Bank Stock Selection based on LASSO-Large Network Connectedness

Huynh Gia Bao Ngoc

A thesis report submitted in partial fulfilment of the
requirements for the award of the degree of
MASTER OF DATA SCIENCE

CHARLES DARWIN UNIVERSITY
COLLEGE OF ENGINEERING, INFORMATION TECHNOLOGY AND ENVIRONMENT
November 2021

DECLARATION

I hereby declare that the work herein, now submitted as an interim report for the degree of Master of Data Science at Charles Darwin University, is the result of my own investigations, and all references to ideas and work of other researchers have been specifically acknowledged. I hereby certify that the work embodied in this interim report has not already been accepted in substance for any degree, and is not being currently submitted in candidature for any other degree

Signature: Huynh Gia Bao Ngoc

Date: 7th November 2021

ABSTRACT

Keywords: LASSO-VAR ,Machine Learning, Network Connectedness, , Stacking Ensemble, Stock Selection

This study uses the network indicator from LASSO- vector autoregressive model as a feature for stock selection. At the same time, a machine learning approach for stock selection is proposed with the Random Forest, Extreme Gradient Boost, Linear Regression, and Artificial Neural Network stacked model to predict the probability of an outperforming stock. The features set with and without connectedness are feed through the learning model to acquire daily portfolios for the winning stocks. The features set with net connectedness show a greater indicative power at both learning and portfolio construction stage. With connectedness, the performance of the learning model is significantly better during crisis periods, whereas the portfolio performance surpasses the benchmark in terms of cumulative returns. This research is the first attempt to use a Large LASSO connectedness for stock selection and study the impact of connectedness as a feature in a high-dimensional data set.

ACKNOWLEDGEMENTS

I would like to thank both of my supervisors – Professor Fen Nee Chong and Professor Bharanidharan Shanmugam for their guidance and patience throughout the course of my research. You vested me with the motivation and the inspiration to walk forward and do better with your continued supports and encouragement.

I would also like to extend my gratitude to my mom and dad, you have always been the drive of my academic journey. From you, I learnt how to push my boundaries further, and to always work harder than the rest.

And last but not least, I could not thank you enough, my friend – Pham Nhat Nhien. You have been along my side, through and through since the start of this project. Your friendship and your brilliance made this journey the most memorable. I wish you the very best in your future academic endeavors and I hope to see you soon.

TABLE OF CONTENTS

DECLARATION I

ABSTRACT II

ACKNOWLEDGEMENTS III

TABLE OF CONTENTS IV

LIST OF TABLES VI

LIST OF FIGURES VII

LIST OF ABBREVIATIONS VIII

LIST OF SYMBOLS IX

BANKING STOCKS SELECTION BASED ON LASSO-LARGE NETWORK CONNECTEDNESS 1

1. INTRODUCTION: 1

1.1. BACKGROUND: 1
1.1.1. *On Network Connectedness:* *1*
1.1.2. *On Stock Selection in Machine Learning:* *3*
1.1.3. *On sectoral stocks and the banking-sector stocks:* *3*
1.2. AIM OF RESEARCH: 3
1.3. STRUCTURE OF PAPER 3

2. METHODOLOGY: 4

2.1. RESEARCH FRAMEWORK: 4
2.2. DATA DESCRIPTIONS: 5
2.3. NETWORK CONNECTEDNESS: 5
2.3.1. *Volatility Measurement:* *5*
2.3.2. *Network Connectedness:* *6*
2.3.3. *Connectedness in sample observation:* *8*
2.4. STOCK SELECTION WITH MACHINE LEARNING: 9
2.4.1. *Features Set:* *10*
2.4.2. *Train and Test Splitting:* *10*
2.4.3. *Cross Validation and Parameters Tuning:* *10*
2.4.4. *Random Forest:* *11*
2.4.5. *XGBoost:* *11*
2.4.6. *Logistic Regression:* *12*
2.4.7. *Stacking with ANN vs Average Combined Forecast :* *12*

3. RESULTS: 14

3.1. LEARNING PERFORMANCE: 14
3.1.1. *Prediction Overview of Random Forest, Extreme Gradient Boost, Logistic Regression :* *14*
3.1.2. *Cross Validation Results:* *15*
3.1.3. *Testing Results* *15*
3.2. PORTFOLIO PERFORMANCE: 16
3.2.1. *Cumulative Returns* *16*
3.2.2. *Fama French 5's Factors Analysis:* *18*

4. CONCLUSIONS 19

4.1. NET CONNECTEDNESS IN STOCK SELECTION: 19
4.2. ENSEMBLE STACKED ANN AND COMBINED FORECAST STOCK SELECTION: 19

4.3. LIMITATIONS: 19

5. FUTURE WORKS: 21

6. REFERENCES 21

APPENDIX A: LITERATURE REVIEW 26

I. INTRODUCTION: 26

II. METHODOLOGY: 26

2.1. RESEARCH QUESTIONS 27

2.2. SEARCH PROCESS: 27

2.2.1. Keywords: 27

2.2.2. Resources: 28

2.3. INCLUSION AND EXCLUSION CRITERIA: 29

2.4. QUALITY ASSESSMENT: 29

2.5. DATA ANALYSIS: 29

III. RESULTS: 29

3.1. SEARCH RESULTS: 29

IV. DISCUSSIONS 31

4.1. CONNECTEDNESS: 31

4.1.1. Volatility Measurements: 31

4.1.2. Types of connectedness: 32

4.1.3. Connectedness in Banking: 35

4.1.4. Investment Strategy in Connectedness: 36

4.2. STOCK SELECTION IN MACHINE LEARNING: 36

4.2.1. Data and Features: 36

4.2.2. Learning Techniques: 36

4.2.3. Investment strategies: 37

4.3. BANKING STOCK SELECTION BASED ON LASSO-LARGE NETWORK CONNECTEDNESS: 37

V. CONCLUSION: 37

VI. REFERENCES: 38

APPENDIX B: FIGURE 40

APPENDIX C: TABLES 45

APPENDIX D: REFERENCES 53

LIST OF TABLES

Table 1 Research Structure 4
Table 2 Bank stocks by country 5
Table 3 Matrix of Variance Decomposition 7
Table 4 Table 4 Parameters of the three models used for Tuning 11
Table 5 Stacked ANN parameters 13
Table 6 Pooled Prediction Correlation of three models (RF,XGB,LR) 15
Table 7 Learning Model Accuracy Scores (of RF, XGB, LR, and stacked ANN) 15
Table 8 Accuracy Scores of Underperform Periods(According to Features Set 1, Features Set 2 results at the same periods are included for comparison) 17
Table 9 Summary of Cumulative Returns (Note: Diff is the differences of features set 2 over features set 1 on performance) 18
Table 10 Fama Fench's 5 Factors Regression (daily average return is the percentage values, and the values initiated with double asterisks are the t-statistics of the independent variables, without asterisks is the coefficient of the independent variable 20
Table 11 Stocks List 23
Table 16 Fama Fench's 5 Factors Regression (daily average return is the percentage values, and the values initiated with double asterisks are the t-statistics of the independent variables, without asterisks is the coefficient of the independent variables 50
Table 15 Summary of cumulative returns (Diff stands for the differences of features set 2 over features set 1 in terms of cumulative return) 51
Table 1A Keywords Table for Question A *(The table is a list of terms also used for the main search keywords. In this case, connectedness is searched alone and terms in investment strategy is added later on for papers that contains investment strategy with connectedness)* 51
Table 2A Keywords Table for Question B *(The table is a list of terms for stock selection in machine learning)* 51
Table 3A Resources table with acronyms 51
Table 4A Types of Connectedness for Researches Included- Question set A *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)* 52
Table 5A Types of Trading Strategy and Factor Types for Researches Included- Question set B *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)* 52
Table 6A Pairwise directional connectedness in a matrix 52
Table 7A Full connectedness matrix 52

LIST OF FIGURES

Figure 1 Research Framework 4
Figure 2 Net Pairwise Connectedness (This figure depicts the top ten percentile important net pairwise connectedness of 93 stocks during the global financial crisis, the left network is recorded on 12/09/2008 before the big crash of 2008 and the right network is recorded on 15/09/2008 during the crash. The network edge colors depict the direction of connectedness, the blue color represent the negative connectedness index while the red color is the negative connectedness value) 8
Figure 3 JPM- Net Connectedness(from 06/08/2006 to 12/10/2020) 9
Figure 4 BUSE- Net Connectedness(from 06/08/2006 to 12/10/2020) 9
Figure 5 Demonstration of the Rolling-Window Training-Validation Splitting of Three Models (Random Forest, Extreme Gradient Boosting, Logistic Regression)... 10
Figure 6 Demonstration of Walk-Forward Training- validation Set... 10
Figure 7 Demonstration of the Training, Validation and Testing Splitting (For RF,XGB,LR) 13
Figure 8 Top percentile stocks.. 14
Figure 1 Process of Systematic Review *(following Kitchenham(2004)'s method)* 27
Figure 2 Search Process Results- Question set A 30
Figure 3 Search Process Results- Question set B 31

LIST OF ABBREVIATIONS

Abbreviations	Meaning
VAR	Vector- autoregression
LASSO	Least absolute shrinkage and selection operator
ANN	Artificial Neural Networks
RF,XGB,LR	Random Forest, Extreme Gradient Boost, Logistic Regression
GEVD	Generalized Error Variance Decomposition
TVP	Time-Varying Parameter
US	United States
FEVD	Forecast Error Variance Decomposition
COVID	Corona Virus
MKT	Market Risk Factor
SMB	Small Minus Big
HML	High Minus Low
RMW	Robust Minus Weak
CMA	Conservative Minus Aggressive

LIST OF SYMBOLS

Symbol	Description	Section
σ_{GKYZ}	Garman Klass Yang Zhang'sVolatility	Volatility
$\boldsymbol{h}, \boldsymbol{l}, \boldsymbol{o}, \boldsymbol{c}$	Open , High, Low, Close Price	Volatility
x_t,$\boldsymbol{x_{t-i}}$	Value vector at time t and value vector at I lag order	Volatility Connectedness
Φ_i	Coefficient matrix at i order	Volatility Connectedness
ϵ_t	Error white noise vector at time t	Volatility Connectedness
p	Lag order	Volatility Connectedness
H	Forecast horizon	Volatility Connectedness
A_i	Moving average correlation coefficient	Volatility Connectedness
σ_{jj}	Standard deviation of the error vector	Volatility Connectedness
Σ	The matrix of vector error	Volatility Connectedness
e_i	Selection vector with i^{th} element unity	Volatility Connectedness
ω	Error forecast variances	Volatility Connectedness
d_{ij}^H	Pairwise connectedness from j to i	Volatility Connectedness
$\theta_{ij}^g(H)$	Variance decomposition factor	Volatility Connectedness
$C_{i\leftarrow j}(H)$	Pairwise connectedness from j to i	Volatility Connectedness
N	Number of variables for VAR(p)	Volatility Connectedness
$\widehat{C}_t(x, H, M_{t-\omega:t})(\hat{\theta})$	Approximation of the rolling window horizon	Dynamic Connectedness: Rolling Window
$\hat{\beta}$	Predicted Coefficients	LASSO-penalty
N	Interation	LASSO-penalty
$\boldsymbol{y_i}$	Dependent Variables	LASSO-penalty
$\boldsymbol{x_i}$	Independent variables	LASSO-penalty
$\boldsymbol{\lambda}$	Tuning Parameter to reduce Cross Validation Error rate	LASSO-penalty
$\boldsymbol{\alpha}$	Tuning Parameter to minimize RSS and Sum of Square	LASSO-penalty
$\boldsymbol{\beta}$	Coeffiction	LASSO-penalty

$L(t)$	The loss function	XGB
y_i	Value for prediction, with y hat is the predicted values	XGB
h_i	The residual of the previous tree	XGB
f_i	The tree of the model	XGB
g_i	Gradient	XGB

Banking stocks selection based on LASSO-Large Network Connectedness

Abstract—**This study uses the network indicator from LASSO- vector autoregressive model as a feature for stock selection. At the same time, a machine learning approach for stock selection is proposed with the Random Forest, Extreme Gradient Boost, Linear Regression, and Artificial Neural Network stacked model to predict the probability of an outperforming stock. The features set with and without connectedness are feed through the learning model to acquire daily portfolios for the winning stocks. The features set with net connectedness show a greater indicative power at both learning and portfolio construction stage. With connectedness, the performance of the learning model is significantly better during crisis periods, whereas the portfolio performance surpasses the benchmark in terms of cumulative returns. This research is the first attempt to use a Large LASSO connectedness for stock selection and study the impact of connectedness as a feature in a high-dimensional data set.**

***Keywords*—LASSO-VAR ,Machine Learning, Network Connectedness, , Stacking Ensemble, Stock Selection**

1. Introduction:

1.1. Background:

1.1.1. On Network Connectedness:

Methods:

The contemporary financial crises of the 21st century has brought us the revelation of network connectedness. All previous events such as the early 90s recession, the year 2000 dot-com bubble and even the global crisis in 2008 have all shown a similar pattern on the interdependence of the financial market.(Diebold, 2012) With an attempt to quantify the observed interdependence, Diebold and Yilmaz (2009) introduced the concept of return and volatility connectedness to depict this market behavior. Return connectedness reflects the market integration over the years with no bursts and a gentle upward trend. Whereas, volatility connectedness displays no trends but its bursts associate with the event of crises. Volatility connectedness is also often referred to as 'fear connectedness' as it tracks the fear-induced trades of investors on the market. Volatility connectedness is chosen as a measure of network connectedness for this study as they are sensitive to the changes, which in turn can have a significant influence on stock models.

Diebold and Yilmaz's network connectedness is the product of variance decomposition (Pesaran & Shin, 1998) from the vector autoregression model for time series data. The authors follow Koop, Perasan and Potters' method (1996) and the subsequent Shin and Perasan (1998) proposal on the refined version of forecast error variance decomposition- the GEVD (generalized variance decomposition), that was later on used for network connectedness. There are also other methods on quantifying the concept of connectedness besides Diebold and Yilmaz (2009) 's approach. Pearson correlation can be applied to depict the indirect connection between asset entities (Reboredo, 2020) . Additionally, Granger Causality has been proven to show the link between assets, however, this connection concept is exclusively directional. Still, Granger's method complements GEVD with the need of identifying assumptions, which is an inevitable step in variance-decomposition and impulse-response analyses(Diebold&Yilmaz, 2014). Diebold and Yilmaz's connectedness, on the other hand, can illustrate the directional and pairwise directional relationships between assets. Thus, connectedness can be observed from a granular level to a bigger scale – which improve the model reflection on the entire the market (Diebold & Yilmaz, 2012).

The topic coverage of academic researches on connectedness varies greatly. The original researches by Diebold and Yilmaz (2009,2012,2014,2015,2017,2018) have been observed to work on multiple asset types in different markets. For instance, in Diebold and Yilmaz's 2014 work the research used high-frequency intraday data for the connectedness among the financial firms with remarkable results. While in the following year research, the authors studied the dynamic connectedness of different countries for the global business cycle. Extensively, for other researches, the scopes are diverse from connectedness of different markets (Abbas, 2019 ;Abubakr, 2021; Ren, 2021) to different asset types (stocks, bonds market) (Papathanasiou, Vasiliou, Magoutas, & Koutsokostas, 2021; Umar, Manel, Riaz, & Gubareva, 2021; Umar, 2021). Nonetheless, connectedness within industries were also analyzed (Choi, McIver, Ferraro, Xu, & Kang, 2021; He, Liu, & Chen, 2018). Additionally, many extensions of VAR(p)- the vector autoregression model were developed to improve the performance of the original method. (Antonakakis, 2017) proposed time varying parameter(TVP) approaches that can enhance the indicative power dynamic connectedness .Different versions of TVP-VAR were developed thereafter to increase the applicability of the time-varying connectedness (Korobilis, 2018). Demirer' s work is also among the new extensions that upscaled the observation pool to the global bank network with the application of LASSO VAR for dimensionality reduction (Demirer, 2018). In order to have VAR to work in a big dimensional space, LASSO-VAR with different penalty functions were employed also in (Wen, 2020).

This research focuses on net connectedness as a feature to be used for bank stocks selection following Demirer (2018)'s and Tiange (2020)'s approach. As stock selection is a data consuming procedure, the sample population has to be big enough so that the outcome won't be biased. LASSO-VAR is ideal to generate the sufficient data for the stock ranking model.

Network Connectedness in investment strategy:

Network analysis has a pronounced implication on trading strategy and portfolio management. Although (Peralta and Zareei, 2016; ;Harrathi, 2016)'s network indicator is not from the direct family of Diebold and Yilmaz's approach, the research presented a portfolio strategy with improved performance from network analysis. On the other hance, investment strategies built with Diebold and Yilmaz's network connectedness have shown some significances on making trading decisions. Some works rely on network connectedness for options and hedging strategy(Maitra, 2020; Mandacı, 2020; Maitra, 2021) while some others rely on network connectedness for portfolio performance (Lee, 2019; Nasreen, 2021). As an example in options and hedging strategy, (Maitra, 2020) presented a detailed investment strategy that help investors optimizing their portfolio in the stock market and commodity futures. In a similar process, (Mandacı, 2020) also uses connectedness method of (Diebold, Yilmaz) to form periodic portfolios for break periods identified by network connectedness. These papers, although all show deep insights on network connectedness, they do not cover to the extent of the direct use of connectedness for investment strategy.

Additionally, among the other researches that use connectedness to work on portfolio management (Lee, 2019 ;Nasreen, 2021), Tae Kuyn Lee's work(2019) is of particular interest as it generated multiple regional trading portfolios from predictive machine learning algorithm with the direct use of the network indicators. Despite this effective approach to obtain portfolio strategies, the research only experimented on a small sample of 15 indices. In fact, the majority of the above mentioned researches are all restricted to a small data set due to the limited nature of the original VAR model and TVP-VAR.

Network connectedness in banking:

Banks are the backbone of the country. Because most banks are either belong to or heavily supported by the government, they are usually thought to be too big to fail. However, as time comes, a collapse of one bank- often time unexpected can be detrimental for the entire market. Banks are the centerpiece of connectedness, as this financial entity is the intermediary of every financial means. here has been a lot of studies on interconnectedness that focus solely on the banking system.Nier (2007) described the vulnerability of the financial market regarding connectedness's risk contagion of the banking system. (Blundell-Wignall, 2014) discussed on network interconnection as it blend into Basel III for market regulation. There are also a considerable amount of researches on the banking systems with Diebold and Yilmaz's connectedness (Diebold & Yilmaz, 2014) (Demirer,2018). These researches report a higher connectedness among the financial institutions than other sectors. Moreover, domestic banks tend to be more connected than the global bank market. As important the banking connectedness can be, there is little known on the topic with investment strategy as a study object.

1.1.2. On Stock Selection in Machine Learning:

In this paper, investment strategy is directly obtained from a choose-the-best-stock system with the help of machine learning. Nowadays, automation and artificial intelligence assist investors greatly in the decision making process. With the substantial development of machine learning, stock analysis has reached a new height.. As an instance, machine learning can help to define the intangible relationships between variables, and even the most abstract ones like sentimental factors. A wide range of factors can be utilized for the selection process with big sample pool, Rasekhschaffe and Jones (2019)'s work is an exemplary model that remarkably implemented a ranking system for the stocks. The study involves more than 5000 stocks into observation, using 194 factors from multiple categories. While different training sets were implemented into multiple models to produce combined forecasts, the stock selection system can produce more reliable outcomes(Rasekhschaffe & Jones, 2019). It is a considerable improvement from the traditional approach where many parts of the evaluation process have to be executed manually, like the aged old method of projecting return with a fixed interest. Which can increase the susceptibility to bias and hardly capture the highly uncertain movements of the market.

Features to filter the best stock can be fundamental, technical, sentimental or a mixture combined factors depending on the research objectives. Aside from fundamental(Vishwanath & Krishnamurti, 2009) and sentimental features(Tiryaki & Ahlatcioglu, 2005) (like market capitalization, news headline respectively), technical variables (like the moving averages, and especially returns, volatilities) (Yu, Chen, & Zhang, 2014) can generate best positive outcomes in feature selection. This research only uses risk adjusted return (best primary features in stock selection) with supplementary network connectedness for the model as a purpose to experiment the effect of connectedness on investment strategy without much noise interferences . As a matter of fact, some factors can be easily computed, however, many others can't be quantified and mostly intangible. While stock selection has been done widely across markets and industries, it is still better to narrow down the research scope to reduce the variance from the unknown factors when there is a lack of resources.

1.1.3. On sectoral stocks and the banking-sector stocks:

Stock allocation policy is crucial to portfolio performance. Vardharaj & Fabozzi proposed three stock allocation schemes (Vardharaj & Fabozzi, 2007) to experiment differences of these schemes on portfolio performance. The publication found that near to three quarter of the cross-sectional returns can be explained by sector allocation policy(in the US market) . Sectoral allocation is among the famous diversification approaches that is frequently used by fund managers. Aside from the emphasises on the importance of network connectedness, this paper can provide a functional selection system for the latter sectoral allocation process. Moreover, connectedness in this case can be a supportive indicator for bank stocks strategy especially in chaotic times. In the past researches about bank stock behaviors, banks has shown to have great market powers.(Asteriou, Pilbeam, & Sarantidis, 2019) In fact, the bigger the bank the more stable and profitable the bank is, and yet the more destructive the bank can be in cases of failures. Nonetheless, it is observed that they also perform much worse than general stocks during periods of crisis.

1.2. Aim of Research:

The research aims to experiment the indicative power of network connectedness on multiple stock selection learning models. Additionally, three learning models chosen (the Random Forest (RF), the Extreme Gradient Boosting (XGB) and Logistic Regression (LR) are ensemble stacked with ANN and compared with the average combined forecast used in (Rasekhschaffe, 2019) 's and (Wolff, 2020)'s work. Effect of network connectedness on both learning efficacy and portfolio performance are thereafter analyzed in this research. The experiment is also conducted with developed countries' bank to fit with Fama French's global regression model for an in depth assessment of the portfolio performance(Fama & French, 2004).

1.3. Structure of Paper

This research includes six parts with an attached systematic literature review (which provides an overall review of current research as well as descriptive and detailed analysis of the research experiments that support the paper objectives) The structure of this paper itself is an elaboration of the experiment techniques, with results discussion in light of banking stocks selection with and without network connectedness. The content of each part can be found in **Table 1.**

Section	Content
Section 1: Introduction	– Provides a chronological development of stock selection and the connectedness methods. – Addresses the rationale for the selected research methods and research gaps
Section 2: Methodology	– Provides a research framework – Provides information about the dataset – Details on LASSO-VAR as a mean to acquire net connectedness – Details on the stock selection process which includes information of the features sets, data splitting process and also the information of the algorithms involved in the research
Section 3: Results Analysis:	– Illustrations and explanation of Machine Learning Results – Illustrations and explanation of the portfolio performance
Section 4: Conclusion	– Conclusion to the research paper.
Section 5: Future work	– Topic expansion and possibilities for future research.

