
**Stock Market prediction
using various Machine Learning models**
for the course

IT402: Soft Computing

Submitted by

**Prasad Jagtap(181IT134)
Mithas Kumar (181IT227)
K Keerthana (181IT221)
Yash Parakh(181IT253)**

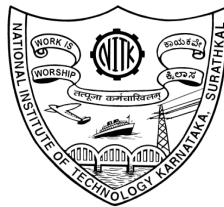
VI SEM B.Tech (IT)

Under the guidance of

**Mrs. Nagamma Patil
Dept of IT, NITK Surathkal**

*in partial fulfillment for the award of the degree
of*
Bachelor of Technology
in

Information Technology



**Department of Information Technology
National Institute of Technology Karnataka, Surathkal.**

April 2021

Acknowledgement


We would like to express our utmost gratitude to our professor, Dr. Nagamma Patil, and our mentor Mr. C. Pandian, who gave us the opportunity to work on this project on the topic ‘Stock Market prediction using various Machine Learning Tools’. It helped us in doing a lot of research and we came across a lot of things related to this topic. We would like to thank the authors of our Base Paper, CNNpred: CNN-based stock market prediction using a diverse set of variables, Ehsan Hoseinzade and Saman Haratizadeh.

Declaration

I hereby declare that the Seminar (IT402) Report entitled.....Stock Market Prediction using Machine Learning.....which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in the Department of Information Technology, is a ***bonafide report of the work carried out by us***. The material contained in this seminar report has not been submitted to any University or Institution for the award of any degree.

Prasad Jagtap (181IT134) 

Mithas Kumar (181IT227) 

Keerthana (181IT221) 

Yash Parakh (181IT253) 

Place: NITK, Surathkal

Date: 14/04/2021

Table of Contents

	Title	Page No.
	Abstract	7
1	Introduction	8
2	Literature Review	9
3	Proposed Methodology	12
4	Results and Analysis	18
5	Conclusion and Future Work	24
	References	25
	Base Paper	27

List of Figures

Fig. No.	Description	Page No.
1	Accuracy Graph of all Algorithms	18
2	Scaled version of Accuracy Graph	18
3	Future price trend prediction of stacked LSTM Model	23
4	Model representation of Base paper	27
5	Return of investment from amount of \$1	28

List of Tables

Table No.	Description	Page No.
1	Logistic Regression : 1-Gram Model	19
2	Logistic Regression : Bi-Gram Model	19
3	Logistic Regression : Tri-Gram Model	19
4	Random Forest: 1-Gram Model	20
5	Random Forest: Bi-Gram Model	20
6	Random Forest: Tri-Gram Model	20
7	Linear Support Vector Machine : 1-Gram Model	21
8	Linear Support Vector Machine : Bi-Gram Model	21
9	Linear Support Vector Machine : Tri-Gram Model	21
10	Non-Linear Support Vector Machine : 1-Gram Model	22
11	Non-LinearSupport Vector Machine : Bi-Gram Model	22
12	Non-Linear Support Vector Machine : Tri-Gram Model	22
13	Naïve Bayes : 1-Gram Model	23

Abstract:

We have used Machine learning techniques to evaluate past data pertaining to the stock market and world affairs of the corresponding time period, in order to make predictions in stock trends. We built a model that will be able to buy and sell stock based on profitable prediction, without any human interactions. The model uses Natural Language Processing (NLP) to make smart “decisions” based on current affairs, articles, etc. With NLP and the basic rule of probability, our goal is to increase the accuracy of the stock predictions.

1. Introduction :

Natural Language Processing is a technique used by a computer to understand and manipulate natural languages. By natural languages, we mean all human-derived languages. Natural language processing or NLP for short is used to analyze text and let machines derive meaning from the input. This human-computer interaction allows us to come up with many different applications to bring man and machine as one.

For example, on google, if we use google translation, that is NLP and so is speech recognition. In our project, we make use of some established NLP techniques to evaluate past data pertaining to the stock market and world affairs of the corresponding time period, in order to make predictions in stock trends.

_____In order to proceed with this objective, we needed to understand what Sentimental Analysis is. Sentimental Analysis is an analytical method that the computer uses to understand a natural language and deduce if the message is positive, neutral or negative. In our case, Sentimental analysis refers to the deduction of the news headlines if they increase the stock or reduce it. By doing so, we end up with the ‘emotional’ status of the data which is what sentimental analysis gives its user.

