Hindawi
Security and Communication Networks
Volume 2023, Article ID 9865041, 1 page
https://doi.org/10.1155/2023/9865041

Retraction

Retracted: Statistical Modeling and Simulation of Online Shopping Customer Loyalty Based on Machine Learning and Big Data Analysis

Security and Communication Networks

Received 26 December 2023; Accepted 26 December 2023; Published 29 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J.-C. Huang, P.-C. Ko, C.-M. Fong, S.-M. Lai, H.-H. Chen, and C.-T. Hsieh, "Statistical Modeling and Simulation of Online Shopping Customer Loyalty Based on Machine Learning and Big Data Analysis," *Security and Communication Networks*, vol. 2021, Article ID 5545827, 12 pages, 2021.

Hindawi
Security and Communication Networks
Volume 2021, Article ID 5545827, 12 pages
https://doi.org/10.1155/2021/5545827

Research Article

Statistical Modeling and Simulation of Online Shopping Customer Loyalty Based on Machine Learning and Big Data Analysis

Jui-Chan Huang,[1] Po-Chang Ko,[2] Cher-Min Fong,[3] Sn-Man Lai,[2] Hsin-Hung Chen,[3] and Ching-Tang Hsieh[4]

[1] *Yango University, Fuzhou 350015, China*
[2] *Department of Intelligent Commerce, National Kaohsiung University of Science and Technology, Kaohsiung City, 80778, Taiwan*
[3] *Department of Business Management, National Sun Yat-Sen University, 70 Lien-hai Rd, Gushan District, Kaohsiung City 804, Taiwan*
[4] *Department of International Business, National Kaohsiung University of Science and Technology, Kaohsiung City 80778, Taiwan*

Correspondence should be addressed to Sn-Man Lai; hjc0718@nkust.edu.tw

Received 6 January 2021; Revised 25 January 2021; Accepted 9 February 2021; Published 18 February 2021

Academic Editor: Chi-Hua Chen

With the increase in the number of online shopping users, customer loyalty is directly related to product sales. This research mainly explores the statistical modeling and simulation of online shopping customer loyalty based on machine learning and big data analysis. This research mainly uses machine learning clustering algorithm to simulate customer loyalty. Call the k-means interactive mining algorithm based on the Hash structure to perform data mining on the multidimensional hierarchical tree of corporate credit risk, continuously adjust the support thresholds for different levels of data mining according to specific requirements and select effective association rules until satisfactory results are obtained. After conducting credit risk assessment and early warning modeling for the enterprise, the initial preselected model is obtained. The information to be collected is first obtained by the web crawler from the target website to the temporary web page database, where it will go through a series of preprocessing steps such as completion, deduplication, analysis, and extraction to ensure that the crawled web page is correctly analyzed, to avoid incorrect data due to network errors during the crawling process. The correctly parsed data will be stored for the next step of data cleaning or data analysis. For writing a Java program to parse HTML documents, first set the subject keyword and URL and parse the HTML from the obtained file or string by analyzing the structure of the website. Secondly, use the CSS selector to find the web page list information, retrieve the data, and store it in Elements. In the overall fit test of the model, the root mean square error approximation (RMSEA) value is 0.053, between 0.05 and 0.08. The results show that the model designed in this study achieves a relatively good fitting effect and strengthens customers' perception of shopping websites, and relationship trust plays a greater role in maintaining customer loyalty.

1. Introduction

With the remarkable progress of machine learning, people are paying more and more attention to its relevance to e-commerce shopping. However, because it is generally still unclear how to recognize big data and optimize neural networks for machine learning, machine learning has not yet had a significant impact on e-commerce practices. However, there are optimistic views that machine learning will have a significant impact on e-commerce and radiology in the next five years.

E-commerce travel websites generate a huge number of user visit records every day, but a large number of visitors are lost. If you can combine user visit information and user consumption information on the site, you can get some useful information from it, which can help companies quickly understand the loss of users; the key reason can also be using the results of the model to predict the user's current status and then prescribe the right medicine, improve the user experience of the enterprise, and retain more users.

Large cloud operators provide machine learning (ML) as a service, enabling customers who have data but do not have ML expertise or infrastructure to train predictive models based on this data. Hunt et al. believes that the existing ML as a service platform requires users to disclose all training data to service operators. He implemented and evaluated Chiron, a machine learning as a service system for privacy protection. His research process is too complicated [1]. Wu et al. believes that EEG and magnetic EEG are the most common noninvasive brain imaging techniques used to monitor brain electrical activity and infer brain function. The main goal of his analysis is to extract information-rich brain spatiotemporal spectral patterns or to infer functional connections between different brain regions. BML is an emerging field, which integrates Bayesian statistics, variational methods, and machine learning techniques, which can solve various problems such as regression, prediction, outlier detection, feature extraction, and classification. His research lacks theoretical practice [2]. Ginneken believes that lung imaging, chest X-ray, and computed tomography have always been one of the key areas in this field. He described how machine learning has become the main technology for solving lung CAD, usually yielding better results than traditional rule-based methods, and how the field is now changing rapidly. He summarized and explained the main differences between rule-based processing, machine learning, and deep learning to illustrate the various applications of CAD in the chest. His research lacks samples [3]. Yu et al. provides a PUF-based lightweight authentication method, which is very practical in the setting of server authentication to the device and the use case of limiting the number of authentication times during the entire life cycle of the device. His solution uses a server-managed challenge/response pair (CRP) locking protocol: unlike the previous method, the adaptive selective challenge adversary with machine learning cannot obtain a new CRP without the implicit permission of the server. He also proposed a degenerate instantiation using a weak PUF, which can provide security protection against opponents with unlimited computing (including any learning opponents) for the life cycle and readout rate of the actual device. His research lacks comparative data [4].

This research mainly uses machine learning clustering algorithm to simulate customer loyalty. Call the k-means interactive mining algorithm based on the Hash structure to perform data mining on the multidimensional hierarchical tree of corporate credit risk and continuously adjust the support thresholds of different levels of data mining according to specific requirements and select effective association rules until satisfactory results are obtained. After conducting credit risk assessment and early warning modeling for the enterprise, the initial preselected model is obtained. The information to be collected is first collected by the web crawler from the target website to the temporary web page database, where it will go through a series of preprocessing steps such as completion, deduplication, analysis, and extraction to ensure that the crawled web page is correctly analyzed, to avoid incorrect data due to network errors during the crawling process. The correctly parsed data will be stored for the next step of data cleaning or data analysis.

