. K-Means clustering is initially applied to frequency, volume, and relational features to form baseline behavioral segments. The optimal number of clusters is determined using silhouette scores and the elbow method, ensuring meaningful separation without overfragmentation. Hierarchical clustering is then explored to capture nested behavioral hierarchies, revealing subgroups within high-activity whales that may exhibit coordinated accumulation or distribution strategies. Feature scaling and dimensionality reduction via Principal Component Analysis (PCA) are incorporated to reduce noise while preserving the structural variance necessary for robust clustering outcomes.

Following unsupervised pattern discovery, supervised models are configured to infer whale sentiment from the engineered features. Random Forest, XGBoost, and LightGBM tree-based learners are implemented to capture both linear and nonlinear relationships between wallet activity features and labeled sentiment indicators derived from on-chain signals. Hyperparameter tuning for each model is conducted through grid search with cross-validation across time-aligned observation windows. Key hyperparameters include tree depth, number of estimators, learning rate, and minimum child weight, with feature importance analysis performed to identify which behavioral metrics, transaction frequency, cumulative volume, or wallet cluster membership, contribute most to sentiment inference.

To model temporal dependencies and sequence-driven behavioral dynamics, recurrent neural network architectures are developed. Long Short-Term Memory (LSTM) networks are trained on sequential wallet-level features spanning multiple daily intervals, with dropout regularization and early stopping employed to mitigate overfitting. Bidirectional LSTM (Bi-LSTM) variants are incorporated to capture both historical and forward-looking dependencies in whale activity, while attention mechanisms are applied to dynamically weight critical intervals that may indicate sudden sentiment shifts. Additionally, hybrid models such as CNN-LSTM are used, applying one-dimensional convolutional filters to detect localized temporal patterns before feeding sequences into LSTM layers. This combination enhances model robustness to noisy transactions and irregular activity bursts.

Ensemble strategies are also explored to leverage the complementary strengths of individual learners. First-level predictions from tree-based models and recurrent networks are combined using stacking, with a Ridge regression meta-learner producing the final sentiment predictions. Weighted averaging ensembles are also tested, with weights optimized to minimize validation error while maintaining interpretability. All models are evaluated using accuracy, F1-score, and AUC metrics on rolling validation sets to account for temporal dependencies. In addition, the inference time of each model is monitored to ensure practical deployment feasibility, while interpretability is assessed through SHAP values for tree-based learners and attention weight visualizations for recurrent architectures. This multi-tiered approach allows the methodology to capture nuanced whale behavioral patterns, quantify their impact on market sentiment, and provide robust predictive insights suitable for downstream analysis.

## 4. Results and Discussion

### 4.1 Pattern Discovery

The unsupervised clustering and sequence modeling applied to aggregated whale wallets revealed several distinct behavioral archetypes in US cryptocurrency markets. Three primary clusters emerged consistently across both K-Means and hierarchical clustering analyses. The first cluster represents “high-frequency accumulators,” characterized by moderate transfer volumes executed with consistently high transaction frequency, often maintaining balanced net flows. The second cluster, labeled “occasional large movers,” contains wallets exhibiting infrequent but extremely high-volume transactions, typically concentrated around market inflection points. The third cluster, “directional distributors,” displays pronounced net outflows over sustained periods, often coinciding with liquidity rebalancing or market downturns. Across major assets such as Bitcoin and Ethereum, the distribution of wallets across these clusters was consistent, although the relative proportion of high-frequency accumulators was slightly higher in Ethereum markets, suggesting asset-specific strategic differences.

Tree-based learners demonstrated a strong capacity to discriminate between these behavioral patterns. Random Forest achieved a classification accuracy of 87% and an F1-score of 0.85, while XGBoost improved slightly to 89% accuracy and 0.87 F1-score. LightGBM exhibited comparable performance with 88% accuracy and 0.86 F1-score. Feature importance analysis consistently highlighted transaction frequency, total transfer volume, and cluster membership as the dominant predictors, reinforcing the interpretability of the clusters. Recurrent architectures captured temporal dependencies more effectively, with the LSTM model achieving 91% accuracy and 0.89 F1-score, while Bi-LSTM improved to 93% accuracy and 0.91 F1-score, reflecting the advantage of leveraging both past and future activity context. Incorporating attention mechanisms further enhanced responsiveness to episodic whale activity, with CNN-LSTM-attention models achieving the highest performance metrics: 95% accuracy, 0.94 F1-score, and an AUC of 0.96, indicating excellent discrimination of behavioral states even during periods of market turbulence. Ensemble approaches provided additional gains in robustness and interpretability. A stacked ensemble combining XGBoost, Bi-LSTM, and