2. Literature Review:

2.1 - Related Work

1. “Machine Learning in Stock Price Trend Forecasting” written by Y. Dai and Y. Zhang at Stanford University, used features like PE ratio, PX volume, PX EBITDA, 10-day volatility, 50-day moving average, etc. to predict the next-day stock price and a long-term stock price.
2. “Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques” written by J. Patel, S. Shah, P. Thakkar, and K. Kotecha for the “Expert Systems with Applications” international journal demonstrated a way to use trend deterministic data to predict stock price movement.
3. “CNNpred: CNN-based stock market prediction using a diverse set of variables” written by Ehsan Hoseinzade and Saman Haratizadeh.
 - CNNpred successfully combines various sources of information for prediction.
 - CNNs filters are designed to better handle financial data.
 - Deep CNN-based framework significantly outperforms shallow ANNs.
 - CNNpred is profitable in 4 out of 5 tested indices in presence of transaction costs.

2.2 - Motivation

Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Predicting how the stock market will move is one of the most challenging issues due to many factors that are involved in the stock prediction, such as interest rates, politics, and economic growth that make the stock market volatile and very hard to predict accurately.

The prediction of shares offers huge chances for profit and is a major motivation for research in this area; knowledge of stock movements by a fraction of a second can lead to high profits. Since stock investment is a major financial market activity, a lack of accurate knowledge and detailed information would lead to an inevitable loss of investment.

2.3 - Problem Statement

_____Implementing a model for stock market prediction to help the investors in the fast-changing stock market using machine learning approach.

Feature extraction from financial data is one of the most important problems in the market prediction domain for which many approaches have been suggested. Among other modern tools, convolutional neural networks (CNN) have recently been applied for automatic feature selection and market prediction.

2.4 - Objectives

- The ultimate goal of our application is to serve retail investors as a third-party investment tool that uses machine learning to help them navigate in the fast-changing stock market
- The project aims to introduce and democratize the latest machine learning technologies for retail investors.
- The aim of the project is to serve as a supplementary quantitative tool for investors to see the market from a different perspective with the help of technology

3. Proposed Methodology:

We have used the Combined_News_DJIA.csv dataset. The Combined_News_DJIA.csv dataset spans from 2008 to 2016. We extended the dataset to include additional data. This additional data is collected from the Guardian's Restful News API for the 2000- 2008 period. We take the 25 most popular headlines for each given day in this period. In addition, we also pull the Dow Jones Index (DJI) of Yahoo Finance's website for the 2000- 2008 period to compare the influence of the data.

There are two channels of data provided in this dataset:

- News data that has historical news headlines from Reddit World News Channel. Only the top 25 headlines are considered for a single date.
- Stock data for Dow Jones Industrial Average (DJIA) over the corresponding time range is used to label the data. The stock data is compiled from Yahoo Finance.

The headlines for each data act as the explanatory data that causes the stock price to either rise (labelled 1) or to fall (labelled 0). We have the top 25 headlines for one single date arranged as one row of the extracted data set.

Since our goal is to predict the tendency of the stock of a specific company, the data that lead the stock's price of the next day to decline or stay the same are labelled "0", while the data that lead the price of the next day to rise are labelled "1". We compare the data between the two pulled dataset and then merge them together to get a more accurate prediction.

With the raw data, we cannot proceed much further until we manipulate the data to suit our analysis and convert the data into vectors that are much easier to

work on. For this, we use Word2Vec. This is a group of related models used to create word embeddings. Word embeddings are sets of language modelling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers. These vectors make up the training and test sets. English is really easy – see all those spaces? That makes it really easy to tokenize – in other words, to determine what’s a word.

So we just use a simple set of rules for English tokenization. This raw data is manipulated using python. We first split the data into lists of words but these lists are flooded with HTML tags and punctuations. We cleaned up the data and removed all HTML tags and punctuations. Then we moved forward with removing stop words. Stop words are words that do not contribute to the meaning or sentiment of the data such as ‘the’, ‘is’, ‘and’, etc.