2. Statistical Modeling and Simulation of Online Shopping Customer Loyalty

2.1. Shopping under the Influence of Big Data and Machine Learning. The development of big data smart shopping and its rapid deployment have led to the generation of large amounts of data at an unprecedented speed [5, 6]. Unfortunately, due to the lack of established mechanisms and standards that benefit from the availability of such data, most of the generated data is wasted without extracting potentially useful information and knowledge [7]. In addition, the highly dynamic nature of smart shopping requires a new generation of machine learning methods that can flexibly adapt to the dynamic nature of data to perform analysis and learn from real-time data [8, 9]. Generally speaking, semisupervision is a necessary condition for smart shopping to meet this challenge. In addition, for the three-level learning framework of smart shopping, the framework should match the hierarchical nature of the big data generated by smart shopping to provide different levels of knowledge abstraction [10].

2.2. Machine Learning Framework for Online Customer Loyalty Measurement. The model to be constructed in this study is a descriptive model, and the clustering method in the unsupervised learning method is used to complete the construction of online customer loyalty measurement model. The connotation of its framework is that when users conduct commodity transactions in the Internet environment, user data and online shopping transaction data are collected and stored through web pages to form a dataset file [11, 12]. For dataset files, machine learning is applied; specifically, clustering algorithm is used for modeling, and online customer loyalty measurement model is constructed [13]. This model forms several clustering subcategories through clustering, and the subcategories contain similar users with similar customer loyalty [14, 15]. Based on the model and clustering results, recommendations based on the loyalty of similar users and similar customers or marketing strategies based on the loyalty of similar users and similar customers can be realized [16]. The machine learning framework for online customer loyalty measurement is shown in Figure 1.

2.3. Machine Learning Clustering Algorithm. Before starting the clustering algorithm, set the experimental dataset and the number of clusters k. The Euclidean distance method is usually used to calculate the distance between the object and the cluster center. The formula is as follows:

$$D\left(X_i, C_j\right)' = \sqrt{\left(X_{i1} - C_{j1}\right)^2 + \left(X_{i2} - C_{j2}\right)^2 + \ldots + \left(X_{im} + C_{jm}\right)^2}. \tag{1}$$

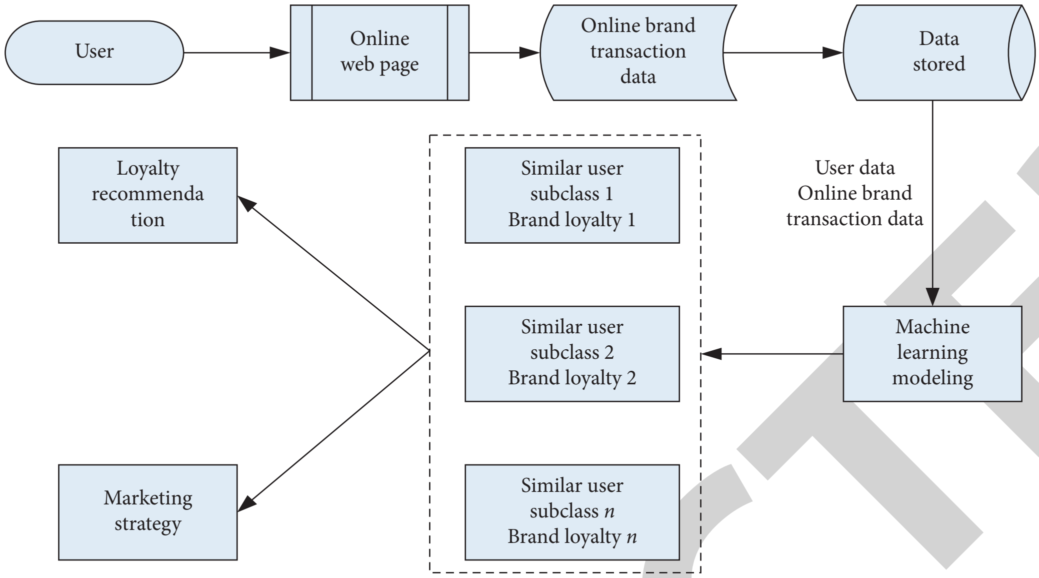

Figure 1: Machine learning framework for online customer loyalty measurement.

If $D(X_i, C_j)$ meets the conditions

$$D(X_i, C_j) = \min\{D(X_i, C_j)'\}, \tag{2}$$

then the data object X_j is put into the C_j class. After all the data objects are put into the corresponding category in turn, the cluster center is readjusted based on each object in the category. Consider

$$\begin{aligned} C_j^* &= \frac{1}{n}\sum_{i=1}^{n_j} x_i^j, \\ J^* &= \sum_{k=1}^{n_j}\sum_{j=1}^{k}\left\| x_k^j - C_j^* \right\|^2. \end{aligned} \tag{3}$$

When judging whether the cluster centers converge, the square error criterion function is selected as the criterion function [17, 18]:

$$F = \sum_{i=1}^{k}\sum_{x \in C_i}\left| x - m_i \right|^2. \tag{4}$$

Among them, F represents the sum of square errors of all objects in the experimental text set [19]. The larger the F value, the higher the accuracy of the algorithm. The calculation methods of the precision, recall, and F metric are expressed as follows:

$$\begin{aligned} P_{i,j} &= \frac{N_1}{N_2}, \\ R_{i,j} &= \frac{N_1}{N_3}, \\ F_{i,j} &= \frac{2PR}{P + R}, \end{aligned} \tag{5}$$

where N_1 refers to the number of texts in cluster i whose category is j and N_2 refers to the number of all texts in cluster i [20, 21]. Let $I(x, y)$ denote the mutual information value of variables x and y; then

$$I(x, y) = \log\frac{P(x, y)}{P(x)P(y)}. \tag{6}$$

Then, its similarity value calculation formula is

$$\text{Sim}(x, y) = \frac{MI(x, y)}{E(i, j)} = \frac{\log(P(i, j)/P(i)P(j))}{\sqrt{\sum_{k=1}^{m}\left(w_{ik} - w_{jk}\right)^3}}. \tag{7}$$

The probability P is equivalently converted to the number of texts containing keywords in the text set TD. The relationship between the two is as follows:

$$P = \frac{TD}{|D|}, \tag{8}$$

where $|D|$ represents the total number of sample information.