Table 1 Research Structure

2. METHODOLOGY:

2.1. Research Framework:

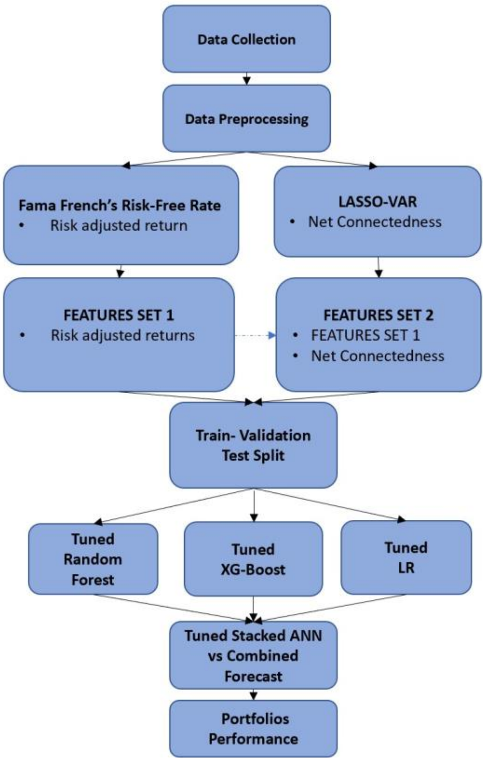


Figure 1 Research Framework

As the research aims to compare the use of the primary feature (historical risk adjusted return) with and without net connectedness (as a network connectedness indicator) for stock selection. This framework (**Figure 1**) proposes a work-flow approach to achieve the research objectives..

For the first steps, open, high, low, close and adjusted close price data of the 93 most capitalized banks in the developed countries cluster (refer to Fama French's developed countries) are retrieved from Yahoo Finance. Main features for this stock selection model can be computed from the five data variables (Open, Low, High, Close, Adjusted Close price) alone. After dates are uniformed, anomalies and missing data are synthesized with interpolation method, we can have the risk adjusted return for features set 1, and adding connectedness for feature set 2.Daily volatility is obtained from the Garman Klass Yang Zhang(Yang & Zhang, 2000)'s formula which is then run through the large LASSO-VAR model for net connectedness as an additional feature for comparison. Another feature- risk adjusted returns, is utilized for labelling while the historical risk adjusted returns are used as feature. Risk adjusted return is the excess return (Fama and French(2004)'s data) by the past one month volatility:

$$R_{adj} = \frac{R - R_f}{\sigma_{GKYZ's\,1M}} \qquad (1)$$

Where by R, R_f are the return and risk-free rate respectively, and $\sigma_{GKYZ's1M}$ is the rolling one month historical volatility computed from the Garman Klass Yang Zhang's volatility.

Risk adjusted return can distinguish the outperforming stocks from the underperformers by setting a median threshold. While the machine learning models can learn to predict the numerical adjusted return, it is better to predict the classification of stock performance to reduce the noisy variances and prevent overfitting (Rasekhschaffe & Jones, 2019). In this case, probability of outperformance is the prediction target to rank stocks for selection.

Once features set are generated, data are split into time rolling windows for a going forward prediction mechanism. Two months test sets are predicted from every 1 year rolling window time frames. Within one frame, training and validation are following a walk forward mechanism whereby two months data are continuously added for the training set after every fold. The train, validation method is applied for all learning algorithm (including ANN) to tune the parameters. Random Forest, XG-Boost, LR uses the optimized parameter to predict the probability of outperformance. Subsequently, the stocks in top 10 percentile of the predicted daily probabilities are chosen for the daily trading long portfolio. As we go on, the combined forecast results of all three algorithms are generated by averaging the probability results. The accuracy of the combined forecast is compared with stacked ANN, along with other three learning models on their accuracy and portfolio performance. Yet, this case is also considered for both features set.

2.2. Data Descriptions:

At the very first step of data collection, the 200 most capitalized bank (in terms of market capitalization) worldwide are chosen. Only the banks (120) in developed countries remain in the sample as a mean to fit with the Fama French's developed market factors(Fama & French, 2004). Eventually, there is 93 banks in the research with available data from 2006 to 2020. While net total directional connectedness following (Diebold & Yilmaz, 2012) is the network indicator to be used in this research framework. The rolling window required to generate this indicator deducts the data availability by two lag days and 200 days deduction of as the size of the window. Monthly volatility and risk adjusted return also move the observation starting date 22 days for monthly volatility and 2 additional days for risk adjusted return. Therefore, a full observation of learning model and portfolio performance is finalized with a learning period from 06/08/2008 to 10/12/2020 (round up to 12 years of data – as every 252 business days is one financial year). The Table 4 and the Figure 2 describe the regional distribution of the stocks selected.

Table 2 Bank stocks by country

Country	Amount
United States	44
Canada	6
Australia	6
United Kingdom	5
France	3
Spain	4
Singapore	3
Netherlands	1
Switzerland	3
Italy	3
Finland	1
Japan	6
Hong Kong	3
Germany	2
Sweden	1
Austria	1
Denmark	1
Total	**93**

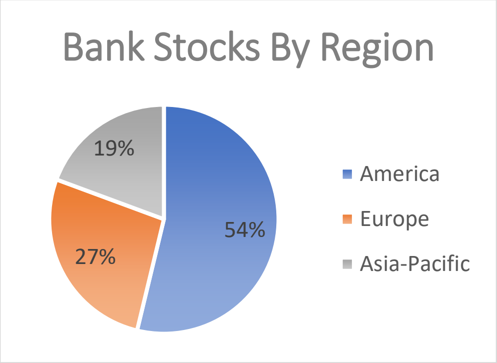


Figure 2 Bank stocks by Region

2.3. Network Connectedness:

Connectedness is the central of modern financial risk measurement and management. It features prominently in key aspects of important risk factors like market risk, credit risk, gridlock risk, systemic risk and business cycle risks,.. etc. Demirer, Diebold, Liu, & Yilmaz, 2018) (2018)'s Connectedness depict the link among financial entities that form a network . While this network has shown some evidences of network effects on investment strategies(Demirer, Diebold, Liu, & Yilmaz, 2018),stock selection with connectedness poses big dimensionality challenge that can render the attempt ineffective. Fortunately, LASSO-VAR can solve the large sample problem to meet the data requirement of stock selection.

2.3.1. Volatility Measurement:

Volatility is the essential factor for the generation of volatility connectedness. In the approaches to retrieve connectedness, volatility can be calculated in multiple ways. Volatility for connectedness has been observed to be from different formulas and computed on many different observation periods. Intraday dat a has been used with log volatility for connectedness to capture the slightest movement of stocks (Diebold & Yılmaz, 2014). With the introduction of high-volume intraday trading, realized volatility has become handy in solving intraday volatility problems. In many cases, high frequency data is often unavailable and very hard to obtain, range-based volatilities can be used as the alternatives for realized volatility. On the regards of range-based techniques, Parkinson's method was used in (Diebold & Yilmaz, 2012) for daily volatility in the research of the directional movement of connectedness. Nonetheless, volatility of Garman Klass has been seen in many connectedness publications(Garman & Klass, 1980). This approach is held on high regards for multiple financial topics, as they are almost as efficient as realized volatility (Diebold, Liu, & Yilmaz, 2017). Albeit the technique is also based on ultra high-frequency sampling, but Garman Klass's approach easier to construct and very robust to microstructure noise. As there is no strict guidance on the measurement of volatility, Garman Klass Yang Zhang's extension volatility was chosen to not only cover the high-frequency data, but also to capture the overnight jumps.(Yang & Zhang, 2000). Beside the fact that GKYZ's volatility is easy to construct like the original Garman Klass's, it is considered to be the top two most efficient volatility measurement (efficiency score=8).

GKYZ upgrades the original Garman Klass by adding the overnight factor- the close price at $t-1$. The range base is now not only intraday but also inter-day at the same time.

$$\widetilde{\sigma_{GKYZ}}^2 = \left(ln\frac{o_i}{c_{i-1}}\right)^2 + \frac{1}{2}\left(ln\frac{h_i}{l_i}\right)^2 - (2ln2-1)\left(ln\frac{c_i}{o_i}\right)^2 \tag{2}$$

Where h_i, l_i, o_i, c_i are the high, low, open , close price of stock at time i respectively. c_{i-1} is the closing price of the previous day.

2.3.2. Network Connectedness:

From Vector Autoregression to Generalized FEVD:

In the year of 1980, the concept of vector autoregressions was firstly introduced by Christopher Sims(Sims, 1980). VAR(p) is a statistical stochastic process model that depicts the relationship between multiples variables as they change over time. p is the lag order; the lagged values are included in the functions along with the error terms. For the parametric set up in this case, the lag order of two $(p = 2)$ and forecast horizon of 10 days $H = 10$ will be employed to obtain the connectedness. As a matter of fact, the forecast horizon of 10 days is also in coherent with Basel requirement of 10-day value at risk (VaR) (Diebold). In order to retrieve network connectedness from VAR(p), variance decomposition needs to be performed for the generalized variance decomposition (Diebold & Yilmaz, 2012). The generalized concept of error variance decomposition was firstly introduced by KPSS (Perasan & Shin, 1998). Granted that the GEVD is invariant to ordering which allows the directional relationship between variables. Moreover, shocks in their model are not orthogonalized, therefore, correlated shocks are possible with the historical observed distribution of the errors. Based on the aforementioned characteristics of GEVD, Diebold and Yilmaz (2012) introduced the asymmetrical and bi-directional relationships of connectedness.

The vector auto regression of the stock prices in p lag order can be described as below:

$$x_t = \sum_{i=1}^{p} \Phi_i \,.\, x_{t-i} + \varepsilon_t \tag{3}$$

Where $\varepsilon_t \sim (0, \Sigma)$ is the error white noise vector, Φ_i are the coefficient matrices, p is the lag order. x_t in this case is the vector of volatilities of N financial institution's stock indices. And x_{t-i} is the vector of past volatility values in p lag order.

Vector autoregression can be transformed into the moving average representation :

$$x_t = \sum_{\mathrm{i}=0}^{\infty} \mathrm{A_i} \varepsilon_{\mathrm{t-i}} \tag{4}$$

$$A_i = \Phi_1 A_{i-1} + \Phi_2 A_{i-2} + \cdots + \Phi_p A_{i-p} \tag{5}$$

Where A_i are the moving average correlation coefficient. With A_0 is NxN identity matrix, and $A_i = 0$ for $i = 0$.

From the moving average representation, variance decomposition can be performed to generate the GEVD. By that, H- step error forecast variances:

$$\omega = \frac{\sum_{h=0}^{H-1}\left(e_i' A_h \mathrm{e_i}\right)^2}{\sum_{h=0}^{H-1}\left(e_i' A_h \Sigma A_h' e_i\right)} \qquad (6)$$

$$And \quad MSE_j(h) = \sum_{h=0}^{H-1}\left(e_i' A_h \Sigma A_h' e_i\right) \qquad (7)$$

Based on the foundation of Koop's research (Koop, Pesaran, & Potter, 1996) , Shin and Perasan introduced the generalized version of the forecast error variance decomposition(Pesaran & Shin, 1998):

$$\theta_{ij}^{g}(H) = \frac{\sigma_{jj}^{-1}\sum_{h=0}^{H-1}\left(e_i' A_h \Sigma\, e_j\right)^2}{\sum_{h=0}^{H-1}\left(e_i' A_h \Sigma A_h' e_i\right)} \qquad (8)$$

Where σ_{jj} is the standard deviation of error vector ε_j, and e_i is the selection vector with i^{th} element unity and zero at the rest. The generalized variance decomposition then become the centerpiece of connected measures for Diebold and Yilmaz's works.

The matrix of variance decomposition is as below:

Table 3 Matrix of Variance Decomposition

	$\boldsymbol{x_1}$	..	$\boldsymbol{x_N}$
$\boldsymbol{x_1}$	d_{11}^{H}	..	d_{1N}^{H}
..	..	..	..
$\boldsymbol{x_N}$	d_{Nj}^{H}	..	d_{NN}^{H}

$$C_{i\leftarrow j}(H) = d_{ij}^{H} = \widetilde{\theta_{\iota J}}(H) = \frac{\theta_{ij}^{g}(H)}{\sum_{j=1}^{N}\theta_{ij}^{g}(H)} \qquad (9)$$

With that $\sum_{j=1}^{N}\widetilde{\theta_{\iota J}^{g}}(H) = 1$ and $\sum_{j,i=1}^{N}\widetilde{\theta_{\iota J}^{g}}(H) = N$. This depicts the pairwise directional relationship between $i\ and\ j$. To be exact, it is the pairwise directional connectedness of the shock from $j\ to\ i$ at the forecast horizon H.

According to Diebold and Yilmaz (2012) , the concept of connectedness can cover from the granular to the macro scale relationships. In another word, connectedness can be measured among firms, industries, and in a larger scope, among countries. The pairwise directional relationship can show the firm-level relationship between companies and on a larger scope, the total directional connectedness can be used to see the relationship of one stock with many others.

Total directional connectedness:

The sum of all elements in one row of the NxN matrix in case of $i \neq j$ is the H-step forecast error variance decomposition of the stock with that row. Therefore, the forecast error variance calculated is the shocks from all other variable elements to that stock.(Diebold & Yilmaz, 2012) Which shows the total directional connectedness 'FROM' others to i:

$$From\ others: C_{i\leftarrow\cdot}^{H} = \sum_{j=1}^{N} d_{ij}^{H} \quad with\ i \neq j \qquad (10)$$

And total directional connectedness 'TO' others:

$$To\ \ others: C_{\cdot\leftarrow j}^{H} = \sum_{i=1}^{N} d_{ij}^{H}\, with \quad i \neq j \qquad (11)$$

Similarly, to pairwise directional connectedness, the net total directional connectedness can be as:

$$Net: C_i^{H} = C_{\cdot\leftarrow i}^{H} - C_{i\leftarrow\cdot}^{H} \qquad (12)$$

The sum row and the sum column difference of the stock is its total directional connectedness.

In this research, net total directional connectedness (also called as net connectedness) is used as an additional feature for stock selection. The indicator consolidates the information of the pairwise and the directional relationships. Since it represents the relationship of one stocks with the others, net connectedness conceptually fits with the stock selection model whereby stock performances (in compared to the market) are predicted.

Dynamic Connectedness:

Dynamic connectedness shows the change of connectedness across the observation period, the rolling window method is implemented to calculate the change of connectedness over time. The window span of 200 days will be applied to calculate the connectedness to observe the change of net connectedness. With the rolling frame of 200 days, $\omega = 200$, we can formulate the rolling window method with the approximation method:

$$\widehat{C_t}\left(x, H, M_{t-\omega:t}\right)\left(\hat{\theta}\right) \qquad (13)$$

With x as the stock vector, H is the forecast horizon, $M_{t-\omega:t}\left(\hat{\theta}\right)$ is the approximation model of connectedness, of which t is the observation period, ω is the window span (Diebold & Yilmaz, 2012).

LASSO and Elastic Net:

The parameter space of VAR grows quadratically with the amount of variables, which can prompt the model to be quickly exhausted with only a few input variables(Demirer et al., 2018; Wen & Wang, 2020). In light of that, LASSO-VAR(the least absolute shrinkage and selection operator) is introduced as an extension of VAR with a special feature -LASSO . LASSO is used to shrink, select and estimate the high-dimensional network of the stocks to be used for the latter selection process. The LASSO elastic net penalty (through a penalized maximum likelihood) is applied for the VAR model (Yuan, Ho, & Lin, 2012), the method can work with large data problems and work extremely well with sparse feature. The elastic net solves the following problem:

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{i=1}^{N}\omega_i l(y_i,\beta_0+\beta^T x_i)+\lambda[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1] \quad (14)$$

Where elastic net penalty $\alpha = 1$ for lasso regression, λ is the grid value can be used for tuning.

2.3.3. Connectedness in sample observation:

Connectedness data retrieved from LASSO-VAR depicts great prominence during the 2008 global financial crisis. The net pairwise connectedness of the 93 stocks sample show the changes of connectedness before and after the main crash of 2008. On the 15th of September 2008, the global crisis hit the climax when Lehman Brothers filed for bankruptcy. Connectedness on this day increased considerably in compare to 3 days earlier before the fateful event. (**Figure 2**)

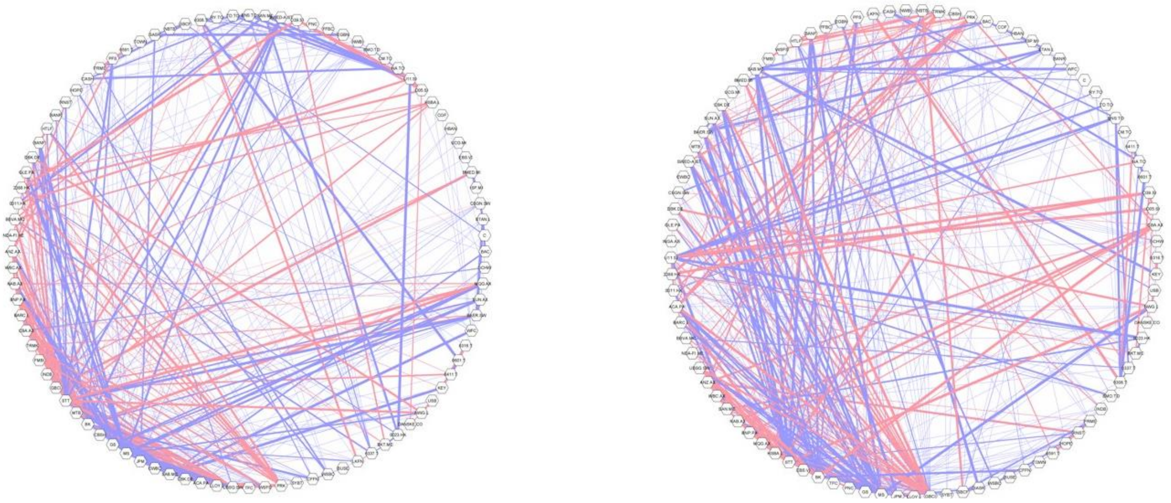


Figure 2 Net Pairwise Connectedness (This figure depicts the top ten percentile important net pairwise connectedness of 93 stocks during the global financial crisis, the left network is recorded on 12/09/2008 before the big crash of 2008 and the right network is recorded on 15/09/2008 during the crash. The network edge colors depict the direction of connectedness, the blue color represent the negative connectedness index while the red color is the negative connectedness value)

The dynamic net connectedness of individual banks implies periods of shocks differently. The smaller banks tend to receive connectedness, while the bigger banks mostly likely to transmit connectedness to others. As the most capitalized bank in the sample, J P Morgan (JPM) mostly show negative connectedness value through out the observation period (**Figure 3**). Which implied that connectedness gives more than to receives connectedness most of the time. The connectedness indices that this bank can send to others are in a great amount. Whereas, the least capitalized bank-like BUSE often receives connectedness but rather in a small amount. Both of these bank show significant change during chaotic periods. The deepest trough of JPM overlap with the 2008 market crisis, whilst connectedness reached to almost -140 of the connectedness value. The second dive is during the European debt crisis in 2011-2012, however, it is not as intense as the global crisis. We can also spot the other troughs on 2014-2015 which happened at the same time with the oil crisis and the Chinese market crash. The brief and volatile crypto crash period is also shown on JPM's graph, although the impact of the most recent crisis – the COVID pandemic is not so significant. On the other hand, BUSE (**Figure 4**) shows a reverse trend on connectedness, however most of its peak still match with the previously discussed crisis events. Significantly, the COVID pandemic crash matches with the deepest and longest trough of the stock. With such reflective characteristic of connectedness on global events, the indicator can be a good feature to assist with the trading decision.

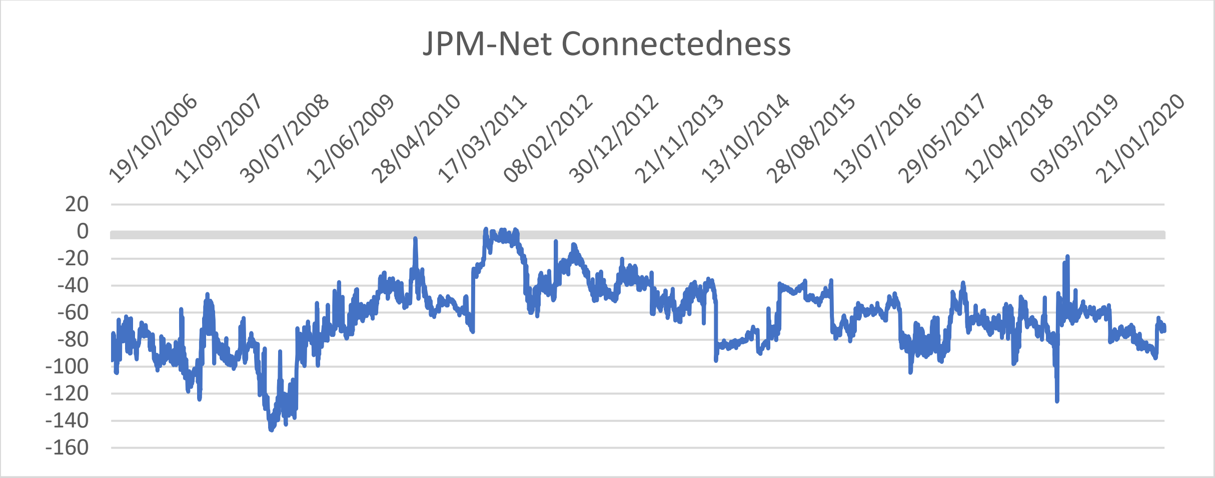


Figure 3 JPM- Net Connectedness(from 06/08/2006 to 12/10/2020)

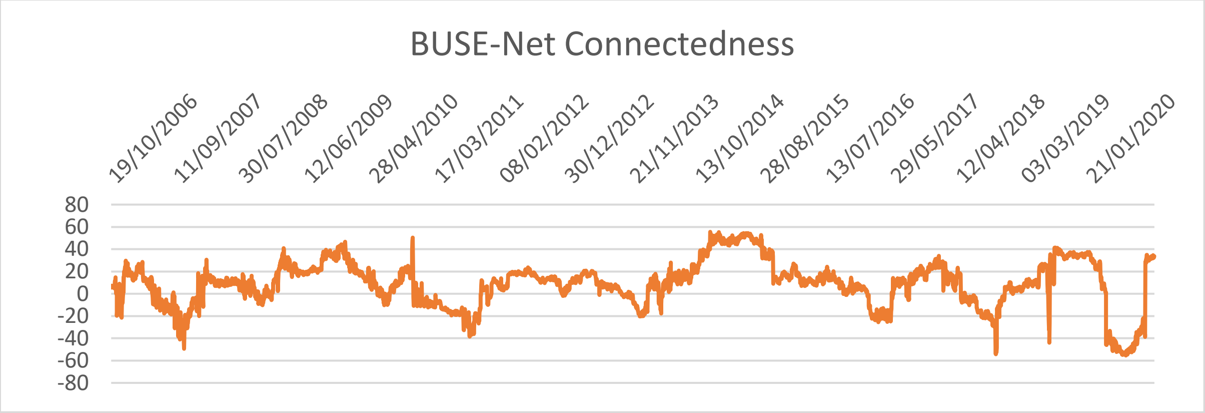


Figure 4 BUSE- Net Connectedness(from 06/08/2006 to 12/10/2020)

2.4. Stock Selection with Machine Learning:

Machine learning is a thriving field, ever since its early days in the late 50s, the field has evolved in an astounding pace. While the amount of learning techniques are increasing substantially. It is hard to put one's finger on the best learning methods, as there are always flaws and merits for each and every approach. Nevertheless, one can still combine forecasts from different classes of algorithm to optimize the results (Rasekhschaffe & Jones, 2019; Wolff & Echterling, 2020). On that account, a stacked model with ANN is used to join the three learning classes chosen for this stock selection model(Random Forest, XG Boost, LR). Two of the models- Random Forest and XG Boost already belong to the ensemble learning line. Ensemble is a learning technique that combines predictions from two or more sub-models. Aside from stacking, bagging and boosting are also among the popular methods of ensemble technique to obtain more prominent outcomes. A bagging method, for instance- the Random Forest, randomly creates tree stumps for parallel learning and aggregate the results through the majority votes. Boosting (XG Boost), on the other hand, trains the weak learners sequentially. From which, results are adjusted by the weights after every model. The end-of-sequence results is the combined results that converts weak learners to a better performing model. Random Forest is good at large dimensional data, works extremely well with missing data problems, albeit struggles with the variance-bias trade off. XGB is good at the variance-bias problem, but it is hyper sensitive to outliers and is not so good at sparse data . Simple linear model like Logistic Regression, although is not an ensemble method, works well on features with linear dependencies, nonetheless, is generally better in capturing linear trend. A compilation of all three models may assist to compliment for each others' weaknesses and simultaneously enhance their strength.

