Yichi Zhang(A22245302). I have discussed with my teammate:
Helene Sy Ivanne(A22339734),
Hwa Ryun Kim(A22230638).

Data

1.One particular exchange or multiple exchanges:

One exchange (eg:Binance)

why:

1.1. Consistency and Control

Using a single exchange controls for heterogeneity in trading rules, tick sizes, and timestamp conventions. This makes the volatility labels and features more consistent.

1.2. Engineering and Reproducibility

A unified data schema simplifies pipeline design and reduces synchronization errors. It also improves reproducibility, since other researchers can easily access the same data source.

1.3. Future Extension

While Phase I focuses on one exchange for internal validity, later stages can test cross-exchange generalization to assess external validity.

2.Specific instruments or multiple instruments

Multiple instruments (top 10/5 coins)

why:

2.1. Market Coverage

Focusing on the top 5–10 coins ensures that the model is trained on the most actively traded instruments. These markets dominate volume and price discovery, making the results more representative.

2.2. Data Quality

Highly liquid instruments reduce noise such as stale quotes or extreme bid–ask spreads. This creates cleaner microstructure features and more reliable short-term volatility labels.

2.3. Generalization

Using multiple assets avoids overfitting to idiosyncratic behavior of a single coin. It allows the model to capture common volatility dynamics across liquid instruments.

2.4. Computational Balance

Restricting the universe to a limited set (top 5/10) keeps the dataset rich but still computationally manageable, ensuring training efficiency.

3.Exact datafields you will be using

level 1 data, best bid/ask and last-trade information capture immediate market movements. These fields reflect spread dynamics, mid-price changes, and trade

aggressiveness, which are direct drivers of realized volatility at the 1-minute horizon.

market sentiment data, sentiment (from news, social media, or sentiment indices) captures exogenous shocks not visible in the order book. Sudden sentiment changes can explain abnormal volatility patterns beyond pure microstructure features.

level 2 data, deeper order book layers provide insight into hidden liquidity, imbalance, and order flow pressure. These microstructure signals are leading indicators of short-term volatility spikes.

funding rate, funding payments link spot and derivative markets. Abrupt shifts in funding often signal changes in trader positioning and leverage buildup, which precede bursts of volatility.

4.The frequency of the data sampling your require
Frequency: 1 minute frequency

5.The length of the data history you will need to backtest it
Data History: 7 years (from 2018-2025) Binance is founded in 2017.

Prediction

1.Target variable or specific conditions
Target variable: volatility, our project goal is to build a high-performance machine learning model that predicts short-term volatility using 1-minute market microstructure data.

2.Prediction time horizon
Prediction time horizon: short-term 15-60 minute or medium-term (1 day). We target 15–60 minutes and 1 day to capture both intraday volatility bursts from order flow and broader daily risk dynamics, ensuring relevance for trading and risk management.

3.Which false positives you are trying to avoid
False Positives: predicted volatility < actual volatility. We aim to avoid cases where predicted volatility is lower than actual volatility, since underestimation leads to insufficient risk buffers, unexpected losses, and poor trading or hedging decisions.

4. Any complicating factors
Complicating Factors: market sentiment, this one acts as a complicating factor because it can trigger sudden volatility spikes not reflected in order book or trade data. News, social media, or unexpected announcements may shift trader behavior rapidly, making volatility harder to predict using microstructure signals alone.