The mutual information value between any two texts can be expressed as

$$\begin{aligned} MI(x, y) &= \log\frac{TD(x, y)/|D|}{(TD(x)/|D|) * (TD(y)/|D|)} \\ &= \log\frac{TD(x, y) * |D|}{TD(x) * TD(y)}. \end{aligned} \tag{9}$$

The similarity value calculation formula between two pieces of information can be expressed as

$$\text{Sim}(i, j) = \frac{MI(i, j)}{E(i, j)} = \frac{\log(TD(x, y) * N/TD(x)TD(y))}{\sqrt{\sum_{k=1}^{m}\left(w_{ik} - w_{jk}\right)^2}}. \tag{10}$$

(1) Online shopping transaction data belongs to a large dataset, with a data volume of one million, which

requires extremely strong scalability and extremely high processing efficiency [22]:

$$\text{Guan}[i] = \max\{w_i | w_i \in d_i, i = 1, 2, 3, \ldots, n\}. \tag{11}$$

(2) Online shopping transaction data processing requires timely, fast execution speed and low computational complexity [23, 24].

(3) Most of the online shopping transaction data are numerical characteristic variables, and nonnumerical characteristic variables can also be converted into numerical data [25]:

$$MI(d_i, C) = \log \frac{TD(\text{Guan}[i], C) * |D|}{TD(\text{Guan}[i]) * TD(C)}. \tag{12}$$

Among them, $i = 1, 2, 3 \ldots n$.

(4) Since the number of clustering subcategories needs to be defined in advance when the algorithm is executed, there are many ways to define it. Usually, statistical methods are used to optimize the number of clustering subcategories, effectively avoiding the shortcomings of the k-means algorithm.

(5) Since the k-means algorithm is affected by the initial centroid, it can be optimized during the model construction process, effectively avoiding the shortcomings of the k-means algorithm.

3. Online Shopping Customer Loyalty Statistical Modeling Experiment

3.1. Architecture Design of the Online Data Collection Model

3.1.1. Framework Design of the Online Data Collection Model. First, extract the information needed by the client based on a certain regular expression.

Second, store the visited URL in the library and compare it with the existing URL every time a new URL is accessed to determine whether it is repeated or not.

3.1.2. Online Data Collection Model's Data Flow Design. The information to be collected is first collected by the web crawler from the target website to the temporary web page database, where it will go through a series of preprocessing steps such as completion, deduplication, analysis, and extraction to ensure that the crawled web page is correctly analyzed, to avoid incorrect data due to network errors during the crawling process. The correctly parsed data will be stored for the next step of data cleaning or data analysis. The web crawler is responsible for a relatively large amount of calculation and network input and output, so it can be physically placed on a different server from the data preprocessing program to reduce the computational load. The web page database and final data storage can also be placed on different servers to reduce network load. The data flow design of the online data acquisition model is shown in Figure 2.

3.1.3. Implementation of Online Data Collection Model for Customer Loyalty Measurement

(1) Crawl URL List according to the Theme. For writing Java programs to parse HTML documents, there are two ways to choose. One way is to apply html parser. It is an HTML parsing library written in java language. Use jsoup to parse web pages in the program, first set the subject keyword and URL and parse HTML from the obtained file or string by analyzing the structure of the website. Secondly, use the CSS selector to find the web page list information, retrieve the data, and store it in Elements.

(2) URL Judgment. Web page duplication is a prominent problem when crawling web pages. The reason is that the number of links to be crawled in the website is huge, and the structure after crawling will be very complicated. Therefore, if the crawler strategy and program are not set before crawling, the crawling process will be chaotic. There are more and more URLs in the queue to be crawled, and there will be a lot of duplicate URLs. Jsoup can effectively solve this problem, using selectors to select, that is, to judge the URL duplication, and effectively remove the URL duplication.

(3) Web Page Analysis and Storage. After the previous operations, the URL address of each web page can be obtained, and the data can be crawled using multiple threads. Because the data captured by this web crawler is the user's detailed information on the product, comments, and so on, through the analysis of the website, the part of the data is transmitted in JSON format, so web page parsing uses Ht tpURLConnection to transmit JSON format data. The URL address to be parsed and the encoding format of the website are passed in as parameters, and the corresponding json object can be obtained by parsing. For the obtained json object, it is necessary to use the json function to parse the product information and comments needed by the user. Finally, the obtained data is stored in a text file, and then it can be used by machine learning algorithms after the following four preprocessing steps:

Completion: check the network collection status. If you do not get the correct content that conforms to the template format, the content of the page needs to be downloaded again, so resubmit the task to the web crawler to download the web page.

Parsing: parse HTML web pages into understandable text, discarding format information that is irrelevant to the subject.

Extract: use the specified template format to extract the required key information from the text content obtained in the previous step and store it according to the field list.

Deduplication: compare the URL where the web page is located with the visited URL. If it has been downloaded, whether to discard the new data or update the existing data according to the rule setting—at the same time,

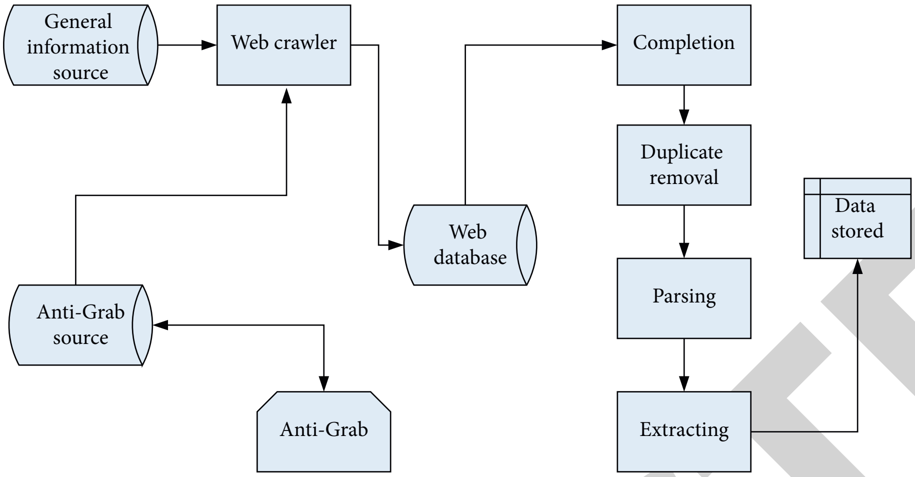

Figure 2: Online data collection model's data flow.

compared with other information, such as the data has been obtained from other URLs—then discard the new data.

