Trading by Charts:
A Multivariate CNN System to Predict Retail Investor Trading

Chaojie Liu
Jay.liu@bristol.ac.uk
University of Bristol

Abstract

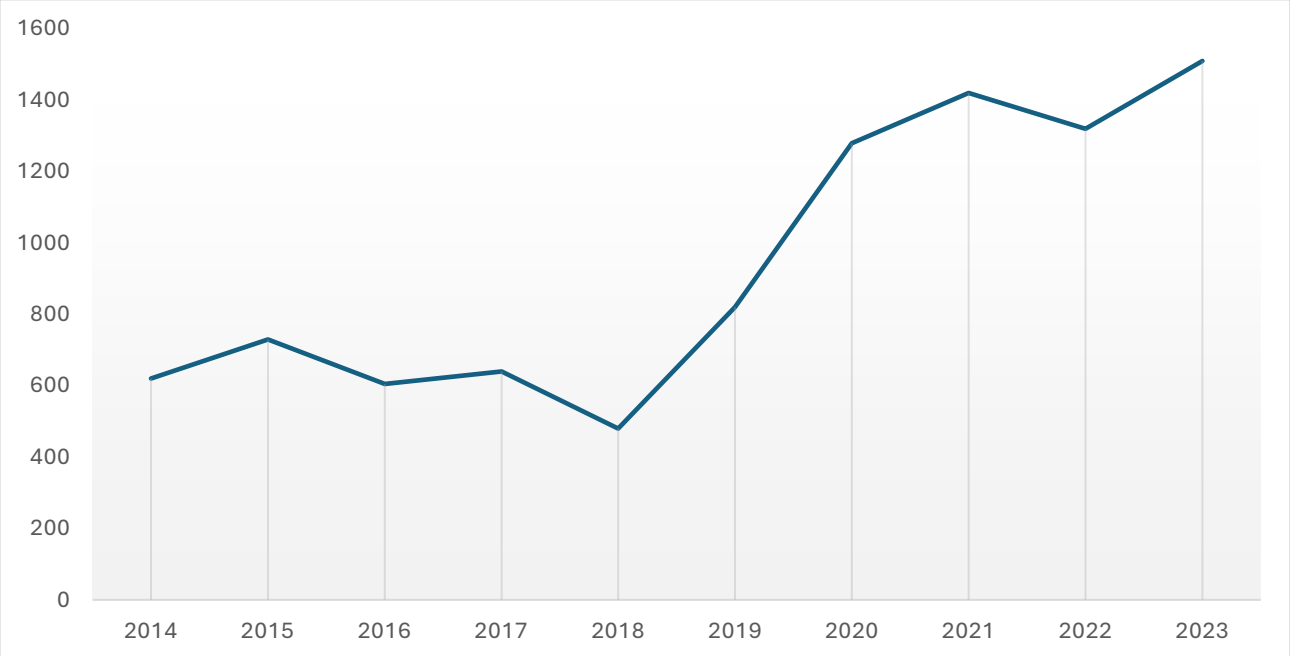
Understanding retail investor behavior is interesting and important but difficult, the prediction of retail investor trading is even much harder. This paper sheds some light on predicting retail investors' holding on stocks by building an accessible time-series prediction system using the CNN technique for pre-COVID and COVID-19 periods. For the first time, a multivariate convolutional neural network (CNN) is built with numerical and graphical data. Focusing on the U.S. stock markets between 2018 and 2020, this study utilizes component stocks of the S&P 500 index as the sample with relative data on stock characteristics, retail investor holding, and retail investor ownership. The pioneering multivariate CNN system performs great in predicting retail investor trading and outperforms the random forests models built which only apply numerical data. The results support previous studies on the performances of deep learning techniques like CNN and investor trading behavior and sentiment. Besides, retail investor holding contains little predictive information for stock price movement. This paper contributes to the economic and financial literature by filling the gap in the predictions of retail investor behavior using cutting-edge machine learning techniques based on novel applications of data. In addition, this prediction system can improve social welfare by helping retail investors make less biased decisions, informing financial institutions to better engage with retail investors, and assisting financial authorities to better monitor and manage risks caused by retail investors in the market.

Keywords: Retail Investment, Investor Behaviour, Convolutional Neural Network, Random Forest, Machine Learning, Deep Learning, Time-Series Prediction

I. Introduction

Individual decision-making has always been an interesting, though sometimes ill-defined, topic in psychology, economics, and other subjects (e.g., Papoport and Wallsten, 1972). To simplify the complexity, the efficient market hypothesis (EMH) is developed, stating that asset prices should reflect all information in the markets (Fama, 1970). EMH implies that investors are rational, and they cannot make abnormal returns as stock prices adjust immediately to new information in the market. If investors cannot make any money, why do they invest? Market anomaly happens as a singular incident or as a recurrence, including calendar anomaly, technical anomaly, and fundamental anomaly (e.g., Latif et al., 2011), triggered by investors over time. Financial and economic scholars, working together with other (social) scientists, are trying to explain the relative decision theories. For example, Camerer (1995) concludes that economists made two major contributions to investor decision-making, including destructive (testing if anomalies can be explained by individual behavioral choice models) and constructive (exploring new theories to explain anomalies or building puzzles of models developed). With the development of computer science, scholars are applying machine learning (ML) techniques to understand and predict individual decision-making behaviors (e.g., Balasubramanian, Ye and Xu, 2022).

Figure 1. Peak Daily Retail Flows ($ Million)



Note: The data is from Vanda Research (2023), showing the daily maximum retail flows in the U.S. stock markets.

Acting as non-professional individuals trading financial assets, retail investors are one of the most important types of investors who can directly affect the financial market, but they do not always

understand the nature of investments. Vanda Research (2023) reports that the daily peak of retail flows into the U.S. stock market is increasingly higher from 2014 to 2023 (Figure 1). In 2023, 61% of American adults are investing in the stock market (Gallup, 2023). However, even though the U.S. has one of the most sophisticated stock markets, retail investors do not always understand the nature of stock investment. For example, the BNY Mellon Retail Investor Survey (2022) reports that retail investors think stocks can provide the largest rates of return in both the short and long run, which is not the case in the market. Though they are not the majority of the financial market, retail investors are found to have significant effects (e.g., Han and Kumar, 2013). Cheng et al. (2021) show that retail investors' attention can affect the market liquidity in the short term. Robinhood traders may affect stock prices significantly and account for a large share of variation in stock returns even though their size is relatively small (Beck and Jaunin, 2023).

Therefore, as a part of individual decision-making in finance, understanding and predicting retail investors' trading behavior is difficult but crucial. Besides understanding retail investment, financial and economic scholars try to find the predictive relationship between stock prices and retail investors' trading behavior. For example, Kelley and Tetlock (2013) apply a unique dataset containing all retail orders in US markets and argue that net buying of both market and limit orders can be used to predict stocks and market orders can be used to predict firm news. Some literature tries to forecast retail order balances. Boehmer et al. (2021) predict retail investor balances based on TAQ data that stocks with more net buying outperform those with negative retail order imbalances (i.e., more sell orders than buy orders) over the next week.

However, very few scholars tried to predict retail investors' trading movements (aggregate behavior) based on stock characteristics (e.g., price, trading volume, price to equity ratio), leaving this specific relationship unclear. Thus, there is a necessity and literature foundation for the potential predictive relationship. Nevertheless, the complexity and variety of human behavior make understanding and predicting retail investor behavior difficult. To handle this problem, I developed a multivariate CNN (Convolutional Neural Network) prediction system to understand, discover, and predict retail investor's trading behavior. To provide a comprehensive and justifiable analysis regarding the predictive relationship, the paper is going to answer the following three main research questions: 1) Can CNNs recognize patterns from price charts and other numerical stock information and predict retail investors' behaviors? 2) How is the result compared with other ML techniques (random forest)? 3) Is it possible to predict stock price by retail investor behavior?

As many retail investors nowadays are trading online through digital financial service providers (e.g., Robinhood, Trading 212, and Etoro), their investment decisions can be made rationally or irrationally for various reasons. Retail investors may trigger "Buy" or "Sell" based on what they see in stock price charts and other information on their investment platforms. The main assumption is that

the stock characteristics will contain some predictive information for the movement of retail investors' holdings, no matter whether they are fundamental traders who trade by analyzing the fundamental values of stocks, technical traders who trade based on their technical analysis via quantitative models, or irrational traders who react by their predictions without faith basis.

This is a pioneering work to understand retail investor trading based on a combination of graphical data (price charts) and numerical stock characteristics, instead of surveys, questionnaires, or other pure numeric data. The sample consists of daily data of all component stocks of the S&P 500 index, representing the most valued 500 stocks in the US markets, and is widely applied in financial studies (e.g., Denis et al., 2003). Other variables include retail investor holding (RIH), stock characteristics, and retail investor ownership of stocks. The data on retail investor decisions are based on the stock popularity of retail investors using the Robinhood platform, which has the largest share of retail investor users in the U.S. market since 2020 (Vanguard, 2021). The stock characteristics data are from Compustat-Capital IQ while the institutional ownership data are from Thomson-Reuters S34 based on 13F forms, both in WRDS databases. After merging, the stock closing price column is converted into price charts of different window lengths with other numerical data, acting as input, for predictions and the RIH acts as the output variable (label). The institutional ownership data is used as a classifier to split the stocks into high, medium, and low retail ownerships. As there may be regime changes before and during the COVID-19 pandemic, which is covered in the data sample, I split the sample into two sets: 2 May 2018 to 12 August 2019 and 13 August 2019 to 13 August 2020.

Moreover, the paper is the first to build a novel time-series prediction system for retail investment using multivariate CNN and other ML techniques. With recent developments in computer science, ML techniques enable this time-series prediction system and may contribute to the understanding of retail investors' trading behavior. Therefore, a special kind of deep learning method and artificial neural network (ANN), convolutional neural network (CNN), is used as the main technique to recognize patterns in stock price charts and other numerical data to link the relationship between these patterns and retail investors' decisions when they invest in stocks on digital finance platforms. Besides, I also applied more interpretable ML techniques (Random Forest-RF, as an example) with numeric data to act as the benchmarks, to compare prediction performances. Lastly, univariate CNN models are built to predict stock prices based on RIH.

