INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN ACCOUNTING, FINANCE AND MANAGEMENT SCIENCES

OPEN ACCESS JOURNAL

Vol 15, Issue 3, (2025) E-ISSN: 2225-8329

Applying Data Mining Methods to Predict the Stock Prices of the Top Five Semiconductor Companies in Taiwan

Liang-Ying Wei*, She-Ting Yeh

Graduate Institute of Digital Technology Innovation and Management Yuanpei University of Medical Technology, Hsin-Chu, Taiwan

*Corresponding Author Email: lywei@mail.ypu.edu.tw

DOI Link: http://dx.doi.org/10.6007/IJARAFMS/v15-i3/26358

Published Online: 03 September 2025

Abstract

Stock trading is a financial tool for increasing passive income. The semiconductor industry has development potential in Taiwan and is a popular stock investment target for investors. This study uses the stock prices of the top five semiconductor companies with the highest annual revenue in Taiwan in 2021 as the data source for stock price prediction. The daily closing prices of the five companies from 2012 to 2021 are collected as experimental data. The data is divided into training data sets and test data sets on an annual basis. Time series analysis is combined with data mining methods such as random forest, neural network, support vector regression, Gaussian process regression, and k-nearest neighbor regression to predict stock prices. MAE (Mean absolute error) and root mean squared error (Root mean squared error) are used to be prediction error evaluation metrics. The changes in error values generated by different prediction models in each test data set were analyzed and compared. Experiments revealed that due to the impact of COVID-19 and the Sino-US trade war, the stock prices of five semiconductor companies rose rapidly from 2020 to 2021. An overall analysis of the stock price prediction results of five algorithms for different companies revealed that the predictive performance of the stock prediction models, ranked from best to worst, was support vector regression (SVR), neural network (NN), random forest (RF), K-nearest neighbor regression (KNN), and Gaussian process regression (GPR). Among them, support vector regression achieved superior predictive performance across different industries and in both low and high stock price volatility stock price predictions. The results of this study will serve as an objective reference for investors when making investments.

Keyword: Semiconductor Industry, Stock Price, Data Mining

Introduction

Today, our lives are filled with numerous technological gadgets that make our lives more convenient. A key component of these products is chips produced by the semiconductor industry. These chips are found in our daily mobile phones, computers, household appliances, cars, and motorcycles. Investing is a financial management method that generates passive income. Common investment options include funds, exchange rates, stocks, real estate, and gold. Stock investment is the most popular among Taiwanese people. As of September 2012, nearly 12 million Taiwanese investors had opened accounts. The potential for development in a country's industry often increases investor interest. The semiconductor industry, with its substantial annual revenue and promising development prospects, is a popular investment choice, sparking a wave of investment since its inception in 1990.

As environmental and current events change, they are reflected in stock market figures. While we cannot guarantee what will happen next, we can use past facts and data mining to predict the likelihood of future events and prepare for them. This study aims to determine whether Taiwan's semiconductor industry stock prices will fluctuate with current events, or whether it is possible to predict tomorrow's stock price drop. Therefore, the objectives of this study are as follows:

1. To examine the reasons for the dramatic stock price fluctuations of Taiwan's top five semiconductor companies by annual revenue over the past decade.
2. To predict the potential for stock price increases and decreases in the future, providing investors with an objective reference for stock purchases and evaluation based on this data.

Related Works

Time Series Analysis

Because historical stock prices are time series data, time series analysis is introduced. A time series is a set of data arranged in chronological order. Its purpose is to understand the correlation between different time points. Time series data properties are categorized into the following four types:

Stability: The average value of data over a period of time remains constant.
Trend: The tendency of data to change, rising or falling slowly over time.
Cyclist: Data changes in a cycle, producing highs and lows with regular fluctuations.
Randomness: Data fluctuates randomly.

Time series with stability, trend, and cyclicality are often used to develop predictive models because they are stable and regular. However, random time series are less suitable for developing predictive models because they lack regular patterns.

Random Forest

The term random forest was first coined by Tin Kam Ho (1995), and Leo Breiman (2001) refined the algorithm. A random forest consists of multiple decision trees, each of which is independent. A random forest randomly selects features from the data collected by multiple decision trees and calculates the average of the results. This yields more accurate data than a single decision tree.

First, the data is divided into training data and prediction data. Suppose there are N samples, each with M features. Then, data are randomly sampled with replacement (indicating the possibility of repeated sampling) to form N data sets. Approximately 36.8% of the initial samples will not appear in the training data set. These unsampled samples are called out-of-bag (OOB) and can be compared with the calculated predictions to obtain the calculated error. The algorithm then randomly selects m features for splitting, where $m < M$, and grows N decision trees. These N decision trees generate predictions. For classification data, majority voting is used, while for regression, averaging is used.

Advantages and Disadvantages of Random Forest:
Advantages:
1. Can efficiently handle large datasets
2. Does not reduce accuracy when handling missing values
3. Can handle imbalanced data
4. Unsampled data can be used as validation data

Disadvantages:
1. Requires large storage capacity
2. Unknown sample features used for decision making

Artificial Neural Networks

An artificial neural network (ANN) was proposed by Warren Sturgis McCulloch and Walter Pitts (1943) as a machine capable of thinking and learning automatically, just like the human brain. The human brain is able to think and learn primarily because it contains numerous cells called neurons. The concept of an ANN stems from this, simulating the operation of neurons in the brain.

The brain's nervous system is composed of approximately 100 billion neurons. These neurons are interconnected, with organs responsible for receiving signals and transmitting them between neurons, ultimately allowing organisms to respond to their environment. An ANN is primarily composed of many nodes, which are analogous to the neurons in the biological brain. The ANN's operation mimics the computational and information-transmission capabilities of neurons, enabling organisms to learn and solve problems automatically.