3.2. *Online Data Cleaning for Customer Loyalty Measurement*

3.2.1. Online Repeated Data Cleaning Algorithm. The data in commodity data files and transaction data files need to be cleaned of repeated records. Therefore, the data file that needs to be cleaned is used as the algorithm input parameter, and the specific functions implemented include the following. First, read the original data from the data file. Then, find and record the label of the duplicate data record. Finally, delete duplicate data records based on the duplicate data record label.

In this study, for the original data files product detail_ original_ data.dat and transaction detail_ original_ data.dat, this algorithm is applied to delete duplicate records, and the cleaned data is stored in the data files product detail. dat and transaction detail. dat. Taking the cleaning file product detail_ original_ data. dat as an example, the duplicated function in the f_clean_duplicate data algorithm is used to identify and mark duplicate records. Then use the clean_data function to clean up duplicate records.

3.2.2. Cleaning of Online Error Data. The cleaning of erroneous data requires the design of the cleaning algorithm based on the error characteristics of the data. For date format errors, it is necessary to determine the correct date format and replace or eliminate the information that caused the date format errors in accordance with this standard. Taking the dataset attribute identifier as the algorithm input parameter, the specific functions implemented include the following. First, the dataset attribute for data cleaning is judged according to the dataset attribute identifier. Then, apply the method of replacing the wrong character. Finally, confirm the data set after erroneous data cleaning. The error is that the data does not conform to the format of the time and date type. Therefore, the core of this algorithm is to replace the wrong characters into characters that conform to the time and date format and use different character replacement methods according to application requirements, including one-time replacement and successive replacement.

In this study, the purchase time (buy date attribute) and comment time (comment date attribute) in the transaction detail_data need to be cleaned. Only after cleaning can it be converted to a date and time data type.

3.2.3. Online Missing Data Cleaning Algorithm for Brand Loyalty Measurement. For missing data, the simplest and most effective cleaning methods are to delete null values and estimate replacement. Null value removal is for records that are missing too much attribute information: estimated replacement is for records that are missing individual attribute information; that is, the data is derived from the original data set, and the calculated average, median, mode, and maximum are used. Replace missing data such as value or minimum value. This algorithm uses the dataset identification, processing method identification, attribute identification, and replacement value as the algorithm input parameters. The specific functions implemented include the following. First, according to the attribute characteristics of the dataset and application requirements, the appropriate processing method is selected for the missing data. Then, the processing subalgorithms are, respectively, called, including deleting the null value record and replacing the null value. Finally, confirm the data set after data cleaning. The core of this algorithm is to determine which processing method to use for missing information. The records missing too much attribute information need to be deleted, while the records missing individual attribute information need to be supplemented and replaced with reasonable information.

In this study, there are a certain number of records with null values in the product detail_data. The characteristics of these records are divided into two categories. One is that the product record data has only the product number, and the other attributes are empty; the other type of product record data contains the number of purchases of the product. These two types of data should adopt different processing methods.

The record with only the product number can no longer effectively reflect the status of the product and needs to be deleted: the record with the number of product purchases is empty, and because there is no purchase record for this product, it needs to be replaced with "0." For the null information in the product dataset, the first type of record must be eliminated first, and then the second type of record must be replaced. It can be seen from the algorithm program that the core of this algorithm is to merge the two datasets according to the connection field and filter the data records according to the filter conditions.

3.2.4. Combining and Filtering Online Data Cleaning of Brand Loyalty Metrics. The online brand loyalty measurement dataset includes product data and transaction data stored in different data source files. It is necessary to establish the relationship between the product and the transaction by specifying the relevant information (product number), merge the data, and establish a merged data set pro_ tran_ data. This algorithm uses the dataset, connection fields, and filter conditions to be merged as input parameters of the algorithm. The specific functions implemented include the following. First, merge the datasets. Then, for the combined dataset, filter records based on filter conditions.

Under normal circumstances, the establishment of a model does not use all the original data but filters out data that meets specific conditions according to the application goal. This research uses commodity brand transaction data as the carrier to complete brand loyalty clustering. According to the research question, combined with the data visualization results of the data cleaning stage, the algorithm filters out specific brand transaction data of specific categories of commodities from the transaction data. That is, the product category is "memory card" and the brand is "SANDISK" commodity transaction data.

3.3. Data Mining and Result Expression. Call the k-means interactive data mining algorithm based on the Hash structure to mine the multidimensional hierarchical tree of corporate credit risk layer by layer and continuously adjust the support thresholds for mining at different levels according to specific requirements and select effective association rules until satisfactory results are obtained.

After conducting credit risk assessment and early warning modeling for the enterprise, the initial preselected model is obtained. Then, the credibility, operational convenience, and security of the preselected model must be further compared before the final model can be obtained. First of all, perform technical evaluation on the preselected models and optimize those models with poor results to obtain backup models. Then, the standby model is further evaluated through business evaluation, and the final early warning model is obtained after optimization. Finally, the constructed enterprise credit risk assessment and intelligent early warning model are applied.

3.4. Algorithm Implementation. In this research, two user feature sets are finally constructed. One is a dataset containing the user number (usrid) and user level; the other is a dataset containing the user number (usrid) and the area where the user is located (max_usrlocnum). In the algorithm f feature_user, data_attribute_flag is the user attribute flag. When the value is 2, it represents the user level attribute, and when the value is 3, it represents the area attribute of the user.

For the memory card transaction records in the dataset protran.cck_data, the user's area is converted to an integer data type to obtain the corresponding numerical information. Although there is no duplicate and conflicting information, it can be counted in the same way as the user level. The statistical result is the dataset of the user's area.

4. Statistical Modeling Analysis of Online Shopping Customer Loyalty

4.1. Reliability Analysis. Reliability reflects the stability of survey samples and is used to test the consistency of the measurement results of the same measurement item after multiple measurements. Reliability is generally expressed by the inherent reliability coefficient Cronbach reliability coefficient, which can measure whether a certain set of items of the sample is measured for the same feature. If the reliability coefficient is above 0.7; although there is still value, it should be measured on the sample. Regarding modification, if it is lower than 0.7, the sample needs to be redesigned. Each variable in this study has corresponding measurement items. It is very important for the follow-up research to measure whether these items are inherently consistent. Use SPSS22.0 software to calculate the Cronbach reliability coefficient and test the internal consistency of the corresponding measurement items of the seven research variables in this article. The test results are shown in Table 1. The results of the reliability analysis are shown in Figure 3. It can be seen from Table 1 that the Cronbach reliability coefficients of each research variable are all above 0.7, reaching the reliability test standard, indicating that the measurement items of each research variable have good internal consistency. The data is relatively reliable and can be analyzed in the next step.