Results suggest that the prediction system built based on a multivariate CNN framework is efficient in predicting time-series RIH over the next 5, 10, 20, 30, and 45 days with different window-length data applied. On one hand, the characteristics of the stocks may contain factors that are considered by retail investors, leading to rational or irrational trading behaviors at an aggregate level. On the other hand, different retail investors (fundamental traders, technical traders, and irrational traders) may react to financial information differently, and irrational traders on Robinhood may

dominate the market and their internal effects are not trading off, so that the aggregated outcome is predictable. In addition, my evidence suggests that market conditions can significantly affect model performances and optimal architectures of ML models, therefore, future applications of ML in predictions should consider regime changes in the environment. The applications of the multivariate CNN framework with graphical and numerical data outperform the RF framework which only has numerical data in this specific task. Moreover, I find that the univariate CNN models are unable to predict stock prices, due to the lack of information or the complex factors of price changes.

By understanding and predicting the retail investor movement on their holdings, this paper not only contributes to the existing economic and financial literature regarding retail investment and the application of new data and methodologies, but also sheds some light on social welfare by informing retail investors to make less biased decisions, helping financial institutions to better engage with retail investors, and enabling financial authorities to better monitor the risks caused by retail investors in the market. More specifically, the prediction system helps retail investors to understand themselves (self-financial education, which is less mentioned in financial education but should be important), make more informed decisions, and reduce bias risks. This is important as retail investors often lack information, compared with institutional investors who have more resources available. With more information to assist decision-making, retail investors may trade more rationally, which brings less uncertainty to the financial markets and therefore benefits all participants in the markets. Additionally, the system may be helpful for financial institutions to make better decisions when they engage with retail investors as there is a lack of systematic understanding of retail investors' reactions to information changes. Nevertheless, the superior resources of institutional investors may draw more benefits from the prediction system as they have more data and computing power, which are extremely important for prediction. This may bring concerns because institutions may exploit profits by trading with or against retail investors. In addition, this prediction system enables financial authorities to predict and forecast potential problems in the markets caused by retail investors to control systematic risks and prevent potential significant negative outcomes (e.g., financial crises).

This paper unfolds as follows. Part II presents literature reviews of two topics: retail investment and applications of ML. Research questions are clearly stated in Part III. Parts IV and V show details about the data and methodologies, respectively. Results and discussions are in Part VI. Part VII provides the robustness check. Ideas for future studies and the limitations in this paper are discussed in Part VIII. The conclusion is provided in Part IX while the reference list is presented at the end (Parts X).

II. Literature Review

As the focus is on building an ML-based prediction system for retail investment trading, this paper should be fitted with two main parts of the literature. One is traditional financial or economic studies about retail investors, including their trading, behavior, and effects in the financial markets. Previous studies in retail investment are comprehensive and precise while this paper is trying to shed light on a new perspective on predicting retail investor trading. Besides, this paper is a work in the literature of applied ML (focusing on CNN) in economics and finance. ML techniques have been widely used in causal inference or prediction tasks in economics and finance (e.g., Chen et al., 2015), but very few scholars tried to use ML techniques to predict retail trading behavior based on graphical data.

a. Retail Investment

The paper contributes to the current financial and economic literature about retail investments and trading decision-making. After considering the costs of investment, retail investors generally do not make net abnormal profits (James, 2000). However, the total trading volume was about 350 billion dollars on the NYSE in 2003 (the number is even higher today), and a large part of it is contributed by retail investors (Subrahmanyam, 2008). If they cannot make money, why do they do so? Scholars argue that some investors may make profits by exploring mispricing opportunities (e.g., Shleifer and Vishny, 1997; Daniel, Hirshleifer and Subrahmanyam, 2001). Some investors made money based on technical analysis. Fritz and Weinardt (2015) argue that technical traders are less likely to exhibit the disposition effect and have the potential to make some profits. Nonetheless, Subrahmanyam (2008) argues that the relationship between stock returns and retail investors' net abnormal returns is weak, but for financial institutions, the story is different as they can make use of pricing problems. The result of this study may illustrate the analysis ability and trading behavior of retail investors in the Robinhood platform, which is representative of the general retail investors in the U.S. and the rest of the world.

This paper can also provide further empirical evidence for previous findings regarding investors' irrational behavior. Overconfidence, a typical irrational behavior, is argued as the main reason for the bad performance of retail investors, which drives investors to overtrade or other irrational trading (e.g., Barber and Odean, 2000; Hvide, 2000). Nevertheless, counterarguments discuss the effects of individual irrational characteristics (e.g., overconfidence) on investment performances. Irrational and overconfident retail investors can bear more risks and end up with greater expected returns in the long run (DeLong et al., 1991; Kyle and Wang, 1997; Hirshleifer, Subrahmanyam and Titman, 2006). The existence of herding and other irrational group behavior contributes to the predictive relationship between stock characteristics and retail investor movement.

Besides, there is a lack of systematic understanding of retail investors' reactions under different market situations with more digital financial services available and more uncertainty in the world since

COVID-19, which is an important part of understanding retail investor behaviors, and this paper aims to provide more evidence about this. The developments of digital devices and financial services have contributed to society and the financial world, but they may affect retail investor behavior as well (Odean, 1999; DeHann and Glover, 2024). With the developments of online trading, investors investing online perform better than those who do not go online before the shift but are worse performers after the shift (Barber and Odean, 2002). Hillenbrand et al. (2020) find that the product information of assets affects retail investment decisions. After the experiences of COVID-19 and the great uncertainties it brings to the world, there may be a regime change in the market and retail investors may behave differently in their investments (Wullweber, 2020) which may bring greater challenges in the predictions. The results of multivariate CNN models in RQ 1 support this argument and provide more evidence from the perspectives of the applications of ML techniques.

Even though there is some behavioral finance literature studying aggregate-level retail investor effects, previous studies mostly focus on the individual level. Retail investor trading is affected by a range of factors (e.g., personal demographic factors by Kannadhasan, 2015; gender by Barber and Odean, 1998; social factors by Baker and Nogsigner, 2002; market situations by Goodfellow, Bohl and Gebka, 2009; others' behavior by Benerjee, 1992; mood by Nofsinger, 2005 and by Goetzmann and Zhu, 2005; weather by Saunders, 1993 and by Hirshleifer and Shumway, 2003) and suffers from great uncertainty (e.g., Mak and Ip, 2017). By questionnaires based on five main constructs of investor behavior (investing horizon, confidence, control, risk preference, and perception of loss), Wood and Zaichkowsky (2004) find that retail investors can be generally classified as risk-intolerant, confident, loss-averse young, and conservative long-term investors. While individual characteristics vary from person to person, this paper sheds light on the prediction of aggregated results of retail investor behavior at a macro (market) level, so the conclusions drawn from this prediction system may be more versatile to understand the aggregated behaviors and outcomes of retail investors' trading. Thereby the public and market participants can better understand retail investor trading, behavior, and potential effects in the financial markets.

Scholars have proved that even though retail investors do not hold the largest proportion of the markets, they may affect the market significantly. Daniel, Hirshleifer and Subrahmanyam (1998) argue that overconfidence and biased self-attribution lead to market under-and-overreactions. Kumar and Lee (2006) find that highly concentrated retail investments lead to stock return movements, especially for those stocks with high costs of arbitrage. Abudy (2020) finds that more retail investors trading in the market are positively related to aggregate stock market liquidity while they do not receive compensation for providing liquidity to the stock markets by trading (Barrot, Kaniel and Sraer, 2016). Beck and Jaunin (2023) argue that the inelastic demand for institutional investors allows Robinhood retail traders to have significant effects on stock prices. Friedman and Zeng (2024) apply trade-level

data to investigate the effects of retail investors on the pricing of public earning news and argue that retail trading facilitates liquidity provision to other traders during the earning announcement period.

Furthermore, there are some works about prediction on retail investments. Kaniel, Saar and Titman (2008) provide evidence regarding the relationship between net retail investments and stock returns. Intense aggregate imbalances can predict stock abnormal returns, respectively, around earnings announcement periods. For less sophisticated investors, trade imbalances are positively correlated with future stock losses (Hvidkjaer, 2008). Kelley and Tetlock (2013) argue that net buying of both market and limit retail orders can be used to predict stocks and market orders can predict firm news, which contributes to market efficiency. Boehmer et al. (2021) propose a novel way to identify retail investment activities based on TAQ trade and other financial data and then make relative predictions on retail investments and their effects.

b. Applications of ML (especially CNN) in Economics and Finance

After its innovation in computer science, ML has been broadly adopted in economics and finance (e.g., Gu, Kelly and Xiu, 2020). Deep learning, a special kind of ML method, is known for its capability to learn features from data with different levels of abstraction or filtration (LeCun, Bengio and Hinton, 2015). There are many financial studies comparing prediction performances of different methods, similar to the RQ 2. In this paper, CNN (one of the artificial neural networks) and other less flexible ML methods are mainly implemented in pattern recognition, feature extraction, and forecasting tasks.

Feature extraction or pattern recognition from data is important in the prediction of finance (e.g., Hoseinzade and Haratizadeh, 2019). ML methods, for example, Artificial Neural Networks (ANNs), are extensively used in pattern recognition tasks in business, biometrics, and many other areas (Abiodun et al., 2019), which are often regarded as efficient and robust (e.g., Bai et al., 2021). Chen et al. (2015) propose a CNN-based method to recognize handwritten character recognition, which is a hard task for humans due to the various appearances of handwritten characteristics. Yang, Jiang and Guo (2019) developed a model for time-series data classification using CNN and RNN (recurrent neural network), both of which are ANNs.

ML predictions for classification problems and regression problems are widely used in finance. Chen et al. (2017) argue that prediction is one of the most significant and challenging tasks involving financial time series data, due to its non-stationary (e.g., Zhang, Lin and Shang, 2017), non-linear, dynamic, noisy, and chaotic natures (e.g., Bezerra and Albuquerque, 2017; Kumar and Thenmozhi, 2014). Gu, Kelly and Xiu (2020) review the canonical studies in empirical asset pricing by applying ML techniques and find that trees and neural networks perform the best as there are gains from allowing non-linear relationships among predictors. Murray, Xia and Xiao (2024) investigate the best

combination of ML model architecture, loss function, weight variable, and input data that output the best predictions for stock prices and returns.