2.4.1. *Features Set:*

This stock selection model uses multiple machine learning methods and combined learning methods to extract rankings from the two given features set (Figure 1) . In both features set, stocks are labelled as out and underperformers depending on the daily median risk adjusted return. The features are trained to predict the stock performance of the next 7 days. For the first features set, daily volatility is calculated with GKYZ's volatility (average of a rolling window of 1 month observation). Risk adjusted return is the excess return from risk-free rate over the previously computed daily volatility (Fama French's 5 factor)(Fama & French, 2004). The latter set uses net connectedness as an additional feature for testing. All features set are standardized before fitting.

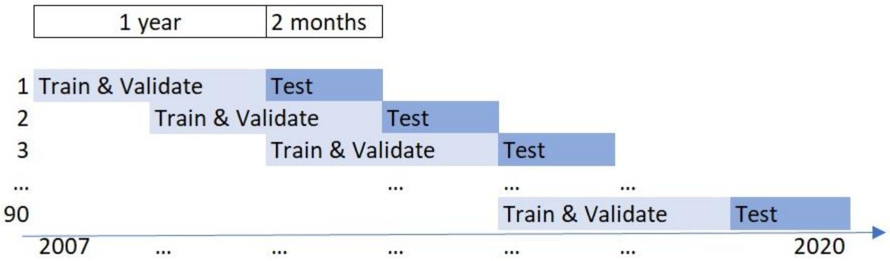


Figure 5 Demonstration of the Rolling-Window Training-Validation Splitting of Three Models (Random Forest, Extreme Gradient Boosting, Logistic Regression)

Subsequently, time series data sets of the features are sent through the learning algorithms to forecast the daily probability of out-performance of each stock. Historical data has been proven to be very effective in forecasting future classification((Lee, Cho, Kwon, & Sohn, 2019)). The data set is also cross-sectional, by which all stocks are included and trained simultaneously. The rationale of this cross-sectional structure choice is referred to the fact that stock is classified to be better or worse than the median market performance. Factors that determine the market performance of one stock should not only be time series, but also cross sectional based on their inter-related relationships.

2.4.2. *Train and Test Splitting:*

Time series data, especially classification data requires much different solution than the typical classification problem. From the very beginning stage, data should be preprocessed and augmented in a way that the analogous structure is remained intact. Ergo, there has been many research works on the structure preserving data for time series (Ausiello, D'Atri, & Protasi, 1980)

Same goes for train test splitting, an intricate problem needs an intricate solution, a time series classification problem needs to be treated as a time series data no less. Therefore, a time rolling learning mechanism is employed for this research data. Training and validation sets are split into 1 year time frames, with a 2 month walk-forward horizon. The 1 year data is trained and cross-validated to rule out the best parameter for the 2 month test set. After each cycle the models are reset and fit to the new training and validation data.

The data spread through out 2007 to the end of 2020, which in turn produces 90 data cycles. As this research is performed on an experimental stance, all data was previously retrieved for the testing purposes. In the real world case scenario, there will need a one year historical data for forecast.

2.4.3. *Cross Validation and Parameters Tuning:*

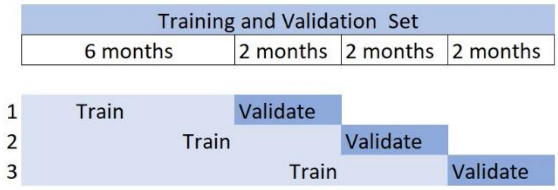


Figure 6 Demonstration of Walk-Forward Training-validation Set

Most learning algorithms need parameter tuning to obtain the best outcomes. Cross validation always goes hand in hand with tuning to ensure that data won't overfit. Same like data augmentation, cross validation needs special treatment for time series data Albeit many researches on the same topic employed a randomly selected stratified sample for the k-fold cross validation with the same data structure. Cross validation for time series should be much more delicate as not to damage the time series structure. For this research, walk-forward cross-validation for every training and validation set is implemented. Data are trained 2 months cumulatively after every fold to predict probabilities on the next two months. Eventually, the average accuracy score of cross validation will help to choose the optimal learning parameter. Parameters are randomly generated to tune the models, there is 6 iterations from these parameter ranges.

Table 4 Table 4 Parameters of the three models used for Tuning

	Params
Random Forest	{'bootstrap':[True], 'n_estimators':[200,800,1000], 'max_depth':[5,10,20],'min_samples_split':[2,4,6],'min_samples_leaf':[1,4,8]
XGBoost	{'learning_rate':[0.01,0.05,0.1,0.3],'max_depth': [3,5,10],'gamma': [0,0.1,0.2,1],'reg_alpha' : [1,20,40],'reg_lambda' : [0,1],'colsample_bytree' : [0.5,0.9],'min_child_weight' : [1,2,6], 'n_estimators': [20,100,180],'seed': [27],'eval_metric':['mlogloss']}
Logistic Regression	{'C':[0.01,1,100],'solver':['newton-cg', 'lbfgs'],'penalty':['l2']}

2.4.4. Random Forest:

Random Forest is the procedure of growing an ensemble of decision trees. The trees of which learn parallelly and have the right to vote to find the popular- also the final classification. Random forest is famous for its unique technique of creating random trees through bootstrap aggregation. Bootstrapping is the method of randomly selecting the features or a features combination at each nodes to grow a tree. Through this technique and from the original data, multiple samples for the trees are created to generate classification. The results from the trees are eventually compiled through aggregation with a voting system. The whole process compiled is the bootstrap aggregation process, or also called as bagging. With the assistance of bagging, Random Forest is much more flexible than decision tree. While the model's bias is significantly reduced in compared to the original decision tree, the model is more prone to over fitting with a high variance.

Similar to decision tree, the design of each trees require a choice of an attribute selection measure. In a random forest classification problem, Gini impurity index is usually used at each leaf nodes to choose the more relevant node for splitting similarly to decision tree. The measure is applied directly to the leaf nodes to find the best path for the final label prediction. To better the model performance, we can tune the number of trees and the design of each trees (for example: the max depth of one tree, minimum sample splits, minimum leaf,.. etc) to optimize prediction.

Particularly, in this work, a range of 200 to 1000 trees with each tree's max depth range from 5 to 20; minimum splits from 2 to 6 and minimum sample leaf from 1 to 8.

2.4.5. XGBoost:

Gradient boosting is a machine learning technique that can be classified as an ensemble method. It also bases on the decision trees models to sort out errors and correct those errors from previous models. This technique employs differentiable loss function and gradient descent optimization algorithm to minimize the loss of gradients. Once values of the loss function are found, they are multiplied with the learning rate and added up continuously until the model fits. The procedure stops when the improvements of residuals subsides or until the maximum specified. The values are then predicted by adding all of the previous scales to produce the end result. The empirical evidence shown from a research on gradient boosting indicates that small steps with the right direction will give better prediction.(Chen et al., 2015) The traditional gradient boosting methods iteratively find out the weak learners and combine them cumulatively to a strong learner. Tianqi Chen in (2016) introduced the Extreme Gradient Boosting algorithm as an improved version of Gradient Boosting. Its performance has outdone the traditional gradient boosting for its scalability and effectiveness(Chen et al., 2015). It can 'automatically apply multi-threaded parallelism for accelerating the execution time' according to (Ding, Nguyen, Bui, Zhou, & Moayedi, 2020). XGB can handle the variance and bias trade-off problem with an appropriate parameter tuning. However, the model does not perform well on sparse and unstructured data, it is also hardly scalable because of its sequential learning characteristic.

The XGB uses the second-order Taylor expansion to formulate the loss function, and the regulation terms like tree depth, weights of leaf nodes are also included in the objective function. The loss function of the model is:

$$L(t) \approx \sum_{n=1}^{n} \left(L(y_i, \widehat{y_{\iota-1}}) + g_i f_i(x_i) + 1/2 h_i f_i^2(x_i) \right) \quad (15)$$

Where: $g_i = f'(t) = \frac{\partial L(y_i, \widehat{y^{t-1}})}{\partial \widehat{(y)^{t-1}}}$ and $h_i = f''(t) = \frac{\partial L(y_i, \widehat{y^{t-1}})}{\partial \widehat{(y)^{t-1}}}$

In order, to find the probability of winning stocks, mlogloss is set as a fix evaluation metric for this experiment. For parameter tuning, tree design are modified within the general range. Learning rate is set from 0.01 to 0.03, regulation parameter like gamma, alpha and delta are tuned within regular range (0 to 1, 1 to 40 and 0 to 1 for each parameter respectively). To elaborate, gamma indicates the condition to split the leaf node using loss function, alpha is the L1 that regulate the weight (it is analogous to LASSO regression which can deal with high dimensionality data), lambda is the L2 – or Ridge Regression is used to reduce overfit.

2.4.6. *Logistic Regression:*

Logistic Regression is a learning model that is extremely useful in cases of solving dichotomous problems(Wright, 1995). The construction of the model is simple, yet effective. Logistic regression is simple in a manner that in its simplest form, it is a logistic function to model binary dependent variables. From the given variables, we assume a linear relationship between the features and the log odds of the event. The variables go through a logistic regression to find the probability of that event. The final product is the probability which rest the case of the classification label. This model learns through a maximum-likelihood function- (or the co-efficient of the model), whereas, the coefficients are continuously adjusted until the best probability is predicted. Simple as it might be, the model is extremely effective in predicting linear relationships. The learning method is also scalable to multinomial regression, and suitable with probability prediction dataset.

Most models require probability calibration to extract the results- which is often non-versatile and insensitive to changes. Therefore, predictive queries that rely on probability would sometimes receive inaccurate outputs. For the reason that probability is the nature core of this learning model, it is a go-to model for stock selection. Moreover, logistic regression can cover the linear relationships among variables without much risk on overfitting. It is indeed potent to be a complementary factor for the stacked approach. The model is optimized by adjusting the C parameter (which is the inverse of regularization strength) with different solvers for optimization (lbfgs and newton-cg). Both lbfgs and newton-cg are following Hessian function, while lbfgs approximates the value, newton-cg compute Hessian to the steps.

2.4.7. *Stacking with ANN vs Average Combined Forecast :*

An ANN is a machine learning technique that synthesize the neural network based upon the biological function of animal brains.(Yegnanarayana, 2009) ANN is constructed with an input layer, some hidden layers and an output layer which contain multiple neurons (nodes). Each nodes are assigned with an activation and a transfer function. In one node, the weight(s) are added to a transfer function to compile the node's input weight. Subsequently, the weight calculated is transferred through the activation function for the output result. At the end of every network, the output is calculated and compared to the true value for errors. Back propagation will use those errors to optimize the predictive output. This procedure is an iterative cycle that send outputs back to the start with the node weights adjusted to bring the error scores to a minimum.

Stacking with ANN has been reporting great results from the previous papers. In this worrk, following the non-parametric learners (SVM) and all the ensemble techniques like bagging, boosting , ANN is used to integrate the results of the three previously tuned models through stacking. The portfolio returns of the ANN stacked model are compared with the combined forecast (the average probability) for a better view on the model performance.

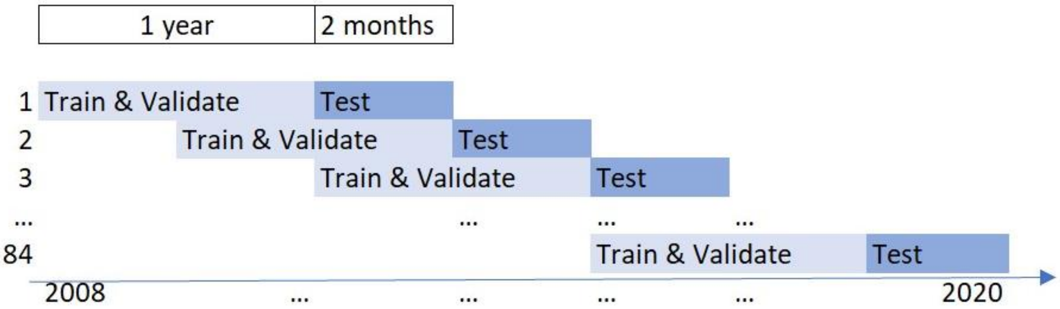


Figure 7 Demonstration of the Training, Validation and Testing Splitting (For RF,XGB,LR)

Features set:

The predictive y hat probability values of the three models are sent through the ANN layers for train test and validation.

Train, validation and test splitting:

The same structure of train, validation, test splitting is employed for the stacking algorithm. Therefore, training data for the stacking model starts from 2008 instead of 2007. There is 84 cycles for the data splitting to forecast every two months forward until 2020. It is noteworthy that probability data are all initially z-score standardized before fitting.

ANN Parameters:

During the tuning period, the ANN model continue to improve by adding more epochs up to peak at epochs 100.For a time-effective approach, epochs for the train and validation section is 30, and 100 epochs is used for testing section. Therefore, ANN in this case tends to perform better than the initial cross-validation results. The model is tuned with the number of layers and type of activation. Relu and sigmoid are the activation function considered for the data set. While ReLU is faster to compute than sigmoid and is able to handle the vanishing gradient problem very well. ReLU tends to blow up activation and to be susceptible the dying ReLU phenomenon. The network architecture is designed as a funnel shape as neurons of each layers are subsequently cut-back to the last layer with only one neuron. Adam optimizer is used for better accuracy, binary cross entropy and accuracy score are the loss parameter and accuracy measurement respectively.

Stock Selection and Portfolio Performance:

Probability of banks' outperformance (of 4 models) in each day is percentile ranked to retrieve the top 10 percentile performers. The stocks belong in this daily top percentiles are gathered to form

Table 5 Stacked ANN parameters

	Params
Stacked ANN	['layers':[32],'activation':'sigmoid']
	['layers':[25],'activation':'relu']
	['layers':[50,25],'activation':'relu']
	['layers':[15],'activation':'sigmoid']

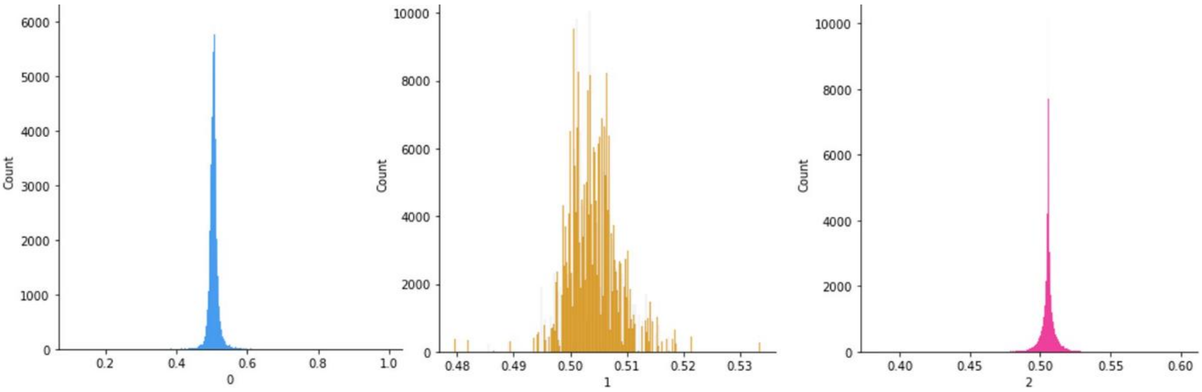
daily long portfolios. The portfolio return is calculated with the risk-weighted return of the chosen stock. This daily returns of the portfolio strategy is compared with the daily equal-weighted returns of all banks involved in the stock selection process as a benchmark. The return performance of the portfolio is subsequently fit with the Fama French's 5 factor model to analyze the performance of this trading strategy.

Figure 8 Top percentile stocks

3. RESULTS:

This part presents the outcomes obtained from following the previously discussed research framework. In order to observe the indicative power of net connectedness, the quality of the learning models and also of the portfolios need to be compared for features set 1 and 2. Similarly, learning performance and portfolio performance needs to be retrieved to analyze the differences between combined forecast and stacked ANN. Therefore, it is plausible to dissect this reading part into two sections, the first one for learning performance and the latter for portfolio performance. This first part focuses largely the classification power of the learning models, when the efficacy of the predicted probability reveals itself in the second section on portfolio performance (as the portfolios were formed on the predicted probabilities).

Figure 9 Distribution of the Random Forest, Extreme Gradient Boost and Logistical Regression (accordingly from left to right)



3.1. Learning Performance:

As we go along the steps of predicting outperform probability of a stock, cross validation results need to be first obtained to find the best parameter. From that, the testing results is used to analyze the significance of the models and the features set. The cross-sectional dataset used to fit the learning models is highly balanced, as it includes all stocks in one day, which is covering the entire 1 year data of the training and validation test.

3.1.1. Prediction Overview of Random Forest, Extreme Gradient Boost, Logistic Regression :

Three models Distribution:

The pooled prediction of the first three models show bell curve shapes in the normal distribution plots.**(Figure 9)** For the fact that the models are trying to predict probability of the binary cases and given the dataset is very balance on the label distribution, the models have the potential to accurately assign the right probabilities to the stocks. Standard deviation of the Random Forest, Extreme Gradient Boosting and Logistic Regression varies, which implies different effects of the learning models. Moreover, the center median is leaning more to the boundary of greater than 0.5, the solid 0-1 label classification results will show signs of overfitting. For the sake of analyzing the pure learning performance of the features sets and the stacked model, no extra regularization method is applied to avoid any biased favors to any particular test cases. As we move on to the final stage for portfolio strategy, this overfitting phenomenon is no longer the concern as stocks in every test cases need to be percentile ranked to the median predicted probabilities.

Pooled Prediction Correlation

The correlation **Table 6** is created from pooling whole predictions from the learning models. The correlations can closely explain the traits of these three models. Because Random Forest and Extreme Gradient Boost are both tree learners in the core, higher correlation between these two are expected. A linear based model like Logistic Regression, hypothetically, is less likely to relate with non-linear models like Random Forest and XGB. Similarly, as observed, correlations of Logistic Regression with other two are quite low. This observation supports the implementation of stacked ANN, similarly to (Wolff & Echterling, 2020), and align with the arguments on the theoretical discussion on part 2.4's opening.

Table 6 Pooled Prediction Correlation of three models (RF,XGB,LR)

	RF	XGB	LR
RF	1	0.4898685	0.3353745
XGB	0.4898685	1	0.42378125
LR	0.3353745	0.42378125	1

3.1.2. *Cross Validation Results:*

The time series cross sectional data, once cross validated with a walk-forward splitting structure, can find its best parameters through for the three learners and the stacked model. The parameter with the best accuracy score is chosen for the testing set. **Table 7** depicts the results of cross validation of the aforementioned three learning models and the additional stacked ANN. Overall, the initial results are close to the observations from Wolfe(2020)'s paper. While they employed the same data structure for fitting, the accuracy scores from the models only range from 50% to 51%, given that a wide range of factors were utilized for the research. Yet, from that, convincing results on the portfolios were documented. Moreover, similar accuracy range were also reported in Lee (2020)'s research, the work focuses on global investment strategy based on network indicators. The authors implemented the indicators(Diebold & Yilmaz's total connectedness and Pearson Correlation indirect connectedness) for a simple features set to predict each index movement individually. The accuracy from Lee's(2020) recorded a higher performance range from 50-60%. Although this method can yield great performance on a small sample, it is computationally expensive for a larger data set regarding both time and space complexities. Especially, it is not suitable for this particular problem, where stocks are predicted to be better or worse than the median market. Predicting stock performance to the median market individually can aggregate to some class imbalance cases, as some stocks are consistently under or out-perform for different time periods. In light of the evidences in previously discussed researches, the efficacy metrics of this study show a reasonable range (50% to 51%) in the performance**.** Considering the effectiveness of features set 1 (without connectedness) and features set 2 (with connectedness) alone, the set with connectedness is modestly better on the account of accuracy. Additionally, for the reason that precision and recall are more balanced, features set 2 have less tendency for the false positives. Which also leads to a better accuracy. The stacked ANN model can enhance the accuracy of the three learners (from 0.1% to 0.3%). Although the discrepancies of ANN between two features set are not very significant.

Table 7 Learning Model Accuracy Scores (of RF, XGB, LR, and stacked ANN)

Models	Without Connectedness				With Connectedness			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
RF	50.202%	50.531%	74.276%	59.627%	50.230%	50.585%	69.703%	58.253%
XGB	50.240%	50.470%	81.383%	61.054%	50.388%	50.588%	81.174%	61.724%
LR	50.388%	50.536%	88.818%	64.161%	50.430%	50.603%	82.957%	62.598%
ANN Stacked	50.555%	50.602%	92.345%	64.880%	50.515%	50.603%	90.854%	64.777%

3.1.3. *Testing Results*

Testing results through out time periods:

Predictions from the model include 90 two-months data for the three learners and 84 two - months data for the stacked ANN set. The data output structure is the result of a rolling window testing mechanism. Subsquently, the accuracy scores of every two months are documented and searched for anomalies. Through the average accuracy scores of the three learners, it is observed that features set 1 produce many underperforming periods . By mapping the important world economy events throughout the prediction periods, it is significant that the crisis periods match with the failing periods of features set 1. The dataset with connectedness, on the other hand, shows a random pattern of underperform periods. There is only 1 case that dataset 2 overlap with a crisis event during 2009 .

Table 8 shows all testing period with the lowest average accuracies (according to features set 1) of Random Forest, Extreme Gradient Boost and Logistic Regression of both data set. Additionally, Stacked ANN accuracy score is included to compare the learning ability. Since the global crisis in 2008, there are many abrupt crash events that took place afterward till now. Of the 93 most capitalized banks observed there are 54% of the banks from North America, 27% from Europe, and 19% from Asia-Pacific (**Figure 2**). Therefore, it is logically rational to focus on the major events from these regions at first, but not to exclude the crisis factors from the other parts of the world. For instance, although China is not among the countries included in the research experiment, Chinese market crash 2015-2016 match with one of the worse performing data. Rationally speaking, crash phenomenon occurs in this country can transfer ample effects to the market worldwide. Due to the fact that China is deeply connected with the other big countries economically (through trading, supply chain) and politically (through regulation and diplomatic policies). By 2014, China was the second most important trading partner with EU, it was also a major player in the Asia-Pacific and North America markets. **Table 8** shows that the dataset with connectedness works better than the data without it. For most of the observations, the features set 2 tends to perform better than its usual performance (the average accuracies through out the period) and surpasses set 1 with a large gap. Outstandingly, during the global crisis 2008, the oil bear market in 2014 following with the Chinese market crash and the recent COVID-19 pandemic, while accuracy scores of set 1 dive to the bottoms, features set 2 thrives (differences at 2.23%, 1.5%,2.77%, 2.17% respectively to the events). One thing worth mentioning that, those events are the worst crises in the past . There is one outlier in the observation , ie the US-China trade-war, however, the magnitude of differences is not substantial (-1.29%). Last but not least, the ANN stacked model remains stable through out time , even during the market crashes. The ANN stacked model for both features sets do not depicts any major differences regarding accuracy, this ensemble method seems to be the most reliable and effective approach according to the results data.