To perform k-means data analysis, we also need to test whether the data meets some of the assumptions of k-means; that is, the normality test of the data and the violation estimation test are required. The data research results are shown in Table 2. The results in Table 2 show that the range of skewness and kurtosis of the data in this study is between −2 and +2, which basically meets the requirements of the normal distribution of sample data. The results of k-means data analysis are shown in Figure 4.

4.2. Overall Model Fit. Before verifying and analyzing the model hypothesis, the fit of the model needs to be tested to determine whether the data and the model fit. The indicators obtained after running the model are shown in Table 3. The results of the model fit test are shown in Figure 5. From this

Table 1: Test results of internal consistency.

Research variables	Reliability factor	Number of items
Commodity characteristics	0.788	4
Shopping	0.813	6
Service quality	0.853	6
Customer satisfaction	0.831	3
Relationship trust	0.834	6
Loyal attitude	0.753	5
Behavioral loyalty	0.796	4

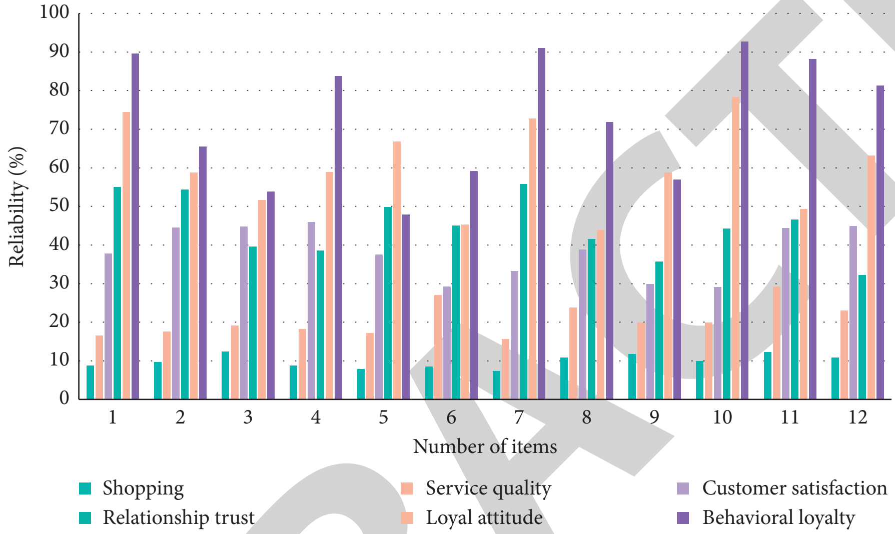

Figure 3: Reliability analysis results.

Table 2: Data research results.

Variable	Minimum	Max	Skewness	Kurtosis
A1	1.000	7	−0.692	0.047
A2	1.000	7	−0.723	0.096
A3	1.000	7	−0.839	0.470
B1	1.000	7	−0.966	0.54
B2	1.000	7	−0.921	0.472
B3	1.000	7	−1.157	1.362

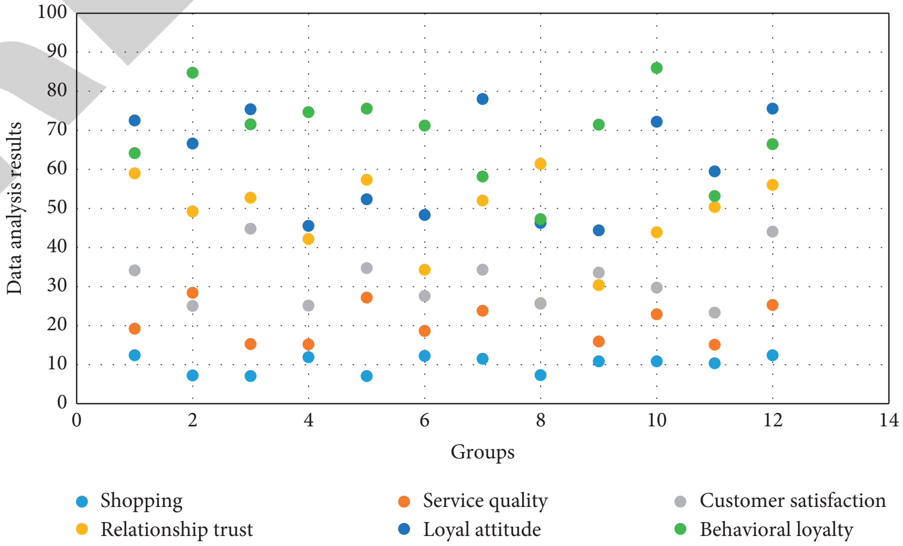

Figure 4: *K*-means data analysis results.

Table 3: Indicators obtained after running the model.

Evaluation index value of each fit	Actual value	Expected value
Chi-square value (x2)	1725.594	x2/df should be less than 3
Degrees of freedom (df)	840	
x2/df	2.054	
GFI	0.877	Greater than or equal to 0.9, indicating good model fit, between 0.8 and 0.9, indicating acceptable model fit
AGFI	0.862	
IFI	0.921	
NFI	0.857	
CFI	0.921	
RMSEA	0.053	Less than 0.05 means good fit, between 0.05 and 0.08

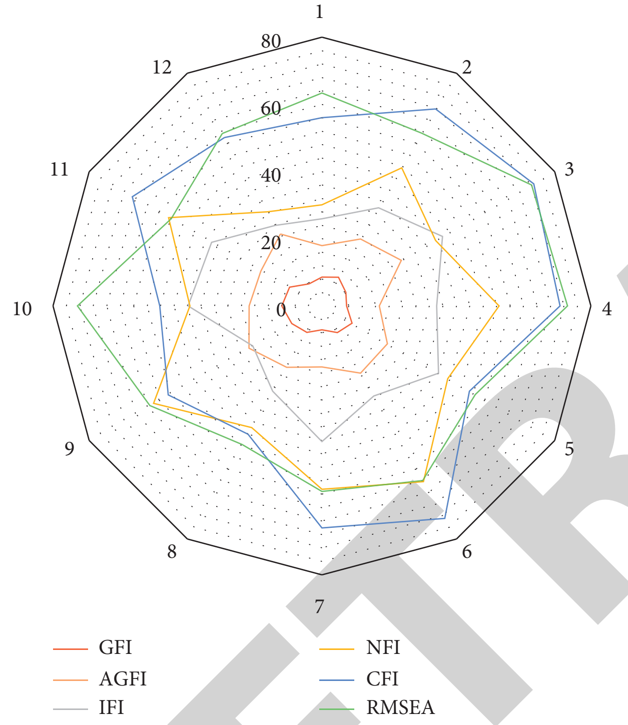

Figure 5: Test results of model fitting degree.