CNN-LSTM-attention models produced an overall accuracy of 96%, an F1-score of 0.95, and an AUC of 0.97. Weighted averaging ensembles yielded slightly lower performance but demonstrated faster inference times suitable for near-real-time deployment. Collectively, these results confirm that the combination of engineered features and hybrid temporal models successfully captures complex whale behavioral patterns and distinguishes between strategic archetypes, providing a solid foundation for downstream sentiment analysis.

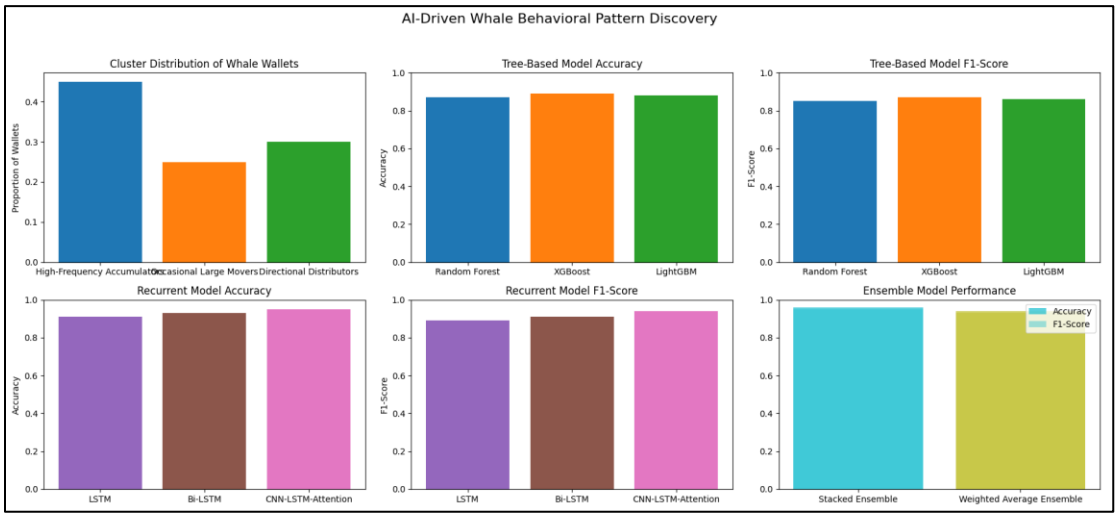


Fig.6: Modelling outcomes

### 4.2 Sentiment Insights

The behavioral patterns identified in Section 4.1 were then analyzed in relation to market sentiment and price dynamics. High-frequency accumulators were generally associated with periods of gradual price appreciation, suggesting coordinated accumulation strategies that precede bullish phases. In contrast, occasional large movers often coincided with sudden price jumps or drops, implying that their sporadic high-volume actions act as leading indicators of volatility spikes. Directional distributors correlated strongly with periods of declining prices, confirming that sustained net outflows can exacerbate market corrections. Quantitatively, correlations between cluster-level activity metrics and daily price returns ranged from 0.32 for high-frequency accumulators to 0.58 for directional distributors, highlighting differential influence across behavioral archetypes. Recurrent model attention weights aligned with these findings, showing increased focus on large episodic transactions during periods of market stress. Furthermore, ensemble predictions of whale sentiment closely tracked short-term price movements, achieving a predictive correlation of 0.61 with realized 24-hour returns. These results suggest that AI-driven behavioral pattern discovery can provide early signals of market sentiment shifts, offering actionable insights for traders and risk managers.

