An analysis and forecasts of online product sales based on BP Neural Network and Pearson Coefficient

Chenghao Wang*
College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, 350117, China
wch19990721@gmail.com
*corresponding author

Jingsi Chen
College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, 350117, China
U1976934@unimail.hud.ac.uk

Ruihao Zeng
International College of Chinese Studies, Fujian Normal University, Fuzhou, 350117, China
xmrhzeng2000@gmail.com

***Abstract*—We construct two models to derive the optimal plan. We successfully utilize our models to determine the best online sales strategy and identify potentially important design features that would enhance product desirability. We establish the Genetic Algorithm-Optimized BP Neural Network Model to find the relationship between time-based measures within data sets and the reputation of products. After quantifying all the indicators, we take BP Neural Network Model as a fitting platform to fit and forecast the reputation-timeline figure. To degrade the influence because of the slow rate of convergence, we introduce Genetic Algorithm to seek optimal global solutions and optimize BP Neural Network Model. Finally, we arrive at the conclusion that the reputation of products has potential correlations with time-based measures within data sets. We establish the Pearson Correlation Model to get the correlation between indexes. Firstly, we quantified all the indicators. According to the model analysis, we find that specific star ratings do incite more reviews. Besides, we use VADER, an NLP Sentiment Analysis Model, to quantify the sentiment descriptors of the reviews. Finally, we conclude that specific quality descriptors of text-based reviews are strongly associated with star rating level but barely associated with the helpful rating level.**

Keywords—Genetic Algorithm-Optimized BP Neural Network Model, Pearson Correlation Model, Sentiment Analysis Model.

I. BACKGROUND

A. Overview

Data are impacting the business ecosystem in various ways. Through the analysis of data, companies can make critical decisions in R & D, business, revenue, etc., and then get better development space.

Previous research indicates that online rating systems are essential in consumer decision making [1]. The consumer feedback mechanism established by Amazon can significantly help us to understand consumer needs and promote product sales.

B. Our work

- Mathematically process the data from Amazon between star ratings, reviews, and helpfulness ratings to help us intuitively understand the factors of product success.
- Establish a model and calculate the trend of product's reputation. Determine the best indicate and help the company determine whether the product succeeded or failed.
- Establish a model, find the relationship between star ratings and reviews, the relationship between specific quality descriptor and rating levels.

II. DATA PROCESSING

A. Overview

Before establishing the model, we process the data provided by the Amazon in three products- hair dryer, microwave and pacifier- to ensure the accuracy of the model results.

We use different methods to quantify the four factors of star ratings, reviews, helpful ratings, and popularity degrees to analyzes. Firstly, using Flesch Reading Ease Readability Formula, length of the review text, whether the review is Amazon Vine Voices, we quantified the reviews into three indexes: readability, depth, and authority. Besides, we use the ratio of helpful votes to total votes in this review to quantify helpful ratings. Then reduce the time dimension of all data to months to calculate sales volume to quantify popularity degree.

B. Review Quantification

Of the four factors mainly used for model analysis, only star rating has been quantified in Datasets, and the remaining reviews, helpfulness ratings, and product sales have not been quantified in Datasets. In this step, we use different methods to quantify these four factors.

1) Reading Ease Readability

We choose Flesch reading-ease test to test the readability of the review. [2]

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

2) Profoundness of review

The longer the review, the more useful information it contains. When more information is available, consumers tend to have more confidence in making decisions, especially when consumers can obtain information without the additional search. The word count is positively related to the profoundness of the review. A long review will help consumers make decisions, so we counted the number of words in each review to indicate the profoundness of the review.

3) Authority of review

We evaluated the authority of the review based on whether the reviewer was Amazon Vine Voices. Amazon gave a definition: Customers are invited to become Amazon Vine Voices based on the trust that they have earned in the Amazon community for writing accurate and insightful reviews [3]. Reviews with this logo represent higher credibility and better quality. Therefore, we believe that this is the most important basis for quantifying authority.

4) Text sentiment of review

We use the front sentiment-based method VADER (for Valence Aware Dictionary for sentiment Reasoning) to quantify reviews. It retains (and even improves on) the benefits of traditional sentiment lexicons like LIWC: it is bigger, yet just as simply inspected, understood, quickly applied (without a need for extensive learning/training) and easily extended [4].