table, we can see that the x2 value is 1725.954, the df degree of freedom value is 840, the x2/df value is 2.054, which meets the requirement of less than 3, and the goodness of fit index (GFI) value is 0.877, which is greater than 0.8. The adjusted goodness of fit index (AGFI) value is 0.862, which meets the requirement of greater than 0.8. The increased fitting index (IFI) value is 0.921, the nonnormative fitting index (NFI) value is 0.857, and the comparative fitting index (CFI) value is 0.921. The values of these indexes all meet the model fitting requirements. At the same time, the approximate root mean square error (RMSEA) value is 0.053, between 0.05 and 0.08, indicating that the model achieves a relatively good fitting effect. In summary, the model proposed in this study fits the data well, and all indicators can better meet the standards set by the fitting index.

The research sample is divided into two samples according to whether the survey object is a student. Among them, there are 213 valid questionnaires for the student group and 168 valid questionnaires for the nonstudent group. By adopting the revised model, using the model to run the student sample and the nontudent sample, respectively, the various fitting indicators after running are shown in Table 4. It can be seen that the two subsample data fit the model well, and the main indicators can pass the test. The comparison results of the survey samples are shown in Figure 6.

4.3. Validity Analysis. Validity refers to the extent to which the survey sample can measure the things needed to be measured; in simple terms, it refers to the validity of the measurement results. This article is mainly to test the construct validity of the survey sample. It refers to the actual measurement of the theoretical structure and characteristics that a test hopes to measure: the degree of measurement or whether the experiment measures the hypothetical theory. In order to test whether the designed measurement item is an explanation of the corresponding measurement variable, this article uses exploratory factor analysis in SPSS22.0 software to perform factor analysis on each research variable and calculate the measurement item of each variable in the variable. In order to test the construction validity of the survey sample, it is generally believed that if the factor loading value is greater than 0.5, it means that the measurement item can effectively explain the variable to be measured. The specific analysis results are shown in Table 5. The results of the validity analysis are shown in Figure 7. It can be seen from Table 5 that the KMO statistics of all research variables are greater than 0.7, which is an acceptable level. From the results of factor analysis of each variable, the factor loading value of each measurement item is basically above 0.7, and the factor loading validity standard greater than 0.5 indicates that each item has a high degree of correlation with the corresponding measurement variable, and the common factor can represent the corresponding measurement item. At the same time, the cumulative explanatory variance of each public factor is more than 73%, which can explain most of the overall situation, indicating that the questionnaire has a good construct validity, and it is effective to use the items of the questionnaire design to measure.

The estimated values of standardized path coefficients for the relationship between variables are shown in Table 6. The

Table 4: Fitting indexes after operation.

	Student sample	Nonstudent sample
Ratio of chi-square to degrees of freedom	1.357	2.263
Normative fit index (NFI)	0.823	0.811
Comparative fit index (CFI)	0.836	0.827
Incremental fit index (IFI)	0.950	0.844
Goodness of fit index (GFI)	0.849	0.830
Tucker Lewis index (TLI)	0.946	0.894
Approximate root mean square residual (RMSEA)	0.040	0.069

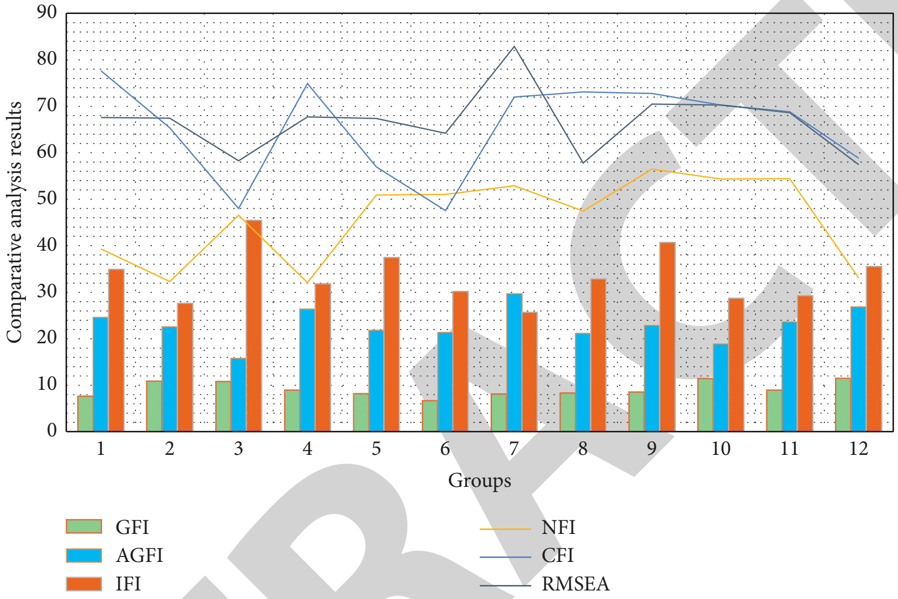

Figure 6: Comparison results of survey samples.

Table 5: Specific analysis results.

Measured variable	Factor loading value	Cumulative explained variance	Statistics	Significance
Service quality	0.784 0.707	74.496%	0.795	0.00
Customer satisfaction	0.739 0.763	76.239%	0.765	0.00
Relationship trust	0.789 0.762	75. 701%	0.773	0.00
Loyal attitude	0.793 0.829	73.784%	0.756	0.00
Behavioral loyalty	0.754 0.796	74.194%	0.744	0.00

first 6 data analysis results are shown in Figure 8. The relationship between customer satisfaction and loyalty is shown in Figure 9. It can be seen from Table 6 that the path coefficients are all positive, and most hypotheses pass the test at the 0.05 significance level. Only the P value of the customer satisfaction and behavior loyalty path coefficient is greater than 0.05, indicating that this path is not significant. Hypothesis 8 fails the test. For this result, this article believes that the shopping website industry is in a fierce market competition environment, and each supermarket will try to do its best in terms of product characteristics, shopping environment, and service quality to satisfy customers. Therefore, for customers and shopping websites that are satisfied, there are many companies that have a certain influence on behavioral loyalty, but it is not enough to keep customers loyal to a supermarket. Among the validated hypotheses, product characteristics have the greatest impact on customer satisfaction and relationship trust, and the path coefficients are 0.299 and 0.292, respectively, indicating that customers pay the most attention to all aspects of the products provided by shopping websites. Generally speaking, customers' demand is their main purpose to go to the