With the training data set, we got to convert them into numeric representation for machine learning. For this, we use the ‘Bag of Words’ model. The Bag of Words model learns a vocabulary from all of the documents, then models each document by counting the number of times each word appears. These values are the feature vectors that are derived from the model.

The Bag of Words model generates feature vectors that only give importance to the number of occurrences of words rather than where they occur and with what words they accompany. To get past this, we use the n-gram model or the skip-gram model. Now, with this model, we can store the order of words in the way they occur in the data.

We use Natural Language Processing (NLP) to interpret and construct the data set. With the manipulated data vectors ready to be trained and tested, the extracted dataset was split in the ratio of 4:1. 80% of the extracted data will be the training data and 20% of the extracted data will be the test data.

Five models in this project to train our data

- Logistic Regression
- Random Forest
- Gaussian Naive Bayes
- Linear Support Vector Machines
- Non-Linear Support Vector Machines

Naïve Bayes:

This model provides a family of probabilistic classifiers that are based on the Bayes theorem with strong independence characteristics within its feature vectors. $P(A/B) = P(B/A)P(A) / P(B)$ where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A | B)$, a conditional probability, is the probability of observing event A given that B is true.

$P(B | A)$ is the probability of observing event B given that A is true.

Naïve Bayes is common to use in Bag of Words. Since we have 25 features (25 top headlines) in each data set (a given day), the step is as below:

Training: Estimate $P(Y | X)$ for all $1:K$ $X_{1:K}$

Testing: Predict $Y = \text{argmax } P(Y | X)$ of Y where $X=1:K$ for all $X=1:K; k=25$, and

$$P(Y | X_{1:K}) = \frac{P(X_{1:K} | Y)P(Y)}{P(X_{1:K})} = \frac{P(X_{1:K} | Y)P(Y)}{\sum_{Y'} P(X_{1:K} | Y')P(Y')}$$

Random Forest:

Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. The RF algorithm grows n decision trees as the weak classifier, each provides different kinds of classification, and then merge all the trees into a forest. Unlike the decision tree or K-NN method, RF doesn't have to take the cross-validation. The step is as follow:

1. Grow many trees. Each tree has m input data, which m is a constant chosen randomly from the K th element in our data set and $m \ll K$ during the trees' growth.
2. At each node test the value of features X , divide the data into two leaves ($Y=1$ if $1:K$ the price increases and $Y=0$ if the price decreases).
3. Each tree grows to the largest extent possible. No pruning for the trees.
4. Estimate the error for the prediction $h(x)$ of i of each tree as the tree grows entirely, that is, the out-of-bag error (OOB) sample of the i -th tree. The accuracy will be $1 - \text{OOB}$.
5. Merge all the trees into a forest and estimate the OOB of the whole forest by the OOB of each tree in the forest. The final prediction is

$$H = \frac{1}{K} \sum_{i=1}^K \alpha_i h_i(x)$$

Support Vector Machine:

An SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data, the algorithm outputs an

optimal hyperplane that categorizes new examples. SVM is a margin-based classification method. It discriminates the data by a separating hyperplane and its margin, i.e.

$$y_i(\beta_0 + \sum_{j=1}^K \beta_j x_{ij}) \leq M \quad \forall i = 1, 2, \dots, n$$

maximize M subject to,

$$\sum_{j=1}^K \beta_j^2 = 1$$

since we have only 2 classes and 25 features in each data, we can get a low testing error if the training error is also small. Therefore it is suitable for classifying our data set. In addition, the kernel method is also taken into our SVM model to sparse the data which cannot be distinguished in original space.

Logistic Regression:

It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Logistic Regression is widely used in classification and data analysis.

It is appropriate when the response takes only one of two values, that is, there are only two classes for the dependent variables. Therefore it is also appropriate to our data set. In HW3 we also use this method to identify whether the mail is spam or not. Here we can use it in a similar way. The probability can be calculated as below:

$$P(Y = 0|x, w) = \frac{1}{1+\exp(w \cdot x)} = \frac{1}{1+\exp(w_0 + \sum_{j=1}^K w_j x_j)}$$

$$P(Y = 1|x, w) = \frac{\exp(w \cdot x)}{1 + \exp(w \cdot x)} = \frac{\exp(w_0 + \sum_{j=1}^K w_j x_j)}{1 + \exp(w_0 + \sum_{j=1}^K w_j x_j)}$$

$$P(Y = 1|x, w) = 1 - P(Y = 0|x, w)$$

where η is the learning rate and $1/2(\lambda||w||)^b$ term used to avoid overfitting (decrease the fluctuation of each iteration of the regression). Upon this, we are able to connect the relation between the specific words that appear frequently and identify what tendency the headline (our data set) has, so that we can predict the price of the stock in the next day. After applying the models, we were left with testing the output data and evaluating the comparisons of the results which we will be covering in the results section.