How Neural Networks Work

After a period of early development, neural networks encountered a bottleneck. This bottleneck was resolved with the emergence of backpropagation neural networks. Although it seemed to receive little attention at the time, it was not until Parker, Paul Werbos, and Ronald J. Williams (1985) re-proposed this architecture that it became mainstream and has remained so to this day. The backpropagation neural network architecture is divided into the following three layers:

1. Input Layer

The input layer is responsible for receiving data and input information in the neural network algorithm architecture and is usually represented by a single layer.

2. Hidden Layer

This layer lies between the input and output layers. The number of nodes in this layer is not fixed; the optimal number is usually determined through experimentation. A large number of hidden layers generally indicates a higher complexity of the problem being addressed by the neural network algorithm, but too many hidden layers can be counterproductive.

3. Output Layer

The output layer is responsible for providing data output in the neural network algorithm architecture and is usually represented by a single layer.

Support Vector Regression

Support vector machines (SVMs) are a new machine learning theory proposed by Boser et al. (1992). They work by finding a hyperplane that maximizes the distance to the nearest sample. The point closest to the sample is the support vector. The greater the separation of the data in the vector, the more accurate the classification.

Support Vector Machine Operation:

Suppose the data points are (x_1, y_1), …, (x_i, y_i), with $x_i \in R^n$ and $y_i \in \{1,-1\}$. Find a hyperplane $f(x)=w^T+b$ such that $y_i=1$ when $f(x)>0$ and $y_i=-1$ when $f(x)<0$. Using $f(x)$, we can distinguish which set the data belongs to. Based on these conditions, the equations are: H1: $w^T+b=1$, H2: $w^T+b=-1$. The distance from H1 to the hyperplane is $|1-b|/\|w\|$, and the distance from H2 to the hyperplane is $|-1-b|/\|w\|$. The resulting margin is $2/\|w\|$.

Support vector regression is derived from support vector machines. Its principle is to find a hyperplane that minimizes the distance between the farthest samples. Support vector regression creates an interval band on both sides of the linear function, with a spacing of ε. All samples within the interval band are not subject to loss; only the support vectors contribute to the model.

Gaussian Process Regression

Gaussian process regression is a matrix-free nonlinear regression method. Suppose the data is (x_i, y_i), where x_i is the real output and y_i is the eigenvector. Assume that for any *n* points, the output of these *n* points will follow a multivariate normal distribution, and the variance matrix of the normal distribution will be the function of the eigenvector.

How Gaussian Process Regression Works:

Modern PC hardware can only process hundreds of thousands of data points simultaneously. To overcome this limitation, Gaussian process regression uses a gradient boosting model approximation method to predict data. In the first round of training, a fixed-size dataset is randomly selected from the training data. A small Gaussian process regression model is constructed and the prediction error of this model is calculated. Then, a new dataset of the same size is selected to predict the error of the previous round. This process is repeated until the specified number of model data is reached. Finally, the previously constructed models are accumulated. Because the data size of each dataset is fixed, problems such as space and time complexity are avoided.

K-Nearest Neighbor Regression

K-Nearest Neighbor classification is a statistical method without a matrix. Its classification method is based on a majority vote of the nearest neighboring data.

How K-Nearest Neighbor Classification Works:

Assume that the green circle represents *k*. When *k* equals 3, the three nearest records are searched, resulting in *k* being classified as a blue triangle. When *k* equals 5, *k* is classified as a red square. K-Nearest Neighbor regression is a derivative of k-Nearest Neighbor classification. When *k* equals 2, the two nearest records are searched, and the average of these two records is the predicted value.

Research Methods

This study focuses on the stock prices of the top five semiconductor companies with the highest annual revenue in Taiwan in 2021, namely TSMC (Taiwan Semiconductor Manufacturing Company Limited), MediaTek (MediaTek Inc.), ASE (ASE Technology Holding Co., Ltd.), UMC (United Microelectronics Corporation), and Novatek (Novatek Microelectronics Corp).

The research data is based on the daily closing prices of five companies from 2012 to 2021. The data was sorted in advance based on years, with the days divided into *t-2, t-1, t*, and *t+1*. The daily closing prices are arranged in the order of *t-2, t-1, t*, and *t+1*. If the stock market is closed on a certain day, the closing price of the company on the next day will be used to make up the difference. The sorted data is divided into two parts. The first half is from January to October, which is used for data training. The second half is from November to December, which is used for data prediction.

This study uses random forest, neural network, support vector regression, Gaussian process regression, and k-nearest neighbor regression analysis to compare the accuracy of five different algorithms. The predictive performances of different algorithms are calculated by evaluation metrics (Mean absolute error, Root mean squared error) and smaller values indicate lower error.

Mean Absolute Error

This is the absolute difference between the predicted value and the actual value, as shown below.

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_1 - x_1|}{n}$$

y_1: predicted value
x_1: actual value
n: number of actual values

Root Mean Squared Error

This is the square root of the difference between the predicted value and the actual value, as shown below.

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_1 - x_1)^2}{n}}$$

y_1: predicted value
x_1: actual value
n: number of actual values

Experiments and Comparisons

TSMC Research Results

As shown in Figure 4-1, TSMC's stock price exhibited a slow but steady upward trend from 2012 to 2018. From 2018 to 2020, although with slightly greater fluctuations, it also continued an upward trend. There was a slight dip in early 2020, followed by a rapid rise. There were significant fluctuations in 2021, but over the long term, it has remained flat.