3.2. Portfolio Performance:

This part presents the portfolio performance formed from the probabilities acquired from the daily stock data. The predicted top daily 10 percentile performing stocks are obtained to form the daily portfolio. Moreover, return performance of the combined forecast with the benchmark model are also included. Lastly, cumulative returns are reported for all test cases with Fama French’s 5 factor analysis.

3.2.1. Cumulative Returns

According to **Figure 9** and **Figure 10**, daily cumulative returns of all the three learners and its combined forecast do not meet the benchmark of the sample. However, returns from `the ANN stacked model of both features set is almost twice in values, in terms of cumulative returns (whole sample). The recession period in 2008-2009 and its residual effect from to 2009 to the end of 2012 (namely the European debt crisis, the US bear market in **Table 8**) seems to have a big effect on all portfolios as there are periods that the cumulative returns reaches negative values. During this period, the stacked ANN model with features set 2 experienced some more volatile episodes than set 1. However, from 2013 to 2015, set 2 surged near to one fifth of set 1’s cumulative return (as of April 2015). The Chinese market crash(2015-2016) takes a toll on features set 1 market performance, by the end of 2016, half of the cumulative returns from 2008 were lost. Although set 2 was also experiencing a sharp drop till mid 2015, and quickly revived onward. Unfortunately, the market crash in 2020 due to coronavirus that started from 20 Feb 2020 to April 2020 incurred great loss to big features set, the results from feature set 2 endured more effects from this pandemic alone.

Table 8 Accuracy Scores of Underperform Periods(According to Features Set 1, Features Set 2 results at the same periods are included for comparison)

Testing Period	Event	Start Date	End Date	Features set 1					Features Set 2					DIff
				RF1	XG1	LR1	AVR1	ANN1	RF2	XG2	LR2	AVR2	ANN2	AVR2-AVR1
2007-2008	**Global Crisis**	10/04/2008	05/08/2008	49.28%	49.13%	49.00%	49.14%	N/A	50.54%	51.77%	51.79%	51.37%	N/A	2.23%
				50.51%	47.67%	50.10%	49.43%	N/A	49.92%	49.18%	49.77%	49.63%	N/A	0.20%
2008-2009	**US-Bear Market**	01/03/2009	01/04/2009	50.10%	49.69%	49.41%	49.74%	50.41%	51.66%	50.54%	51.00%	51.07%	49.59%	1.33%
2009-2010	**Europe debt crisis (1)**	13/10/2009	30/11/2009	49.77%	50.13%	49.44%	49.78%	50.69%	50.59%	50.03%	50.61%	50.41%	48.82%	0.63%
	US-Flash Crash + (1)	11/03/2010	04/05/2010	49.74%	49.67%	50.56%	49.99%	49.10%	49.92%	51.05%	50.64%	50.54%	50.92%	0.55%
2010-2011	**-1**	05/05/2010	01/07/2010	49.08%	49.39%	50.61%	49.69%	50.95%	49.28%	50.61%	50.61%	50.17%	50.59%	0.48%
2011-2012	**Black Monday**	14/09/2011	04/11/2011	49.92%	50.00%	49.67%	49.86%	50.36%	49.26%	50.56%	50.51%	50.11%	50.26%	0.25%
		26/12/2011	13/02/2012	49.18%	49.74%	50.08%	49.67%	49.28%	49.54%	50.03%	50.21%	49.92%	50.56%	0.26%
2012-2013(YB)		14/02/2012	03/04/2012	49.87%	49.26%	50.21%	49.78%	50.15%	49.23%	50.46%	51.00%	50.23%	50.69%	0.45%
2013-2014	**Bear Oil Market**	26/06/2014	22/08/2014	48.41%	48.62%	50.61%	49.22%	50.23%	48.82%	50.54%	50.51%	49.96%	50.56%	0.74%
2014-2015		19/10/2014	05/12/2014	49.13%	49.51%	49.64%	49.43%	50.97%	50.49%	51.74%	50.56%	50.93%	50.61%	1.50%
2015-2016	**Chinese Market Crash**	04/09/2015	03/02/2016	48.80%	46.98%	47.06%	47.61%	50.51%	50.28%	50.56%	50.31%	50.38%	50.64%	2.77%
				49.36%	50.69%	49.49%	49.85%	50.69%	50.26%	50.64%	50.82%	50.57%	50.69%	0.73%
				48.69%	50.64%	50.64%	49.99%	50.67%	49.95%	50.64%	50.44%	50.34%	50.67%	0.35%
2016-2017	**US-Market Sells Off**	07/11/2016	26/12/2016	49.64%	49.03%	49.54%	49.40%	50.79%	51.51%	50.31%	52.07%	51.30%	50.56%	1.89%
		06/04/2017	02/06/2017	49.41%	49.28%	50.61%	49.77%	50.31%	49.16%	49.23%	49.69%	49.36%	50.46%	-0.41%
2017-2018	**US-China Trade-war**	08/01/2018	18/04/2018	49.36%	50.54%	49.21%	49.70%	50.33%	49.92%	49.97%	48.67%	49.52%	50.44%	-0.18%
				51.20%	50.54%	50.54%	50.76%	50.54%	48.54%	49.31%	50.56%	49.47%	50.51%	-1.29%
2018-2019	**Crypto-Crash**	18/06/2018	10/10/2018	49.64%	48.54%	50.54%	49.57%	50.28%	49.87%	49.95%	50.28%	50.03%	50.54%	0.46%
				49.05%	48.52%	50.05%	49.21%	50.49%	50.59%	50.67%	50.69%	50.65%	50.56%	1.44%
		29/11/2018	16/01/2019	48.72%	48.59%	50.54%	49.28%	50.51%	50.31%	50.54%	50.00%	50.28%	50.54%	1.00%
2020	**COVID**	17/03/2020	11/05/2020	48.13%	47.47%	47.80%	47.80%	50.54%	50.15%	49.77%	49.97%	49.97%	50.54%	2.17%
Whole Sample														
Average							50.31%	50.37%				50.35%	50.44%	0.04%
Min							47.61%	48.77%				48.69%	48.82%	-2.29%
Max							52.44%	50.97%				51.49%	50.99%	2.77%

Note: The table depicts the all underperforming periods of the three learners (RF,XGB,LR) using features set1 by the average accuracy score of the three learners of every testing periods. The score is compared against the features set 2 in the same period. While stacked ANN is not included in the average score as it is already an ensemble method itself.

At the end of the observation period, the dataset uses net connectedness is still 11.97% more profitable than the dataset without connectedness(**Table 9**). The three learning model are overall quite indifferent for both features set, except for the Random Forest with a 14.3% profit gap in favor of features set 1.

3.2.2. Fama French 5's Factors Analysis:

A factor regression analysis is implemented for to detect if the portfolio is attributable to common factors. Moreover, we can find the long term effect of the factor exposures to the machine learning derived portfolios.In this case, Fama and French (2014) 's 5 factors model is used for the whole

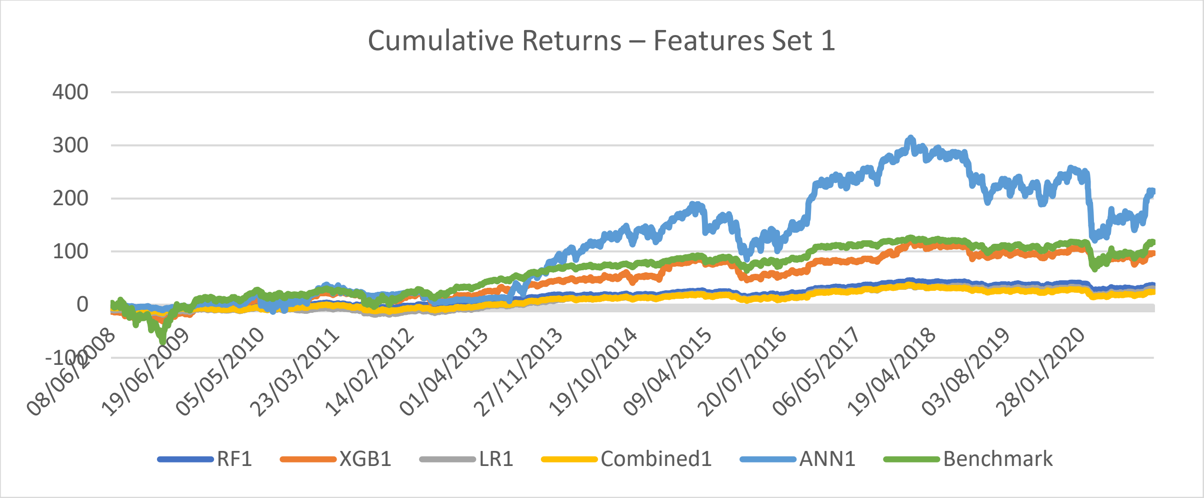


Figure 9 Cumulative Returns of Features Set 1

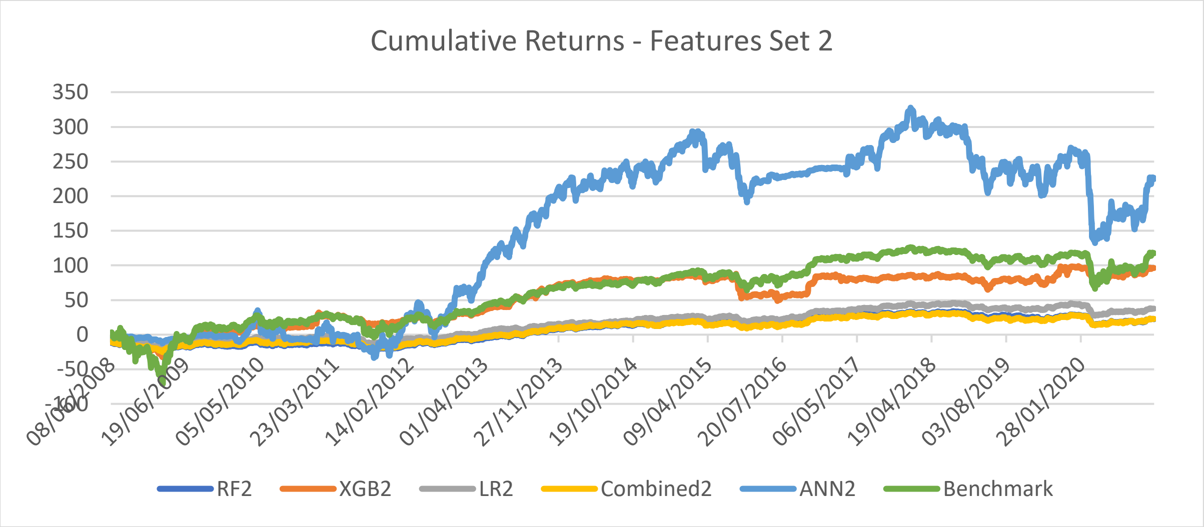


Figure 10 Cumulative Returns of Features Set 2

Table 9 Summary of Cumulative Returns (Note: Diff is the differences of features set 2 over features set 1 on performance)

	RF	XGB	LR	Combined	ANN	Benchmark
Features Set1	36.73974	96.86174	30.45821	23.9263	212.7951	117.5522
Features Set2	22.40561	96.08957	37.44842	22.56447	224.7732	117.5522
Diff	-14.3341	-0.77217	6.990208	-1.36182	11.97815	

sample analysis. The factors consist of the market risk factor (MKT), small minus big factor(SMB), high minus low factor (HML), robust minus weak factor(RMW), and conservative minus aggressive factor (CMA). **Table 10** shows the summary of the market factor in regression with the daily returns yield from the machine learning models. Most factors are already preset by the choice of sample at the beginning of the research. Stereotypically speaking, most bank with high market capitalization have high book to market value and low operating profitability. (Explain) Therefore, a long portfolio selected from 93 large capital banks can be presumed to have a high coefficient with market risk factor(MKT), positive coefficient with the value factor (HML) and negative with the size factor(SMB) and profitability factor(RMW). However, for the last factor- the investment factor, coefficients may varies. Although most big banks manage their investment conservatively, some emerging banks are quite robust with their investment approach.

The adjusted R-squares depict low fitting points for the five factor model, while datasets with ANN stacked produce the lowest Adjusted R-Square. The Alpha of all models with both features set show positive alpha values, therefore all models perform better than the regression model predicted. For both features set, the table also shows that the strategy from the stacked ANN model is significantly correlated to the market risk factor, while the other strategies are lower but still comparatively significant. It is reasonable in a sense that these are the long strategies that have not gone through proper portfolio diversification. Moreover, the research performs on a population concentrated in banking stocks, a high market risk factor describes finely the sample chosen. The size factor (SMB) also describes the chosen sample with negative values obtained from every models and features sets. As a matter of fact, the initial selection of banks for the experiment contains only large market capital banks. From the table, we can see the positive beta for the value factor(the HML), all investment strategies from the models tend to expose to value stocks. In the long run, the portfolios with high book to market value stocks can be more profitable than the low-book to market value stocks. The results from the profitability factor (RMW) shows that the models focus on weak operating profitability firms. Notably, the ANN stack model has the most negative significance. However, the investment factor (CMA) depicts different coefficients of the models. While the two datasets of ANN stacked model describe a great positive tendency toward CMA factor and the other models also remain positively correlated, random forest with net connectedness poses a negative position with CMA. All in all, the stacked ANN model in both cases with and without connectedness perform better than the bench mark in the long run, following the Fama French theory (Fama & French, 2004). As a matter of fact, they depict a significant positive coefficient for most factors and resolve the negative CMA coefficient issue in the benchmark, except for the profitability and the size factor which are ingrained in the whole sample characteristic. Lastly, from **Table 10** Fama French's factor remains comparative between the features set.

4. CONCLUSIONS

4.1. Net connectedness in stock selection:

This research shows that net connectedness is a good indicator to predict performance of stocks. The accuracy of the features set with connectedness considerably enhanced during crises and market crashes. Without this network indicator, the efficacy wane in the cases of adversity. This outcome align with the documented results of Lee(2020)'s work on predicting indices movement.

In the cases of portfolio performance, net connectedness with stacked ANN can yield better profit. Although there are small improvements during adverse events, the features set with net connectedness still suffers from profit loss. In the long run, net connectedness investment strategy is comparable with the set without connectedness when considering Fama French's regression analysis.

4.2. Ensemble Stacked ANN and Combined Forecast stock selection:

The learning results imply that stacked ANN is stable and reliable through crises for both features set. The portfolio performance of stacked ANN does not only outperform benchmark but also surpasses the combined forecast on the prospect of portfolio performance. Especially, the combination of features set 2 and stacked ANN can yield better profit. Moreover, the factor analysis for this model in terms of profitability shows positive results better than benchmark in the long term.

4.3. Limitations:

A more diverse sample pool could not obtained due to the limitation of the current VAR model and

Table 10 Fama Fench's 5 Factors Regression (daily average return is the percentage values, and the values initiated with double asterisks are the t-statistics of the independent variables, without asterisks is the coefficient of the independent variable

					Without Connectedness					With Connectedness	
	RF1	XG1	LR1	Combined1	ANN1	RF2	XG2	LR2	Combined2	ANN2	Benchmark
Daily avg. ret	0.0118	0.03115	0.01149	0.0063	0.0605	0.0096	0.0301	0.01	0.013	0.0638	0.0333
R squared	0.257	0.202	0.24	0.253	0.172	0.268	0.212	0.269	0.267	0.186	0.399
Alpha	0.0106	0.0301	0.0109	0.0074	0.0435	0.008	0.0304	0.0117	0.0069	0.0457	0.0385
	**1.719	**1.807	**1.741	**1.203	**0.951	**1.352	**2.139	**1.96	**1.135	**0.987	**0.02
MKT	0.1364	0.3181	0.1283	0.135	0.8287	0.1337	0.2407	0.1365	0.1358	0.95	0.5277
	**17.233	**14.917	**16.05	**17.075	**14.127	**17.545	**13.19	**17.835	**17.438	**16.018	**0.026
SMB	-0.0358	-0.062	-0.0371	-0.0357	-0.2527	-0.0348	-0.1401	-0.0182	-0.0404	-0.1875	-0.2774
	**0.017	**-1.319	**-2.108	**-2.048	**-1.955	**-2.071	**-3.485	**-1.077	**-2.356	**-1.435	**0.057
HML	0.0892	0.1557	0.088	0.0816	0.3581	0.0763	0.2632	0.0693	0.0754	0.1265	0.6994
	**4.577	**2.966	**4.473	**4.191	**2.479	**4.066	**5.859	**3.679	**3.93	**0.866	**0.063
RMW	-0.2461	-0.6531	-0.2535	-0.2539	-1.3294	-0.2723	-0.4977	-0.2917	-0.2764	-1.4272	-0.8691
	**-8.25	**-8.123	**-8.411	**-8.514	**-6.011	**-9.478	**-7.234	**-10.113	**-9.413	**-6.383	**0.097
CMA	0.0325	0.1367	0.0228	0.0443	0.9742	0.0211	-0.0805	0.021	0.0188	1.2298	-0.5236
	**1.081	**1.689	**0.753	**1.476	**4.376	**0.73	**-1.162	**0.723	**0.635	**5.465	**0.097
Obs	2952	2952	2952	2952	2952	2952	2952	2952	2952	2952	2952

the time constraint of the research project. Therefore, a more realistic stock model could not be created with a wider range of features and learning methods. Unfortunately, for parameters optimization, hardware limitation on the time and space complexity prevent the research to go distances on a full-fledge random search tuning.

5. FUTURE WORKS:

Different methods of calculating volatility can be used to enhance the accuracy of prediction. As a matter of fact, that realized volatility from Diebold's method calculated from the 5-mins intervals intra day data, or the log difference of the open high low closing price for Yang Zhang Satchell‘s approach is proven to be more precise in reflecting the market fluctuation. For the future endeavors, the research will find possibilities to expand the scope to predict volatility for volatility spillovers and find the most effective machine learning method to predict volatility. Moreover, the research scope can expand to different asset types and more complex, yet well-built portfolio model.

6. REFERENCES

An, P., Li, H., Zhou, J., Li, Y., Sun, B., Guo, S., & Qi, Y. (2020). Volatility spillover of energy stocks in different periods and clusters based on structural break recognition and network method. *Energy, 191*, 116585.

Asteriou, D., Pilbeam, K., & Sarantidis, A. (2019). The Behaviour of Banking Stocks During the Financial Crisis and Recessions. Evidence from Changes-in-Changes Panel Data Estimations. *Scottish Journal of Political Economy, 66*(1), 154-179.

Audrino, F., & Tetereva, A. (2019). Sentiment spillover effects for US and European companies. *Journal of Banking & Finance, 106*, 542-567.

Ausiello, G., D'Atri, A., & Protasi, M. (1980). Structure preserving reductions among convex optimization problems. *Journal of Computer and System Sciences, 21*(1), 136-153.

Ball, S., Banerjee, A., Berry, C., Boyle, J. R., Bray, B., Bradlow, W., . . . Denniston, A. (2020). Monitoring indirect impact of COVID-19 pandemic on services for cardiovascular diseases in the UK. *Heart, 106*(24), 1890-1897.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4): Springer.

Brauneis, A., & Mestel, R. (2018). Price discovery of cryptocurrencies: Bitcoin and beyond. *Economics Letters, 165*, 58-61.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1*(4).

Coffman, J., & Weaver, A. C. (2012). An empirical performance evaluation of relational keyword search techniques. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 30-42.

Demirer, M., Diebold, F. X., Liu, L., & Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics, 33*(1), 1-15.

Diebold, F. X., Liu, L., & Yilmaz, K. (2017). *Commodity connectedness*. Retrieved from

Diebold, F. X., & Yilmaz, K. (2011). Equity market spillovers in the Americas. *Financial stability, monetary policy, and central banking, 15*, 199-214.

Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting, 28*(1), 57-66.

Diebold, F. X., & Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics, 182*(1), 119-134.

Ding, Z., Nguyen, H., Bui, X.-N., Zhou, J., & Moayedi, H. (2020). Computational intelligence model for estimating intensity of blast-induced ground vibration in a mine based on imperialist competitive and extreme gradient boosting algorithms. *Natural Resources Research, 29*(2), 751-769.

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of economic perspectives, 18*(3), 25-46.

Gabauer, D., & Gupta, R. (2018). On the transmission mechanism of country-specific and international economic uncertainty spillovers: Evidence from a TVP-VAR connectedness decomposition approach. *Economics Letters, 171*, 63-71.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of business*, 67-78.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University, 33*(2004), 1-26.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15.

Koop, G., Pesaran, M. H., & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics, 74*(1), 119-147.

Lee, T. K., Cho, J. H., Kwon, D. S., & Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Systems with Applications, 117*, 228-242.

Lundgren, A. I., Milicevic, A., Uddin, G. S., & Kang, S. H. (2018). Connectedness network and dependence structure mechanism in green investments. *Energy Economics, 72*, 145-153.

Maitra, D., Guhathakurta, K., & Kang, S. H. (2021). The good, the bad and the ugly relation between oil and commodities: An analysis of asymmetric volatility connectedness and portfolio implications. *Energy Economics, 94*, 105061.

Mensi, W., Maitra, D., Vo, X. V., & Kang, S. H. (2021). Asymmetric volatility connectedness among main international stock markets: A high frequency analysis. *Borsa Istanbul Review, 21*(3), 291-306.

Nasreen, S., Tiwari, A. K., & Yoon, S.-M. (2021). Dynamic Connectedness and Portfolio Diversification during the Coronavirus Disease 2019 Pandemic: Evidence from the Cryptocurrency Market.

Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters, 58*(1), 17-29.

Pieters, G., & Vivanco, S. (2017). Financial regulations and price inconsistencies across Bitcoin markets. *Information Economics and Policy, 39*, 1-14.

Qarni, M. O., Gulzar, S., Fatima, S. T., Khan, M. J., & Shafi, K. (2019). Inter-markets volatility spillover in US bitcoin and financial markets. *Journal of Business Economics and Management, 20*(4), 694-714.

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal, 75*(3), 70-88.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, 1-48.

Tiryaki, F., & Ahlatcioglu, M. (2005). Fuzzy stock selection using a new fuzzy ranking and weighting algorithm. *Applied Mathematics and computation, 170*(1), 144-157.

Vardharaj, R., & Fabozzi, F. J. (2007). Sector, style, region: Explaining stock allocation performance. *Financial Analysts Journal, 63*(3), 59-70.

Vishwanath, R., & Krishnamurti, C. (2009). *Investment management: A modern guide to security analysis and stock selection*: Springer.

Wen, T., & Wang, G.-J. (2020). Volatility connectedness in global foreign exchange markets. *Journal of Multinational Financial Management, 54*, 100617.

Wolff, D., & Echterling, F. (2020). Stock Picking with Machine Learning. *Available at SSRN 3607845*.

Wright, R. E. (1995). Logistic regression.