There are many applications of multivariate CNN models in time-series classification and regression tasks (e.g., Liu, Hsaio and Tu, 2019; Cavalli and Amoretti, 2021). However, very few of them tried to apply both graphical and numerical data in a multivariate CNN model at the same time. Agrawal et al. (2019) apply ANNs to predict financial market performances. Besides, Hoseinzade and Haratizadeh (2019) develop a CNN-based model to apply data in different markets (S&P 500, NASDAQ, etc) and make relevant predictions, whose performances are significantly better than the baseline algorithms. Lee et al. (2019) propose a stock price prediction framework using Deep Q-Network and CNN.

Some prediction studies are using ML approaches for retail investors. For example, Yang et al. (2019) suggest deep learning methods to forecast the operational risks of retail investors. Chen, Ho and Liu (2019) apply statistical and ML models (Naïve Bayes, Support Vector Machines, and Random Forest, etc) to predict investment performance based on investors' personalities. Few scholars tried to apply ML techniques to predict retail investor trading behavior.

III. Research Questions

The re-emergence of market participation by retail investors brings the market attention to regulatory reforms, which protect not only retail investors from risks of stock investments but also the financial market from retail investors. Fisch (2022) challenges these arguments and overstates the harms and benefits of retail trading. But the concern of understanding retail investors, from the perspective of either risk management or social welfare, is unwavering. To develop a reliable predicting system for retail investor trading, it is important to have a comprehensive and justifiable understanding of the system itself and how it can be applied to the broader stock market beyond its sample data, which can contribute to market participants' welfare. Hence, this paper aims to answer the following three main research questions (RQs):

a. RQ 1: Can CNNs recognize patterns from price charts and other numerical stock information and predict retail investors' behaviors?

If so, how reliable is the forecasting system based on CNNs? The core aim of this study is to build the prediction system, therefore, I set the first part of the research questions focused on system building. In this part, I am going to test if the system built is reliable and applicable to predict the retail investor movements in the stock market with out-of-sample test data.

b. RQ 2: How is the result compared with other ML techniques (e.g., RF)?

This RQ is trying to understand if the multivariate CNN models with graphical and numerical data can provide superior results. After building and testing the models, this paper compares the overall prediction performances of the CNN-based prediction system and the numerical data-based RF. This provides a comprehensive view regarding the performances of the multivariate CNN model developed.

c. RQ 3: Is it possible to predict stock price by retail investor behavior?

In other words, I am interested in testing if it is possible to use the RIH charts to predict price patterns as CNN performs better with graphical data. If the stock characteristics can be used to predict retail investor movement, there is a possibility that retail investor movement may contain predictive information for stock information. This is similar to the study by Boehmer et al. (2021) but they use traditional data to recognize retail investor imbalance and make relative predictions on stock characteristics based on the imbalance. The factor RIH, which can be regarded as a different measure of imbalance, may be used as an alpha in asset pricing models or a factor in stock price prediction models.

IV. Data

The full dataset of this paper contains all stocks in the S&P 500 index with three main parts: the RIH (Retail Investors Holding), stock characteristics, and retail investor ownership (RIO).

The S&P 500 index's component stocks contribute to a significant proportion of the U.S. market values. As it is representative, many financial scholars utilize it as a small but precise sample to conduct research in the U.S. market (see e.g., Denis et al., 2003). The sample requires minimum computational costs and time of training, which enables the public to apply or adjust hyper-parameters in the system if needed. The original RIH data is at an hourly frequency over the sample period, initially provided by the Robinhood online trading platform, which started providing API access to stock popularity data of retail investors using the platform from 2 May 2018 to 13 August 2020. Thanks to Robintrack (www.robintrack.com) for saving all the data provided by Robinhood API during that period, RIH data is directly downloaded from this website. RIH data, even though scholars may name it differently, has been widely applied in economic and financial studies (e.g., Beck and Jaunin, 2023). As the Robinhood platform has the largest share of the U.S. retail investment market, the sample is normally regarded as representative and insightful when studying retail investor behaviors in the U.S. market (Boehmer et al., 2021). RIH data can be a measure of retail investor trading which can directly reflect the aggregate-level retail investor trading behavior, i.e., RIH increases if they are "joining" or

RIH decreases if they are "leaving" the stock. In the main CNN models to predict retail investor trading behavior, RIH is the output variable.

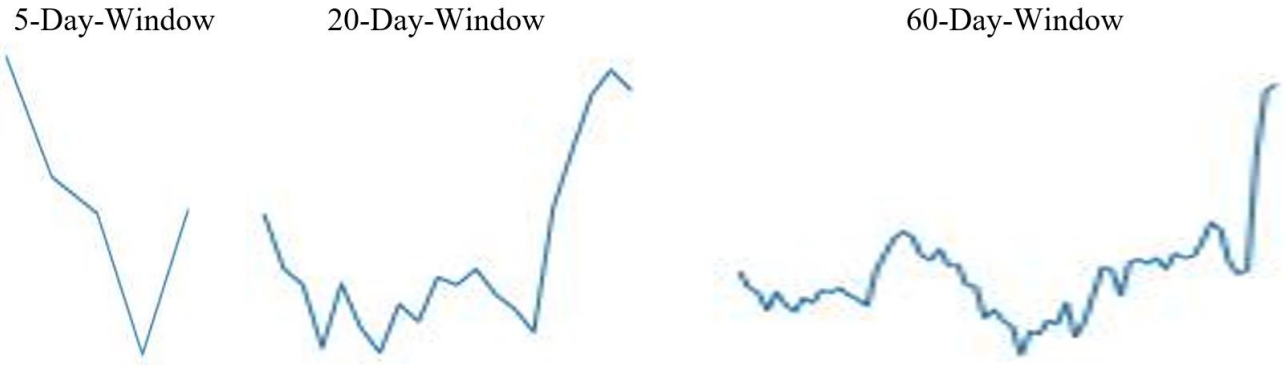
Stock characteristics are correlated with retail investor movement. For example, Berger (2021) builds an aggregate investor sentiment measure and finds that small, young, and volatile firms tend to have large sensitivities to retail investors while high retail sentiment periods lead to bad market returns in the following period. The stock characteristics data of S&P 500 component stocks is from Compustat-Capital IQ in the WRDS database. Compustat is a very standard and widely applied database in financial research regarding securities. I downloaded daily data for all available stocks listed in North America (USA and Canada) over the sample period, then matched them with S&P 500 stocks. The RIO data is calculated from the institutional ownership data from the WRDS Thomson-Reuters S34 dataset, based on 13F forms filled by public companies quarterly. High retail trading proportion stocks have strong lottery features that attract speculative retail investor (Han and Kumar, 2013). As there is a mismatch of the time frequencies for RIH data and stock data, I convert the hourly RIH data into daily RIH data considering the forecasting window and computing costs when training the model. Then, I cleaned out all stocks that were not included in the RIH data as the RIH variable is one of the most important variables in the sample. Moreover, I also deleted stocks with RIH data but without any stock characteristic data over the whole time period to make sure the input and output variables can match with each other.

Due to COVID-19, retail investor behavior and risk and trading preferences may be significantly affected by the pandemic, which might make them more sensitive to price changes, new information, or other factors. This regime change should not be ignored as the sample ranges before the spread of COVID-19 and afterward. Thus, I split the sample into two sets: 2 May 2018 to 12 August 2019 and 13 August 2019 to 13 August 2020 to understand if there are significant differences in the aggregate-level retail investor trading behavior or the results of predictions. During both periods, I randomly select 60% of the data as the training sample, 20% as the validation sample, and 20% as the test sample.

Input variables in the sample include stock price, dummy variable of high RIO (1 if the RIO is greater than or equal to 60%, 0 otherwise), dummy variable of low RIO (1 if the RIO is greater than or equal to 30% but less than 60%, 0 otherwise), trading volume, founding year, dividend yield, price to equity ratio (P/E), price to book value ratio (P/B), market value. Among which, dividend yield, P/E, P/B, and market value can be regarded as the measure for growth (young) or value (old) firm ----which has more or less herding behaviors, respectively (see e.g., Yao, Ma and He, 2014). Investors have been proven to behave differently when investing in old companies compared with old companies (Berger, 2021).

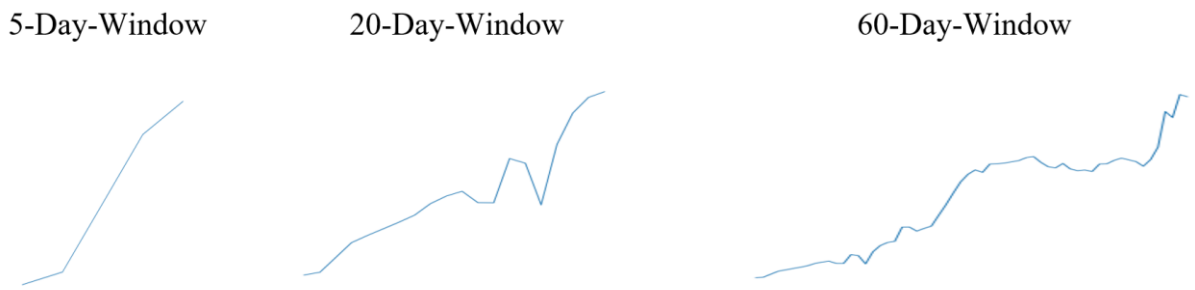
Then, I need to convert some numerical data (e.g., stock prices and RIH) into graphical data, which is generated into several windows (5-day, 20-day, and 60-day) so that the most suitable window for different lengths of predictions can be found. The graphical data should have some superior advantages over numerical data, especially when it is applied with the CNN model in predicting the retail investor behavior who trades with fluctuations they see on stock price charts. For example, the CNN model may detect more non-linear relationships and interactions from the stock price charts that are not detectable by CNN or other ML techniques using numerical data (Bogousslavsky, Fos and Muravyev, 2024). Besides, many technical traders and irrational investors are trading based on their analysis, their predictions, or their feelings on price charts. Other numerical data is also applied at the same time as they are important information to refer for retail investor's decision-making. Scaling is applied to both numerical and graphical data to avoid misprediction and incorrect results (Ahsan et al., 2021).

Figure 2. Examples of Graphical Data of Stock Close Price



Note: Examples of graphical close price data (Price Charts) of Apple Inc. generated with different window lengths. These price charts with different window lengths are used to predict the time-series movements of RIH in the future five to 45 days, so the immediate and long-term relationships can be discovered.