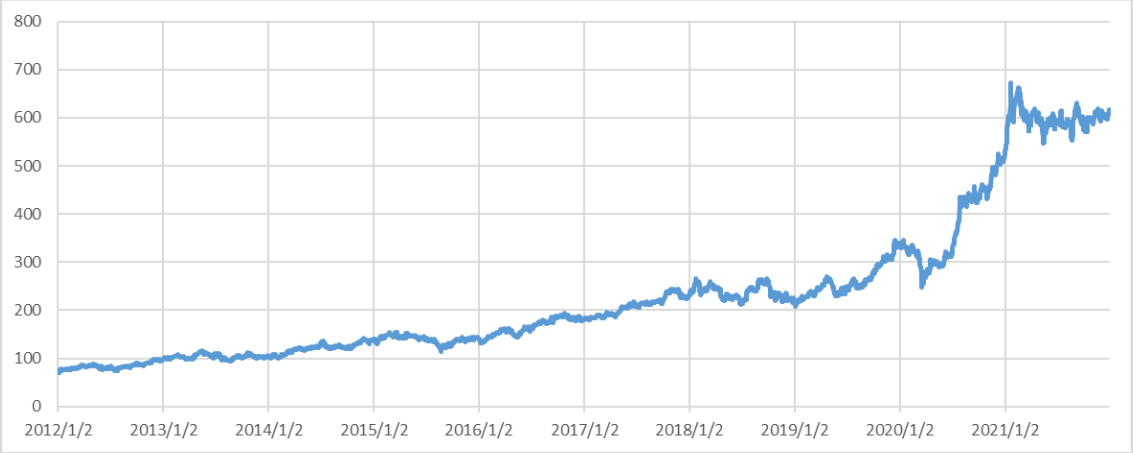

Figure 4-1 TSMC's stock price fluctuations from 2012 to 2021

TSMC's fluctuation chart from 2012 to 2021 is divided into ten years with one year as the unit for comparison, as shown in Figure 4-2.

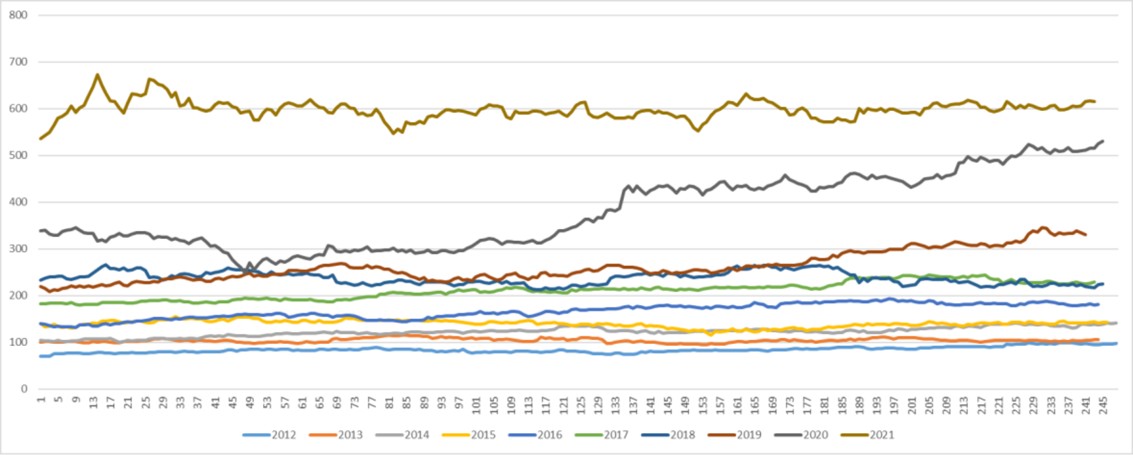

Figure 4-2 Comparison of TSMC's stock price fluctuations from 2012 to 2021

Table 4-1

TSMC's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
stock price difference	29.4	21.1	41	39.5	61.5	64.5	54	137	282	137

Table 4-2
TSMC MAE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	5.6848	1.1907	4.3701	1.7934	2.5732
NN	5.7868	1.2448	1.8427	2.6389	1.7067*
SVR	0.7886*	1.0181*	1.4642*	1.5249*	1.7110
GPR	10.3970	1.9987	11.8032	1.7872	9.5151
KNN	5.5795	1.6113	4.3250	2.1220	3.9268
	2017	2018	2019	2020	2021
RF	5.6031	3.1301	20.6369	39.4512	6.5254
NN	2.1996	3.3266	14.8441	13.4213	6.5416
SVR	2.1518*	2.8736*	3.6284*	5.8288*	4.4551*
GPR	13.2584	17.0901	49.6572	77.5513	4.7300
KNN	5.5375	5.4295	19.8625	42.9390	7.1220

* Best prediction among 5 algorithms

Table 4-3
TSMC RMSE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	6.5563	1.3951	4.9118	2.2379	3.1140
NN	6.3280	1.5443	2.3073	2.9558	2.1474*
SVR	1.1183*	1.2322*	1.9539*	1.8867*	2.2560
GPR	10.7683	2.3184	12.1184	2.3532	9.8347
KNN	6.4336	1.6113	5.0336	2.5686	4.3673
	2017	2018	2019	2020	2021
RF	6.7676	4.0686	24.5129	44.0723	7.8237
NN	2.9432	4.1988	17.6538	15.6644	7.8709
SVR	2.8978*	3.6515*	4.8467*	7.2956*	5.5870*
GPR	14.2190	17.6400	50.8880	79.2435	5.9242
KNN	6.3487	6.5419	23.8646	47.8071	7.1220

* Best prediction among 5 algorithms

MediaTek Research Results

As shown in Figure 4-3, MediaTek's stock price showed a slow but steady upward trend from 2012 to mid-2014, then slowly declined until 2016. There was a slight increase from mid-2017 to early 2018, and from 2019 to early 2021, it saw a rapid but volatile rise. Despite a mid-year decline, it rebounded by the end of the year.