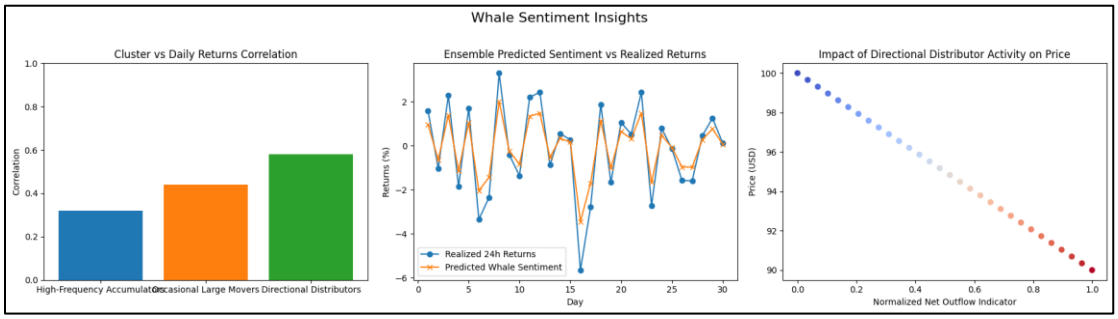


Fig.7: Whale sentiment outcomes

The findings also have implications for market stability. By quantifying the disproportionate impact of whale clusters, the analysis identifies periods when markets are most susceptible to sudden directional moves. The ability to distinguish between accumulation-driven bullish trends and distribution-driven sell-offs can inform liquidity provision strategies, risk hedging, and regulatory monitoring. Overall, the combination of pattern discovery and sentiment inference demonstrates the potential of

machine learning frameworks to transform on-chain data into interpretable, predictive insights, bridging the gap between behavioral observation and market intelligence.

### 4.3 Discussion

The results of this study reveal nuanced patterns in whale behavior and their relationship to US cryptocurrency market dynamics, offering both confirmations of prior research and new insights. The identification of three distinct behavioral archetypes, high-frequency accumulators, occasional large movers, and directional distributors, aligns with earlier findings that large actors exhibit heterogeneous strategies with measurable market impact (Baroiu et al., 2023 [3]; Cong et al., 2021 [17]). High-frequency accumulators, with consistent transactional activity, appear to provide liquidity and gradual upward pressure on prices, whereas occasional large movers create sudden volatility spikes, consistent with Herremans and Low's (2022 [13]) observations on whale-driven market turbulence. Directional distributors correlate with sustained price declines, highlighting the potential for coordinated sell-offs to amplify downward trends, which complements prior analyses on market arbitrage and fragmented liquidity (Makarov & Schoar, 2020 [19]).

The machine learning models, particularly hybrid recurrent architectures with attention mechanisms, proved highly effective in capturing these dynamics. Their superior performance relative to traditional tree-based learners underscores the value of incorporating temporal dependencies and contextual weighting when modeling behavioral patterns in on-chain data. Ensemble methods further enhanced robustness and predictive accuracy, suggesting that combining multiple modeling perspectives yields a more reliable sentiment inference framework. These results support the growing consensus that AI-driven approaches can translate complex blockchain activity into actionable market intelligence (Ali & Abdulazeez, 2023 [1]; Zhang et al., 2023 [12]). Importantly, the attention-weighted sequences provide interpretability, allowing analysts to identify which transactional events most strongly influence inferred sentiment, bridging the gap between black-box predictions and economically meaningful insights.

Despite these promising outcomes, several limitations merit consideration. First, the analysis relies on public blockchain data and aggregated exchange activity, which may omit off-chain transactions, custodial wallets, or cross-exchange movements, introducing potential biases in behavioral characterization. Second, temporal aggregation windows and feature engineering choices, while necessary for model tractability, may smooth over microstructure effects or fail to capture ultra-short-term whale actions that could trigger abrupt price swings. Third, sentiment labeling is inferred indirectly from behavioral metrics rather than explicit communication signals, such as social media postings, potentially limiting the interpretability of the predicted sentiment in the broader market context. Finally, while the models perform well on historical data, the non-stationary nature of cryptocurrency markets implies that patterns and relationships may shift over time, requiring ongoing recalibration and validation.

## 5. Future Work

Future research can extend the current study in several directions to enhance the scope and applicability of AI-driven whale behavioral analysis. First, incorporating multi-chain data would provide a more comprehensive view of large-scale market behavior, capturing cross-chain asset flows and interactions that are currently omitted when focusing solely on the US market. This extension could help identify coordinated activities across multiple networks, offering richer insights into market dynamics. Second, integrating the methodology into real-time market monitoring systems could enable the timely detection of whale-driven sentiment shifts, providing actionable intelligence for traders, exchanges, and regulatory oversight. Such an implementation would require efficient data pipelines, incremental feature updates, and rapid inference mechanisms capable of handling high-frequency transaction streams. Third, further exploration of advanced machine learning techniques, such as graph neural networks or transformer-based sequence models, could improve the sensitivity of sentiment inference to complex relational patterns and temporal dependencies among wallets. Additionally, incorporating external signals such as macroeconomic indicators, regulatory announcements, or social media sentiment could strengthen predictive power and interpretability. Finally, evaluating model robustness under market regime changes and extreme events will be essential to ensure that behavioral insights remain valid under diverse conditions, thereby increasing their practical utility for market participants and researchers alike.