Yang, D., & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business, 73*(3), 477-492.

Yegnanarayana, B. (2009). *Artificial neural networks*: PHI Learning Pvt. Ltd.

Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. *Procedia computer science, 31*, 406-412.

Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). An improved glmnet for l1-regularized logistic regression. *The Journal of Machine Learning Research, 13*(1), 1999-2030.

Table 11 Stocks List

Ticker	Name	Market Capitalization	Country
JPM	JPMorgan Chase	488,893,000,000	United States
BAC	Bank of America	357,213,000,000	United States
WFC	Wells Fargo	190,866,000,000	United States
MS	Morgan Stanley	179,427,000,000	United States
C	Citigroup	143,406,000,000	United States
RY.TO	Royal Bank Of Canada	142,454,000,000	Canada
SCHW	Charles Schwab	139,712,000,000	United States
CBA.AX	Commonwealth Bank	133,368,000,000	Australia
GS	Goldman Sachs	127,797,000,000	United States
TD.TO	Toronto Dominion Bank	121,390,000,000	Canada
HSBA.L	HSBC	108,405,000,000	United Kingdom
MQG.AX	Macquarie	96,852,001,500	Australia
USB	U.S. Bancorp	89,121,071,104	United States
PNC	PNC Financial Services	83,608,870,912	United States
BNP.PA	BNP Paribas	81,450,549,248	France
TFC	Truist Financial	79,449,079,808	United States
BNS.TO	Scotiabank	75,419,328,512	Canada
COF	Capital One	73,555,279,872	United States
WBC.AX	Westpac Banking	69,674,164,224	Australia
NAB.AX	National Australia Bank	66,084,303,655	Australia
BMO.TO	Bank of Montreal	65,546,137,600	Canada
SAN.MC	Santander	63,512,920,064	Spain
D05.SI	DBS	58,454,736,896	Singapore
ANZ.AX	ANZ Bank	57,771,939,417	Australia
INGA.AS	ING	57,179,029,504	Netherlands
UBSG.SW	UBS	55,818,297,344	Switzerland
ISP.MI	Intesa Sanpaolo	55,673,331,712	Italy
NDA-FI.HE	Nordea Bank	52,719,791,735	Finland
CM.TO	CIBC	50,701,774,848	Canada
8316.T	Sumitomo Mitsui Financial Group	48,241,459,200	Japan
BK	Bank of New York Mellon	45,351,161,856	United States
LLOY.L	Lloyds Banking Group	45,009,866,752	United Kingdom
BBVA.MC	Banco Bilbao Vizcaya Argentaria	44,675,006,464	Spain
BARC.L	Barclays	43,581,313,024	United Kingdom
ACA.PA	Crédit Agricole	43,253,817,344	France
O39.SI	OCBC Bank	38,047,156,397	Singapore
8411.T	Mizuho Financial Group	36,191,977,472	Japan
NWG.L	NatWest Group	35,183,005,696	United Kingdom
0011.HK	Hang Seng Bank	32,798,741,762	Hong Kong
2388.HK	Bank of China (Hong Kong)	31,904,874,331	Hong Kong
U11.SI	UOB	31,867,148,598	Singapore

STT	State Street Corporation	31,294,013,440	United States
UCG.MI	UniCredit	29,819,087,234	Italy
GLE.PA	Société Générale Société	26,931,077,315	France
DBK.DE	Deutsche Bank	26,322,493,440	Germany
NA.TO	National Bank of Canada	25,726,072,592	Canada
CSGN.SW	Credit Suisse	24,012,253,184	Switzerland
HBAN	Huntington Bancshares	22,960,508,928	United States
8591.T	ORIX	22,850,541,568	Japan
SWED-A.ST	Swedbank	22,668,140,544	Sweden
KEY	KeyCorp	20,915,812,352	United States
MTB	M&T Bank	19,725,049,856	United States
EBS.VI	Erste Group Bank	17,886,908,225	Austria
DANSKE.CO	Danske Bank	14,454,364,083	Denmark
STAN.L	Standard Chartered	14,272,246,784	United Kingdom
BAER.SW	Julius BÃ¤r	14,237,941,396	Switzerland
SUN.AX	Suncorp	11,621,976,817	Australia
EWBC	East West Bancorp	11,125,363,712	United States
8308.T	Resona Holdings	9,703,317,873	Japan
8601.T	Daiwa Securities Group	8,886,962,180	Japan
CBK.DE	Commerzbank	8,430,365,755	Germany
CBSH	Commerce Bancshares	8,230,009,856	United States
BMED.MI	Banca Mediolanum	8,009,868,033	Italy
GBCI	Glacier Bancorp	5,355,588,608	United States
BKT.MC	Bankinter	5,301,172,596	Spain
8337.T	Chiba Bank	4,973,784,064	Japan
SAB.MC	Banco Sabadell	4,710,816,798	Spain
0023.HK	Bank of East Asia	4,707,853,001	Hong Kong
INDB	Independent Bank	2,557,478,912	United States
WSFS	WSFS Financial	2,454,242,304	United States
FRME	First Merchants Corporation	2,285,708,288	United States
FFBC	First Financial Bank	2,266,230,528	United States
TOWN	TowneBank	2,262,498,304	United States
WSBC	WesBanco	2,240,403,712	United States
FMBI	First Midwest Bancorp	2,187,111,424	United States
SASR	Sandy Spring Bank	2,183,318,016	United States
HTLF	Heartland Financial USA	2,053,779,584	United States

RNST	**Renasant Corp**	2,051,216,512	United States
TRMK	**Trustmark**	2,045,401,216	United States
PRK	**Park National Corp**	2,043,557,760	United States
BANF	**BancFirst**	1,990,674,944	United States
SBCF	**Seacoast Banking**	1,972,694,528	United States
BANR	**Banner Bank**	1,917,446,528	United States
EGBN	**Eagle Bancorp**	1,862,989,184	United States
PFS	**Provident Financial Services**	1,858,709,120	United States
HOPE	**Hope Bancorp**	1,797,997,696	United States
LKFN	**Lakeland Financial Corp**	1,789,223,040	United States
CASH	**Meta Financial Group**	1,711,383,808	United States
NWBI	**Northwest Bank**	1,708,546,944	United States
CFFN	**Capitol Federal Savings Bank**	1,615,321,856	United States
NBTB	**NBT Bancorp**	1,576,121,728	United States
SYBT	**Stock Yards Bancorp**	1,566,275,456	United States
BUSE	**First Busey**	1,408,945,792	United States

APPENDIX A: LITERATURE REVIEW

Banking stocks selection based on LASSO-Large Network Connectedness A Literature Review

Abstract –
Background: Although there is a great significance of connectedness on portfolio implication, there has not been many works that directly use connectedness for investment strategy, especially on a bigger scale.
Aim: With an intention to study the indicative power of connectedness on a large data sample, a bank stock selection approach using data from the LASSO-Large Network connectedness is implemented in this research paper. This systematic review aims to collect and analyze the past related researches on the topic.
Method: The standard systematic review method was employed with exhaustive search of all related research papers from 2008.
Results: Through the systematic review, we can confirm the importance of connectedness in the banking industry. Moreover, the review also validated the necessity of using LASSO- VAR and its combination with the stock selection model for investment strategy from a large data sample.
Conclusions: Although there are many approaches toward connectedness, LASSO-VAR is the most appropriate model for the research, regarding the size of data sample. On the same regard, stock selection with machine learning is versatile and effective for the research. The review successfully covered the entire topic, however, there pose some limitations. Despite stock selection is suitable with the main research's aim, other techniques are also available but they are not within the scope of this review.

I. INTRODUCTION:

Connectedness can help portfolio managers to derive a profitable investment strategy, especially during crises (Lundgren, Milicevic, Uddin, & Kang, 2018). Although researches on connectedness has been increasingly prevalent, there are only a few of them that can construct an investment strategy directly from this network indicator and there is hardly one done with a large sample size. On the other hand, machine learning for stock selection is a scalable technique that can rank stock based on multiple features. From which, a trading strategy can be obtained. In order to study the importance of connectedness on investment performance, this research utilizes connectedness as a feature for stock selection**.** Considering the novelty of the subject, it is an absolute necessity to perform a systematic literature review to compile the fair, essential knowledge for both investment strategies with connectedness and machine learning in stock selection . The objectives of this systematic review is to shed lights on the research gaps and further more provides the ground knowledge the main research.

II. METHODOLOGY:

This (Kitchenham, 2004) systematic literature review follows the Kitchenham's 2004 guideline for the procedures.
A systematic literature review consists of 5 stages: research identification, selection of primary studies, study quality assessment, data extraction and monitoring and data synthesis as the final stage.

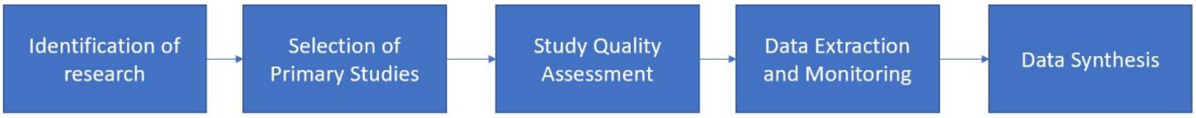

Identification of research
Selection of Primary Studies
Study Quality Assessment
Data Extraction and Monitoring
Data Synthesis

Figure 9 Process of Systematic Review *(following Kitchenham(2004)'s method)*

At the first step of identifying the research, it is required that researchers have some ground understanding on the topic. From which, a subset of research questions can be formed for the in depth. Moreover, the need for a systematic literature review should also be clarified as a rationale for review purpose from the beginning stage. Subsequently, a selection of primary are compiled through the proposed search process. This set of selected researches have to meet the preset inclusion criteria, while the exclusion criteria filter the irrelevant researches . As a result, the selected researches are the most relevant with the topic for further reviews. Furthermore, this set of researches goes through the third stage for quality assessment under a scoring system, resulting graded researches to find the best and most related works. (Kitchenham et al., 2009) This final data are then extracted and synthesized for the report.

2.1. Research Questions

For this review on bank stock selection based on LASSO-VAR model, there are two main concepts we need to review on. Firstly, we review connectedness and connectedness used for investment strategy and secondly, on machine learning for stock selection as an approach to the investment strategy.

Addresses: (List the questions and propose how to resolve the questions)

A. **Investment strategy with connectedness:**

Q1: What are the approaches to connectedness following Diebold and Yilmaz (Diebold & Yilmaz, 2011) method?

Q2: What are the approaches to investment strategies with connectedness in the past ?

Q3: What are their results and limitation?

B. **Stock Selection in Machine Learning**:

Q1: What are the learning techniques that the past papers used for stock selection?

Q2: What are their limitation?

With respect to the first question set, there has been many techniques used to quantify the connection between firms through connectedness, namely, the CoVar approach (An et al., 2020), Granger Causality (Audrino & Tetereva, 2019), Pearson Correlation (Benesty, Chen, Huang, & Cohen, 2009). However, this review centers around the vector autoregression approach of Diebold and Yilmaz and its direct variations. Through exhaustive search, text screening and data extraction, answers to the aforementioned questions will be revealed.

2.2. Search Process:

The search procedure starts with finding every possible keywords for the search engine. From that, more related past studies are covered and risks of missing data are substantially reduced. Moreover, research data is extracted from trustworthy sources to ensure information collected for review is precise and up to date.

2.2.1. Keywords:

By using thesaurus finders from reputative dictionaries like Oxford and Merry-Am Webster, we can determine the key words that have the potential to answer the review questions. A pilot search is performed with the original keywords to expand the key words bank, which, tremendously helps expanding the coverage of the search engine (Coffman & Weaver, 2012). As a result, a final list of thesaurus terms are generated to expand the search results. These thesauri are the synonym and most frequently-used terms of the concept to capture as many relevant results as possible. Truncation and wildcard methods are also applied to the search engine for the optimization the search process. Proximity operator are also used to increase matching-results chances.

Table 12 Keywords Table for Question A *(The table is a list of terms also used for the main search keywords. In this case, connectedness is searched alone and terms in investment strategy is added later on for papers that contains investment strategy with connectedness)*

Connectedness	**Investment Strategy**
Connection, Relationship, Interdependence	Portfolio, Strategy, Investment
Volatility Connectedness ,Volatility Spillover	Trading Portfolio, Trading Strategy, Trading Investment
Volatility Transmission	Portfolio Implication, Strategy Implication, Investment Implication
Spillover Effects	
Contagion Risks	
Vector Autoregression	
VAR(p)	

Table 13 Keywords Table for Question B *(The table is a list of terms for stock selection in machine learning)*

Stock Selection	**Machine Learning**
Ranking, Selection	Models, Algorithms, Learners
Stock Ranking, Stock Selection	Learning Models, Learning Algorithm
Portfolio	Artificial Intelligence

2.2.2. Resources:

Databases for the search is a mixture of subject specific information technology databases like ACM, and multi-disciplinary databases like IEEE, GGL, SD, CSL. Other publications refers to data from trusted associations and authorities like (IMF, WTO, etc.). Data extracted from these sources are journal articles, publications, report from different sites and association

Table 14 Resources table with acronyms

SOURCE	ACRONYMS
• IEEExplore	IEEE
• ACM Digital library:	ACM
• Google scholar (scholar.google.com)	GGL
• Citeseer library (citeseer.ist.psu.edu)	CSL
• ScienceDirect (www.sciencedirect.com)	SD
• Other Publications	OR

The process to retrieve research data from these sources is done completely manually. The procedure utilizes the same series keywords that are cultivated to search commands according to the source requirements. Publications retrieved from these sites are dated from 2008 onward to prevent outdated information for question set A. As a matter of fact, Diebold and Yilmaz's connectedness only came into fruition in 2008. Data from question set B can range from 2000-2020. However, the everchanging pace of machine learning motivates a timeframe from 2010 to ensure all data knowledge are up to date. Data is then screened for duplication and further put through the inclusion and exclusion criteria, and lastly the quality assessment to assess the quality of the key researches.

2.3. Inclusion and Exclusion Criteria:

The inclusion and exclusion criteria are the main screening criteria. Therefore, the decision to whether retain or discharge data depends on these set of standards, this helps to filter out the non-relevant researches and save only the essential papers. To include a research, these criteria must be considered:

A. Investment strategies

1. Diebold and Yilmaz method (Diebold & Yilmaz, 2011) is the core technique to calculate the volatility connectedness of the research. Which also consists of direct extensions of the method to improve the model

B. Stock Selection

1. The research uses predictive machine learning for the stock selection model.

As this review in fact includes a subset of two review topics, the inclusion and exclusion criteria should be assigned exclusively for each group. For topic A, any topics that are outside the scope of Diebold and Yilmaz connectedness are excluded. However, extensions to the VAR(p) (with generalized variance decomposition) model that were directly proposed by the authors are exempted from exclusion. As a matter of fact that we want to find out the distribution of connectedness and also the popularity of investment strategy with connectedness among the current researches. There should not be a cut-throat screening criteria but a rather broader screening benchmark. For topic B, only stocks selection methods that involve machine learning are within the review scope. To achieve the screening goal, an effective screening approach is vital to filter the researches. The screening procedures are initialized with title/abstract screening and afterward the full-text screening. This two steps screening protocol helps to improve the time efficiency and space complexity of the process.

2.4. Quality Assessment:

Quality Assessment is a grading procedure for the chosen set of researches. For each up tick of these questions, the research will get one point. The total points will reflect the quality of the research, and give rooms for further analysis and implications on research gaps. The selection criteria for the systematic review is as below (under question form, Yes for 1 point):

A. Investment Strategies with Connectedness:

1. Does the paper provide the investment strategy with connectedness?
2. Is the actual portfolio performance were calculated and analyzed?
3. Does the paper use machine learning for the investment strategy?
4. Does the sample size in the research greater than 80?
5. Is the investment strategy based on stock selection?

B. Stock Selection:

1. Does the paper have positive outcome?
2. Do they include net connectedness in stock selection in the past? (No)

These assessment criteria are set in an incremental way that the paper with the n score will automatically have all the n traits of the preceding quality criteria. For instance, if the quality score of question set A is 2, it means that the paper meets the quality criteria one and two of question set A.

2.5. Data Analysis:

Data once finalized through the selection and quality assessment will be extracted and synthesized to answer the initial research questions. The systematic review is expected to provide the quality information of connectedness approaches, especially on the investment strategy. And extensively, quality information on the stock selection are also collected. The review cannot only build up the knowledge body of connectedness but also provide great insights on stock selection researches on machine learning. Additionally, the research reflect the gap that gives implication for substantial research contribution.

III. RESULTS:

3.1. Search Results:

A. Connectedness:

From the search results, there are more than 2000 publications were retrieved from the databases. Specifically, there are 235 researches from the IEEE, 7 from ACM, 1270 from GGL, 13 from CSL, 799 from SD and 165 citations are from other miscellaneous sources. Duplication files are screened and omitted to avoid misleading data report. All 785 non-identical researches are shortlisted to 431 articles by applying the inclusion/exclusion criteria. According to the table (), researches on connectedness focus largely on the original VAR(p) model, following with the TVP-VAR and lastly LASSO-VAR. Due to the fact that analyses on connectedness and the portfolio implication are often observation based and done manually, it is hard to produce the same quality of analyses on LASSO-VAR with the same level of granularity.

From the inclusion/exclusion process, among the 431 papers regarding connectedness, there are only 52 papers that provide portfolio implications. Within which, only 20 papers provide actual portfolio strategies and there is only one paper use machine learning for the trading strategy with connectedness. The one paper is Lee's work in 2020, it is a rare find that employed machine learning with the network indicators (Pearson correlation and Diebold and Yilmaz) (Ball et al., 2020), the paper documented the improvements of the trading portfolio with these indicators during crises. Although this research only worked on a small data sample of 15 indices, it propels this research to take initiatives on implementing machine learning on a larger scale. As a fact, there has not been any papers work on a high-dimensional network and no paper have used stock selection as an investment strategy with connectedness.

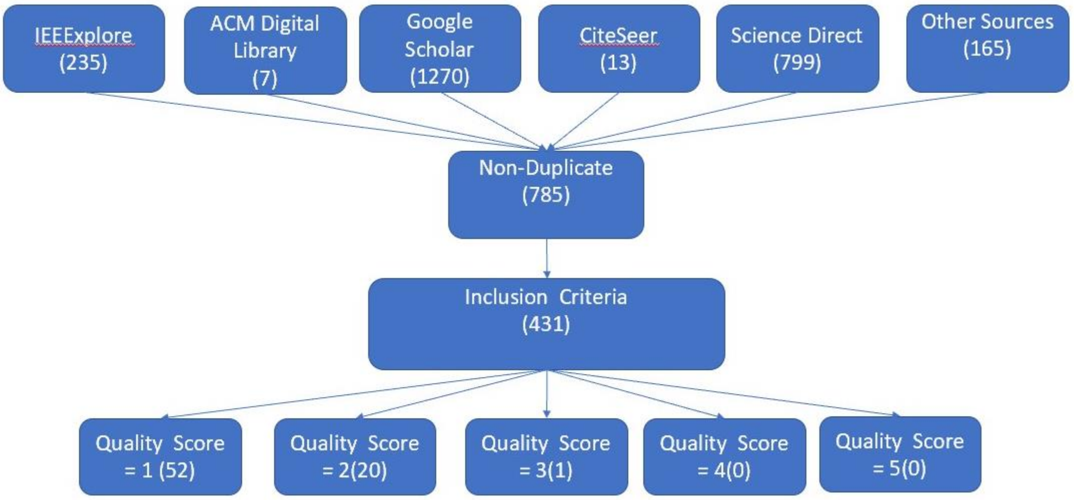


Figure 10 Search Process Results- Question set A

Table 15 Types of Connectedness for Researches Included- Question set A *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)*

Classic VAR	TVP-VAR	LASSO-VAR
308	92	31

B. Machine Learning for Stock Selection:

Researches from different sources are filtered with the duplicates, and only the relevant topics that meet the inclusion criteria remain for quality assessment. In this case there is 194 papers related to the topic, of which, most paper produce profitable portfolios from the learning models. However, there has not been any papers in stock selection included connectedness as a learning factor. While the investment strategy they provide varies on the approaches (in terms of the long and long-short term portfolios), most of the papers rely on a mixture of learning factors. Whereas, sentimental factors cannot stand alone during the prediction process, because the sentimental factors are often not deterministic on stocks' performance but rather a supporting factor instead. There are some papers rely on only fundamental or technical factors, otherwise most prefer a mixed factor learning approach.

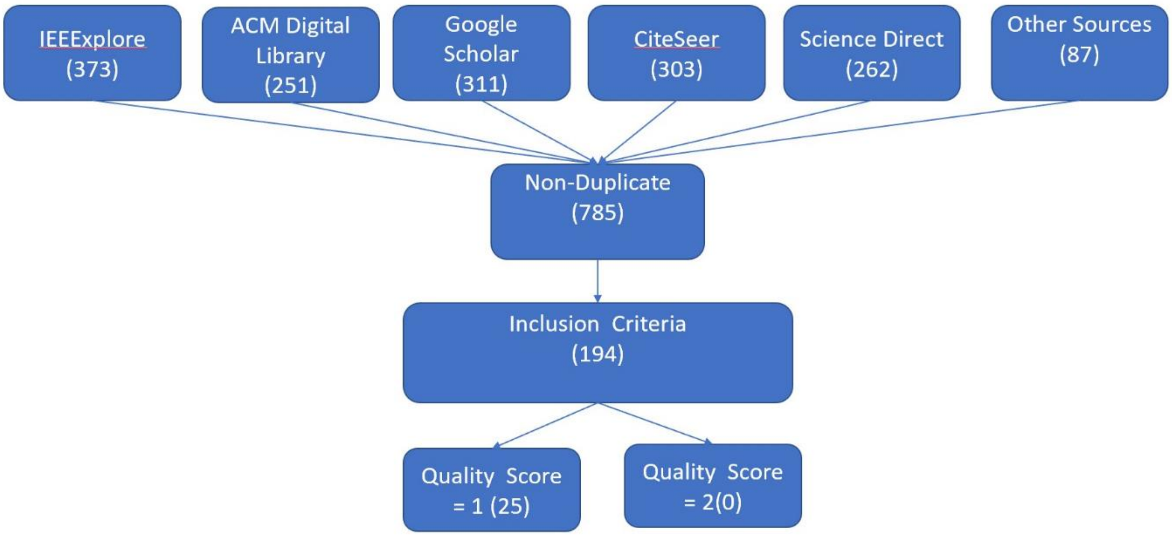


Figure 11 Search Process Results- Question set B

Table 16 Types of Trading Strategy and Factor Types for Researches Included- Question set B *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)*

	Long Portfolios	Long-Short Portfolios	Both
Fundamental Factors	42	8	6
Technical Factors	13	6	2
Sentimental Factors	0	0	0
Mixed Factors	74	21	12

With these short-listed researches, beside the insightful knowledge that we can obtained from, . Their methods can act as a benchmark for comparison and they also provide guidances along the way so that the research process can become much more feasible.

IV. DISCUSSIONS

4.1. Connectedness:

4.1.1. Volatility Measurements:

There are multiple approaches to calculate volatility connectedness. On the first publication about connectedness with generalized decomposition variance (KPPS), Diebold and Yilmaz (Diebold & Yilmaz, 2012) used the high-frequency intra-day trading data to compute realized volatility for the VAR(p) model. The daily realized-volatility in this method is the sum of squared log price changes over 78 5-minute intervals during the trading day.