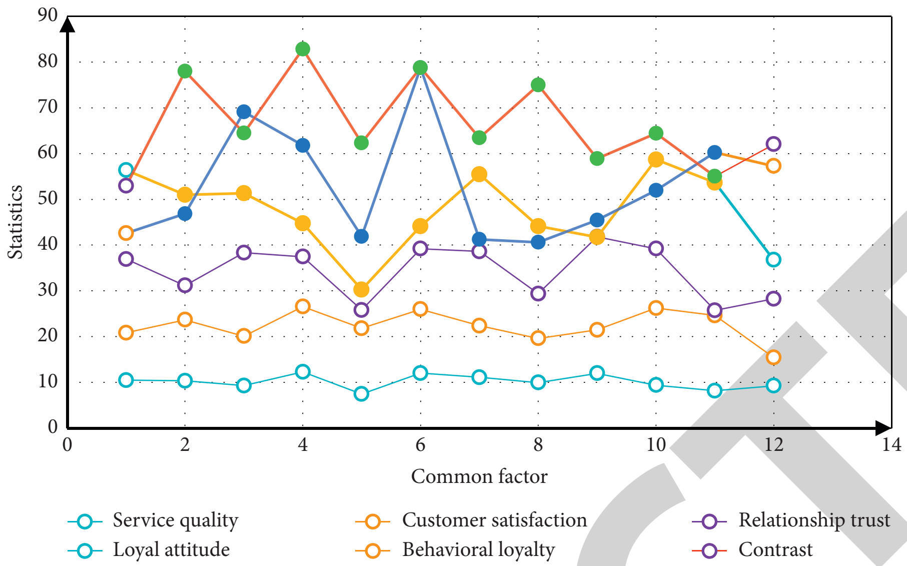

Figure 7: Results of validity analysis.

Table 6: Standardized path coefficient estimates for the relationship between variables.

Label	Path relationship	Estimated value	C. R.	Test result
H1	Customer satisfaction-product characteristics	0.299	3.522	Stand by
H2	Relationship trust-commodity characteristics	0.292	4.098	Stand by
H3	Customer satisfaction-shopping environment	0.258	2.081	Stand by
H4	Relationship trust-shopping environment	0.215	3.524	Stand by
H5	Customer satisfaction-service quality	0.271	3.079	Stand by
H6	Relationship trust-quality of service	0.286	3.813	Stand by
H7	Loyal attitude-customer satisfaction	0.415	6.538	Stand by
H8	Behavioral loyalty-customer satisfaction	0.122	1.165	Not support

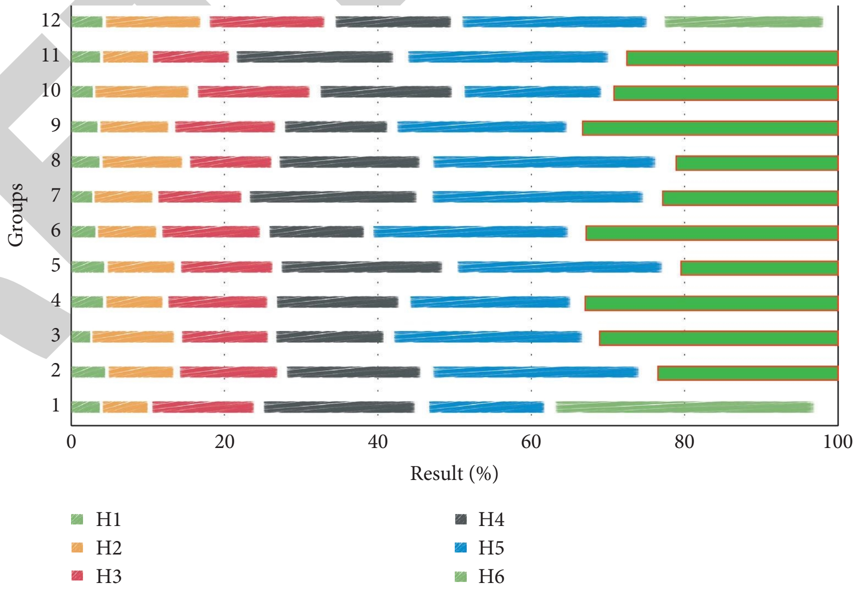

Figure 8: Results of the first six data analysis.

shopping website; the shopping environment is a specific scenario, customers are easily satisfied in a good shopping environment, and relationship trust requires a longer process, and the path coefficients are 0.286 and 0.271, respectively, indicating that the shopping website quality of service in the shopping environment gives customers a

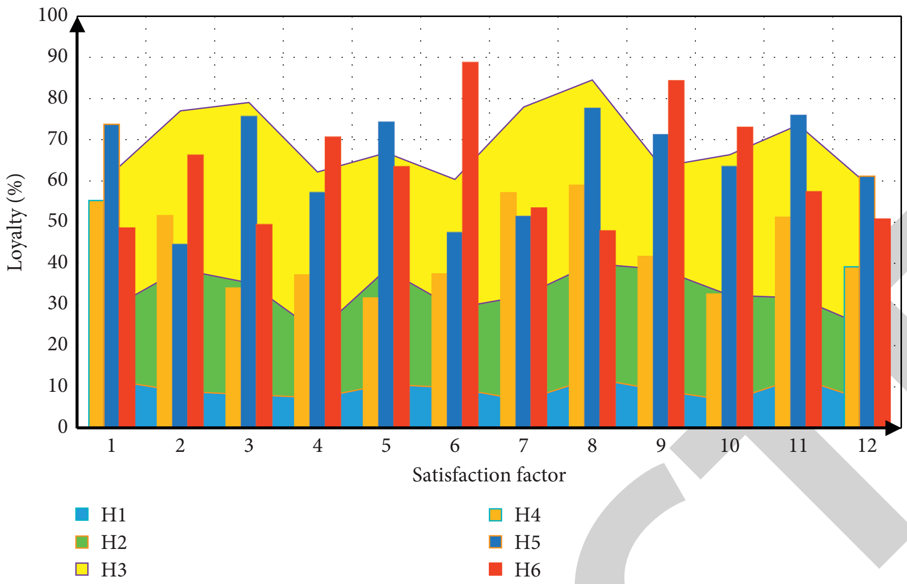

Figure 9: Relationship between customer satisfaction and loyalty.

deeper impression, because the service is provided by the staff, and the good relationship between the customer and the staff is conducive to the formation of trust in the shopping website; customer satisfaction has a positive impact on attitude loyalty, the path coefficient is 0.415, and relationship trust has a greater impact on attitude loyalty. The path coefficient is 0.585. Relationship trust will also have an impact on behavior loyalty. The path coefficient is 0.388, indicating that shopping websites must more importantly satisfy customers. It is necessary to enhance customers' trust in the relationship of shopping websites, which has a greater effect on maintaining customer loyalty.