Extract the lexical features from the text of the review, and get the intensity contained in each review according to the calculation.

C. Helpfulness Rating Quantification

When reading online reviews, consumers are not only concerned about the number and content of reviews, but also the credibility of reviews. The review's vote mechanism can help consumers identify credibility.

Because the identity of the two parties in the communication is not clear, when the consumer reads the review, a credible review will be considered more useful, provide a higher reference value, and get a larger percentage of helpful votes. We think that calculating the percentage of helpful votes can well reflect the role of a review.

Helpfulness rating reflects the effect of review very well, we consider it as one of the important indicators of the model.

III. Model: Genetic Algorithm- Optimized BP Neural Network Model

A. Model Building

In this model, We use the data processed in the Structural equation model, Introduced Number of vine and Word of review to reflect the authority and depth of reviews. Based on a total of 6 influencing factors, build a model based on BP neural network[5].

Based on the cleaned and time-reduced data, the top 15 products sold in each category were extracted as analysis objects.

We use MATLAB to build this model. According to two assumptions:

1) the Reading ease readability of reviews, the Profoundness of reviews, the Authority of reviews, star rating, helpful rating, and the historical sales volume all have different degrees of positive impact on product reputation, and reputation is directly related to product.

2) There is some complex linear relationship between sales volume, product sales volume and time.

A Bayesian Regulation neural network is constructed to verify the relationship, and genetic algorithms are used to optimize the neural network. The initial weight of the data is from the result of Model 1.

(1) Parameter setting and Hidden layer node

Set the initial number of hidden neurons to 3, and the function used to train the network performance is Bayesian Regulation.

A three-layer BP neural network (as shown in the figure) is selected to establish the prediction model.

In order to obtain the appropriate number of hidden layer nodes, we first set a small number of hidden layer nodes in the network. According to the principle of error minimization, the number of hidden layers is continuously increased under the same training sample. After comparison, the number of hidden layer nodes is determined as 5. [6]

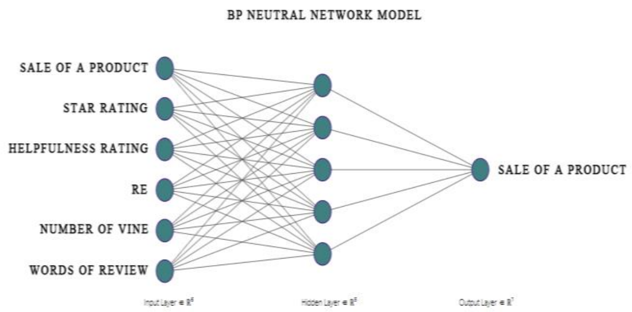

Fig. 1. Model Structure

(2) Training and Simulation

The MATLAB neural network toolbox functions are used to program the construction, training and simulation of the BP neural network model. The sample is divided into 2 parts: 85% is selected as the training sample, and the remaining 15% is used as the test verification sample.

(3) Data Normalization

Considering the fluctuation range of the data, the data is converted to the [0,1] interval using a normalization method.

$$X_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Where X_i is input data, X_{min} is data minimum, X_{max} is data maximum.

(4) Genetic Algorithm Optimization

We use GA for function optimization and find the global optimal solution of the problem to improve the accuracy of the prediction[7].

A population P is initialized with a random number, the initial population is q = 30, the number of iterations is 100 times, the integer is used for encoding, and the population has a crossover phenomenon.

The individual code length S equation is: $S = a \cdot b + b \cdot c + b + c$, where a is the number of nodes in the input layer, b is the number of nodes in the hidden layer, and c is the number of nodes in the output layer[8].

Using roulette method to select individuals with good adaptability to form new species group.

The crossover probability (0.5,0.7,0.75), and the mutation probability is (0.001, 0.003, 0.05). After the initial test run, the cross probability increases, the population diversity shows an upward trend, and as the population size increases, the risk of the algorithm producing a locally optimal solution decreases.