Long Short-Term Memory:

The data used in this model has date, close, open, high, low, volume for a specific stock price (TATA motors in this case). The data contains the mentioned information for every specific day from 2016 to 2021 (exactly 5 years).

These networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. It has an advantage over traditional neural networks due to its capability to process the entire sequence of data. Its architecture comprises the cell, input gate, output gate and forget gate.

The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. The cell of the model is responsible for keeping track of the dependencies between the elements in the input sequence. The input gate controls the extent to which a new value flows into the cell, the forget gate controls the extent to which a value remains in the cell, and the output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit.

4. Results and Analysis:

This is the Accuracy graph of our algorithms. The stacked LSTM model is judged based on its accuracy only. The Naive Bayes algorithm wasn't able to run in a 2-gram or 3-gram model.

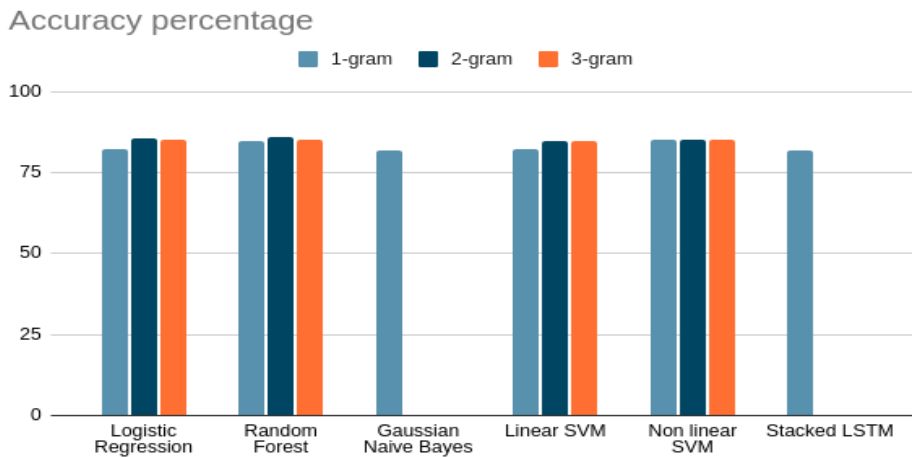


Fig1. Accuracy Graph of all algorithms

Here is the zoomed-in version of the graph for better understanding:

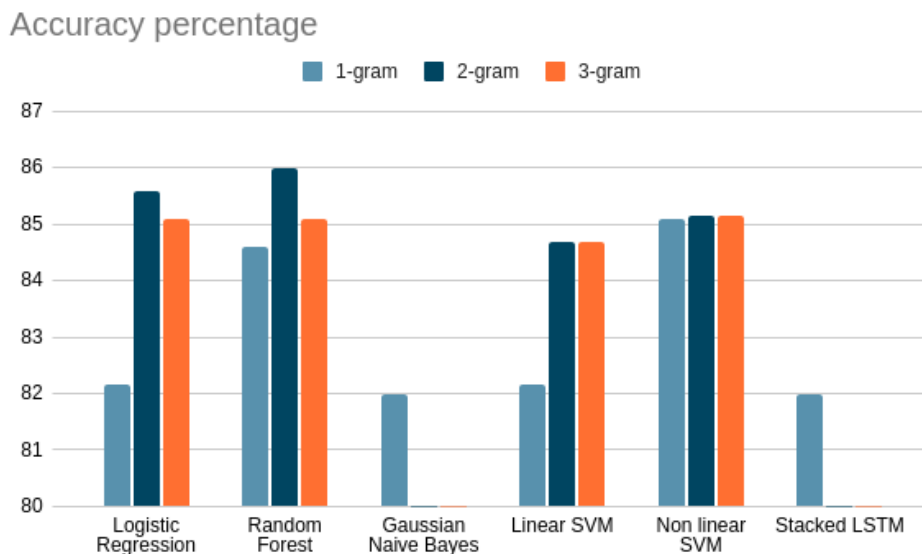


Fig2. Scaled version of accuracy graph

Logistic Regression Model:

Predicted	0	1
Actual		
0	149	37
1	30	162

Table 1: 1-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	159	27
1	27	165

Table 2: Bi-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	142	44
1	12	180

Table 3: Tri-Gram Model Prediction Accuracy

Random Forests Model:

Predicted	0	1
Actual		
0	144	42
1	16	176

Table 4: 1-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	143	43
1	10	12

Table 5: Bi-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	130	56
1	0	192

Table 6: Tri-Gram Model Prediction Accuracy

Linear Support Vector Machine:

Predicted	0	1
Actual		
0	151	35
1	32	160

Table 7: 1-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	160	26
1	32	160

Table 8: Bi-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	145	41
1	17	175

Table 9: Tri-Gram Model Prediction Accuracy

Non - Linear Support Vector Machine:

Predicted	0	1
Actual		
0	130	56
1	0	192

Table10: 1-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	130	56
1	0	192

Table 11: Bii-Gram Model Prediction Accuracy

Predicted	0	1
Actual		
0	120	66
1	0	192

Table 12: Tri-Gram Model Prediction Accuracy

Naïve Bayes Model:

Predicted	0	1
Actual		
0	155	31
1	37	155

Table 13: 1-Gram Model Prediction Accuracy

Stacked LSTM Model:

This is the trend deterministic model derived from Stacked LSTM where it determines the future of the stock price



Fig3. Future price trend prediction of Stacked LSTM mode

5. Conclusion and Future Work:

Social Media can sometimes be deceiving when delivering the right frame of speech. Here we have used Twitter feeds and news articles around the web that has influenced the stock market of a company. Through this project, it helped us understand the basics of Natural Language Processing. Even though you can't bet your money on the stock from this project, this work can be treated as a solid understanding of the basics of Natural Language Processing.

Using the same model for different text data is also feasible. It was interesting to know about how to go from text data to vectors of numbers and applying Machine learning techniques that can help to influence the stock market of a company. It helped us gain a wider sense of the power of NLP in various applications. From reading about machine learning models in class to implementing them with real data and observing the performance of a model, tuning the parameters, performing exploratory data analysis set a great learning curve for future projects.

This project leaves room for future work and ways to accomplish them:

1. The number of features used in the data set can be expanded. Right now we have gathered the top 25 News headlines, It is important to have more features that help the model learn better.
2. We are looking at the stock of one company. We can expand it to work for multiple companies at once and we can also include real time- time series analysis.
3. Perform multi-class classification for various parameters of stock trading.

References:

1. F. Xu and V. Keelj, "Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction," 2014 IEEE 16th Conference on Business Informatics, Geneva, 2014, pp. 60-67. doi: 10.1109/CBI.2014.37
2. L. Bing, K. C. C. Chan and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," 2014 IEEE 11th International Conference on e-Business Engineering, Guangzhou, 2014, pp. 232-239. doi: 10.1109/ICEBE.2014.47
3. D. Rao, F. Deng, Z. Jiang and G. Zhao, "Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure," 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2015, pp. 381-384. doi: 10.1109/IHMSC.2015.114
4. Z. Jiang, P. Chen and X. Pan, "Announcement Based Stock Prediction," 2016 International Symposium on Computer, Consumer and Control (IS3C), Xi'an, 2016, pp. 428-431. doi: 10.1109/IS3C.2016.114
5. W. Bouachir, A. Torabi, G. A. Bilodeau and P. Blais, "A bag of words approach for semantic segmentation of monitored scenes," 2016 International Symposium on Signal, Image, Video and Communications (ISIVC), Tunis, 2016, pp. 88-93. doi: 10.1109/ISIVC.2016.7893967
6. D. Sehgal and A. K. Agarwal, "Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework," 2016 International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad, 2016, pp. 251-255. doi: 10.1109/SYSMART.2016.7894530
7. R. Zhao; K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," in IEEE Transactions on Fuzzy Systems , vol.PP, no.99, pp.1-1 doi: 10.1109/TFUZZ.2017.2690222
8. V. U. Thompson, C. Panchev and M. Oakes, "Performance evaluation of similarity measures on similar and dissimilar text retrieval," 2015 7th