Feature Construction

1. List the features you will be considering for your signal

2. **Write the spec for each feature, be as detailed as possible**
3. **Explain what is the value range you expect from each feature**

the above 3 questions can be answered by the tables below:

Level	Feature Name	Formula / Specification	Theoretical Rationale	Expected Value Range
1: Basic	OFI_raw	buy_volume - sell_volume	(Order Flow Imbalance) Measures the net pressure from aggressive market orders. A high absolute value signifies strong directional pressure, which often precedes increased volatility.	$[-10^2, 10^3]$
1: Basic	L1_Imbalance	(close_bid_size - close_ask_size) / (close_bid_size + close_ask_size)	(Limit Order Book Imbalance) Captures the imbalance of passive liquidity at the best bid and ask. A severe imbalance may indicate a weakening of support/resistance, leading to higher price volatility.	[–1,1]
1: Basic	VWAP_deviation	(close_price - vwap) / close_price	(VWAP Deviation) Shows how far the current price has strayed from the volume-weighted average price. Large deviations may lead to mean-reversion movements, a form of volatility.	[–0.05,0.05]
1: Basic	Mean_Spread	mean_spread	(Average Bid-Ask Spread) Represents market liquidity and transaction costs. A widening spread signals that liquidity providers perceive higher risk, which is strongly correlated with rising volatility.	0.01%–0.1%
1: Basic	VPIN_proxy	50-minute rolling sum of `	buy_volume - sell_volume	[0,1]
2: Derived	OFI_std_30m	30-min rolling standard deviation of	(OFI Stability) Measures the stability of order flow. A high value indicates that order	typically an order of

		OFI_raw	flow is erratic and unpredictable, suggesting impending market instability and higher volatility.	magnitude smaller than the underlying feature
2: Derived	Spread_ma_15m	15-min moving average of Mean_Spread	(Spread Trend) Captures the recent trend in transaction costs. A rising trend suggests a gradual withdrawal of liquidity and a potential shift into a higher volatility regime.	values fluctuate around the mean of the base feature
2: Derived	OFI_roc_10m	(OFI_raw_t - OFI_raw_{t-10}) / OFI_raw_{t-10}	(Rate of Change of OFI) Measures the 'acceleration' of order flow pressure. A rapid increase in OFI can be a more potent predictor of imminent volatility than a high but stable OFI level.	typically an order of magnitude smaller than the underlying feature
3: Interaction	Liquidity_Weighted_OFI	OFI_raw / Mean_Spread	(Liquidity-Weighted OFI) Tests the hypothesis that order flow has a greater price impact in illiquid markets. An imbalance during periods of wide spreads is expected to be amplified, causing more volatility.	10^2–10^4
3. Interaction	Trade_Intensity	(buy_volume + sell_volume) / (buy_trade_count + sell_trade_count)	(Average Trade Size) Differentiates between markets dominated by large institutional trades versus smaller retail trades. High intensity may signal informed trading, which can lead to volatility.	[1, 1000]
3. Interaction	Price_Impact_per_Volume	15-min realized volatility / 15-min sum of `	OFI_raw	10^{-6}–10^{-3}

4: Cross-Asset	BTC_vol_lag_15m	Realized volatility of Bitcoin, lagged by 15 minutes	(Volatility Spillover) Tests the hypothesis that volatility in the market-leading asset (BTC) spills over to other assets (ETH) with a time lag.	0.1%–2%
4: Cross-Asset	BTC_OFI_lag_15m	OFI_raw of Bitcoin, lagged by 15 minutes	(Sentiment Spillover) Strong order flow pressure in BTC can influence broad market sentiment and liquidity, thus acting as a leading indicator for volatility in ETH.	$[-10^2, 10^3]$
4: Cross-Asset	ETH_BTC_corr_60m	60-min rolling correlation between ETH and BTC returns	(Correlation Regime) Changes in correlation signal shifts in the market regime. A breakdown of correlation (decoupling) or a sudden tightening can precede periods of high volatility.	[–1,1]

4. How will you check for feature quality and collinearity

feature quality and collinearity: GARCH → XGBoost or Transformer. We will first benchmark all features against a simple GARCH model to test whether they add incremental predictive power. Then, to detect and control collinearity, we use correlation matrices and VIF checks before modeling. Finally, model-specific diagnostics from XGBoost (feature importance, SHAP) and Transformers (attention weights) ensure that redundant or unstable features are pruned, leaving only those that consistently improve out-of-sample volatility forecasts.