Figure 3. Examples of Graphical Data of Retail Investor Holding (RIH)



Note: Examples of graphical RIH data of Apple Inc. generated with different window lengths. These price charts with different window lengths are used to predict the time-series movements of stock close price in the future five to 45 days, so the immediate and long-term relationships can be discovered.

V. Methodology

This paper focuses on applying the novel multivariate CNN model to predict RIH based on graphical and numerical input data. To provide comprehensive and justifiable results, I also apply other less flexible but more interpretable ML techniques, RF. Then, I evaluate and compare the prediction performances using Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared. For the last RQ (RQ 3), I developed a univariate CNN model to provide some initial insights on predicting stock prices based on retail investor trading. This section will provide details regarding the models applied in the execution and training process.

To be consistent across three RQs, prediction tasks are treated as time-series regression problems using MSE as the target loss function, where $\hat{y}$ is the final output estimated from the prediction function, and y is the actual value from data. Instead of predicting whether more retail investors will hold the stock or not as a classification problem (i.e. will RIH increase or decrease for the next period), time-series regression prediction is more straightforward and precise as we can see how the RIH will fluctuate due to information changes on stocks (e.g., Liu, Hsaio and Tu). In RQ 1 and RQ 2, y and $\hat{y}$ are about RIH. In RQ 3, y and $\hat{y}$ are about stock prices. Together with the loss function (MSE), MAE and R^2 functions are applied to evaluate the prediction performances.

$$Loss\ Function\ (MSE)\ to\ Minimise: L_{(y_i,\hat{y}_i)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Rediduals\ (RSS)}{Total\ Sum\ of\ Squares\ (TSS)}$$

It is true that ML techniques, especially ANNs and CNNs, perform great in pattern recognition and prediction, but there are potential issues at the same time. For example, Abiodun et al. (2019) argue that problems with ANNs still need to be solved, including but not limited to data analysis problems (e.g., speed, accuracy, volume), scaling, monetary and time cost of training. Srivastava et al. (2014) argue that overfitting is a series problem in deep neural networks with many parameters. Thus, in this paper, some tricks and sacrifices regarding data processing and model training will be needed and specified later in the Methodology section.

a. Multivariate CNN Models for RQ 1

Deep learning techniques are known to work well in complex pattern recognition tasks. CNN is regarded as one of the best ML and deep learning techniques to extract features from graphical data. CNN has been largely adopted in image data tasks (e.g., Hoseinzade and Haratizadeh, 2019). Hoseinzade and Haratizadeh (2019) have proved that CNN is extremely successful in time-series

predictions. In recent years, scholars have been using CNN with stock chart images to make relative predictions and build profitable investment portfolios (e.g., Lee et al., 2019; Jiang, Kelly and Xiu, 2023). As CNN has been widely applied in economic and financial studies, I will not discuss too many technical details regarding modeling but will focus on how graphical data and numerical data are applied together in the multivariate CNN model and the selection of hyperparameters.

One of the major challenges is to apply both numerical and graphical data at the same time in a multivariate CNN model. To the best of my knowledge, this is the first to make this practice, at least in financial studies, though there are many practices for applying numerical data in multivariate CNN models (e.g., Cavalli and Amoretti, 2021) and graphical data in multivariate CNN models (e.g., Liu, Hsaio and Tu, 2018). As the stock price charts are generated with different window lengths, the daily numerical data (input and output data) need to be generated into different window lengths as well. Under each stock, the first date of the window is indexed as 1 and the rest is done in the same manner for both graphical and numerical data so that they can be matched together by the index. The whole dataset after cleaning contains five main parts: stock name, index, graphical input data array, numerical input data array, and the numerical target/output data array. The graphical data first passes through 2D convolutional layers, with the output of a 1D vector after flattening. The numerical data first passes through its dense layers (fully connected layers) using ReLu. Then, the first step processed outputs from graphical data and numerical data is combined and passed through the additional dense layers using ReLu. Finally, these outputs are worked through the output layer, a dense layer, with the linear activation function for regression tasks.

To save computational costs and time, not all hyperparameters are tuned and the following parameters are decided based on knowledge or previous studies. Compared with other deep learning networks, instead of fully connecting all the layers, the CNN architecture allows the convolution layers to be sparsely connected with the input layer. This setting allows filters to recognize features more efficiently (Khalil et al., 2021). Batch normalization is applied as it can accelerate the training by decreasing internal covariate shifts (Ioffe and Szegedy, 2015). ReLu (Rectified Linear Unit), which is computationally efficient and able to handle the vanishing gradient issue, is applied as the activation function. The same padding is used to keep the output of the convolutional layer has the same dimensions as the input. Adam (adaptive moment estimation) is applied as the optimizer to adjust the weights of networks to minimize MSE (the loss function) as it has a balanced performance while Adamax (a variant of Adam optimizer) works better when specific adjustments are needed and RMSprop (root mean square propagation) is more suitable on simpler tasks (Choi et al., 2019). The number of epochs, denoting how many times the dataset is passed through the network, is set at a medium value of 30 as a large number may improve the performances but bring the risks of overfitting.

As it is nearly impossible to decide the optimal values of hyperparameters manually, the hyperparameters described below are all tuned with different values by the Optuna function, a next-generation function for hyperparameter tuning (Akiba et al., 2019) to find the best combination of hyperparameters for the tasks (Tuba et al., 2021). Srivastava et al. (2014) argue that overfitting can be solved by the dropout function, which is to drop units and their connections randomly during training from the network. Batch size (32 or 64) describes the number of data used to train in a single pass, a larger batch size may increase the accuracy of the gradient estimated but also increase the computational costs of training the network. Convolutional layers (1 to 3) are set before the fully connected layers and output layers to capture complex features, while more convolutional layers may lead to overfitting and higher computational costs. The number of fully connected layers (2 to 3) affects the level of representativeness of abstractions from the data. The number of filters, determining the feature types that can be extracted from each convolutional layer, is 64 or 128 to capture more details. Kernel size (tuning between 2-2 or 3-3 as the image size is 8-8) is the size of filters, smaller size may capture more detailed features while larger size can capture more global features (Argawal and Mittal, 2020). The learning rate (allowing between 10^{-4} to 10^{-5}) is one of the most important hyperparameters that illustrates the size of changes in the model in response to the updates of weights in the model according to the estimated error each time. Gradient clipping can be used to prevent the risk of exploding gradients that may make the model unstable or unable to converge (Zhang et al., 2019).

The main prediction system (RQ 1) is built following the steps below. For pre and COVID periods, hyperparameters (batch size, number of convolutional and fully connected layers, number of filters in the convolutional layers, kernel size, dropout rate, learning rate, learning rate, gradient clipping, and optimizer) are tuned by Optuna to find the best combination for the CNN model on various window lengths (5-day, 20-day, and 60-day) of graphical data over different lengths of prediction target periods (next 5-day, 10-day, 20-day, 30-day, and 45-day). Then, I train, validate, and test the CNN models with the most appropriate combinations of hyperparameters. As a result, the prediction system with CNN models is packed and can be easily applied with simple input information, including the length of prediction, level of retail ownership (high, medium, or low), stock prices, and other stock characteristics if any.

b. Random Forests (RF) for RQ 2

In order to answer RQ2, RF models are set with numerical data, under similar procedures and environments (i.e., same splits of dataset, same scaling of dataset, and same prediction periods applied), to compare the prediction performances with the CNN prediction system built. Compared with neural networks (including multivariate CNN), RF can provide more interpretability while sacrificing some complexity in time series prediction tasks (Zhao et al., 2019). Compared with traditional financial

econometric models like autoregressive integrated moving averages (ARIMA), Kane et al. (2014) argue that RF performs better in time-series prediction tasks though RF techniques are less interpretable compared with ARIMA models. RF has become popular and has been applied in plenty of studies about financial predictions (e.g., Khaidem, Saha and Dey, 2016; Sadorsky, 2021). Zhang, Zhong and Hu (2022) applied RF with time-series data to predict the risks in the financial system.

Considering the brevity, I will only discuss the selection of hyperparameters which ensures the balance, complexity, and generalization of models. All possible variables (stock characteristics) are applied so that RF has the greatest potential to prune and form the most suitable models (e.g., Tyralis and Papacharalampous, 2017). Estimators are the number of trees used in the forest, more estimators may increase model performance as they can decrease variance while they lift more burdens on the computation at the same time. Max depth describes how complex every single tree is. Deeper trees may extract more patterns but bring concerns regarding overfitting. In addition, a minimum sample split requires the minimum number of samples needed when splitting an internal node. The higher requirements help to solve overfitting problems by preventing overweighting on specific features. Min samples leaf is the least samples needed in a leaf node, which could decrease the effects of noises. Max features regulate the highest number of features that can be considered in a split, controlling the randomness introduced in building trees. Lower max features can avoid overfitting but may miss important patterns. Bootstrap denotes whether bootstrap is applied in the sample when building trees. Max samples describe the maximum number of samples that can be used to train each estimator, which controls randomness in training. OOB (out-of-bag) score is whether the out-of-bag samples are used to estimate the accuracy. Min weight fraction leaf describes the minimum weighted fraction of all weights required for one single leaf node, which avoids the model being too skewed. Max leaf nodes denote the maximum number of leaf nodes allowed in one single tree, preventing the models built from being too complex and overfitting.

c. Univariate CNN Models for RQ 3

The univariate CNN model is trying to see if the graphical data of RIH can be used to predict close-price movements of stocks. The sample is the same, all component stocks of the S&P 500 index with efficient data available. The graphical data of RIH is also generated into 5, 20, and 60-day window lengths to predict the time series of close prices over the next 5, 10, 20, 30, and 45 days (similar to RQ 1). Besides, all hyperparameters described in RQ 1 are also tuned in RQ 3. Consistent with RQ 1, hyperparameters tuned include the number of filters, kernel size, number of convolutional layers, number of dense layers, dense units, dropout rate, learning rate, gradient clipping, and batch sizes.

VI. Results & Discussions

a. Can CNNs recognize patterns from price charts and other numerical stock information and predict retail investors' behaviors?