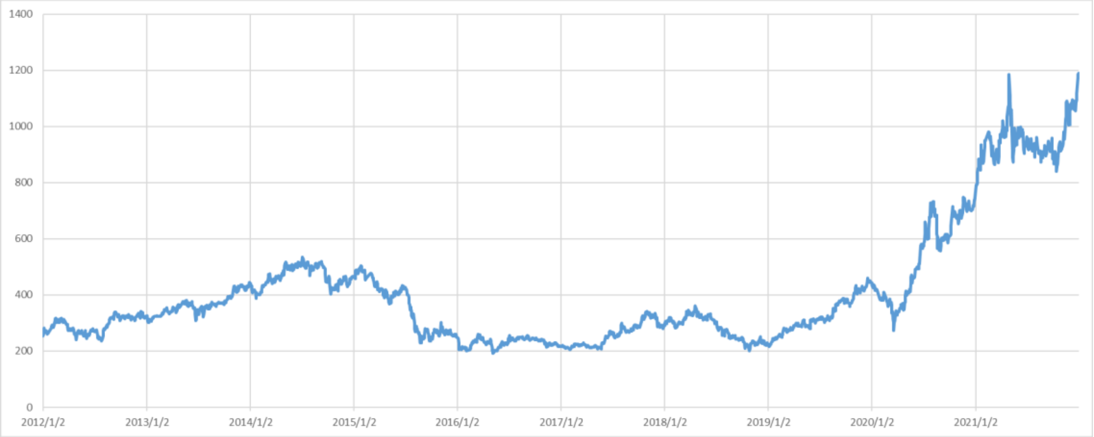

Figure 4-3 MediaTek's stock price fluctuations from 2012 to 2021

The fluctuation chart of MediaTek from 2012 to 2021 is divided into ten years for comparison, as shown in Figure 4-4.

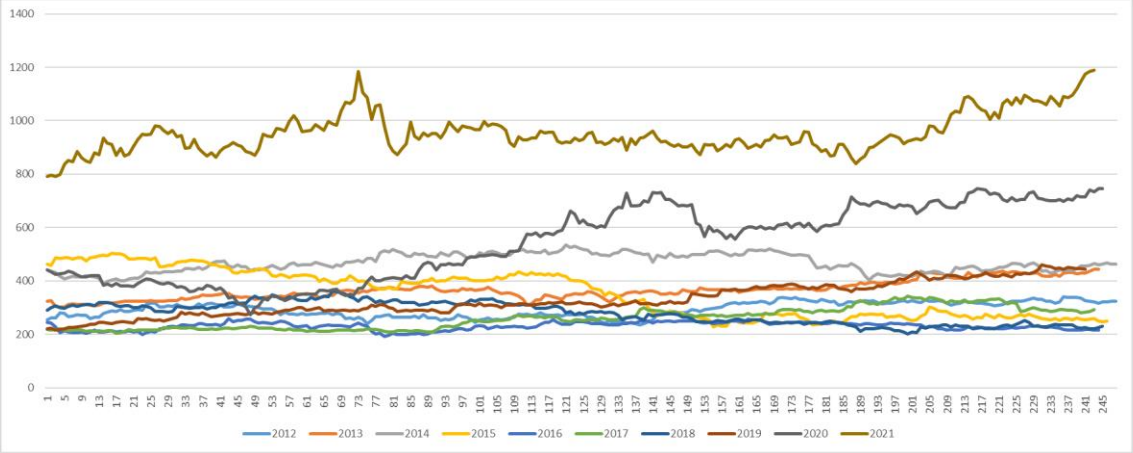

Figure 4-4 Comparison of MediaTek's stock price fluctuations from 2012 to 2021

Table 4-4

MediaTek's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
stock price difference	104	143	147	275	68	138	160	243.5	473	398

Table 4-5

MediaTek MAE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	5.0883	32.8250	5.8650	6.8694	4.6180
NN	4.2650	8.1595	6.2718	4.9613*	4.5613
SVR	3.8813*	4.3569*	5.7432*	5.0887	2.8544*
GPR	18.1198	55.7080	31.6186	116.1985	13.2619
KNN	5.8846	37.7000	7.7750	11.6220	5.3780
	2017	2018	2019	2020	2021
RF	5.5627	9.0391	28.2805	13.5435	32.0370
NN	5.1514*	4.1634	18.5592	24.2898	29.7336
SVR	5.2524	3.9487*	5.4716*	9.2929*	25.1157*

GPR	21.0538	63.1824	80.2290	104.4962	100.8158
KNN	9.8250	12.8205	20.7625	20.4390	78.0000

* Best prediction among 5 algorithms

Table 4-6

MediaTek RMSE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	6.4075	33.8476	8.2268	8.6446	5.6195
NN	5.3862*	9.5688	8.8037	6.6561	5.4200
SVR	5.4145	5.8123*	8.0833*	5.8437*	3.6393*
GPR	19.3245	56.1566	33.3468	116.4830	13.9054
KNN	7.6254	38.8112	9.9649	15.2329	6.6337
	2017	2018	2019	2020	2021
RF	8.4757	10.5237	32.1849	17.1311	44.7620
NN	7.3686*	5.4569	21.3243	26.9919	38.4250
SVR	7.4854	5.2345*	6.9866*	12.5813*	31.8689*
GPR	24.3967	63.4660	81.3223	105.7977	113.4649
KNN	13.1525	15.5126	25.6090	24.1832	89.0037

* Best prediction among 5 algorithms

ASE Holdings Research Results

As shown in Figures 4-5, ASE Holdings experienced a slow upward trend from 2012 to early 2015, followed by a decline until mid-2015, a decline from early 2018 to 2019, an increase until 2020, and a subsequent decline. It then experienced a rapid but volatile rise at the end of 2020, followed by a sharp decline at the end of 2021.