Moreover, in cases that high-frequency data was not available, many other approaches were also employed as an alternative. Rogers and Satchell (Pieters & Vivanco, 2017) and Garman Klass's (Brauneis & Mestel, 2018) method have been seen on connectedness analysis for bitcoin and forex (Qarni, Gulzar, Fatima, Khan, & Shafi, 2019), and commodity (Diebold et al., 2017) respectively. We have also seen Demirer (Demirer et al., 2018)uses LASSO-VAR with Parkinson's volatility for large bank connectedness. Yang Zhang's volatility and the Garman Klass's extensions were also applied in Nasreen in 2019. Besides these range-based volatility approaches, some researchers work on the traditional standard deviations of different period (4,8,12 weeks observation) (TKLee) or refers to the daily log change of price as volatility (Wen, T., Wang, G.-J., 2020. Volatility connectedness in global foreign exchange markets.)

As observed, because there is no strict guidance on volatility measurement, the choice of volatility approach rather bases on the availability of data. In a research on cryptocurrency connectedness, Shuyue Yi (2018) tested the robustness of different types of volatilities on the data set, the author found that connectedness indicator is not sensitive to different volatility measurements.(Rogers and Satchell, Parkinson, Garman Klass and Yang Zhang).

4.1.2. Types of connectedness:

There are two main variants that are the direct extensions of the Diebold and Yilmaz's connectedness, the time varying parameter vector autoregression and the LASSO-VAR. Although there are also some other types of connectedness proposed, but they are ruled out as these papers are not the extensions introduced by Diebold and Yilmaz. The descriptive Table 4 above depicts the distribution of the researches among different types of VAR. The classic model takes up the majority of the publications. As we go along, a bigger data set requires a smarter, more effective regimes, the LASSO-VAR model (Demirer et al., 2018) was therefore introduced to reduce the dimensionality of dataset. Which take the third place for is coverage in the connectedness research population. The second runner is TVP-VAR with 92 publications out of the found 431 publications. The time varying parameter version of the model is more sensitive the dynamic connectedness which can resolve the problems of the rolling window horizon of the classic VAR(p). It is a more applicable method as many real-world problems require a versatile, easy to compute time series data set.

4.1.2.1. The original VAR(p)

The popularity of the classic VAR(p) model with generalized variance decomposition is gaining popularity overtime. The review takes into account the 2009's research as a part of the 2012 research as it was yet to implement to commonly used generalized variance decomposition, although the concept and topology of connectedness has been formed in this 2009 research. While the later research on 2012, presents connectedness in a well-grounded development. This is the classic volatility connectedness model that has big influences on many portfolio analysis problems. The central concept of GVD, the original connectedness and its topology from VAR(p) are included here. The next parts discuss on other VAR versions are modified and extended to improve the classic VAR.

Generalized Variance Decomposition:

The VAR(p) is a statistical stochastic process model that depicts the relationship between multiples variables as they change over time.

The vector auto regression of the stock prices in p lag order can be described as below:

$$x_t = \sum_{i=1}^{p} \Phi_i \,.\, x_{t-i} + \varepsilon_t \qquad (2)$$

Where $\varepsilon_t \sim (0, \Sigma)$ is the error white noise vector, Φ_i are the coefficient matrices, p is the lag order. x_t in this case is the vector of volatilities of N financial institution's stock prices. And x_{t-i} is the vector of past volatility values in p lag order.

Vector autoregression can be transformed into the moving average representation:

$$x_t = \sum_{\mathrm{i}=0}^{\infty} \mathrm{A}_\mathrm{i} \varepsilon_{\mathrm{t}-\mathrm{i}} \qquad (3)$$

$$\text{And} \quad A_i = \Phi_1 A_{i-1} + \Phi_2 A_{i-2} + \cdots + \Phi_p A_{i-p} \qquad (4)$$

Where A_i are the moving average correlation coefficients. With A_0 is a NxN identity matrix, and $A_i = 0$ for $i = 0$.

From the moving average representation, the moving average coefficients are transformed to variance decompositions. By that, H- step error forecast variances can be described as below:

$$\omega = \frac{\sum_{h=0}^{H-1} \left(e_i' A_h \mathrm{e}_\mathrm{i}\right)^2}{\sum_{h=0}^{H-1} \left(e_i' A_h \Sigma A_h' e_i\right)} \qquad (5)$$

$$And \quad MSE_j(h) = \sum_{h=0}^{H-1} \left(e_i' A_h \Sigma A_h' e_i\right) \qquad (6)$$

Based on the foundation of Koop et al., Shin and Perasan introduced the generalized version of the forecast error variance decomposition:

$$\theta_{ij}^{g}(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} \left(e_i' A_h \Sigma\, e_j\right)^2}{\sum_{h=0}^{H-1} \left(e_i' A_h \Sigma A_h' e_i\right)} \qquad (7)$$

Where σ_{jj} is the standard deviation of error vector ε_j, and e_i is the selection vector with i^{th} element unity and zero at the rest of the matrix. The generalized variance decomposition then become the centerpiece of connected measures for Diebold and Yilmaz's works.

Connectedness:

From the forecast error variance decomposition, we can have the matrix as below:

Table 17 Pairwise directional connectedness in a matrix

	x_1	..	x_N
x_1	d_{11}^H	..	d_{1N}^H
..	..	..	..
x_N	d_{Nj}^H	..	d_{NN}^H

Connectedness from j to i can be calculated following this formula:

$$C_{i\leftarrow j}(H) = d_{ij}^H = \widetilde{\theta_{ij}}(H) = \frac{\theta_{ij}^g(H)}{\sum_{j=1}^N \theta_{ij}^g(H)} \qquad (8)$$

With that $\sum_{j=1}^N \widetilde{\theta_{ij}^g}(H) = 1$ and $\sum_{j,i=1}^N \widetilde{\theta_{ij}^g}(H) = N$. This depicts the pairwise directional relationship between $i \; and \; j$. To be exact, it is the pairwise directional connectedness of the shock from $j \; to \; i$ at the forecast horizon H.

Table 18 Full connectedness matrix

	x_1	..	x_N	***From Others***
x_1	d_{11}^H	..	d_{1N}^H	$\sum_{j=1}^{N} d_{1j}^H$ ***with j*** $\neq$ **1**
..	..	..	..	..
x_N	d_{Nj}^H	..	d_{NN}^H	$\sum_{j=1}^{N} d_{Nj}^H$ ***with j*** $\neq$ **N**
To Others	$\sum_{i=1}^{N} d_{i1}^H$ ***with i*** $\neq$ **1**	..	$\sum_{i=1}^{N} d_{iN}^H$ ***with i*** $\neq$ **1**	$\frac{1}{N}\sum_{i=1,j=1}^{N} d_{ij}^H$ ***with i*** $\neq$ **j**

The covariance decomposition of vector N is the NxN matrix of the table above, at the upper left corner of the table. Each off diagonal elements of the matrix depicts the pairwise directional relationship factors in N. The matrix can be inferred to as $D^H = [d_{\{ij\}}^H]$.

Pairwise directional connectedness:

The off diagonal elements in matrix D^H can define the weight of the from and to relationship the network. The pairwise directional connectedness can be depicted as:

$$C_{i\leftarrow j}^H = d_{ij}^H \qquad (9)$$

And moreover, in general $C_{i\leftarrow j}^H \neq C_{i\leftarrow j}^H$. Which shows $N^2 - N$ pairwise directional connected measures.

Net pairwise connectedness:

The net pairwise connection between two elements of matric NxN can be written as:

$$C_{ij}^{H} = C_{i\leftarrow j}^{H} - C_{i\rightarrow j}^{H} \tag{10}$$

The amount of net pairwise directional connectedness measures would be $(N^2 - N)/2$

Total directional connectedness:

The sum of all element in one row of the NxN matrix in case of $i \neq j$ is the H-step forecast error variance decomposition of the stock with that row. The forecast error variance comes is the shocks received from all other elements in that row. Which shows the total directional connectedness from others to i:

$$From\ others: C_{i\leftarrow \cdot}^{H} = \sum_{j=1}^{N} d_{ij}^{H} \ \ with\ i \neq j \tag{11}$$

And otherwise, to others:

$$To\ \ others: C_{\cdot\leftarrow j}^{H} = \sum_{i=1}^{N} d_{ij}^{H}\, with\ \ i \neq j \tag{12}$$

Net total directional connectedness:

Similarly to pairwise directional connectedness, the net total directional connectedness can be as:

$$C_{i}^{H} = C_{\cdot\leftarrow i}^{H} - C_{i\leftarrow \cdot}^{H} \tag{13}$$

The sum row and the sum column difference of the stock I GVD is its total directional connectedness.

Grand Total Connectedness:

The total connectedness can be computed as below formula:

$$C^{H} = 1/N \sum_{i,j=1}^{N} d_i\, j^{H} \ \ with\ i \neq j \tag{14}$$

4.1.2.2. LASSO-VAR

Since the introduction of LASSO-VAR (Demirer et al., 2018), the model has been gaining great popularity for its ability to solve the space and time complexity problem of VAR(p). There are in fact many different LASSO-VAR models with different penalty function for LASSO. The details below are regarding to the LASSO-VAR extension of (Demirer et al., 2018) with Yilmaz as a co-author. The VAR(p) model with LASSO can perform very well in respect of high dimensional data matrices.

There are actually multiple ways to work with high dimensional VAR(p), one can purely shrink it with Bayesian analysis or Ridge Regression. Another way is selecting factors by employing information criterion like AIC or SIC. LASSO (the least square shrinkage and selection operator), in fact, is a combination of the two methods- shrink and select.

LASSO is designed to solve the penalized estimation problem:

$$\hat{\beta} = argmin_{\beta}(\sum_{t=1}^{T}(y_t - \sum_i \beta_i x_{it})^2 + \lambda \sum_{i=1}^{K} |\beta_i|^q) \tag{15}$$

With a simple extension of the adaptive elastic net (Zhou and Zhang,2009):

$$\hat{\beta} = argmin_{\beta}\left(\sum_{t=1}^{T}(y_t - \sum_i \beta_i x_{it})^2 + \lambda \sum_{i=1}^{K} \omega_i \left(\frac{1}{2}|\beta_i| + \frac{1}{2}\beta_i^2\right)\right) \tag{16}$$

Under conditions that $\omega_i = 1/\left|\widehat{\beta_{\iota,OLS}}\right|$with λ is the selected equation by equation of the 10-fold cross validation. The model is consistent for the best Kullback-Liebler approximation to the true DGP.

4.1.2.3. TVP-VAR

The time varying parameter model for vector autoregression is exceptionally compatible for research works that mainly rely on dynamic connectedness. Instead of using the rolling window horizon, the variances of this model are ‘allowed’ to be time-varying with the Kalman Filter estimation, the forgetting (Koop et al., 1996) factor are also included. Chances of losing observations, distorted parameters are ruled out with this model.

The below part will discuss the TVP-VAR of (Gabauer & Gupta, 2018):

$$Y_t = \beta_t Y_{t-1} + \epsilon_t \tag{17}$$

$$\beta_t = \beta_{t-1} + v_t \tag{18}$$

Whereby $\epsilon_t \mid F_{t-1} \sim N(0, S\_t)$ and $v_t \mid F_{t-1} \sim N(0, R_t)$

In this case, there exists not only the classic VAR(p) model with the values at lag 1 order, the coefficient matrix (which is now time-varying), and the white noise vector error (that with the time varying variance-covariance matrix $NxN,\ S_t$). But also, the time varying coefficient equation, where β_t is an NxN_p dimensional time varying coefficient matrix value of which depends on β at $t-1$ on an error matrix v_t (NxN_p) with a variance-covariance matrix $NxNp.$

These functions are used to estimate the generalized variance decomposition of Diebold (Diebold & Yilmaz, 2012), which is based on the generalized impulse function of Koop(Koop et al., 1996) and the generalized forecast of variance decomposition Shin and Perasan (Pesaran & Shin, 1998). This time the moving average presentation has the moving average correlation coefficient with values of β:

$$Y_t = \beta_t Y_{t-1} + \epsilon_t \tag{19}$$

$$Y_t = A_i \epsilon_t \tag{20}$$

$$A_{0,t} = I \tag{21}$$

$$A_{i,t} = \beta_{1,t} A_{i-1,t} + \ldots + \beta_{p,t} A_{i-p,t} \tag{22}$$

Where $\beta_t = [\beta_{1,t}, \beta_{2,t}, \ldots \beta_{p,t}]'$. With the applied generalized impulse response function similarly to the original VAR(p), we can calculate the generalized forecast error variance decomposition for time-varying connectedness

Besides the time-varying parameter for the vector autoregression model, the research also includes a large Bayesian prior model to process the larger dataset, however, the data sample could only reach to 35 financial institutions.

Lastly, the volatility connectedness concept is certainly gaining its popularity in the research community. While there are many ways to use the VAR(p), LASSO-VAR is the most appropriate technique for the high dimensional characteristic of this review problem.

4.1.3. Connectedness in Banking:

Well-established works has been done in respects of connectedness in banking. In fact, on the first publications about connectedness, Diebold and Yilmaz (2007, 2008, 2012, 2018) focus largely on banking and other financial institutions' connectedness. Banks collapse is among the fears of investors as a flashback from previous major financial crises (etc. 2000, 2008, 2015). As banks play the central role in every financial activity, they transmit ample effects to the market once they fail. Many regulative bodies for banks have been formed to control the interdependence of banks and maintain the financial stability of the entire market. For instance, Basel III was introduced to limit bank exposures to one or many other counterparties, this Basel framework requires banks to measure and set boundaries to large exposures in relation to their capital.

According to Diebold and Yilmaz (2012), total static connectedness of a group of finance stocks are generally considered to be higher than other stocks group. However, a group of global banks together can have a lower total connectedness than the domestic ones (Diebold & Yilmaz, 2012). Regardless, dynamic total connectedness of global banks hit the troughs during adverse events. Connectedness in the global banking shows prominent patterns during financial crises. Whereas, prominent relationships between United States and other countries were depicted during the global financial crisis (Demirer et al., 2018). Therefore, the strong indication of connectedness in banks can provide significant investment implication for the banking industry.

4.1.4. Investment Strategy in Connectedness:

Although connectedness can provide implications on portfolio strategy, most papers only discuss on the hypothetical implication of the concept itself (Nasreen, Tiwari, & Yoon, 2021). There are still some papers extensively work on a more realistic approach where actual portfolio were formed and analyzed. Of those papers, many focus on the hedging, and portfolio diversification strategy. Particularly, D Maitra paper (Mensi, Maitra, Vo, & Kang, 2021) and Guhatharkutar (Maitra, Guhathakurta, & Kang, 2021), VaR (Value at risk) or the Conditional VaR were used as risks measurement to allocate weights for the optimal portfolio. The strategies were provided periodically according to the crisis signal given by the Diebold and Yilmaz connectedness indicator. Connectedness in these cases are not the main component that determine the portfolio, but rather a part of the analysis. In this category of investment strategy with connectedness, T.K.Lee's (2020) is a rare find that employed network indicators directly into machine learning for regional allocation strategy. The learning results from this research show a significant improvement of the portfolios (with network indicators) during crises. However, similarly to other papers that use connectedness for investment strategy, the study sample of this work only restricted to a handful indices. Firstly, due to the reason that VAR(p) has a limited capability on sample size and computational cost, and secondly, because analyses for investment strategies are often labor intensive. Despite LASSO-VAR has been shown to be very effective with high-dimensional data, producing meaningful results from the model remain a bigger data challenge.

4.2. Stock Selection in Machine Learning:

Stock selection is a method of choosing stocks for trading portfolios based on a variety of different factors. The traditional stock selection method works on a factor by factor basis, whereby rankings are retrieved based on analysis of each factor chosen. With the application of machine learning for stock selection, all factors (or in data science terms features) can be the input the learning model to predict the stock ranking in multiple days ahead. This technique cannot only perform timely trading portfolios but also can cover a wide range of factors and stocks. Stock selection in machine learning can be helpful to measure the importance of connectedness in investment strategies with a bigger scale and with less computational cost.

4.2.1. Data and Features:

Data used for stock selections usually comes in a big amount. The amount of stocks used for selection can be as little as 100 (Kompas, 2002) as or as many as 5907 Rasekhschaffe, K.C., Jones, R.C., 2019. Machine Learning for Stock Selection. Depending on the research objectives that data sample can be adjusted accordingly.

Factors involve into stock selection belong to these main categories: fundamental indicators, technical indicators, sentimental indicators. Many experiments were conducted on one type of indicator ()()(). Some others combine all factors in all three aforementioned categories for learning. However, with such a wide range of factors available, researchers may prone to bias during the feature selection process.

4.2.2. Learning Techniques:

As the majority of stock selection problems aim to answer whether a stock is going to outperform or underperform in the observing market({Sorensen, 2000;Tiryaki, 2005;Fu, 2018), the criteria that determine a good stock is crucial to produce a profitable portfolio. The determinant of outperformance usually refers to returns(Brinson et al, 1991), some papers use risk adjusted returns to enhance prediction robustness. Nonetheless, Sharpe ratio have been seen in (Israelsen, 2005;Ledoit & Wolff, 2018) as a classic performance indicator. For stock selections using purely the fundamental factors, holding period returns are often used to determine stock performance. From the given label, the learning models sometimes rely on the label of stock performance to predict the likelihood of a winning stock(Sanders & Hambrick, 2007)(Huang, 2012).

Machine learning is a well-developed field with a big library of learning methods. Which opens up many possibilities to conduct stock selection from. For this learning problem, classification based answers are sought for. in Rasekhschaffe (2019)'s work, a wide selection of classification models were implemented to retrieve the best performing stock. The author used Ada Boost, Ensemble Bagging SVM with Artificial Neural Network to extract the probability of daily winning stock. All probability results from the learning model were averaged to generate the combined average learning results. The combined average learning probabilities showed a superior performance to the other models as an outcome of ensemble learning. On a very similar approach, Dominik Wolff(2020) employed different regression models combine with bagging, boosting (Random Forest and Gradient Boost respectively) methods and the neural networks. Lastly, the average combined forecast were obtained and reported to produce better portfolio performance same as Rasekhschaffe's (2019).

4.2.3. Investment strategies:

There are two main paths to construct a trading portfolio with the learning results retrieved. Following the first path, we can select the top performers to form a long portfolio for the investment strategy. In order to add in some diversification, some papers also proposed a long-short strategy, which takes the top performers for the long strategy and short the bottom performers to enhance profitability especially for volatile periods. While long portfolio can out perform its long-short counterpart, it is less efficient than the long-short portfolios during highly volatile periods . (Song, 2017)

4.3. Banking Stock Selection based on LASSO-Large Network Connectedness:

Although many researches has highlighted the importance of connectedness in investment strategies (Harrathi, 2016;Galinskaitė, 2018;Mandacı, 2020;Nasreen, 2021), there have not been any papers that work on this problem in a larger scale. In order to study the indicative power of connectedness on the investment strategy obtained from a high dimensional data sample, LASSO-VAR can be used to operate variance decomposition on large dimensional data. Additionally, stock selection with machine learning can assist to create stock investment strategy with connectedness from the LASSO-VAR model.

V. CONCLUSION:

This literature review sheds light on the possibilities of expanding the scope of connectedness on the larger scale and especially using the high-dimension data of connectedness for investment strategy. Although there are many techniques toward connectedness, LASSO-VAR is the most appropriate model for the research, regarding the size of data sample. On the same regard, stock selection with machine learning is versatile and effective for the research. The review successfully covered the entire topic, however, there pose some limitations. Despite stock selection is suitable with the main research's aim, other techniques are also available but they are not within the scope of this review. Similarly to other VAR extensions that are not included in this paper.

VI. REFERENCES:

An, P., Li, H., Zhou, J., Li, Y., Sun, B., Guo, S. & Qi, Y. 2020. Volatility spillover of energy stocks in different periods and clusters based on structural break recognition and network method. Energy, 191, 116585.

Audrino, F. & Tetereva, A. 2019. Sentiment spillover effects for US and European companies. Journal of Banking & Finance, 106, 542-567.

Ball, S., Banerjee, A., Berry, C., Boyle, J. R., Bray, B., Bradlow, W., Chaudhry, A., Crawley, R., Danesh, J. & Denniston, A. 2020. Monitoring indirect impact of COVID-19 pandemic on services for cardiovascular diseases in the UK. Heart, 106, 1890-1897.

Benesty, J., Chen, J., Huang, Y. & Cohen, I. 2009. Pearson correlation coefficient. Noise reduction in speech processing. Springer.

Brauneis, A. & Mestel, R. 2018. Price discovery of cryptocurrencies: Bitcoin and beyond. Economics Letters, 165, 58-61.

Coffman, J. & Weaver, A. C. 2012. An empirical performance evaluation of relational keyword search techniques. IEEE Transactions on Knowledge and Data Engineering, 26, 30-42.

Demirer, M., Diebold, F. X., Liu, L. & Yilmaz, K. 2018. Estimating global bank network connectedness. Journal of Applied Econometrics, 33, 1-15.

Diebold, F. X., Liu, L. & Yilmaz, K. 2017. Commodity connectedness. National Bureau of Economic Research.

Diebold, F. X. & Yilmaz, K. 2011. Equity market spillovers in the Americas. Financial stability, monetary policy, and central banking, 15, 199-214.

Diebold, F. X. & Yilmaz, K. 2012. Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of forecasting, 28, 57-66.

Gabauer, D. & Gupta, R. 2018. On the transmission mechanism of country-specific and international economic uncertainty spillovers: Evidence from a TVP-VAR connectedness decomposition approach. Economics Letters, 171, 63-71.

Kitchenham, B. 2004. Procedures for performing systematic reviews. Keele, UK, Keele University, 33, 1-26.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. & Linkman, S. 2009. Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51, 7-15.

Koop, G., Pesaran, M. H. & Potter, S. M. 1996. Impulse response analysis in nonlinear multivariate models. Journal of econometrics, 74, 119-147.

Lundgren, A. I., Milicevic, A., Uddin, G. S. & Kang, S. H. 2018. Connectedness network and dependence structure mechanism in green investments. Energy Economics, 72, 145-153.

Maitra, D., Guhathakurta, K. & Kang, S. H. 2021. The good, the bad and the ugly relation between oil and commodities: An analysis of asymmetric volatility connectedness and portfolio implications. Energy Economics, 94, 105061.

Mensi, W., Maitra, D., Vo, X. V. & Kang, S. H. 2021. Asymmetric volatility connectedness among main international stock markets: A high frequency analysis. Borsa Istanbul Review, 21, 291-306.

Nasreen, S., Tiwari, A. K. & Yoon, S.-M. 2021. Dynamic Connectedness and Portfolio Diversification during the Coronavirus Disease 2019 Pandemic: Evidence from the Cryptocurrency Market.

Pesaran, H. H. & Shin, Y. 1998. Generalized impulse response analysis in linear multivariate models. Economics letters, 58, 17-29.

Pieters, G. & Vivanco, S. 2017. Financial regulations and price inconsistencies across Bitcoin markets. Information Economics and Policy, 39, 1-14.

Qarni, M. O., Gulzar, S., Fatima, S. T., Khan, M. J. & Shafi, K. 2019. Inter-markets volatility spillover in US bitcoin and financial markets. Journal of Business Economics and Management, 20, 694-714.