Table 1 presents the performances of the multivariate CNN models for the period between 2 May 2018 and 12 August 2019. Results show that the multivariate CNN models can accurately make out-of-sample predictions on RIH before the COVID-19 pandemic based on price charts and other numerical data with different window lengths, with test r-squares of 0.86 to 0.97. In other words, the CNN models built can extract efficient patterns from stock price charts and other numerical stock characteristics. Then they can link those patterns with the movements of RIH and make relative movements. From the test MSE and MAE perspectives, the models perform well, and the losses are low and constant.

Filters are set as 64 across all the trials and prediction tasks, suggesting it is a moderate size that captures many features but is not too computationally costly. The kernel size changes from (2, 2) for short prediction periods to (3, 3) for longer periods, indicating that longer periods contain more complex patterns that need to be captured by larger kernel sizes, which is aligned with Li et al. (2019). Interestingly, the evidence shows that more convolutional layers, dense units, and dense layers do not necessarily perform better. In other words, more factors play important roles in the accuracy of predictions and over-complex network architectures may cause over-fitting or inefficiency problems without performance improvements. The dropout rates are increasing with the lengths of forecasts, suggesting that they are effective in mitigating overfitting problems in longer-term predictions. With longer prediction lengths, learning rates decrease corresponding to complexity.

Before COVID-19 (stable period), the models perform better with less complex architecture or less aggressive tuning. However, during the unstable period (COVID-19 period in Table 2), models need to be built more carefully to balance between model complexity and potential overfitting issues due to the unstable market conditions. These market conditions are reflected in the process of modeling as there are more noses and shifts in the dataset.

Therefore, the multivariate CNN models generally perform well in the time-series predictions for RIH based on graphical and numerical stock characteristics over pre- and COVID periods. Evidence from these models indicates that if the market is stable, the models can achieve lower errors and higher r-squares (prediction accuracy) when predicting using similar or less complex architectures. The increased uncertainty worldwide due to the pandemic did spread to the financial markets, leading to higher errors and lower prediction accuracies as there were more unpredictable changes in the financial markets during the COVID-19 period. These conclusions are consistent with previous financial studies regarding the risks brought by global events in the financial markets using traditional risk measures like volatility and VaR (Value-at-Risk) (Dhaene et al., 2006; Wullweber, 2020).

Table 1. Multivariate CNN Models for Pre-COVID Period

Data Window Length	5	20	20	20
Prediction Length	5	5	10	20
Number of Trials	100	100	100	100
Filters	64	64	64	64
Kernel Size	(2, 2)	(2, 2)	(3, 3)	(3, 3)
Convolutional Layers	1	3	1	1
Dense Layers	2	3	2	2
Dense Units	512	512	256	512
Dropout Rate	0.10705756	0.12923119	0.16709925	0.18162824
Learning Rate	0.00024596	0.00019005	0.00030673	0.00014213
Batch Size	32	32	32	32
Clipnorm	None	None	None	None
Best Validation MSE	0.01227511	0.02093476	0.04649147	0.03433068
Test MSE	0.01376898	0.06847247	0.02236040	0.02613123
Test MAE	0.05849719	0.02093476	0.06604411	0.07328673
Test R^2	0.93404990	0.93853039	0.92468584	0.93135941
Continued:				
Data Window Length	60	60	60	60
Prediction Length	10	20	30	45
Number of Trials	100	100	100	100
Filters	64	128	64	128
Kernel Size	(2, 2)	(2, 2)	(3, 3)	(3, 3)
Convolutional Layers	2	2	2	1
Dense Layers	3	3	3	3
Dense Units	512	512	512	512
Dropout Rate	0.10033642	0.11322465	0.17594485	0.12864903
Learning Rate	0.00044655	0.00029482	0.00014121	0.00018232
Batch Size	32	32	32	32
Clipnorm	None	None	None	None
Best validation loss	0.00756842	0.03071809	0.03083617	0.01589363
Test MSE	0.01108817	0.03262032	0.03671031	0.01180057
Test MAE	0.04466520	0.07894426	0.09687176	0.05690427
Test R^2	0.96705741	0.87264123	0.86247122	0.95263809

Note: This table presents the hyperparameters tuned and model performances for different lengths of prediction based on different window lengths of price charts in the multivariate CNN models for the pre-COVID period (2 May 2018 to 12 August 2019). Test MSE, MAE, and R-square are used to measure the model performances as this is for time-series predictions.

Table 2. Multivariate CNN Models for the COVID Period

Data Window Length	5	20	20	20
Prediction Length	5	5	10	20
Number of Trials	100	100	100	100
Filters	64	64	64	64
Kernel Size	(2, 2)	(2, 2)	(3, 3)	(3, 3)
Convolutional Layers	3	1	2	1
Dense Layers	2	3	2	3
Dense Units	512	512	512	512
Dropout Rate	0.10002411	0.10228834	0.11627218	0.1003674
Learning Rate	0.00034749	0.00027455	0.00025962	0.00022848
Batch Size	32	32	32	32
Clipnorm	None	None	None	None
Best Validation MSE	0.03933112	0.0335716	0.03167419	0.03416783
Test MSE	0.03846738	0.06357579	0.0458964	0.04389987
Test MAE	0.07394379	0.0846988	0.08236904	0.081453
Test R^2	0.85726351	0.83163944	0.84035748	0.80927323
Continued:				
Data Window Length	60	60	60	60
Prediction Length	10	20	30	45
Number of Trials	100	100	100	100
Filters	64	64	64	64
Kernel Size	(3, 3)	(3, 3)	(2, 2)	(2, 2)
Convolutional Layers	1	1	1	1
Dense Layers	2	3	2	2
Dense Units	512	512	512	512
Dropout Rate	0.10760396	0.13686995	0.1017124	0.1358614
Learning Rate	0.0002638	0.00020428	0.00044819	0.00015705
Batch Size	32	32	32	32
Clipnorm	None	None	None	None
Best validation loss	0.02430349	0.02906801	0.02811958	0.03119752
Test MSE	0.0707968	0.03966645	0.04515128	0.05144647
Test MAE	0.10126211	0.08942027	0.08543306	0.09112121
Test R^2	0.84601009	0.85665327	0.85369778	0.85277516

Note: This table presents the hyperparameters tuned and model performances for different lengths of prediction based on different window lengths of price charts in the multivariate CNN models for the COVID period (13 August 2019 to 13 August 2020). Test MSE, MAE, and R-square are used to measure the model performances as this is for time-series predictions.

From the perspectives of ML application and modeling, the key takeaway is that market conditions can significantly affect model performances and optimal architectures. When building prediction models for the financial market using ML techniques, regime changes should also be considered just as qualified scholars do when they are making predictions using traditional finance models like autoregressive integrated moving averages (Crawford and Fratantoni, 2003). Dynamic modeling (e.g., hyperparameter tuning) for ML models is necessary to adapt to the changes in market situations.

Even though the models show their capabilities in predicting retail investor holdings on stocks based on open-market information, more analyses are needed to understand why they can do so. On one hand, the characteristics of the stocks may contain factors that are considered by most of the retail investors in the sample, leading to rational or irrational trading behaviors at an aggregate level. Price charts may play an important role as many investors may trade based on what they see. As CNN techniques are proven to perform well in feature extractions from graphical data, the CNN models may pair those patterns in the datasets with the RIH movements.

On the other hand, different retail investors (fundamental traders, technical traders, and irrational traders) may react to financial information differently, and one type of them may dominate the market and their internal effects are not trading off, so that the aggregated outcome is predictable.

Under EMH, fundamental traders believe that the stock price should directly reflect the real value of the firm (Fama, 1970). Fundamental analysis provides insights regarding the overall performance of companies. Retail investors who believe in EMH and do fundamental analysis will trade based on changes in the real value of the firm, which can be reflected in the stock price. Even though sometimes the stock price derives from the real value of the firm, scholars (e.g., Chen, Goldstein and Jiang, 2007) find that stock prices, especially in sophisticated financial markets (e.g., U.S. and U.K.), are generally aligned with the value of firms, which are captured by financial ratios of the stock. Fundamental investors buy stocks at a relatively low price compared with the fundamental value they estimated and sell when they find the stocks are overvalued by the market. However, the fundamental value of firms does not change significantly over a short period, as a result, fundamental traders tend to trade less as well (Breaban and Noussair, 2015).

Besides, many other retail investors (precisely, technical traders) are currently trading based on their technical analysis (Edwards, Magee and Bassetti, 2018). Treynor and Ferguson (1985) argue that, based on public information, technical analysis is applied to generate valuable insights that are not appreciated by the public. More specifically, technical traders are trying to predict price trends based on previous stock prices, trading volume, or other public data (Park and Irwin, 2007). Therefore, the patterns in prices and other stock information may be linked with these technical investors' trading behavior.

Lastly, retail investors may trade irrationally (i.e., irrational traders) and their behaviors can be affected by complicated factors (e.g., networking effects). Brown and Jennings (1989) argue that retail investors have rational expectations regarding the relationship between prices and signals from past prices. However, not all retail investors are rational. In other words, some retail investors, who do not trade based on fundamental or technical analysis, may trade by unreasonable signals or based on factors like other investors' movements, causing a herding event or networking effect that will be directly reflected in a sharp increase or decrease in stock prices and more retail investors will react based on this observation.

In this case, the aggregated outcomes are less complex and predictable as the irrational traders tend to receive similar information and have similar emotional reactions due to human nature. For example, Shefrin and Statman (1984) argue that irrational investors have disposition effects, who tend to sell too quickly if they make profits and tend to hold too long when they suffer losses, which is evident by Odean (1998). A recent famous example is the story of stock GameStop (GME) which was heavily shorted, especially by hedge funds (Malz, 2021). The GME's share price, unexpectedly, experienced an extremely high run-up from the lowest close price of $2.8 per share to the highest price of $347.51 per share (January 2021 to May 2021). This episode is due to the high popularity of trading in the stock by retail investors using the Robinhood investment platform and their sentiment to band together and resist institutional investors. This provides the foundation for relationships between stock characteristics and irrational investors' trading, which can be detected by ML and especially deep learning methods that are good at extracting non-linear patterns from data (e.g., LeCun, Bengio and Hinton, 2015).

Therefore, as RIH measures the popularity of stocks traded in Robinhood, the aggregated outcomes predicted by the CNN models are based on the movements of a specific type of investors who share the largest proportion of the popularity of the stocks. Robinhood, as a commission-free online trading portfolio for the public, comprises a significant part of retail investors who have fewer resources and less financial literacy (Prasad, Kiran and Sharma, 2021). In other words, irrational investors may dominate the Robinhood platform so that the prediction of RIH based on stock characteristics can be realized.

b. How is the result compared with other ML techniques (random forest)?