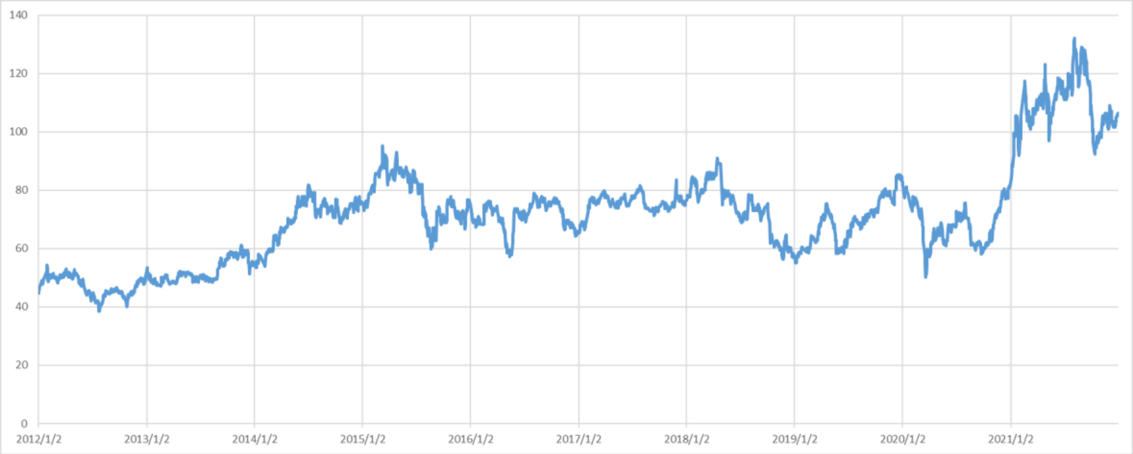

Figure 4-5 ASE Holding's stock price fluctuations from 2012 to 2021

The fluctuation chart of ASE Technology Holding from 2012 to 2021 is divided into ten years with one year as the unit for comparison, as shown in Figure 4-6.

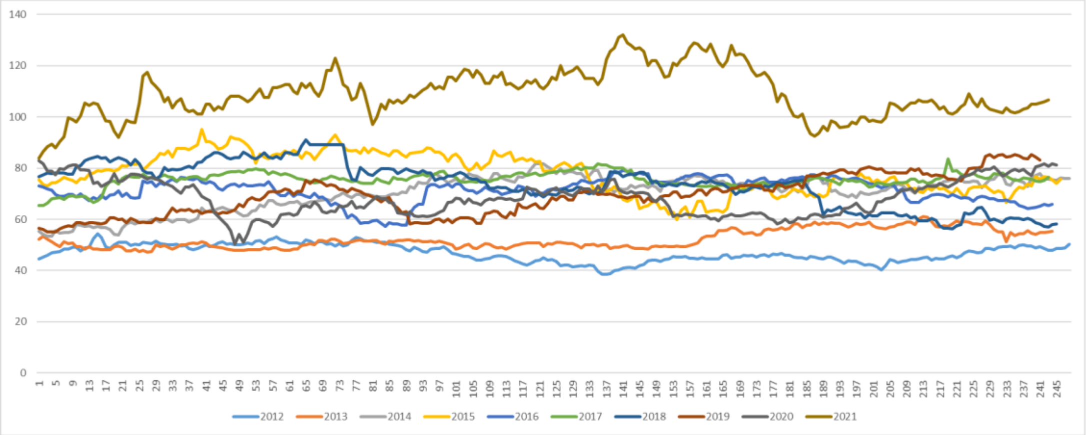

Figure 4-6 Comparison of ASE Holding's stock price fluctuations from 2012 to 2021

Table 4-7

ASE Holding's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
stock price difference	15.92	14.05	28.31	35.39	21.83	18.1	34.6	30.4	32.6	48.1

Table 4-7

ASE Holding's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016
RF	0.6651	0.9454	1.2023	1.2790	1.3276
NN	0.5927	0.9362	0.9340*	1.3938	0.8096
SVR	0.5198*	0.8588*	0.9830	1.2613*	0.7271*
GPR	1.5073	2.5954	1.9865	8.3843	4.9090
KNN	0.8600	1.2497	1.6555	1.8437	2.4771
	2017	2018	2019	2020	2021
RF	0.8030*	2.3953	2.8930	1.3119	1.2369*
NN	0.8298	1.2724	2.1182	1.6913	2.0186
SVR	0.8129	0.8509*	0.8307*	1.0962*	1.2505
GPR	1.1876	16.6970	8.3709	6.4902	8.1848
KNN	0.8925	2.1897	3.1300	1.9293	2.6585

* Best prediction among 5 algorithms

Table 4-9

ASE Holdings RMSE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	0.7602	1.3053	1.5759	1.6405	1.7385
NN	0.6945	1.3186	1.2479*	1.7924	1.0853
SVR	0.6448*	1.1888*	1.3099	1.5636*	0.9954*
GPR	1.8777	2.9365	2.3094	8.6262	5.3056
KNN	1.0476	1.5597	2.1364	2.4502	3.4643
	2017	2018	2019	2020	2021
RF	1.3126*	2.8084	3.8432	1.5370	1.6347*
NN	1.3278	1.5566	2.7235	1.9328	2.3849
SVR	1.4034	1.2380*	1.1777*	1.3398*	1.6636
GPR	1.6674	16.7766	8.8148	7.3104	8.3870
KNN	1.5046	2.5783	4.0697	2.3845	3.3749

* Best prediction among 5 algorithms

UMC Research Results

As shown in Figures 4-7, UMC's stock price remained flat from 2012 to mid-2020, then rose sharply until 2021. From 2021 to the end of the year, it continued to rise, but with significant fluctuations