Taking prediction error as performance function of individual fitness value. Call fitness function to calculate fitness and find the optimal fitness:

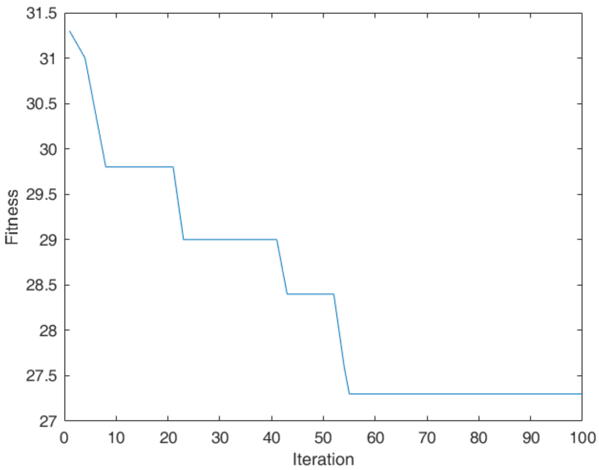

Fig. 2. Fitness calculation

B. BP Neural Network Training and Prediction

The BP neural network is re-established by the determined parameters, and the optimal individuals obtained after optimization by the genetic algorithm are re-assigned to the initial weight thresholds of each connection layer of the BP neural network.

Then, the normalized sample data is sent to a BP neural network for learning operations. Give the best individual of Genetic Algorithm to BP Neural network and fitting nonlinear function again. Repeated training, and finally get a half-year sales forecast based on the training model.

We got results for one of those three products, the top-5-sales of microwaves are shown below:

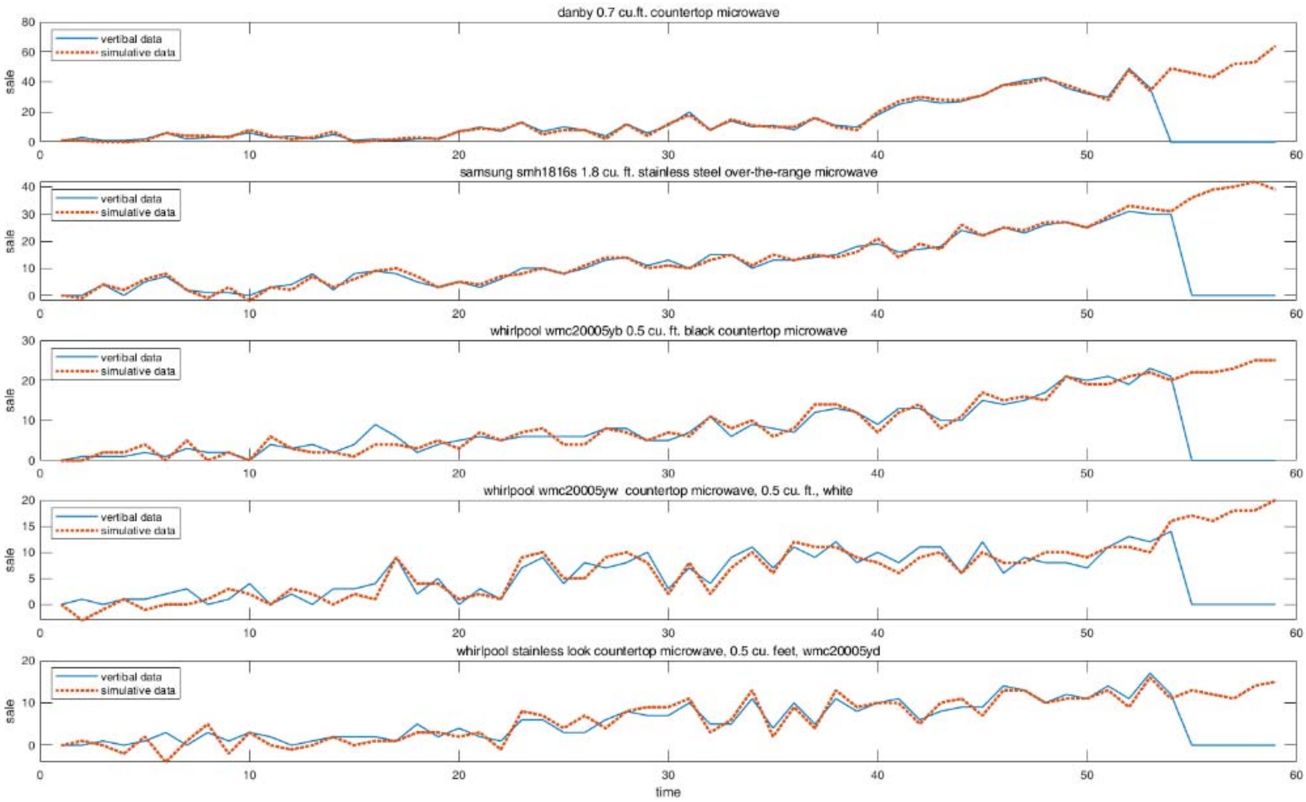

Fig. 3. Fitting spline of top-5 microwave products

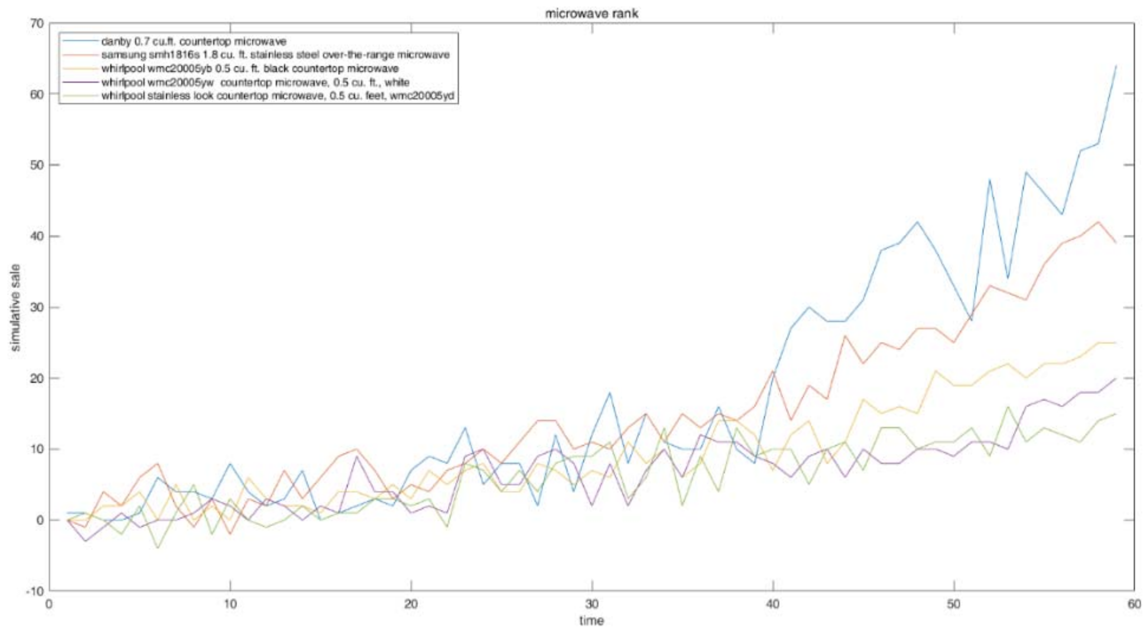

Fig. 4. Sales trend of microwave

It showed that the reputation has potential correlations with time-based measures which are reflected on the Figure 3. Figure 3 also indicates that the reputation is increasing in the online marketplace from the blooming curves.

Also we can find in the Figure 4 that the "philips avent bpa free soothie pacifier, 0-3 months, 2 pack, packaging may vary" of pacifier, the "danby 0.7 cu.ft. countertop microwave" of microwave and the "andis 1875-watt fold-n-go ionic hair dryer, silver/black (80020)" of hair dryer are the most potentially successful three products among all products because of their rapidly growing simulative sale curves.

IV. Model: Pearson Correlation Model

In order to get the relationship of specific star rating and the number of reviews, the relationship of specific text-based quality descriptors and rating levels. We establish the Pearson correlation model to measure the strength of the association between two sets of variables. We use the distance the data and the average to represent the star rating specificity, and the review length to represent the number of reviews.