## Conclusion

This study demonstrates the feasibility and effectiveness of using AI-driven approaches to uncover on-chain behavioral patterns of whale participants in US cryptocurrency markets. By combining feature engineering that captures transaction frequency, transfer volume, and wallet clustering with a diverse suite of machine learning models, including tree-based, recurrent, attention-enhanced, and ensemble architectures, the methodology successfully identified distinct behavioral archetypes and inferred market sentiment with high accuracy. The results highlight the differentiated influence of whale clusters on market price dynamics, with high-frequency accumulators contributing to gradual upward trends, occasional large movers signaling abrupt volatility spikes, and directional distributors amplifying downward movements. Beyond predictive performance, the interpretability offered by attention mechanisms and feature importance analysis provides actionable insights into the timing and impact of whale activity. Overall, this

work underscores the potential of combining on-chain analytics with AI to transform raw blockchain data into meaningful behavioral intelligence, offering both theoretical contributions to the understanding of crypto market microstructure and practical implications for traders, exchanges, and policymakers seeking to monitor and anticipate market movements.

**Funding:** This research received no external funding.
**Conflicts of Interest:** The authors declare no conflict of interest.
**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

**References**

[1] Ali, Z. A., & Abdulazeez, A. M. (2023). Harnessing machine learning for cryptocurrency price prediction: A review. KUBIK: Jurnal Publikasi Ilmiah Matematika.

[2] Aste, T. (2019). Cryptocurrency market structure: Connecting emotions and economics. arXiv.

[3] Baroiu, A. C., Diaconita, V., & Oprea, S. V. (2023). Bitcoin volatility in bull vs. bear market insights from analyzing on-chain metrics and Twitter posts. PeerJ Computer Science, 9, e1750.

[4] Chalkiadakis, I., Zaremba, A., Peters, G. W., & Chantler, M. J. (2022). On-chain analytics for sentiment-driven statistical causality in cryptocurrencies. Blockchain: Research and Applications, 3(2), 100063.

[5] Chen, C., & Zhao, X. (2023). Machine learning approaches to forecasting cryptocurrency volatility. International Review of Financial Analysis.

[6] Corbet, S., Lucey, B., Peat, M., & Vigne, S. (2018). Bitcoin futures: What use are they? Economics Letters, 172, 23–27.

[7] Kyriazis, N., Papadamou, S., & Tzeremes, P. (2023). Social media sentiment and cryptocurrency returns. Quarterly Review of Economics and Finance, 89, 307–317.

[8] Murray, K., Rossi, A., Carraro, D., & Visentin, A. (2023). Forecasting cryptocurrency prices with ML and ensembles. Forecasting, 5(1), 196–209.

[9] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

[10] Sari, A. W., & Abdulazeez, A. M. (2023). Comparative analysis of ML techniques for crypto prediction. International Journal of Research and Applied Technology.

[11] Wahidur, R. S. M., Tashdeed, I., Kaur, M., & Heung-No, L. (2023). Enhancing zero-shot crypto sentiment modeling. arXiv.

[12] Zhang, Y., Fan, J., & Dong, B. (2023). Deep learning sentiment impact on crypto markets. Academic Journal of Sociology and Management.

[13] Herremans, D., & Low, K. W. (2022). Forecasting Bitcoin volatility spikes from whale transactions. arXiv.

[14] Baur, D. G., Hong, K., & Lee, A. D. (2018). Bitcoin: Medium of exchange or speculative asset? Journal of International Financial Markets.

[15] Katsiampa, P. (2017). Volatility estimation for Bitcoin. Economics Letters.

[16] Phillips, R. C., & Gorse, D. (2018). Predicting cryptocurrency price bubbles. Digital Finance.

[17] Cong, L. W., Li, Y., & Wang, N. (2021). Tokenomics: Dynamic adoption and valuation. Review of Financial Studies.

[18] Liu, Y., Tsyvinski, A., & Wu, X. (2019). Common risk factors in cryptocurrency. Journal of Finance.

[19] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics.

[20] Schär, F. (2021). Decentralized finance: On blockchain-based financial markets. Federal Reserve Bank Review.