APPENDIX B: FIGURE

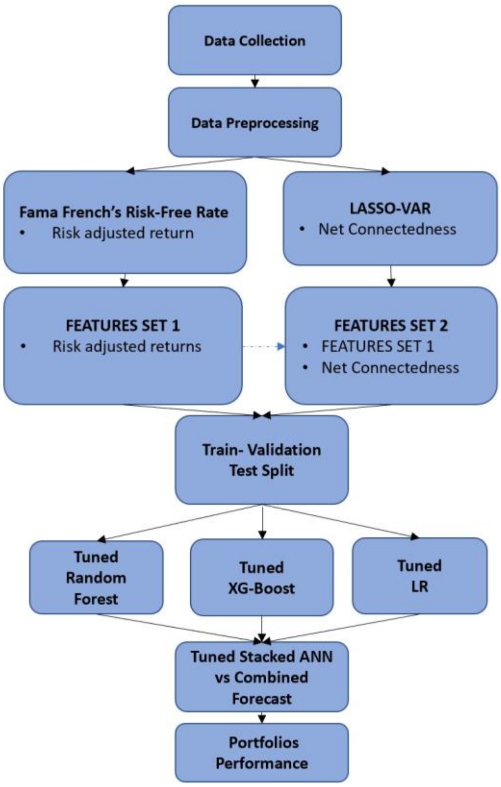


Figure 12 Research Framework

.

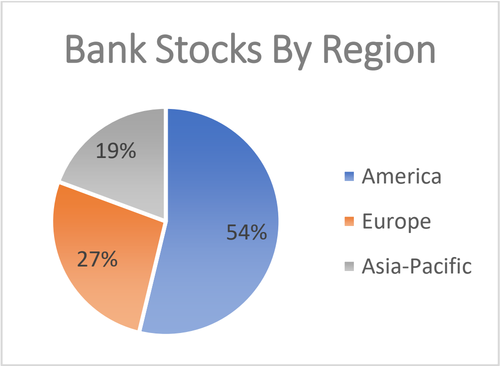


Figure 2 Bank stocks by Region

:

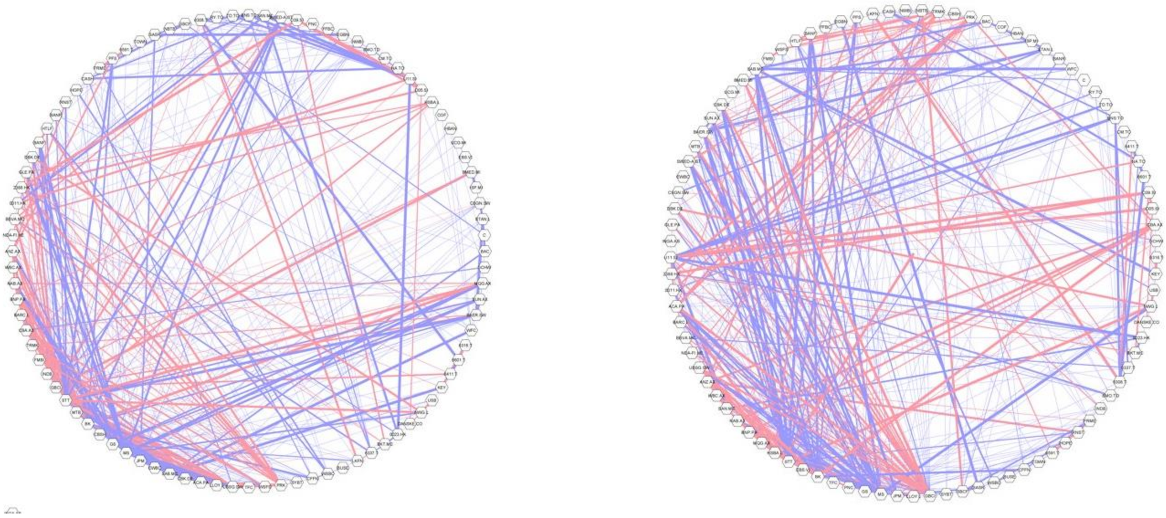


Figure 3 Net Pairwise Connectedness (This figure depicts the top ten percentile important net pairwise connectedness of 93 stocks during the global financial crisis, the left network is recorded on 12/09/2008 before the big crash of 2008 and the right network is recorded on 15/09/2008 during the crash. The network edge colors depict the direction of connectedness, the blue color represent the negative connectedness index while the red color is the negative connectedness value)

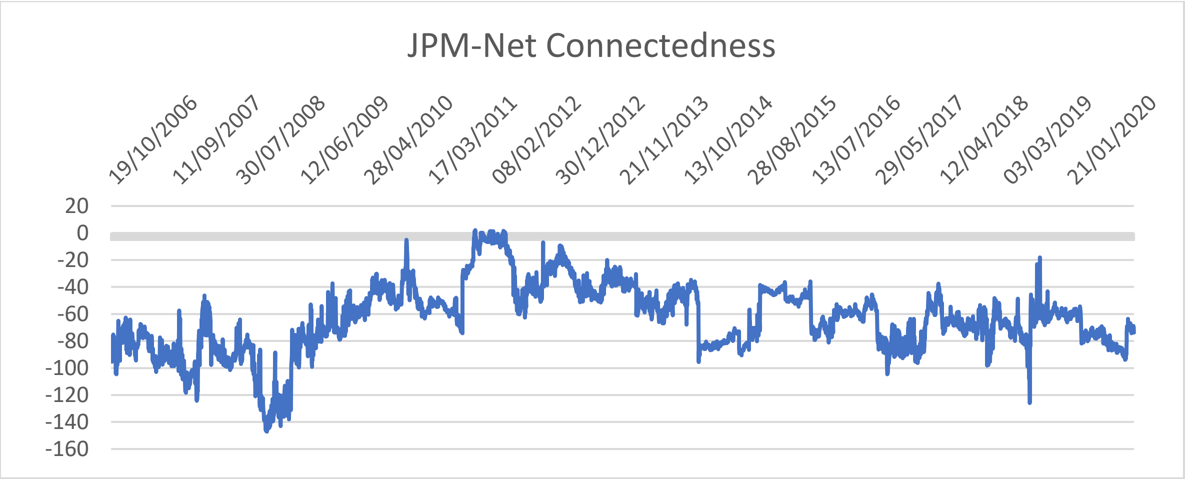


Figure 4 JPM- Net Connectedness(from 06/08/2006 to 12/10/2020)

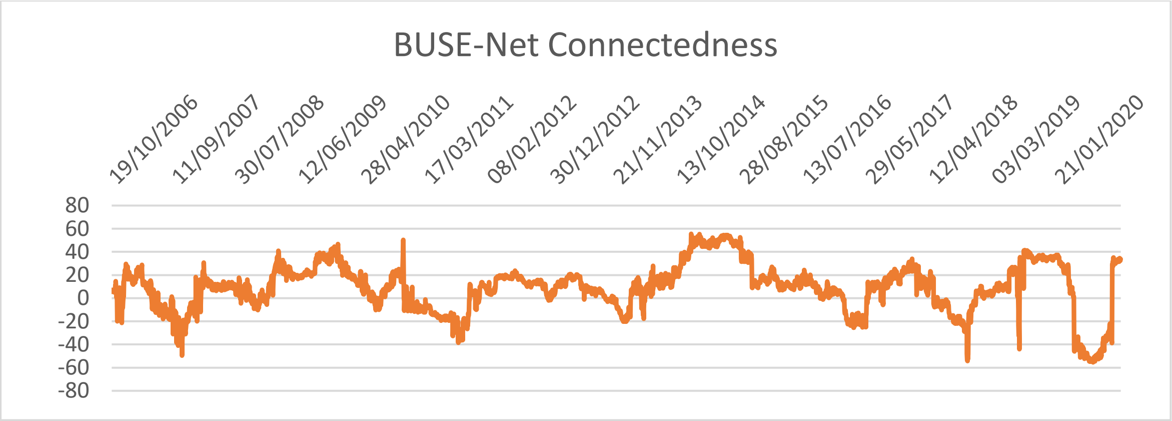


Figure 5 BUSE- Net Connectedness(from 06/08/2006 to 12/10/2020)

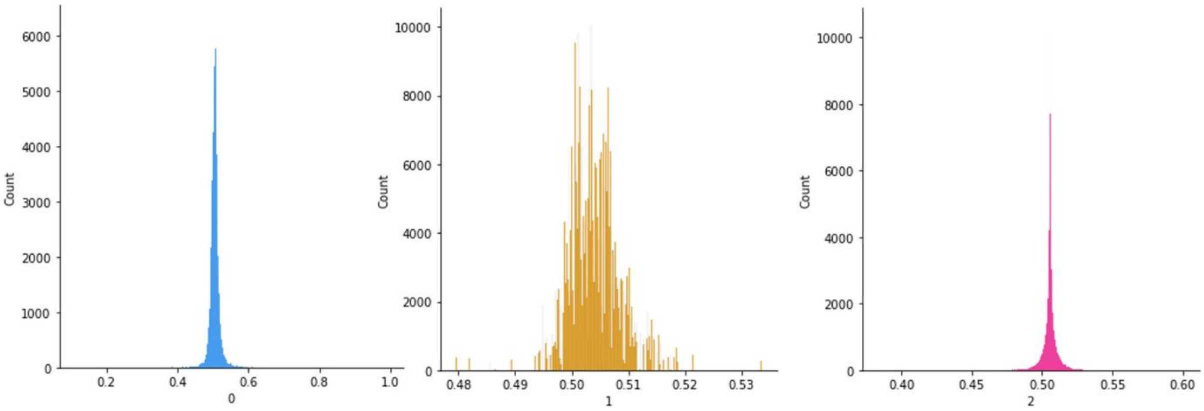


Figure 6 Distribution of Random Forest, Extreme Gradient Boosting, Logistic Regression

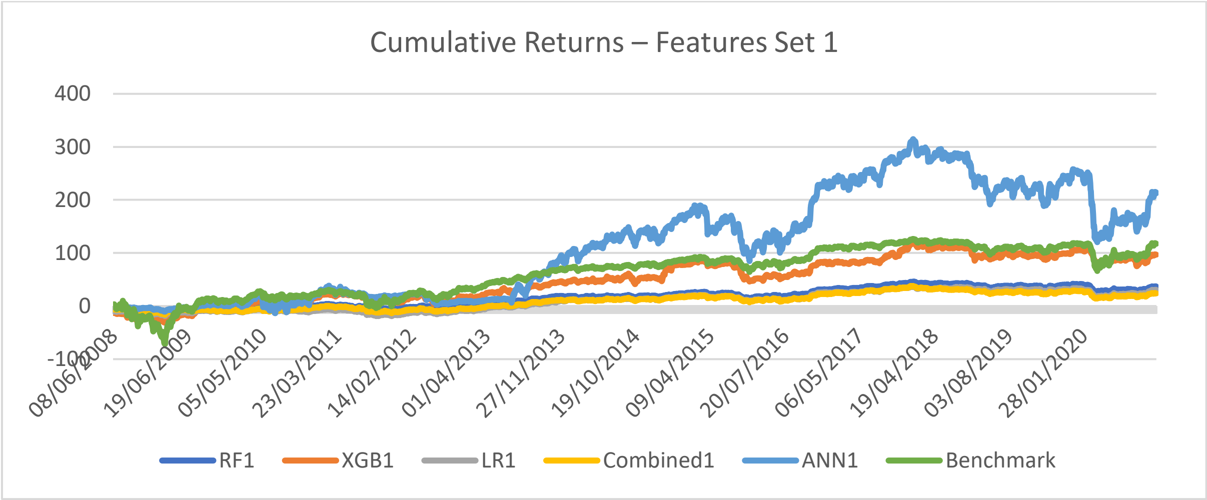


Figure 7 Cumulative Returns of Features Set 1

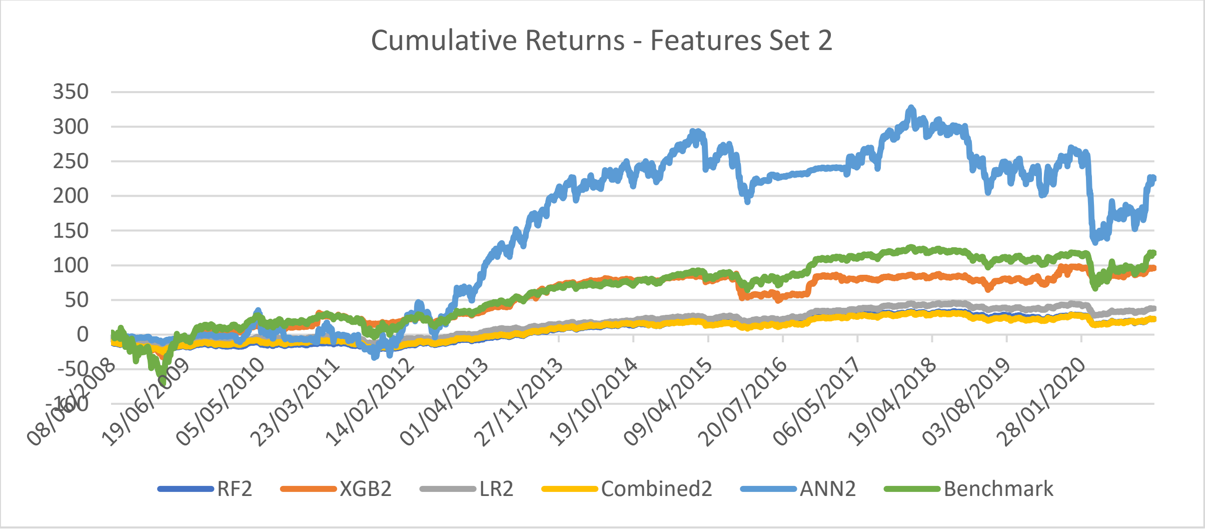


Figure 8 Cumulative Returns of Features Set

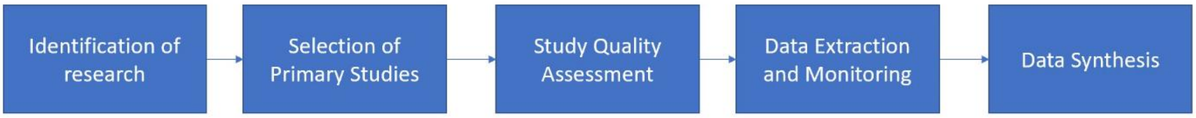

Identification of research
Selection of Primary Studies
Study Quality Assessment
Data Extraction and Monitoring
Data Synthesis

Figure 13A Process of Systematic Review *(following Kitchenham(2004)'s method)*

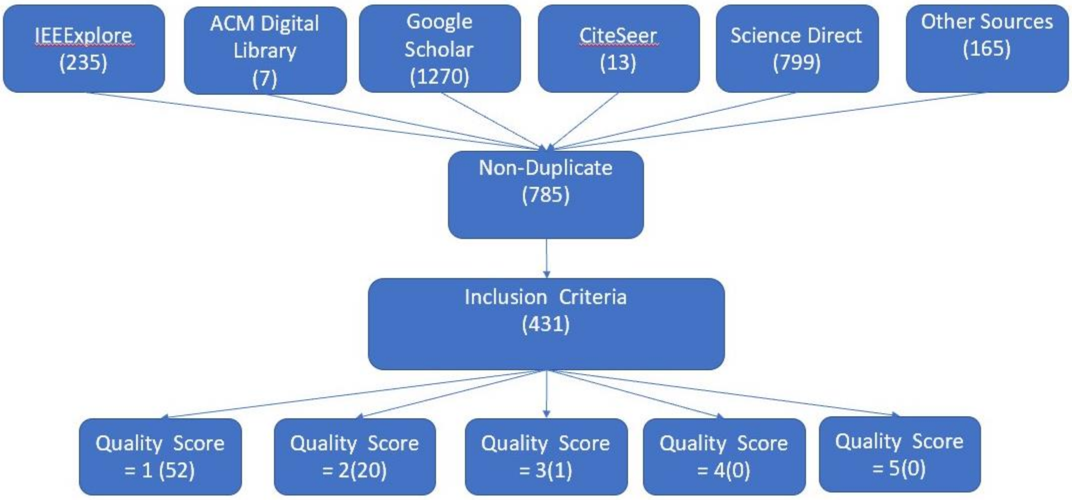


Figure 14A Search Process Results- Question set A

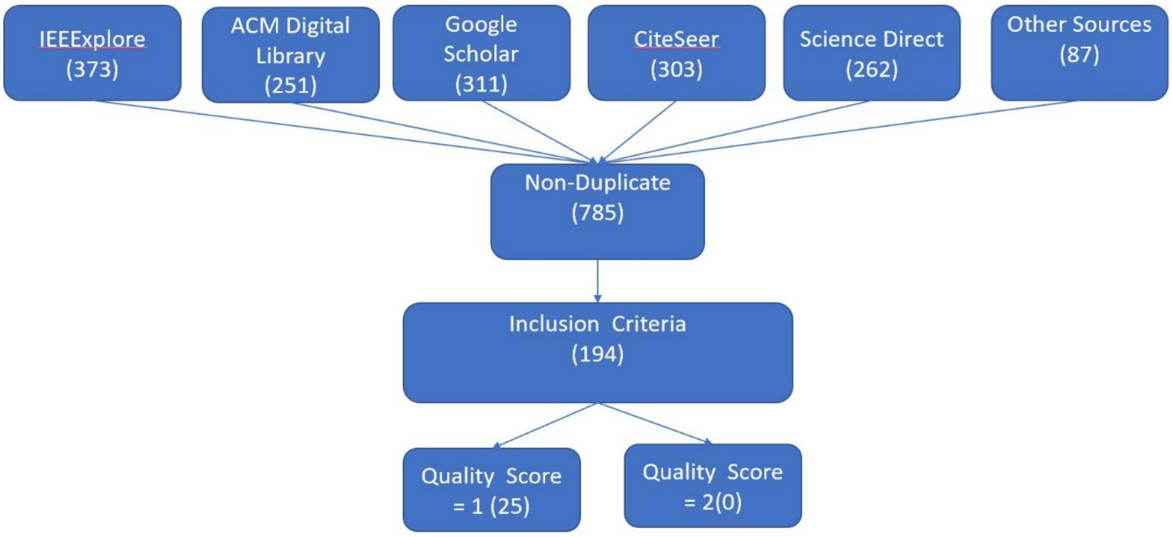


Figure 15A Search Process Results- Question set B

APPENDIX C: TABLES

Table 1 Research Paper Structure

Section	Content
Section 1: Introduction	– Provides a chronological development of stock selection and the connectedness methods. – Addresses the rationale for the selected research methods and research gaps
Section 2: Methodology	– Provides a research framework – Provides information about the dataset – Details on LASSO-VAR as a mean to acquire net connectedness – Details on the stock selection process which includes information of the features sets, data splitting process and also the information of the algorithms involved in the research
Section 3: Results Analysis:	– Illustrations and explanation of Machine Learning Results – Illustrations and explanation of the portfolio performance
Section 4: Conclusion	– Conclusion to the research paper.
Section 5: Future work	– Topic expansion and possibilities for future research.

Table 2 Bank stocks by country

Country	Amount
United States	44
Canada	6
Australia	6
United Kingdom	5
France	3
Spain	4
Singapore	3
Netherlands	1
Switzerland	3
Italy	3
Finland	1
Japan	6
Hong Kong	3
Germany	2
Sweden	1
Austria	1
Denmark	1
Total	**93**

:

Table 3 Matrix of Variance Decomposition

	x_1	..	x_N
x_1	d_{11}^H	..	d_{1N}^H
..	..	..	..
x_N	d_{Nj}^H	..	d_{NN}^H

Table 4 Pooled Prediction Correlation of three models (RF,XGB,LR)

	RF	XGB	LR
RF	1	0.4898685	0.3353745
XGB	0.4898685	1	0.42378125
LR	0.3353745	0.42378125	1

Table 5 Learning Model Accuracy Scores (of RF, XGB, LR, and stacked ANN)

Models	Without Connectedness				With Connectedness			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
RF	50.202%	50.531%	74.276%	59.627%	50.230%	50.585%	69.703%	58.253%
XGB	50.240%	50.470%	81.383%	61.054%	50.388%	50.588%	81.174%	61.724%
LR	50.388%	50.536%	88.818%	64.161%	50.430%	50.603%	82.957%	62.598%
ANN Stacked	50.555%	50.602%	92.345%	64.880%	50.515%	50.603%	90.854%	64.777%

Table 6 Stocks List

Ticker	*Name*	*Market Capitalization*	*Country*
JPM	*JPMorgan Chase*	*488,893,000,000*	*United States*
BAC	*Bank of America*	*357,213,000,000*	*United States*
WFC	*Wells Fargo*	*190,866,000,000*	*United States*
MS	*Morgan Stanley*	*179,427,000,000*	*United States*
C	*Citigroup*	*143,406,000,000*	*United States*
RY.TO	*Royal Bank Of Canada*	*142,454,000,000*	*Canada*
SCHW	*Charles Schwab*	*139,712,000,000*	*United States*
CBA.AX	*Commonwealth Bank*	*133,368,000,000*	*Australia*
GS	*Goldman Sachs*	*127,797,000,000*	*United States*
TD.TO	*Toronto Dominion Bank*	*121,390,000,000*	*Canada*
HSBA.L	*HSBC*	*108,405,000,000*	*United Kingdom*
MQG.AX	*Macquarie*	*96,852,001,500*	*Australia*
USB	*U.S. Bancorp*	*89,121,071,104*	*United States*
PNC	*PNC Financial Services*	*83,608,870,912*	*United States*
BNP.PA	*BNP Paribas*	*81,450,549,248*	*France*
TFC	*Truist Financial*	*79,449,079,808*	*United States*
BNS.TO	*Scotiabank*	*75,419,328,512*	*Canada*
COF	*Capital One*	*73,555,279,872*	*United States*
WBC.AX	*Westpac Banking*	*69,674,164,224*	*Australia*
NAB.AX	*National Australia Bank*	*66,084,303,655*	*Australia*
BMO.TO	*Bank of Montreal*	*65,546,137,600*	*Canada*
SAN.MC	*Santander*	*63,512,920,064*	*Spain*
D05.SI	*DBS*	*58,454,736,896*	*Singapore*
ANZ.AX	*ANZ Bank*	*57,771,939,417*	*Australia*
INGA.AS	*ING*	*57,179,029,504*	*Netherlands*
UBSG.SW	*UBS*	*55,818,297,344*	*Switzerland*