The data of the three products were input into the model and the Pearson correlation coefficients were all positive. Among them, the hair dryer has the highest Pearson correlation coefficient, reaching 0.4547. We get specific star ratings incite more reviews. Then, we use the VADER model text-based in the NLP Sentiment analysis to quantify the emotional polarity of the reviews, which is divided into The star rating and helpful rating were brought into the model for calculation, and the results were all positive. The emotional polarity of the hairdryer and the star rating Pearson correlation coefficient were the highest, reaching 0.538. It can be concluded that specific quality descriptors of text-based reviews strongly associated with rating levels.

A. Model Building

In statistics, the Pearson correlation model is a measure of the linear correlation between two variables.

1) Definition of Pearson correlation coefficient

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables[9]:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

The above equation defines the overall correlation coefficient ρ. Estimating the covariance and standard deviation of the sample, we can get the Pearson correlation coefficient r.

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}.$$

Estimate the standard score mean of the sample point (X_i, Y_i) to get the equivalent expression:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_X}\right)\left(\frac{Y_i-\overline{Y}}{\sigma_Y}\right).$$

where $\frac{X_i-\overline{X}}{\sigma_X}$ is the standard score, $\overline{X}$ is the sample mean, σ_X is the sample standard deviation

2) Model Interpretation

The Pearson correlation coefficient ranges from -1 to 1. A coefficient value of 1 means that X and Y can be well described by a straight-line equation, all data points fall well on a straight line, and Y increases with X. A value of the coefficient of −1 means that all data points lie on a straight line, and Y decreases as X increases. A coefficient value of 0 means that there is no linear relationship between the two variables, which can well reflect the relationship of the data.[10]

B. Star ratings and Reviews

After statistics, we got the average product star rating Sra:

TABLE I. AVERAGE PRODUCT RATING

Product name	Hair dryer	Microwave	Pacifier
Sra	4.1881	4.0332	4.3400

We use the distance between the data and the average to indicate star rating specificity Ss, The calculation equation is as follows:

$$Ss = |Sr - Sra|$$

Then use the established Pearson correlation model to perform correlation analysis on Specificity of star ratings and reviews length. The reviews length uses the average of all review length under different Specific of star ratings. The Pearson correlation coefficient is shown in the table.

TABLE II. CORRELATION ANALYSIS RESULT 1

Products name	Hairdryer	Microwave	Pacifier
r	0.4547	0.4434	0.2219

Fit the results above:

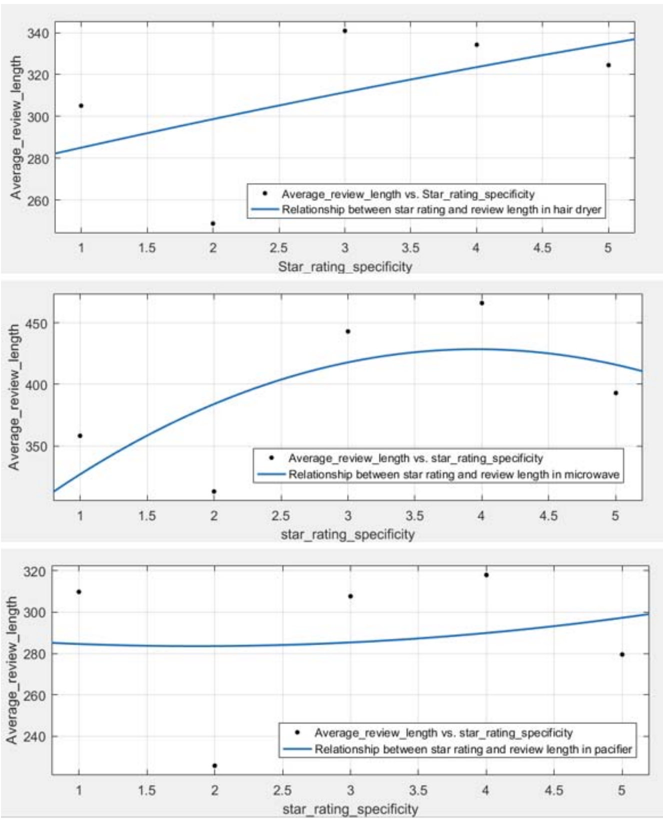

Fig. 5. Tthe fitted curve of three products

It can be seen that for the three products, the correlation between Specificity of star ratings and reviews length is high, and a positive correlation is obtained by fitting the results.