International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, 2015, pp. 577-584

9. C. Sreejith, M. Indu and P. C. R. Raj, "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-4. doi: 10.1109/ICCCNT.2013.6726777
10. M. Kaya, G. Fidan and I. H. Toroslu, "Sentiment Analysis of Turkish Political News," 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, 2012, pp. 174-180. doi: 10.1109/WI-IAT.2012.115

Base Paper:

The paper which our team chose as the base paper is [CNNpred: CNN-based stock market prediction using a diverse set of variables](#). The paper was authored by Ehsan Hoseinzade (Simon Fraser University) & Saman Haratizadeh (University of Tehran) in March 2019. The paper was published in Expert Systems with Applications, Volume 129, 1 September 2019, Pages 273-285.

Here are some analyses that we derived from the base paper:

They have used a 3D-CCNpred model for predicting the stock price. It is should be technically better in comparison to 2D-CCNpred.

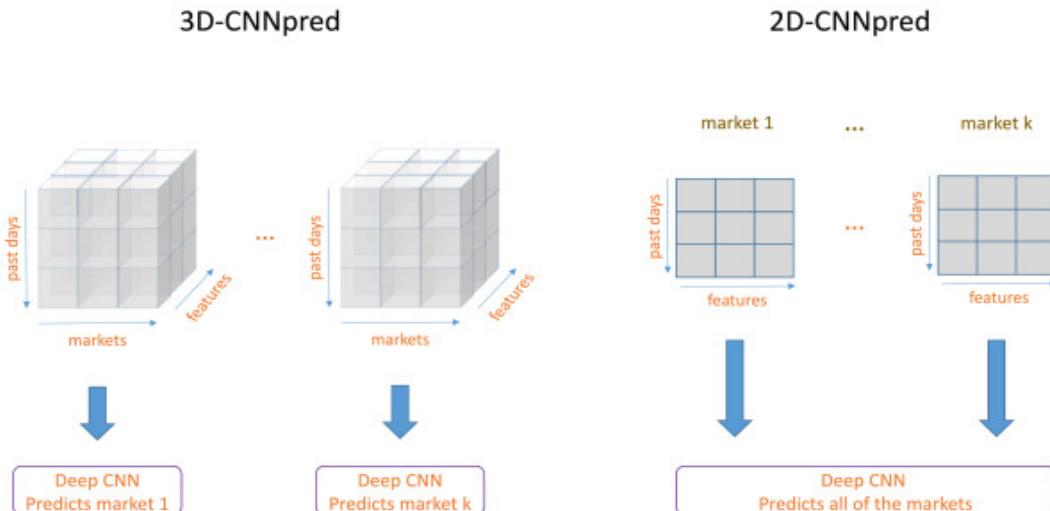


Fig 4: Model representation of Base paper

They have mentioned that the ANN model has 65% accuracy.

The Base paper has mentioned the following:

CNNpred successfully combines various sources of information for prediction.

CNNs filters are designed to better handle financial data.

Deep CNN-based framework significantly outperforms shallow ANNs.

CNNpred is profitable in 4 out of 5 tested indices in presence of transaction costs.

They have mentioned that the ANN model has 65% accuracy.

This is the Final Result of the base paper while Trading for \$1:

Strategy	Rate of costs	S&P 500	DJI	NASDAQ	NYSE	RUSSELL
Buy and hold	%0	1.1794	1.2387	1.2985	1.1338	1.2134
	%0.1	1.1784	1.2378	1.2975	1.1328	1.2124
2D-CNNpred	%0	1.2378	1.2740	1.2595	1.1808	1.2312
	%0.1	1.2356	1.2718	1.2575	1.1787	1.2291
3D-CNNpred	%0	1.2364	1.2191	1.3606	1.1456	1.2604
	%0.1	1.2343	1.2170	1.3585	1.1445	1.2582

Fig 5: Return of investment from amount of \$1