Table 3 presents the performances of using numerical data with RF to predict the RIH based on stock characteristics. Consistent with the findings of the multivariate CNN models in RQ 1, shorter prediction periods tend to perform better. Generally, even though the RF models can make some good predictions for RIH over some periods in the future (i.e., the following 5, 10, and 20 days), the performances are significantly lower than the performances by the multivariate CNN models. One may

argue that this is due to the differences in the sample period, but the fact is that even compared with performances by multivariate CNN in the COVID-19 period (the most unstable period), the overall performances of RF models are still worse. Therefore, it is reasonable to conclude that the multivariate CNN model, based on both graphical and numerical data, outperforms the RF models with numerical data in this specific time-series prediction task.

Nevertheless, interestingly, Kumar and Thenmozhi (2006) find that RF performs better than neural networks and discriminant analysis in predicting stock index fluctuations, a classification problem. These contrasting facts may indicate that it is hard to make a justified conclusion regarding the performances of different ML techniques. The combination of graphical and numerical data may benefit the multivariate CNN models so that more patterns can be extracted, compared with RF models using numerical data only. The performances are determined by a range of factors including data applied, tasks defined, and architectures used. Thus, it is important to compare different ML techniques when trying to make predictions.

Table 3. Random Forest Models for Both pre-COVID and COVID Periods

Prediction Length	5	10	20	30	45
Number of Trials	100	100	100	100	100
Estimators	100	200	200	200	200
Max Depth	None	10	10	10	10
Min Samples Split	5	2	2	5	5
Min Samples Leaf	4	1	1	4	4
Max Features	0.75	Auto	Auto	Auto	Auto
Bootstrap	True	True	True	False	False
Max Samples	0.75	0.75	0.75	None	None
Oob Score	False	False	False	False	False
Min Weight Fraction Leaf	0.01	0.01	0.01	0.0	0.0
Max Leaf Notes	20	None	20	10	10
Test MSE	0.10849843	0.13458694	0.18984327	1.98573423	2.38978717
Test MAE	0.16574234	0.19852314	1.29350829	1.087312556	1.57349058
Test R^2	0.86327423	0.81761405	0.7086693	0.13298094	-0.77988210

Note: This table shows the best hyperparameters found by tuning in RF models and their performances for different lengths of prediction horizons. The data ranges from 2 May 2018 to 13 August 2020 (full sample period), covering both pre-COVID and COVID-19 periods.

Table 4. Univariate CNN Model to Predict Stock Price Before COVID

Data Window Length	5	20	20	20
Prediction Length	5	5	10	20
Number of Trials	100	100	100	100
Filters	128	128	128	128
Kernel Size	(2, 2)	(3, 3)	(2, 2)	(3, 3)
Convolutional Layers	3	3	2	3
Dense Layers	3	3	3	3
Dense Units	256	512	256	512
Dropout Rate	0.19964969	0.10621485	0.1818255	0.12364725
Learning Rate	0.00028521	0.00027265	0.00018013	0.00033332
Batch Size	64	64	64	32
Clipnorm	1	1	None	1
Best Validation MSE	2.40419769	2.38905025	2.37669468	2.31487799
Test MSE	0.13734864	0.13618591	0.13553624	0.13646883
Test MAE	0.28608149	0.28491801	0.284942	0.28518656
Test R^2	-2152.4766	-1978.026	-1856.0105	-1771.4948
Continued:				
Data Window Length	60	60	60	60
Prediction Length	10	20	30	45
Number of Trials	100	100	100	100
Filters	64	128	128	64
Kernel Size	(3, 3)	(3, 3)	(2, 2)	(3, 3)
Convolutional Layers	1	2	1	3
Dense Layers	3	3	2	2
Dense Units	256	256	512	512
Dropout Rate	0.11728207	0.16332556	0.19027454	0.10455531
Learning Rate	0.00047374	0.00049694	0.00039510	0.00010281
Batch Size	32	64	64	64
Clipnorm	None	1	None	1
Best Validation MSE	2.29347467	2.34607530	2.34782696	2.22212839
Test MSE	0.13561931	0.13439223	0.13482787	0.13390299
Test MAE	0.28451496	0.28270304	0.28382364	0.27995661
Test R^2	-1572.0978	-1538.3334	-1287.7703	-1011.5887

Note: This table presents the hyperparameters tuned and model performances for different lengths of prediction based on different window lengths of RIH charts in the univariate CNN models for the pre-COVID period (2 May 2018 to 12 August 2019). Test MSE, MAE, and R-square are used to measure the model performances as this is for time-series predictions.

Table 5. Univariate CNN Model to Predict Stock Price During COVID

Data Window Length	5	20	20	20
Prediction Length	5	5	10	20
Number of Trials	100	100	100	100
Filters	128	64	64	64
Kernel Size	(2, 2)	(2, 2)	(2, 2)	(3, 3)
Convolutional Layers	3	3	1	3
Dense Layers	2	3	3	3
Dense Units	512	512	512	512
Dropout Rate	0.18498491	0.12647734	0.10664773	0.12620943
Learning Rate	0.00039117	0.00014189	0.00017456	0.00012899
Batch Size	32	64	64	64
Clipnorm	None	None	1	1
Best Validation MSE	2.53506565	0.23077881	0.22999623	0.22946684
Test MSE	0.24082744	0.26394188	0.24994528	0.23718926
Test MAE	0.3300198	0.32789856	0.32932186	0.32553157
Test R^2	-2481.1570	-2620.9893	-2265.4048	-2374.8445
Continued:				
Data Window Length	60	60	60	60
Prediction Length	10	20	30	45
Number of Trials	100	100	100	100
Filters	64	128	64	64
Kernel Size	(2, 2)	(2, 2)	(3, 3)	(2, 2)
Convolutional Layers	2	2	3	1
Dense Layers	2	3	3	3
Dense Units	512	512	512	512
Dropout Rate	0.19804415	0.11088118	0.14573051	0.16457965
Learning Rate	0.00019138	0.00033215	0.00015317	0.00011133
Batch Size	32	32	64	1
Clipnorm	None	1	None	64
Best Validation MSE	2.49292278	2.47085595	2.44986868	2.41192506
Test MSE	0.28180960	0.26291627	0.25468320	0.24594986
Test MAE	0.35219872	0.34776357	0.34459248	0.33970031
Test R^2	-1920.8431	-1566.2758	-1134.5947	-756.4354

Note: This table presents the hyperparameters tuned and model performances for different lengths of prediction based on different window lengths of RIH charts in the univariate CNN models for the COVID period (13 August 2019 to 13 August 2020). Test MSE, MAE, and R-square are used to measure the model performances as this is for time-series predictions.

c. Is it possible to predict stock price by retail investor behavior?

The short answer is no, at least the univariate CNN model using graphical RIH as input data cannot predict the stock close price over the whole sample period (pre-COVID in Table 4 and COVID period in Table 5). Prediction accuracies of all the univariate CNN models for the RQ are bad while the test MSE and MAE are relatively stable. This does not necessarily suggest that retail investor trading contains no information about stock prices. A well-known fact is that stock prices are too complex to predict even though many scholars and practitioners in the industry have been trying to do so since the existence of stock markets (e.g., Firtz and Weinhartz., 2015; Kelly and Xiu, 2020).

VII. Robustness Check

To check the robustness of the findings above, focusing on RQ 1, I first extend the number of trials in the multivariate CNN model when predicting RIH for the pre-COVID period (Table 6). Compared with the next 5-day prediction using 5-day-window data in RQ 1, the performance of the model applied more trials in the tuning process is better. This may be because more trials allow the architecture to tune more times with different kinds of combinations of hyperparameters, but the risk of overfitting increases as well.

Table 6. More Trials on Multivariate CNN

Data Window Length	5
Prediction Length	5
Number of Trials	200
Filters	64
Kernel Size	(3, 3)
Convolutional Layers	2
Dense Layers	3
Dense Units	512
Dropout Rate	0.10005535
Learning Rate	0.00049966
Batch Size	32
Clipnorm	None
Best Validation MSE	0.01098521
Test MSE	0.01147238
Test MAE	0.01245235
Test R^2	0.97985983

VIII. Further Studies and Limitations

This paper is performed to the best of my knowledge and capabilities, however, there are some potential issues and limitations that remain. More studies may be motivated and improved based on this.

First and foremost, the sample size applied in this study is not satisfying as my computing power is very limited. More data are needed to train, validate, and test to ensure that the models built are robust and general enough to work across different stocks. The Robinhood dataset contains RIH information for about 8000 stocks, therefore, it is possible to utilize all the information there and link it with stock characteristics in the financial markets, which are publicly available. The ML techniques are known to work well within the sample and have many concerns regarding overfitting (Cavalli and Amoretti, 2021). It is necessary to test the models generalized in a different and new dataset from the one used in training and validating, to make sure that the models can keep performing well when applying in the practices.

In addition, there are potential concerns about data leakage in time-series prediction, which may make the prediction accuracy incorrectly high, as discussed in a very recent paper by Rosenblatt et al. (2024). If data is spread across the train, valid, and test datasets in the time series, the ML techniques may learn from that information and make the predictions look more accurate than they were. This false accuracy rate is not desirable, and it is important to avoid training data containing future information. In this paper, the sample is randomly split but rolling windows of data are applied to handle this issue (Nayak and Ojha, 2020). However, the problem might still exist as the problem may start from splitting the data. Future studies can either apply the rolling window split or the expanding window split for the whole sample to solve the problem (Rosenblatt et al., 2024).

In the future, more methods can be applied. It might be interesting to understand retail investors on Robinhood and other online trading platforms through surveys or questionnaires to provide more evidence regarding the relationship between stock characteristics and retail investor trading. It could also be interesting to discover patterns in stock price charts that may be extracted by CNN and other ML techniques, checking if they can be linked with technical analysis. As CNN has its limitations (Lipton, Berkowitz and Elkan, 2015), other deep learning techniques may be applied to compare relative performances.