5. Conclusion

For missing data, the simplest and most effective cleaning methods are to delete null values and estimate replacement. Null value removal is for records that are missing too much attribute information: estimated replacement is for records that are missing individual attribute information; that is, the data is derived from the original dataset, and the calculated average, median, mode, and maximum are used. Replace missing data such as value or minimum value. The algorithm uses the dataset identification, processing method identification, attribute identification, and replacement value as the algorithm input parameters. The core of this algorithm is to determine which processing method to use for missing information. The records missing too much attribute information need to be deleted, while the records missing individual attribute information need to be supplemented and replaced with reasonable information.

In this study, there are a certain number of null records in the commodity dataset. The characteristics of these records are divided into two categories. One is that the product record data has only the product number, and the other attributes are empty; the other type of product record data contains the number of purchases of the product. These two types of data should adopt different processing methods. The record with only the product number can no longer effectively reflect the status of the product and needs to be deleted: the record with the number of product purchases is empty because there is no purchase record for this product, and it needs to be replaced and supplemented. For the null information in the product data set, the first type of record must be eliminated first, and then the second type of record must be replaced. It can be seen from the algorithm program that the core of this algorithm is to merge the two datasets according to the connection field and filter the data records according to the filter conditions.

Call the *k*-means interactive data mining algorithm based on the Hash structure to mine the multidimensional hierarchical tree of corporate credit risk layer by layer and continuously adjust the support thresholds for mining at different levels according to specific requirements and select effective association rules until satisfactory results are obtained. After conducting credit risk assessment and early warning modeling for the enterprise, the initial preselected model is obtained. Then, the credibility, operational convenience, and security of the preselected model must be further compared before the final model can be obtained. First of all, perform technical evaluation on the preselected models and optimize those models with poor results to obtain backup models. Then, the standby model is further evaluated through business evaluation, and the final early warning model is obtained after optimization. Finally, the constructed enterprise credit risk assessment and intelligent early warning model are applied.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] T. Hunt, C. Song, R. Shokri et al., "Privacy-preserving machine learning as a service," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 123–142, 2018.

[2] J. X. Wang, J. L. Wu, and H. Xiao, "Physics-informed machine learning for predictive turbulence modeling: using data to improve RANS modeled Reynolds stresses," *Physical Review Fluids*, vol. 2, no. 3, pp. 1–22, 2016.

[3] V. B. Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning," *Radiological Physics and Technology*, vol. 10, no. 1, pp. 1–10, 2017.

[4] M. D. Yu, M. Hiller, J. Delvaux et al., "A lockdown technique to prevent machine learning on PUFs for lightweight authentication," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 3, pp. 146–159, 2017.

[5] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: approaches and challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.

[6] S. M. Copp, P. Bogdanov, M. Debord et al., "Base motif recognition and design of DNA templates for fluorescent silver clusters by machine learning," *Advanced Materials*, vol. 28, no. 16, pp. 5839–5845, 2016.

[7] N. Taherkhani and S. Pierre, "Centralized and localized data congestion control strategy for vehicular ad hoc networks using a machine learning clustering algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 11, pp. 1–11, 2016.

[8] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: pushing the limits for machine learning, artificial intelligence, and computer vision," *IEEE Consumer Electronics Magazine*, vol. 6, no. 2, pp. 48–56, 2017.

[9] H. H. Kim and N. R. Swanson, "Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods," *International Journal of Forecasting*, vol. 34, no. 2, pp. 339–354, 2016.

[10] R. A. Taylor, J. R. Pare, A. K. Venkatesh et al., "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach," *Academic Emergency Medicine*, vol. 23, no. 3, pp. 269–278, 2016.

[11] G. Wang, M. Kalra, and C. G. Orton, "Machine learning will transform radiology significantly within the next 5 years," *Medical Physics*, vol. 44, no. 6, pp. 2041–2044, 2017.

[12] Y. Chen, T. Chen, Z. Xu, N. Sun, and O. Temam, "DianNao family," *Communications of the Acm*, vol. 59, no. 11, pp. 105–112, 2016.

[13] G. Valdes, T. D. Solberg, M. Heskel, L. Ungar, and C. B. Simone, "Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy," *Physics in Medicine and Biology*, vol. 61, no. 16, pp. 6105–6120, 2016.

[14] P. Plawiak, T. Sosnicki, M. Niedzwiecki, Z. Tabor, and K. Rzecki, "Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1104–1113, 2016.

[15] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1108–1120, 2016.

[16] R. N. Bryan, "Machine learning applied to alzheimer disease," *Radiology*, vol. 281, no. 3, pp. 665–668, 2016.

[17] J. Joung, "Machine learning-based antenna selection in wireless communications," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2241–2244, 2016.

[18] Y. Li, H. Li, F. C. Pickard et al., "Machine learning force field parameters from ab initio data," *Journal of Chemical Theory and Computation*, vol. 13, no. 9, pp. 4492–4503, 2017.

[19] I. Orsolic, D. Pevec, M. Suznjevic, and L. Skorin-Kapov, "A machine learning approach to classifying YouTube QoE based on encrypted network traffic," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22267–22301, 2017.

[20] M. W. Gaultois, A. O. Oliynyk, A. Mar et al., "Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties," *Apl Materials*, vol. 4, no. 5, pp. 199–205, 2016.

[21] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2017.

[22] Z. Obermeyer and E. J. Emanuel, "Predicting the future-big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016.

[23] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[24] R. H. Byrd, G. M. Chin, W. Neveitt et al., "On the use of stochastic hessian information in optimization methods for machine learning," *Siam Journal on Optimization*, vol. 21, no. 3, pp. 977–995, 2016.

[25] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.