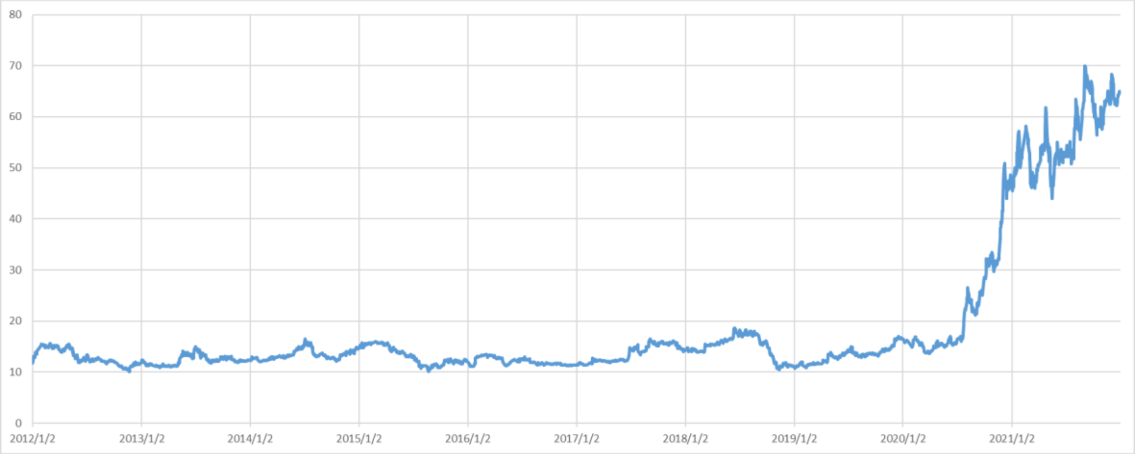

Figure 4-7 UMC’s stock price fluctuations from 2012 to 2021

The fluctuation chart of UMC from 2012 to 2021 is divided into ten years with one year as a unit for comparison, as shown in Figure 4-8.

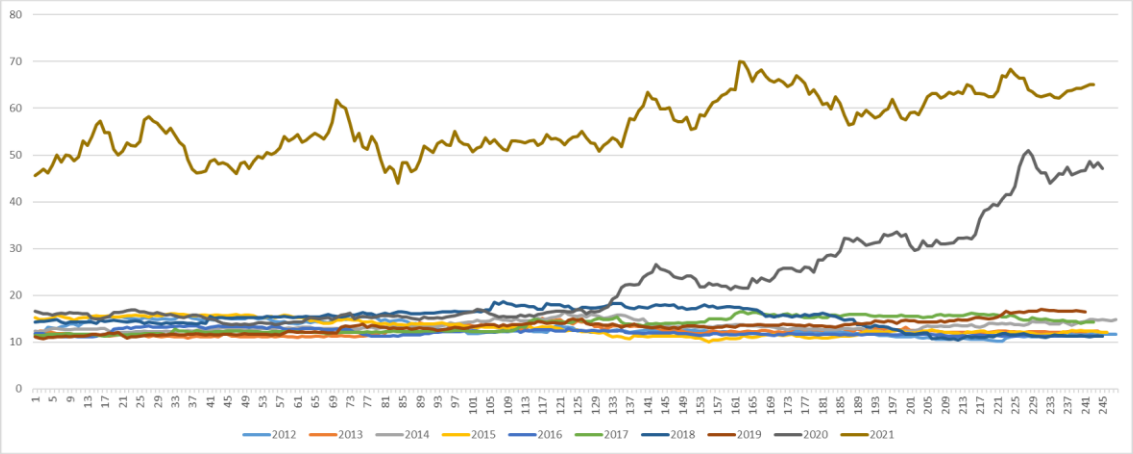

Figure 4-8 Comparison of UMC’s stock price fluctuations from 2012 to 2021

Table 4-10

UMC’s stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
stock price difference	5.55	4.15	4.5	6	2.4	5.2	8.2	6.3	37.25	26.05

Table 4-10

UMC's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016
RF	0.2667	0.0969	0.2535	0.1952	0.0607
NN	0.2646	0.1013	0.2040	0.1665	0.0588
SVR	0.1436*	0.0817*	0.1995*	0.1661*	0.0539*
GPR	2.1650	0.5875	0.2997	1.7092	0.8898
KNN	0.3346	0.1988	0.3413	0.2530	0.1073

	2017	2018	2019	2020	2021
RF	0.2019	0.5867	1.1719	8.9681	1.4474
NN	0.1684	0.4565	0.5920	6.0600	1.2760
SVR	0.1640*	0.1458*	0.1721*	1.0425*	0.8736*
GPR	0.4762	4.5556	1.8924	8.4357	4.9123
KNN	0.2850	0.5821	1.5144	8.8963	2.2341

* Best prediction among 5 algorithms

Table 4-12

UMC RMSE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	0.3364	0.1197	0.3043	0.2567	0.0869
NN	0.2984	0.1199	0.2580	0.2182	0.0787
SVR	0.1982*	0.1079*	0.2530*	0.2108*	0.0774*
GPR	2.2039	0.5969	0.3483	1.7235	0.8944
KNN	0.3952	0.2312	0.4193	0.3265	0.1502

	2017	2018	2019	2020	2021
RF	0.2838	0.6623	1.4701	10.8810	1.7654
NN	0.2186	0.4986	0.6628	7.4323	1.4381
SVR	0.2167*	0.1835*	0.2520*	1.4290*	1.2144*
GPR	0.5809	4.5634	2.0594	9.1607	5.1314
KNN	0.3691	0.6567	1.8538	10.8206	2.7977

* Best prediction among 5 algorithms

Novatek Research Results

As shown in Figure 4-9, Novatek's stock price remained flat from 2012 to 2019, then rose sharply from 2019 to early 2021, followed by a sharp decline with significant volatility.