IX. Conclusions

Many economics and finance scholars have put effort into understanding individual decision-making behavior, which is known to be complex while significant for the development of society and social welfare. Retail investors of stocks are found to affect not only the stock or financial markets but also

wider society (e.g., Subrahmanyam, 2008). However, not all retail investors know the nature of their investments which brings great uncertainty and risks in the markets (BNY Mellon Retail Investor Survey, 2022). Therefore, it is necessary to predict retail investor trading behavior and understand its potential effects on the markets. Previous researchers (e.g., Barber and Odean, 2019) have worked on this for a long time and contributed to retail investment, a significant part of the financial literature. Scholars (e.g., Boehmer et al., 2021) have worked on a sub-area of retail investment studies of estimating and predicting the effects of retail investors (e.g., trading imbalance) on the financial market (e.g., stock price, liquidity, etc). Few of them tried to estimate and predict the retail investor's movements or imbalance.

As a pioneering work, this paper tries to shed some light by building a widely accessible time-series prediction system for retail investor trading behavior using publicly available data. Combining graphical and numerical data and applying them to a multivariate CNN model is a bold trial but the results are satisfying. With different window lengths input data (5, 20, and 60), the prediction system built is found to be effective in predicting RIH for the next 5, 10, 20, 30, and 45 days. This study provides further evidence regarding the extra uncertainty brought by the COVID-19 pandemic into the financial markets. Besides, evidence suggests that it is necessary to consider market conditions before the application of different ML models which may be affected by regime changes. This study also argues that irrational investors are dominating the Robinhood platform, suggesting more actions may be taken from the perspective of financial education. With numerical data, random forest models are found to perform worse compared with the multivariate CNN models. Moreover, I also find that the univariate CNN models built with RIH cannot predict stock close prices.

This paper tries to shed some light on financial literature by applying novel data and methodologies to the understanding of retail investments. In addition, various financial market participants may draw benefits from this predicting system as well. First and foremost, retail investors can benefit by understanding their peers' reactions to new information on stock characteristics. Besides, financial institutions could learn retail investor behavior and better engage with retail investors, trading with or against retail investors. Last but not least, financial authorities may utilize this prediction framework to build relative warning systems to avoid market crises due to significant fluctuations in retail investment.

There are some potential concerns related to this study as well, including sample size, data leakage, and methodologies applied. More insights should be drawn from future studies to help the public further understand retail investor behavior, an interesting and long-arguing but sometimes ill-defined topic.

X. Reference List

Abiodun, O. I., Kiru, M. U., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., & Gana, U. (2019). Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. *IEEE Access*, *7*, 158820–158846. https://doi.org/10.1109/access.2019.2945545

Abudy, M. M. (2020). Retail investors' trading and stock market liquidity. *The North American Journal of Economics and Finance*, *54*, 101281.

Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *The Visual Computer*, *36*(2), 405-412.

Agrawal, S., Thakkar, D., Soni, D., Bhimani, K., & Patel, C. (2019). Stock market prediction using machine learning techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(2), 1099-1103.

Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Allsopp, L., & Hey, J. D. (2000). Two experiments to test a model of herd behaviour. Experimental Economics, 3(2), 121–136. doi:10.1007/bf01669304

Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, *120*, 108102. Gu, S.,

Bakar, S., & Yi, A. N. C. (2016). The impact of psychological factors on investors' decision making in Malaysian stock market: a case of Klang Valley and Pahang. *Procedia Economics and Finance*, *35*, 319-328.

Bakar, S., & Yi, A. N. C. (2016). The impact of psychological factors on investors‟ decision making in malaysian stock market: a case of Klang Valley and Pahang. Procedia Economics and Finance, 35, 319–328.

Baker, H. K., & Nofsinger, J. R. (2002). Psychological biases of investors. Financial Services Review, 11

Baker, H. K., & Nofsinger, J. R. (2002). Psychological biases of investors. *Financial services review*, *11*(2), 97-116.

Baker, H. K., & Ricciardi, V. (2014). How biases affect investor behaviour. *The European Financial Review*, 7-10.

Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, *47*(3), 448-465.

Barber, B. M., & Odean, T. (1999). The courage of misguided convictions. *Financial Analysts Journal*, *55*(6), 41-55.

Barber, B. M., & Odean, T. (2000a). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance*, *55*(2), 773-806.

Barber, B. M., & Odean, T. (2000b). The courage of misguided convictions: The trading behavior of individual investors. SSRN Electronic Journal. doi:10.2139/ssrn.219175

Barber, B. M., & Odean, T. (2001). Boys will be Boys: Gender, Overconfidence, and Common Stock Investment. The Quarterly Journal of Economics, 116(1), 261–292. doi:10.1162/003355301556400

Barber, B. M., & Odean, T. (2002). Online Investors: Do the Slow Die First? *Review of Financial Studies*, *15*(2), 455–488. https://doi.org/10.1093/rfs/15.2.455

Barber, B. M., Lin, S., & Odean, T. (2023). Resolving a paradox: Retail trades positively predict returns but are not profitable. *Journal of Financial and Quantitative Analysis*, 1-35.

Barrot, J. N., Kaniel, R., & Sraer, D. (2016). Are retail traders compensated for providing liquidity?. *Journal of Financial Economics*, *120*(1), 146-168.

Bellofatto, A., D'Hondt, C., & De Winne, R. (2018). Subjective financial literacy and retail investors' behavior. *Journal of Banking & finance*, *92*, 168-181.

Berger, D. (2022). Investor sentiment: A retail trader activity approach. *Review of Accounting and Finance*, *21*(2), 61-82.

Bezerra, P.C.S., Albuquerque, P.H.M. Volatility forecasting via SVR–GARCH with mixture of Gaussian kernels. Comput Manag Sci 14, 179–196 (2017). https://doi.org/10.1007/s10287-016-0267-0

BNY Mellon. (2024). *U.S. Retail Investor: Insights & Implications | BNY Mellon*. Bnymellon.com. https://www.bnymellon.com/us/en/insights/the-state-of-the-us-retail-investor.html

Boehmer, E., Jones, C. M., Zhang, X., & Zhang, X. (2021). Tracking retail investor activity. *The Journal of Finance*, *76*(5), 2249-2305.

Bogousslavsky, V., Fos, V., & Muravyev, D. (2024). Informed trading intensity. *The Journal of Finance*, *79*(2), 903-948.

Breaban, A., & Noussair, C. N. (2015). Trader characteristics and fundamental value trajectories in an asset market experiment. *Journal of Behavioral and Experimental Finance*, *8*, 1-17.

Brooks, C., & Williams, L. (2022). When it comes to the crunch: Retail investor decision-making during periods of market volatility. *International Review of Financial Analysis*, *80*, 102038.

Camerer, C. (1995). Individual decision making. *The handbook of experimental economics*, *1*, 587-704.

Cavalli, S., & Amoretti, M. (2021). CNN-based multivariate data analysis for bitcoin trend prediction. *Applied Soft Computing*, *101*, 107065.

Cheng, F., Chiao, C., Wang, C., Fang, Z., & Yao, S. (2021). Does retail investor attention improve stock liquidity? A dynamic perspective. *Economic Modelling*, *94*, 170-183.

Chen, H., Xiao, K., Sun, J., & Wu, S. (2017). A double-layer neural network framework for high-frequency forecasting. *ACM Transactions on Management Information Systems (TMIS)*, *7*(4), 1-17.

Chen, L., Wang, S., Fan, W., Sun, J., & Naoi, S. (2015, November). Beyond human recognition: A CNN-based framework for handwritten character recognition. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 695-699). IEEE.

Chen, T. H., Ho, R. J., & Liu, Y. W. (2019). Investor personality predicts investment performance? A statistics and machine learning model investigation. *Computers in Human Behavior*, *101*, 409-416.

Cheng, F., Chiao, C., Wang, C., Fang, Z., & Yao, S. (2021). Does retail investor attention improve stock liquidity? A dynamic perspective. *Economic Modelling*, *94*, 170–183. https://doi.org/10.1016/j.econmod.2020.10.001

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*.

Crawford, G. W., & Fratantoni, M. C. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics*, *31*(2), 223-243.

Corter, J. E., & Chen, Y. J. (2006). Do investment risk tolerance attitudes predict portfolio risk?. *Journal of business and psychology*, *20*, 369-381.

Daniel, K. D., Hirshleifer, D., & Subrahmanyam, A. (2001). Overconfidence, Arbitrage, and Equilibrium Asset Pricing. *The Journal of Finance*, *56*(3), 921–965. https://doi.org/10.1111/0022-1082.00350

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *the Journal of Finance*, *53*(6), 1839-1885.

deHaan, E., & Glover, A. (2024). Market Access and Retail Investment Performance. *The Accounting Review*, 1-27.

deHaan, E., & Glover, A. (2024). Market Access and Retail Investment Performance. *The Accounting Review*, 1-27.

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics*, *118*(3), 909–968. https://doi.org/10.1162/00335530360698469

Dhaene, J., Vanduffel, S., Goovaerts, M. J., Kaas, R., Tang, Q., & Vyncke, D. (2006). Risk measures and comonotonicity: a review. *Stochastic models*, *22*(4), 573-606.

Durham, G., & Geweke, J. (2013). Improving asset price prediction when all models are false. *Journal of Financial Econometrics*, *12*(2), 278-306.

Edmans, A., Garcia, D., & Norli, O. (2006). Sports Sentiment and Stock Returns. *SSRN Electronic Journal*, *62*(4). https://doi.org/10.2139/ssrn.677103

Evans, D. A. (2006). Subject perceptions of confidence and predictive validity in financial information cues. *The Journal of Behavioral Finance*, *7*(1), 12-28.

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383-417.

Farrell, M., Green, T. C., Jame, R., & Markov, S. (2022). The democratization of investment research and the informativeness of retail investor trading. *Journal of Financial Economics*, *145*(2), 616-641.

Fisch, J. (2022). GameStop and the Reemergence of the Retail Investor. *Boston University Law Review*, 1799. https://scholarship.law.upenn.edu/faculty_scholarship/2817/

Fong, K. Y., Gallagher, D. R., & Lee, A. D. (2014). Individual investors and broker types. *Journal of Financial and Quantitative Analysis*, *49*(2), 431-451.

Friedman, H. L., & Zeng, Z. (2024). Retail investor trading and market reactions to earnings announcements. *Available at SSRN 3817979*.

Fritz, F., & Weinhardt, C. (2015). The role of technical analysis in retail investor trading. *Available at SSRN 2734884*.