ISP.MI	***Intesa Sanpaolo***	*55,673,331,712*	*Italy*
NDA-FI.HE	***Nordea Bank***	*52,719,791,735*	*Finland*
CM.TO	***CIBC***	*50,701,774,848*	*Canada*
8316.T	***Sumitomo Mitsui Financial Group***	*48,241,459,200*	*Japan*
BK	***Bank of New York Mellon***	*45,351,161,856*	*United States*
LLOY.L	***Lloyds Banking Group***	*45,009,866,752*	*United Kingdom*
BBVA.MC	***Banco Bilbao Vizcaya Argentaria***	*44,675,006,464*	*Spain*
BARC.L	***Barclays***	*43,581,313,024*	*United Kingdom*
ACA.PA	***Crédit Agricole***	*43,253,817,344*	*France*
O39.SI	***OCBC Bank***	*38,047,156,397*	*Singapore*
8411.T	***Mizuho Financial Group***	*36,191,977,472*	*Japan*
NWG.L	***NatWest Group***	*35,183,005,696*	*United Kingdom*
0011.HK	***Hang Seng Bank***	*32,798,741,762*	*Hong Kong*
2388.HK	***Bank of China (Hong Kong)***	*31,904,874,331*	*Hong Kong*
U11.SI	***UOB***	*31,867,148,598*	*Singapore*
STT	***State Street Corporation***	*31,294,013,440*	*United States*
UCG.MI	***UniCredit***	*29,819,087,234*	*Italy*
GLE.PA	***Société Générale Société***	*26,931,077,315*	*France*
DBK.DE	***Deutsche Bank***	*26,322,493,440*	*Germany*
NA.TO	***National Bank of Canada***	*25,726,072,592*	*Canada*
CSGN.SW	***Credit Suisse***	*24,012,253,184*	*Switzerland*
HBAN	***Huntington Bancshares***	*22,960,508,928*	*United States*
8591.T	***ORIX***	*22,850,541,568*	*Japan*
SWED-A.ST	***Swedbank***	*22,668,140,544*	*Sweden*
KEY	***KeyCorp***	*20,915,812,352*	*United States*
MTB	***M&T Bank***	*19,725,049,856*	*United States*
EBS.VI	***Erste Group Bank***	*17,886,908,225*	*Austria*
DANSKE.CO	***Danske Bank***	*14,454,364,083*	*Denmark*
STAN.L	***Standard Chartered***	*14,272,246,784*	*United Kingdom*
BAER.SW	***Julius BÃ¤r***	*14,237,941,396*	*Switzerland*
SUN.AX	***Suncorp***	*11,621,976,817*	*Australia*
EWBC	***East West Bancorp***	*11,125,363,712*	*United States*
8308.T	***Resona Holdings***	*9,703,317,873*	*Japan*
8601.T	***Daiwa Securities Group***	*8,886,962,180*	*Japan*
CBK.DE	***Commerzbank***	*8,430,365,755*	*Germany*
CBSH	***Commerce Bancshares***	*8,230,009,856*	*United States*
BMED.MI	***Banca Mediolanum***	*8,009,868,033*	*Italy*
GBCI	***Glacier Bancorp***	*5,355,588,608*	*United States*
BKT.MC	***Bankinter***	*5,301,172,596*	*Spain*
8337.T	***Chiba Bank***	*4,973,784,064*	*Japan*
SAB.MC	***Banco Sabadell***	*4,710,816,798*	*Spain*

0023.HK	***Bank of East Asia***	*4,707,853,001*	*Hong Kong*
INDB	***Independent Bank***	*2,557,478,912*	*United States*
WSFS	***WSFS Financial***	*2,454,242,304*	*United States*
FRME	***First Merchants Corporation***	*2,285,708,288*	*United States*
FFBC	***First Financial Bank***	*2,266,230,528*	*United States*
TOWN	***TowneBank***	*2,262,498,304*	*United States*
WSBC	***WesBanco***	*2,240,403,712*	*United States*
FMBI	***First Midwest Bancorp***	*2,187,111,424*	*United States*
SASR	***Sandy Spring Bank***	*2,183,318,016*	*United States*
HTLF	***Heartland Financial USA***	*2,053,779,584*	*United States*
RNST	***Renasant Corp***	*2,051,216,512*	*United States*
TRMK	***Trustmark***	*2,045,401,216*	*United States*
PRK	***Park National Corp***	*2,043,557,760*	*United States*
BANF	***BancFirst***	*1,990,674,944*	*United States*
SBCF	***Seacoast Banking***	*1,972,694,528*	*United States*
BANR	***Banner Bank***	*1,917,446,528*	*United States*
EGBN	***Eagle Bancorp***	*1,862,989,184*	*United States*
PFS	***Provident Financial Services***	*1,858,709,120*	*United States*
HOPE	***Hope Bancorp***	*1,797,997,696*	*United States*
LKFN	***Lakeland Financial Corp***	*1,789,223,040*	*United States*
CASH	***Meta Financial Group***	*1,711,383,808*	*United States*
NWBI	***Northwest Bank***	*1,708,546,944*	*United States*
CFFN	***Capitol Federal Savings Bank***	*1,615,321,856*	*United States*
NBTB	***NBT Bancorp***	*1,576,121,728*	*United States*
SYBT	***Stock Yards Bancorp***	*1,566,275,456*	*United States*
BUSE	***First Busey***	*1,408,945,792*	*United States*

Table 7 Testing Accuracy Scores of Features Set 1 (the underperforming periods)

Testing Period	Event	Start Date	End Date	Features set 1					Features Set 2					DIff
				RF1	XG1	LR1	AVR1	ANN1	RF2	XG2	LR2	AVR2	ANN2	AVR2-AVR1
2007-2008	**Global Crisis**	10/04/2008	05/08/2008	49.28%	49.13%	49.00%	49.14%	N/A	50.54%	51.77%	51.79%	51.37%	N/A	2.23%
				50.51%	47.67%	50.10%	49.43%	N/A	49.92%	49.18%	49.77%	49.63%	N/A	0.20%
2008-2009	**US-Bear Market**	01/03/2009	01/04/2009	50.10%	49.69%	49.41%	49.74%	50.41%	51.66%	50.54%	51.00%	51.07%	49.59%	1.33%
2009-2010	**Europe debt crisis (1)**	13/10/2009	30/11/2009	49.77%	50.13%	49.44%	49.78%	50.69%	50.59%	50.03%	50.61%	50.41%	48.82%	0.63%
	US-Flash Crash + (1)	11/03/2010	04/05/2010	49.74%	49.67%	50.56%	49.99%	49.10%	49.92%	51.05%	50.64%	50.54%	50.92%	0.55%
2010-2011	**-1**	05/05/2010	01/07/2010	49.08%	49.39%	50.61%	49.69%	50.95%	49.28%	50.61%	50.61%	50.17%	50.59%	0.48%
2011-2012	**Black Monday**	14/09/2011	04/11/2011	49.92%	50.00%	49.67%	49.86%	50.36%	49.26%	50.56%	50.51%	50.11%	50.26%	0.25%
		26/12/2011	13/02/2012	49.18%	49.74%	50.08%	49.67%	49.28%	49.54%	50.03%	50.21%	49.92%	50.56%	0.26%
2012-2013(YB)		14/02/2012	03/04/2012	49.87%	49.26%	50.21%	49.78%	50.15%	49.23%	50.46%	51.00%	50.23%	50.69%	0.45%
2013-2014	**Bear Oil Market**	26/06/2014	22/08/2014	48.41%	48.62%	50.61%	49.22%	50.23%	48.82%	50.54%	50.51%	49.96%	50.56%	0.74%
2014-2015		19/10/2014	05/12/2014	49.13%	49.51%	49.64%	49.43%	50.97%	50.49%	51.74%	50.56%	50.93%	50.61%	1.50%
2015-2016	**Chinese Market Crash**	04/09/2015	03/02/2016	48.80%	46.98%	47.06%	47.61%	50.51%	50.28%	50.56%	50.31%	50.38%	50.64%	2.77%
				49.36%	50.69%	49.49%	49.85%	50.69%	50.26%	50.64%	50.82%	50.57%	50.69%	0.73%
				48.69%	50.64%	50.64%	49.99%	50.67%	49.95%	50.64%	50.44%	50.34%	50.67%	0.35%
2016-2017	**US-Market Sells Off**	07/11/2016	26/12/2016	49.64%	49.03%	49.54%	49.40%	50.79%	51.51%	50.31%	52.07%	51.30%	50.56%	1.89%
		06/04/2017	02/06/2017	49.41%	49.28%	50.61%	49.77%	50.31%	49.16%	49.23%	49.69%	49.36%	50.46%	-0.41%
2017-2018	**US-China Trade-war**	08/01/2018	18/04/2018	49.36%	50.54%	49.21%	49.70%	50.33%	49.92%	49.97%	48.67%	49.52%	50.44%	-0.18%
				51.20%	50.54%	50.54%	50.76%	50.54%	48.54%	49.31%	50.56%	49.47%	50.51%	-1.29%
2018-2019	**Crypto-Crash**	18/06/2018	10/10/2018	49.64%	48.54%	50.54%	49.57%	50.28%	49.87%	49.95%	50.28%	50.03%	50.54%	0.46%
				49.05%	48.52%	50.05%	49.21%	50.49%	50.59%	50.67%	50.69%	50.65%	50.56%	1.44%
		29/11/2018	16/01/2019	48.72%	48.59%	50.54%	49.28%	50.51%	50.31%	50.54%	50.00%	50.28%	50.54%	1.00%
2020	**COVID**	17/03/2020	11/05/2020	48.13%	47.47%	47.80%	47.80%	50.54%	50.15%	49.77%	49.97%	49.97%	50.54%	2.17%
Whole Sample														
Average							50.31%	50.37%				50.35%	50.44%	0.04%
Min							47.61%	48.77%				48.69%	48.82%	-2.29%
Max							52.44%	50.97%				51.49%	50.99%	2.77%

Note: The table depicts the all underperforming periods of the three learners (RF,XGB,LR) using features set1 by the average accuracy score of the three learners of every testing periods. The score is compared against the features set 2 in the same period. While stacked ANN is not included in the average score as it is already an ensemble method itself.

Table 19 Fama Fench's 5 Factors Regression (daily average return is the percentage values, and the values initiated with double asterisks are the t-statistics of the independent variables, without asterisks is the coefficient of the independent variables

	Without Connectedness					With Connectedness					
	RF1	XG1	LR1	Combined1	ANN1	RF2	XG2	LR2	Combined2	ANN2	Benchmark
Daily avg. ret	0.0118	0.03115	0.01149	0.0063	0.0605	0.0096	0.0301	0.01	0.013	0.0638	0.0333
R squared	0.257	0.202	0.24	0.253	0.172	0.268	0.212	0.269	0.267	0.186	0.399
Alpha	0.0106	0.0301	0.0109	0.0074	0.0435	0.008	0.0304	0.0117	0.0069	0.0457	0.0385
	**1.719	**1.807	**1.741	**1.203	**0.951	**1.352	**2.139	**1.96	**1.135	**0.987	**0.02
MKT	0.1364	0.3181	0.1283	0.135	0.8287	0.1337	0.2407	0.1365	0.1358	0.95	0.5277
	**17.233	**14.917	**16.05	**17.075	**14.127	**17.545	**13.19	**17.835	**17.438	**16.018	**0.026
SMB	-0.0358	-0.062	-0.0371	-0.0357	-0.2527	-0.0348	-0.1401	-0.0182	-0.0404	-0.1875	-0.2774
	**0.017	**-1.319	**-2.108	**-2.048	**-1.955	**-2.071	**-3.485	**-1.077	**-2.356	**-1.435	**0.057
HML	0.0892	0.1557	0.088	0.0816	0.3581	0.0763	0.2632	0.0693	0.0754	0.1265	0.6994
	**4.577	**2.966	**4.473	**4.191	**2.479	**4.066	**5.859	**3.679	**3.93	**0.866	**0.063
RMW	-0.2461	-0.6531	-0.2535	-0.2539	-1.3294	-0.2723	-0.4977	-0.2917	-0.2764	-1.4272	-0.8691
	**-8.25	**-8.123	**-8.411	**-8.514	**-6.011	**-9.478	**-7.234	**-10.113	**-9.413	**-6.383	**0.097
CMA	0.0325	0.1367	0.0228	0.0443	0.9742	0.0211	-0.0805	0.021	0.0188	1.2298	-0.5236
	**1.081	**1.689	**0.753	**1.476	**4.376	**0.73	**-1.162	**0.723	**0.635	**5.465	**0.097
Obs	2952	2952	2952	2952	2952	2952	2952	2952	2952	2952	2952

Table 20 Summary of cumulative returns (Diff stands for the differences of features set 2 over features set 1 in terms of cumulative return)

	RF	XGB	LR	Combined	ANN	Benchmark
Features Set1	36.73974	96.86174	30.45821	23.9263	212.7951	117.5522
Features Set2	22.40561	96.08957	37.44842	22.56447	224.7732	117.5522
Diff	-14.3341	-0.77217	6.990208	-1.36182	11.97815	

Table 21A Keywords Table for Question A *(The table is a list of terms also used for the main search keywords. In this case, connectedness is searched alone and terms in investment strategy is added later on for papers that contains investment strategy with connectedness)*

Connectedness	**Investment Strategy**
Connection, Relationship, Interdependence	Portfolio, Strategy, Investment
Volatility Connectedness ,Volatility Spillover	Trading Portfolio, Trading Strategy, Trading Investment
Volatility Transmission	Portfolio Implication, Strategy Implication, Investment Implication
Spillover Effects	
Contagion Risks	
Vector Autoregression	
VAR(p)	

Table 22A Keywords Table for Question B *(The table is a list of terms for stock selection in machine learning)*

Stock Selection	**Machine Learning**
Ranking, Selection	Models, Algorithms, Learners
Stock Ranking, Stock Selection	Learning Models, Learning Algorithm
Portfolio	Artificial Intelligence

Table 23A Resources table with acronyms

SOURCE	ACRONYMS
• IEEExplore	IEEE
• ACM Digital library:	ACM
• Google scholar (scholar.google.com)	GGL
• Citeseer library (citeseer.ist.psu.edu)	CSL
• ScienceDirect (www.sciencedirect.com)	SD
• Other Publications	OR

Table 24A Types of Connectedness for Researches Included- Question set A *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)*

Classic VAR	TVP-VAR	LASSO-VAR
308	92	31

Table 25A Types of Trading Strategy and Factor Types for Researches Included- Question set B *(From the inclusion/exclusion criteria, the researches that are included for the review use these types of connectedness)*

	Long Portfolios	Long-Short Portfolios	Both
Fundamental Factors	42	8	6
Technical Factors	13	6	2
Sentimental Factors	0	0	0
Mixed Factors	74	21	12

Table 26A Pairwise directional connectedness in a matrix

	x_1	..	x_N
x_1	d_{11}^H	..	d_{1N}^H
..	..	..	..
x_N	d_{Nj}^H	..	d_{NN}^H

Table 27A Full connectedness matrix

	x_1	..	x_N	***From Others***
x_1	d_{11}^H	..	d_{1N}^H	$\sum_{j=1}^{N} d_{1j}^H \quad \text{with } j \neq 1$
..	..	..	..	..
x_N	d_{Nj}^H	..	d_{NN}^H	$\sum_{j=1}^{N} d_{Nj}^H \quad \text{with } j \neq \text{N}$
To Others	$\sum_{i=1}^{N} d_{i1}^H \quad \text{with } i \neq 1$	..	$\sum_{i=1}^{N} d_{iN}^H \quad \text{with } i \neq 1$	$\frac{1}{N} \sum_{i=1,j=1}^{N} d_{ij}^H \text{ with } i \neq \text{j}$

APPENDIX D: REFERENCES

An, P., Li, H., Zhou, J., Li, Y., Sun, B., Guo, S., & Qi, Y. (2020). Volatility spillover of energy stocks in different periods and clusters based on structural break recognition and network method. *Energy, 191*, 116585.

Asteriou, D., Pilbeam, K., & Sarantidis, A. (2019). The Behaviour of Banking Stocks During the Financial Crisis and Recessions. Evidence from Changes-in-Changes Panel Data Estimations. *Scottish Journal of Political Economy, 66*(1), 154-179.

Audrino, F., & Tetereva, A. (2019). Sentiment spillover effects for US and European companies. *Journal of Banking & Finance, 106*, 542-567.

Ausiello, G., D'Atri, A., & Protasi, M. (1980). Structure preserving reductions among convex optimization problems. *Journal of Computer and System Sciences, 21*(1), 136-153.

Ball, S., Banerjee, A., Berry, C., Boyle, J. R., Bray, B., Bradlow, W., . . . Denniston, A. (2020). Monitoring indirect impact of COVID-19 pandemic on services for cardiovascular diseases in the UK. *Heart, 106*(24), 1890-1897.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4): Springer.

Brauneis, A., & Mestel, R. (2018). Price discovery of cryptocurrencies: Bitcoin and beyond. *Economics Letters, 165*, 58-61.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1*(4).

Coffman, J., & Weaver, A. C. (2012). An empirical performance evaluation of relational keyword search techniques. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 30-42.

Demirer, M., Diebold, F. X., Liu, L., & Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics, 33*(1), 1-15.

Diebold, F. X., Liu, L., & Yilmaz, K. (2017). *Commodity connectedness*. Retrieved from

Diebold, F. X., & Yilmaz, K. (2011). Equity market spillovers in the Americas. *Financial stability, monetary policy, and central banking, 15*, 199-214.

Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting, 28*(1), 57-66.

Diebold, F. X., & Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics, 182*(1), 119-134.

Ding, Z., Nguyen, H., Bui, X.-N., Zhou, J., & Moayedi, H. (2020). Computational intelligence model for estimating intensity of blast-induced ground vibration in a mine based on imperialist competitive and extreme gradient boosting algorithms. *Natural Resources Research, 29*(2), 751-769.

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of economic perspectives, 18*(3), 25-46.

Gabauer, D., & Gupta, R. (2018). On the transmission mechanism of country-specific and international economic uncertainty spillovers: Evidence from a TVP-VAR connectedness decomposition approach. *Economics Letters, 171*, 63-71.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of business*, 67-78.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University, 33*(2004), 1-26.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15.

Koop, G., Pesaran, M. H., & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics, 74*(1), 119-147.

Lee, T. K., Cho, J. H., Kwon, D. S., & Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Systems with Applications, 117*, 228-242.

Lundgren, A. I., Milicevic, A., Uddin, G. S., & Kang, S. H. (2018). Connectedness network and dependence structure mechanism in green investments. *Energy Economics, 72*, 145-153.

Maitra, D., Guhathakurta, K., & Kang, S. H. (2021). The good, the bad and the ugly relation between oil and commodities: An analysis of asymmetric volatility connectedness and portfolio implications. *Energy Economics, 94*, 105061.

Mensi, W., Maitra, D., Vo, X. V., & Kang, S. H. (2021). Asymmetric volatility connectedness among main international stock markets: A high frequency analysis. *Borsa Istanbul Review, 21*(3), 291-306.

Nasreen, S., Tiwari, A. K., & Yoon, S.-M. (2021). Dynamic Connectedness and Portfolio Diversification during the Coronavirus Disease 2019 Pandemic: Evidence from the Cryptocurrency Market.

Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters, 58*(1), 17-29.

Pieters, G., & Vivanco, S. (2017). Financial regulations and price inconsistencies across Bitcoin markets. *Information Economics and Policy, 39*, 1-14.

Qarni, M. O., Gulzar, S., Fatima, S. T., Khan, M. J., & Shafi, K. (2019). Inter-markets volatility spillover in US bitcoin and financial markets. *Journal of Business Economics and Management, 20*(4), 694-714.

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal, 75*(3), 70-88.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society,* 1-48.

Tiryaki, F., & Ahlatcioglu, M. (2005). Fuzzy stock selection using a new fuzzy ranking and weighting algorithm. *Applied Mathematics and computation, 170*(1), 144-157.

Vardharaj, R., & Fabozzi, F. J. (2007). Sector, style, region: Explaining stock allocation performance. *Financial Analysts Journal, 63*(3), 59-70.

Vishwanath, R., & Krishnamurti, C. (2009). *Investment management: A modern guide to security analysis and stock selection*: Springer.

Wen, T., & Wang, G.-J. (2020). Volatility connectedness in global foreign exchange markets. *Journal of Multinational Financial Management, 54,* 100617.

Wolff, D., & Echterling, F. (2020). Stock Picking with Machine Learning. *Available at SSRN 3607845*.

Wright, R. E. (1995). Logistic regression.

Yang, D., & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business, 73*(3), 477-492.

Yegnanarayana, B. (2009). *Artificial neural networks*: PHI Learning Pvt. Ltd.

Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. *Procedia computer science, 31*, 406-412.

Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). An improved glmnet for l1-regularized logistic regression. *The Journal of Machine Learning Research, 13*(1), 1999-2030.

An, P., Li, H., Zhou, J., Li, Y., Sun, B., Guo, S. & Qi, Y. *2020. Volatility spillover of energy stocks in different periods and clusters based on structural break recognition and network method. Energy, 191, 116585.*

Audrino, F. & Tetereva, A. *2019. Sentiment spillover effects for US and European companies. Journal of Banking & Finance, 106, 542-567.*

Ball, S., Banerjee, A., Berry, C., Boyle, J. R., Bray, B., Bradlow, W., Chaudhry, A., Crawley, R., Danesh, J. & Denniston, A. *2020. Monitoring indirect impact of COVID-19 pandemic on services for cardiovascular diseases in the UK. Heart, 106, 1890-1897.*

Benesty, J., Chen, J., Huang, Y. & Cohen, I. *2009. Pearson correlation coefficient. Noise reduction in speech processing. Springer.*

Brauneis, A. & Mestel, R. *2018. Price discovery of cryptocurrencies: Bitcoin and beyond. Economics Letters, 165, 58-61.*

Coffman, J. & Weaver, A. C. *2012. An empirical performance evaluation of relational keyword search techniques. IEEE Transactions on Knowledge and Data Engineering, 26, 30-42.*

Demirer, M., Diebold, F. X., Liu, L. & Yilmaz, K. *2018. Estimating global bank network connectedness. Journal of Applied Econometrics, 33, 1-15.*

Diebold, F. X., Liu, L. & Yilmaz, K. *2017. Commodity connectedness. National Bureau of Economic Research.*

Diebold, F. X. & Yilmaz, K. *2011. Equity market spillovers in the Americas. Financial stability, monetary policy, and central banking, 15, 199-214.*

Diebold, F. X. & Yilmaz, K. *2012. Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of forecasting, 28, 57-66.*

Gabauer, D. & Gupta, R. *2018. On the transmission mechanism of country-specific and international economic uncertainty spillovers: Evidence from a TVP-VAR connectedness decomposition approach. Economics Letters, 171, 63-71.*

Kitchenham, B. *2004. Procedures for performing systematic reviews. Keele, UK, Keele University, 33, 1-26.*

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. & Linkman, S. *2009. Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51, 7-15.*

Koop, G., Pesaran, M. H. & Potter, S. M. *1996. Impulse response analysis in nonlinear multivariate models. Journal of econometrics, 74, 119-147.*

Lundgren, A. I., Milicevic, A., Uddin, G. S. & Kang, S. H. *2018. Connectedness network and dependence structure mechanism in green investments. Energy Economics, 72, 145-153.*

Maitra, D., Guhathakurta, K. & Kang, S. H. *2021. The good, the bad and the ugly relation between oil and commodities: An analysis of asymmetric volatility connectedness and portfolio implications. Energy Economics, 94, 105061.*

Mensi, W., Maitra, D., Vo, X. V. & Kang, S. H. *2021. Asymmetric volatility connectedness among main international stock markets: A high frequency analysis. Borsa Istanbul Review, 21, 291-306.*

Nasreen, S., Tiwari, A. K. & Yoon, S.-M. *2021. Dynamic Connectedness and Portfolio Diversification during the Coronavirus Disease 2019 Pandemic: Evidence from the Cryptocurrency Market.*

Pesaran, H. H. & Shin, Y. *1998. Generalized impulse response analysis in linear multivariate models. Economics letters, 58, 17-29.*

Pieters, G. & Vivanco, S. *2017. Financial regulations and price inconsistencies across Bitcoin markets. Information Economics and Policy, 39, 1-14.*

Qarni, M. O., Gulzar, S., Fatima, S. T., Khan, M. J. & Shafi, K. *2019. Inter-markets volatility spillover in US bitcoin and financial markets. Journal of Business Economics and Management, 20, 694-714.*