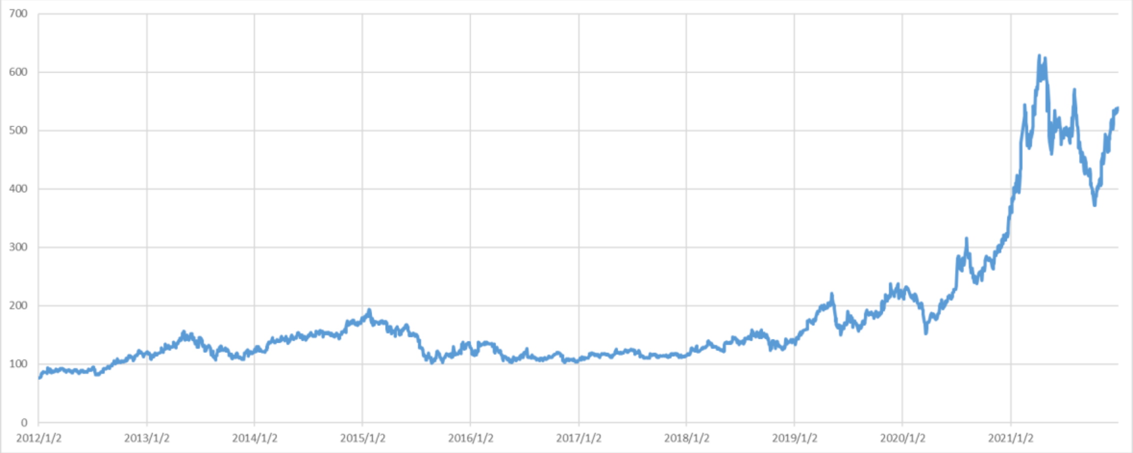

Figure 4-9 Novatek stock price fluctuations from 2012 to 2021

The fluctuation chart of Novatek from 2012 to 2021 is divided into ten years for comparison, as shown in Figure 4-10.

The fluctuation chart of Novatek from 2012 to 2021 is divided into ten years for comparison, as shown in Figure 4-10.

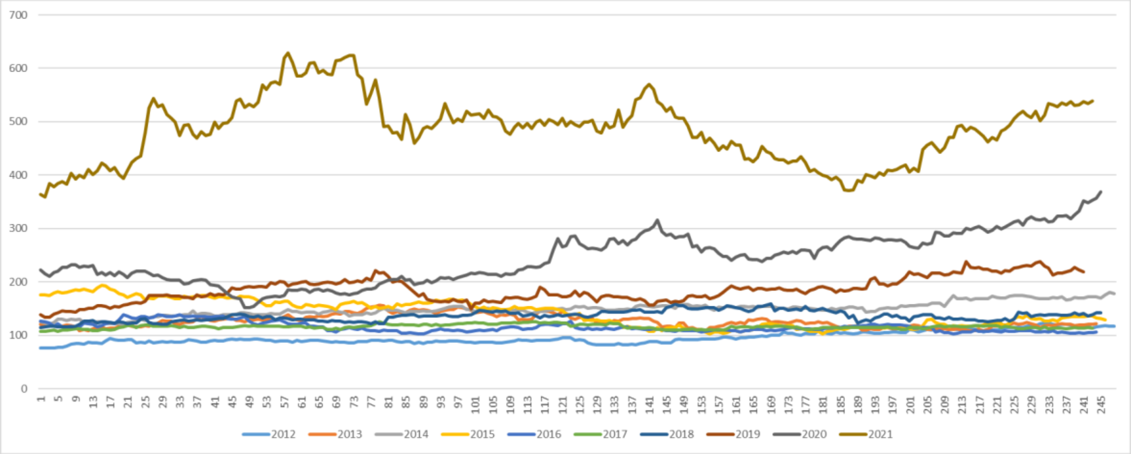

Figure 4-10 Comparison of Novatek’s stock price fluctuations from 2012 to 2021

Table 4-13

Novatek's stock price difference from 2012 to 2021

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
stock price difference	47.1	49.5	60	92	35.5	18.5	44	104.5	217	269.5

Table 4-14

Novatek MAE values from 2012 to 2021

	2012	2013	2014	2015	2016
RF	6.6650	2.3601	15.5322	3.9733	1.6083
NN	4.8075	2.3193	2.5756	2.6924	3.6605
SVR	1.4197*	1.7743*	2.4797*	2.5418*	1.3181*
GPR	17.1583	14.1890	20.0741	24.2180	11.0087
KNN	6.7436	3.7750	15.8750	4.2073	1.7683
	2017	2018	2019	2020	2021
RF	1.2888	2.3285	12.2492	18.7274	10.4493
NN	1.2278	2.2349	9.6885	12.8652	8.4763*
SVR	1.1918*	2.1525*	4.0522*	5.2193*	8.9953
GPR	2.4915	5.8474	33.6198	51.9139	24.6211
KNN	1.5500	2.8462	11.6625	23.5000	18.2195

* Best prediction among 5 algorithms

Table 4-15

RMSE values of Novatek from 2012 to 2021

	2012	2013	2014	2015	2016
RF	7.6460	2.9243	16.0353	5.1361	2.3592
NN	5.8014	2.8083	3.7737	3.5011	4.1110
SVR	1.8463*	2.2245*	3.5418*	3.2724*	2.0774*
GPR	17.4988	14.7189	20.4200	24.8021	11.2215
KNN	7.8324	4.5097	16.3864	5.4117	2.4157
	2017	2018	2019	2020	2021
RF	1.5544	3.2961	13.8967	26.0462	13.3229
NN	1.5576	3.1852*	11.1976	16.1811	11.1909*
SVR	1.4915*	3.2203	5.7604*	6.8623*	11.6226
GPR	3.0420	7.0888	34.2395	54.9783	30.5264
KNN	1.8858	4.0000	13.6457	29.0388	22.8447

* Best prediction among 5 algorithms

Average and Best Prediction Times

Table 4-16

MAE average value

	TSMC	MediaTek	ASE	UMC	Novatek
RF	9.0959	14.3729	1.4060	1.3249	7.5182
NN	5.3553	11.0116	1.2597	0.9348	5.0548
SVR	2.5445*	7.1006*	0.9191*	0.3043*	3.1145*
GPR	19.7788	60.4684	6.0313	2.5923	20.5142
KNN	9.8455	21.0207	1.8886	1.4747	9.0147

* Best prediction among 5 algorithms

Table 4-17

RMSE average

	TSMC	MediaTek	ASE	UMC	Novatek
RF	10.5460	17.5823	1.8156	1.6167	9.2217
NN	6.3614	13.5402	1.6064	1.1224	6.3308
SVR	3.2726*	9.2950*	1.2525*	0.4143*	4.1920*
GPR	20.5308	62.7664	6.4012	2.7263	21.8536
KNN	11.1699	24.5729	2.4570	1.8020	10.7971