Gallup Inc. (2023). *Global Research*. Gallup.com. https://www.gallup.com/analytics/318875/global-research.aspx

Goetzmann, W. N., & Zhu, N. (2005). Rain or Shine: Where is the Weather Effect? *European Financial Management*, *11*(5), 559–578. https://doi.org/10.1111/j.1354-7798.2005.00298.x

Goodfellow, C., Bohl, M. T., & Gebka, B. (2009). Together we invest? Individual and institutional investors' trading behaviour in Poland. *International Review of Financial Analysis*, *18*(4), 212-221.

Guermat, C., & Harris, R. D. F. (2002). Forecasting value at risk allowing for time variation in the variance and kurtosis of portfolio returns. *International Journal of Forecasting*, *18*(3), 409–419. https://doi.org/10.1016/s0169-2070(01)00122-4

Han, B., & Kumar, A. (2013). Speculative retail trading and asset prices. *Journal of Financial and Quantitative Analysis*, *48*(2), 377-404.

Hillenbrand, C., Saraeva, A., Money, K., & Brooks, C. (2020). To invest or not to invest?: The roles of product information, attitudes towards finance and life variables in retail investor propensity to engage with financial products. *British Journal of Management*, *31*(4), 688-708.

Hirshleifer, D., & Shumway, T. (2003). Good Day Sunshine: Stock Returns and the Weather. *The Journal of Finance*, *58*(3), 1009–1032. https://doi.org/10.1111/1540-6261.00556

Hirshleifer, D., Subrahmanyam, A., & Titman, S. (2006). Feedback and the success of irrational investors☆. *Journal of Financial Economics*, *81*(2), 311–338. https://doi.org/10.1016/j.jfineco.2005.05.006

Hoseinzade, Ehsan, and Saman Haratizadeh, 2019, CNNPRED: CNN-based stock market prediction using a diverse set of variables, Expert Systems with Applications 129, 273–285.

Hsieh, S. F., Chan, C. Y., & Wang, M. C. (2020). Retail investor attention and herding behavior. *Journal of Empirical Finance*, *59*, 109-132.

Hu, Guosheng, Yuxin Hu, Kai Yang, Zehao Yu, Flood Sung, Zhihong Zhang, Fei Xie, Jianguo Liu, Neil Robertson, Timpathy Hospedales, and Qiangwei Miemie, 2018, Deep stock representa-

tion learning: From candlestick charts to investment decisions, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2706–2710 (IEEE).

Hvide, H. K. (2002). Pragmatic beliefs and overconfidence. *Journal of Economic Behavior & Organization, 48*(1), 15–28. https://doi.org/10.1016/s0167-2681(01)00221-9

Hvidkjaer, S. (2008). Small Trades and the Cross-Section of Stock Returns. *Review of Financial Studies, 21*(3), 1123–1151. https://doi.org/10.1093/rfs/hhn049

Ioffe, Sergey, and Christian Szegedy, 2015, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in Proceedings of the 32nd International Conference on International Conference on Machine Learning 37, 448–456.

James, K. R. (2000). The price of retail investing in the UK. *Financial Services Authority occasional paper, 6*.

Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2000). Losing Sleep at the Market: The Daylight Saving Anomaly. *American Economic Review, 90*(4), 1005–1011. https://doi.org/10.1257/aer.90.4.1005

Kaniel, R., Liu, S., Saar, G., & Titman, S. (2012). Individual investor trading and return patterns around earnings announcements. *The Journal of Finance, 67*(2), 639-680.

Kaniel, R., Saar, G., & Titman, S. (2008). Individual investor trading and stock returns. The Journal of finance, 63(1), 273-310.

Kannadhasan, M. (2015). Retail investors' financial risk tolerance and their risk-taking behaviour: The role of demographics as differentiating and classifying factors. *IIMB Management Review, 27*(3), 175-184.

Kannadhasan, M. (2015). Retail investors' financial risk tolerance and their risk-taking behaviour: The role of demographics as differentiating and classifying factors. *IIMB Management Review, 27*(3), 175-184.

Kelley, E. K., & Tetlock, P. C. (2013). How wise are crowds? Insights from retail orders and stock returns. *The Journal of Finance, 68*(3), 1229-1265.

Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

Kumar, A., & Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, *61*(5), 2451-2486.

Kumar, M., & Thenmozhi, M. (2006, January). Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*.

Kumar, M., & Thenmozhi, M. (2014). Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting and Finance*, *5*(3), 284. https://doi.org/10.1504/ijbaaf.2014.064307

Kyle, A. S., & Wang, F. A. (1997). Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test? *The Journal of Finance*, *52*(5), 2073–2090. https://doi.org/10.1111/j.1540-6261.1997.tb02751.x

Latif, M., Arshad, S., Fatima, M., & Farooq, S. (2011). Market efficiency, market anomalies, causes, evidences, and some behavioral aspects of market anomalies. *Research journal of finance and accounting*, *2*(9), 1-13.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, Jinho, Raehyun Kim, Yookyung Koh, and Jaewoo Kang, 2019, Global stock market prediction based on stock chart images using deep q-network, IEEE Access 7, 167260–167277.

Lee, Jinho, Raehyun Kim, Yookyung Koh, and Jaewoo Kang, 2019, Global stock market predictionbased on stock chart images using deep q-network, IEEE Access 7, 167260–167277.

Liao, L., Wang, Z., Xiang, J., Yan, H., & Yang, J. (2021). User interface and firsthand experience in retail investing. *The Review of Financial Studies*, *34*(9), 4486-4523.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Liu, C. L., Hsaio, W. H., & Tu, Y. C. (2018). Time series classification with multivariate convolutional neural network. *IEEE Transactions on industrial electronics*, *66*(6), 4788-4797.

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 510-519).

Locke, J., Lowe, A., & Lymer, A. (2013). Interactive data and retail investor decision-making: an experimental study. *Accounting & Finance*, *55*(1), 213–240. https://doi.org/10.1111/acfi.12048

Long, J. B. D., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1991). The Survival of Noise Traders in Financial Markets. *The Journal of Business*, *64*(1), 1. https://doi.org/10.1086/296523

Mak, M. K., & Ip, W. H. (2017). An exploratory study of investment behaviour of investors. *International Journal of Engineering Business Management*, *9*, 1847979017711520.

Malz, A. M. (2021). The GameStop episode: what happened and what does it mean?. *Journal of Applied Corporate Finance*, *33*(4), 87-97.

Michels, J. (2024). Retail investor trade and the pricing of earnings. *Review of Accounting Studies*, 1-36.

Muradoglu, G., & Harvey, N. (2012). Behavioural finance: the role of psychological factors in financial decisions. *Review of behavioural Finance*, *4*(2), 68-80.

Murray, Scott, Houping Xiao, and Yusen Xia, 2021, Charting by machines, Working paper, Georgia State University.

Nayak, S. K., & Ojha, A. C. (2020). Data leakage detection and prevention: Review and research directions. *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, 203-212.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, *6*(3), 144-160.

Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *SSRN Electronic Journal*, *5*. https://doi.org/10.2139/ssrn.94142

Odean, T. (1999). Do Investors Trade Too Much? *American Economic Review*, *89*(5), 1279–1298.

Prasad, S., Kiran, R., & Sharma, R. K. (2021). Influence of financial literacy on retail investors' decisions in relation to return, risk and market analysis. *International Journal of Finance & Economics*, *26*(2), 2548-2559.

Rapoport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual review of psychology*.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, *15*(1), 1829.

Saunders, E. M. (1993). Stock prices and the Wall Street weather. *American Economic Review*, *83*(5), 1337–1345.

Schwartz, R. A. (1970). Efficient capital markets: A review of theory and empirical work: Discussion. *The Journal of Finance*, *25*(2), 421-423.

Shefrin, H., & Statman, M. (1985). The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. *The Journal of Finance*, *40*(3), 777–790. https://doi.org/10.1111/j.1540-6261.1985.tb05002.x

Shiva, A., Narula, S., & Shahi, S. K. (2020). What drives retail investors‟ investment decisions? Evidence from no mobile phone phobia (Nomophobia) and investor fear of missing out (I-FoMo). *Journal of Content, Community and Communication*, *10*(6), 2-20.

Shleifer, A., & Vishny, R. W. (1997). The Limits of Arbitrage. *The Journal of Finance*, *52*(1), 35. https://doi.org/10.2307/2329555

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. https://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 2014, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15, 1929–1958.

Subrahmanyam, A. (2008). Behavioural finance: A review and synthesis. *European Financial Management*, *14*(1), 12-29.

Tuba, E., Bačanin, N., Strumberger, I., & Tuba, M. (2021). Convolutional neural networks hyperparameters tuning. In *Artificial intelligence: theory and applications* (pp. 65-84). Cham: Springer International Publishing.

Vanda Research. (2023). *Home - Vanda*. Www.vandaresearch.com. https://www.vandaresearch.com/

van der Beck, P., & Jaunin, C. (2021). The equity market implications of the retail investment boom. *Swiss Finance Institute Research Paper*, (21-12).

Website (cardify.ai). (July 15, 2021). Market share of retail investment companies in the United States from January 2020 to June 2021 [Graph]. In Statista. Retrieved August 30, 2024, from https://www.statista.com/statistics/1257441/retail-investing-market-share-usa/

Wen, F., Xu, L., Ouyang, G., & Kou, G. (2019). Retail investor attention and stock price crash risk: evidence from China. *International Review of Financial Analysis*, *65*, 101376.

Wood, R., & Zaichkowsky, J. L. (2004). Attitudes and Trading Behavior of Stock Market Investors: A Segmentation Approach. *Journal of Behavioral Finance*, *5*(3), 170–179. https://doi.org/10.1207/s15427579jpfm0503_5

Wullweber, J. (2020). The COVID-19 financial crisis, global financial instabilities and transformations in the financial system. *Global Financial Instabilities and Transformations in the Financial System (September 7, 2020)*.

Yang, Y., Kolesnikova, A., Lessmann, S., Ma, T., Sung, M. C., & Johnson, J. E. (2018). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *arXiv preprint arXiv:1812.06175*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014*, *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, C., Zhong, H., & Hu, A. (2022). A Method for Financial System Analysis of Listed Companies Based on Random Forest and Time Series. *Mobile Information Systems*, *2022*(1), 6159459.

Zhang, J., He, T., Sra, S., & Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications*, *477*, 161-173.