C. Specific Quality Descriptors of Review and Rating

We use the VADER-based text sentiment recognition method to get the text sentiment value. Correlation analysis was performed on the review sentiment and rating levels using the Pearson correlation model. Among them, we divide the rating level into star rating level and helpful rating level, result is shown in the table:

TABLE III. CORRELATION ANALYSIS RESULT 2

Products name	Hair dryer	Microwave	Pacifier
R_s	0.5381	0.5626	0.4914
R_h	0.0905	0.1263	0.1277

For the three products, review sentiment has a higher positive correlation with star rating level. The positive correlation between review sentiment and helpful rating level is low.

D. Model Conclusion

In conclusions, specific star ratings do incite more reviews. Specific quality descriptors of text-based reviews are strongly associated with star rating level, barely associated with helpful rating level.

V. CONCLUSION

The first model showed that the reputation of all three products have potential correlations with time-based measures. And it also indicates that the each of the product's reputation is increasing in the online marketplace from the blooming curves.

What's more, we can find that the "philips avent bpa free soothie pacifier, 0-3 months, 2 pack, packaging may vary" of pacifier, the "danby 0.7 cu.ft. countertop microwave" of microwave and the "andis 1875-watt fold-n-go ionic hair dryer, silver/black (80020)" of hair dryer are the most potentially successful three products among all products because of their rapidly growing simulative sale curves.

In last model, we can find out that specific star ratings do incite more reviews. And specific quality descriptors of text-based reviews are strongly associated with star rating level, but barely associated with helpful rating level.

REFERENCES

[1] Sulin Ba, Paul Pavlou, "Evidence OF the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior," September 2002MIS Quarterly 26(3):243-268

[2] Flesch, R. F.. How to write plain English: A book for lawyers and consumers. Harpercollins. 1979

[3] Amazon Site Features https://www.amazon.com/gp/help/customer/display.html?nodeId=GPEWN3RSQPEU2HST

[4] Hutto, C. J., & Gilbert, E. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text" January 2015 Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social MediaAt: Ann Arbor, MI

[5] Mei-Hua Zhou, Fu-Dong Wang, Hong-Hong Zhang. "Study of enterprises marketing risk early warning system based on BP neural network model" September 2006Journal of China University of Mining and Technology 16(3):371-375

[6] Feng Li, Cheng Liu. "Application Study of BP Neural Network on Stock Market Prediction" Conference: 9th International Conference on Hybrid Intelligent Systems (HIS 2009), August 12-14, 2009, Shenyang, China

[7] Shifei Ding,Chunyang Su, & Junzhao Yu. “An optimizing BP neural network algorithm based on genetic algorithm” August 2011Artificial Intelligence Review 36(2):153-162

[8] Lijun Gao, Yanting Li, Lu Zhang, Feng Lin, Maode Ma. “Research on Detection and Defense Mechanisms of DoS Attacks Based on BP Neural Network and Game Theory,” March 2019IEEE Access PP(99):1-1

[9] Ozer, D. J. “Correlation and the Coefficient of Determination” March 1985Psychological Bulletin 97(2):307-315

[10] Jeremy Adler,Ingela Parmryd. “Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient is Superior to the Mander's Overlap Coefficient” August 2010Cytometry Part A 77(8):733-42

[11] Sharma Kartik; Aggarwal Ashutosh; Singhania Tanay; Gupta Deepak; Khanna Ashish (2019). Hiding Data in Images Using Cryptography and Deep Neural Network. Journal of Artificial Intelligence and Systems, 1, 143–162.

[12] Emre Altinkaya, Kemal Polat, Burhan Barakli (2019). Detection of Alzheimer’s Disease and Dementia States Based on Deep Learning from MRI Images: A Comprehensive Review. Journal of the Institute of Electronics and Computer, 1, 39-53.