* Best prediction among 5 algorithms

Table 4-18

Best prediction times

	TSMC	MediaTek	ASE	UMC	Novatek	sum
RF	0	0	4	0	0	4
NN	2	4	2	0	3	11
SVR	18	16	14	20	17	85
GPR	0	0	0	0	0	0
KNN	0	0	0	0	0	0

Conclusions

Based on the closing price fluctuations of five semiconductor companies from 2012 to 2021, we can see that although stock prices fluctuated significantly from 2020 to 2021, they were all experiencing rapid growth. The main factors influencing semiconductor stock price fluctuations over the past two years are speculated to be as following: In the early stages of the COVID-19 outbreak, people's living and working habits changed. Office workers began working from home, and students began learning remotely. Computers played a key role in both of these processes. This increased demand for chips led to supply chain shortages, benefiting major semiconductor companies across the upstream, midstream, and downstream sectors. Further, The US sanctions on semiconductors for SMIC, China's largest wafer foundry, also benefited major semiconductor companies.

Furthermore, Table 4-18 shows that support vector regression (SVR) performs best, so the following analysis will use SVR as the benchmark. From Tables 4-2 and 4-3, TSMC's forecast results for 2016 showed the lowest MAE and RMSE values for the neural network algorithm, slightly lower than support vector regression. In the remaining years, support vector regression had the lowest MAE and RMSE values. From Tables 4-2 and 4-3, TSMC's forecast results for 2019 and 2020 showed that, with the exception of support vector regression, the MAE and RMSE values for the other four algorithms were higher than those for the same algorithms in the other eight years. From Table 4-5, MediaTek's forecast results for 2015 and 2017 showed the lowest MAE values for the neural network algorithm, slightly lower than support vector regression. In the remaining years, support vector regression had the lowest MAE values. From Table 4-6, MediaTek's forecast results for 2012 and 2017 showed the lowest RMSE values for the neural network algorithm, slightly lower than support vector regression. In the remaining years, support vector regression had the lowest RMSE values. From Tables 4-5 and 4-6, MediaTek's prediction results for 2019 and 2020 showed significantly higher MAE and RMSE values for the other four algorithms compared to those for support vector regression. From Tables 4-5 and 4-6, MediaTek's prediction results for 2021 showed significantly higher MAE and RMSE values for all five algorithms. From Tables 4-8 and 4-9, ASE's prediction results for 2014 showed the lowest MAE and RMSE values for the neural network algorithm, slightly lower than those for support vector regression. From Tables 4-8 and 4-9, ASE's prediction results for 2017 and 2021 showed the lowest MAE and RMSE values for the random forest algorithm, slightly lower than those for support vector regression. From Table 4-15, Novatek's prediction results for 2018 showed the lowest RMSE value for the neural network algorithm, slightly lower than that for support vector regression. From Tables 4-14 and 4-15, Novatek's prediction results for 2021 showed the lowest MAE and RMSE values for the neural network, slightly lower than those for support vector regression.

In terms of algorithm performance comparison, Table 4-18 shows that the best prediction times for the algorithms are ranked in order: support vector regression performs best, followed by the neural network, then random forest, while Gaussian process regression and k-nearest neighbor regression perform poorly. As shown in Tables 4-16 and 4-17, the error value for Gaussian process regression is even larger than that for k-nearest neighbor regression.

In subsequent studies, this paper used only five algorithms for forecasting. Researchers may use a wider range of algorithms. This study covers the period 2012 to 2021. Semiconductor industry stock price fluctuations have been particularly volatile since 2020, and this is expected to continue after 2022. Subsequent researchers may extend the research timeframe. This paper used the stock prices of the previous three days to predict the next day's stock price. Subsequent researchers could also try using the previous day to predict the next day, or the previous two days to predict the next day, and potentially obtain different results.

References

Bayer F. M., Bayer D. M., Pumi G. (2017), Kumaraswamy autoregressive moving average models for double bounded environmental data, Journal of Hydrology, 555, 385-396.

Berry, M. J. A., & Linoff, G. (1997). Data Mining Techniques: for Marking, Sales, and Customer Support. New York: John Wiley & Sons Inc.

Bollerslev T. (1986) Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics. 31 307-327.

Boser B., Guyon I., Vapnik V. (1992) "A training algorithm for optimal margin classifiers", Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 5, 144-152.

Box G., Jenkins G. (1976) Time series analysis: Forecasting and control, San Francisco: Holden-Day.

Breiman L. (2001). Random Forests. Machine learning.

Caruana R., Niculescu-Mizil A., Crew G., Ksikes A. (2004) Ensemble selection from libraries of models, International conference on machine learning.

Chang P. L., Tsai C. T. （2000）, "Evolution of technology development strategies for Taiwan's semiconductor industry： Formation of research consortia", Industry and Innovation, Sydney, 7, 185.

Engle R. F. (1982) Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation. Econometrica. 50(4) 987-1008.

Fayyad, U. M., Shapi, G. P., Smyth, P., Uthursamy, R. (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press/The MIT Press.

Hippert H. S., Pedreira C. E., Castro R. (2001) Neural networks for short-term load forecasting: a review and evaluation. *"IEEE Trans Power Syst",* 16 44-55.

Ho T. K. (1995). Random decision forests. Proceedings of 3rd International conference on document analysis and recognition

Huarng K. H. (2001) Effective lengths of intervals to improve forecasting in fuzzy time series, Fuzzy Sets and Systems. 123 155-162.

Parker, D. B. (1985) "Learning-logic： Casting the cortex of the human brain in silicon." Technical Report TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA.

Sutskever, I. (2012) Training recurrent neural networks, University of Toronto, Ph.d. thesis .