DLINE JOURNALS

Journal of Intelligent Computing

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2026: 17 (1)
https://doi.org/10.6025/jic/2026/17/1/14-25

Data Mining–Driven Customer Retention in Mobile Communication Systems

Wei Yue
Physical Education Department of Henan University of Urban Construction
Pingdingshan, Henan, 467036. China
18637591599@163.com

ABSTRACT

The rapid expansion of mobile communication services and the intensification of market competition have made customer retention a critical determinant of profitability and sustainability for telecommunications operators. With the proliferation of operational support systems and customer interaction platforms, mobile operators generate massive volumes of heterogeneous data that, if properly analyzed, can yield actionable insights for strategic decision making. This paper presents a comprehensive, systematically revised study of the application of data mining techniques to customer value analysis, segmentation, and retention in the mobile communications sector. Building on customer lifetime value (CLV/LTV) theory and hybrid data mining algorithms, a communication data mining model is developed to support customer relationship management (CRM) and decision support systems. An expectation maximization (EM)–based hybrid algorithm is employed to handle incomplete and mixed type data, enabling robust customer classification and churn related analysis. Experimental results using SAS Enterprise Miner demonstrate satisfactory predictive accuracy and practical relevance. The findings confirm strong associations between customer satisfaction, loyalty, and churn risk, and highlight the importance of identifying high value yet churn prone customer segments. This study contributes to the literature by integrating value theory, probabilistic modeling, and data mining within the specific operational context of mobile communication enterprises, particularly in emerging markets.

Keywords: Data Mining, Customer Retention, Mobile Communications, Customer Lifetime Value, Hybrid Algorithms

Received: 28 August 2025, Revised 27 November 2025, Accepted 4 December 2025

1. Introduction

Following China's accession to the World Trade Organization (WTO), regulatory liberalization and market-oriented reforms significantly intensified competition in the telecommunications industry. [1] Among various subsectors, mobile communications rapidly emerged as the most dynamic and profitable segment, surpassing

traditional fixed line services and was closely followed by internet based communication services. [2] A defining global trend in telecommunications development has been the transition from wired to wireless networks, driven by advances in mobile technology and changing consumer preferences.

Industry reports from organizations such as the International Telecommunication Union (ITU) and International Data Corporation (IDC) indicate that while overall global telecommunications growth moderated during the 1990s, mobile communications experienced explosive expansion. [3] In China, this trend was even more pronounced: mobile subscriber growth rates far exceeded those of fixed line telephony, and mobile services gradually became the primary revenue driver of the telecommunications sector. [14] As mobile penetration increased, traditional revenue sources such as long distance and local fixed line services declined in relative importance, while mobile service revenues steadily rose.

By the mid-2000s, the number of mobile users had begun to outpace fixed line subscribers, and projections suggest continued dominance of mobile services in the coming decades. In such a competitive and saturated market, acquiring new customers is increasingly costly, making customer retention and value maximization central strategic objectives for mobile operators. Consequently, advanced analytical methods capable of extracting knowledge from large scale customer data have become indispensable. [5, 6, 7]

Recent research also highlights the integration of big data analytics with customer relationship management (CRM) systems. By combining structured transactional data with unstructured behavioral and sentiment data, organizations can develop highly personalized and proactive retention strategies. These advances reinforce the notion that data mining and predictive analytics are no longer optional tools, but foundational components of modern retention management.

2. Review of the Earlier Research

Telecommunications stand as the fastest-growing industry globally. As this sector expands, the challenges and competition also escalate. To navigate these challenges, telecommunications providers are seeking to understand customer needs better, and the utilization of information has become crucial for survival [8]. For more established mobile operators, the extensive historical data accumulated through operating and support systems is undoubtedly a significant asset. Data mining systems are among the most effective approaches for fully leveraging these vital resources to achieve the aforementioned objectives [9]. Nevertheless, we believe that research on the application of data mining within China's mobile communications sector remains limited. This viewpoint stems from observed gaps in prior studies, which have predominantly focused on telecommunications rather than the mobile communications industry. Thus, building on previous work, this article aims to examine in greater depth data mining applications in mobile communication operations. Historically, research has been more concentrated on data warehouses, while the data mining that relies on them has not received adequate focus. Our national research in this domain remains in its nascent stages, whereas other countries have made considerable strides in this field. We aspire for this research to be integrated with China's unique circumstances to enhance and advance the utilization of data mining technology in mobile communication operations within the country [10]. In summary, this article examines the implementation and application of data mining technologies to enhance the competitiveness of mobile communication operators in China.

Data mining has emerged as a transformative discipline across a broad spectrum of human endeavors by

uncovering previously unknown yet actionable patterns often termed “knowledge” from vast and complex datasets. Owing to its analytical power and versatility, data mining has become an indispensable tool in numerous application domains, including banking, retail, healthcare, insurance, and bioinformatics, where it supports strategic decision making, risk mitigation, and service optimization [11]

In higher education, student retention is a critical performance metric, commonly measured as retention rates, which reflect the proportion of students who continue their enrollment from one academic year to the next. High retention rates are typically the result of institutional efforts to implement targeted academic support mechanisms and pedagogical innovations aimed at reducing attrition and preventing study deferrals. To effectively tackle this persistent challenge, it is essential to examine both the underlying determinants of student persistence and the methodological frameworks that enable the construction of robust predictive models. In this regard, Educational Data Mining (EDM) and Learning Analytics (LA) have gained widespread adoption as powerful approaches for extracting meaningful insights from educational data. These techniques not only enhance instructional strategies but also facilitate the early identification of at risk students. A growing body of literature demonstrates the successful application of predictive methods including machine learning (ML), statistical analysis, and deep learning (DL) within LA ecosystems. A systematic review of recent academic publications underscores the efficacy of these advanced analytical techniques in addressing complex student retention issues and informing proactive intervention strategies [12]

Parallel developments are evident in the commercial sector, where customer retention has become a cornerstone of sustainable business growth and long-term profitability. The ability to anticipate and mitigate customer churn empowers organizations to refine engagement tactics, elevate satisfaction levels, and strengthen financial outcomes. Machine learning plays a pivotal role in this domain, with models such as logistic regression, decision trees, random forests, and gradient boosting algorithms (e.g., XGBoost, LightGBM) consistently delivering high predictive accuracy. Moreover, deep learning architectures including Long Short-Term Memory networks (LSTMs) and artificial neural networks have shown promise in capturing temporal and nonlinear patterns in customer behavior. These models draw upon heterogeneous data sources, such as transaction histories, digital footprints, sentiment signals from social media or surveys, and macroeconomic indicators, to generate early warnings of potential churn. Crucially, the effectiveness of these predictive systems is heavily dependent on rigorous feature engineering and comprehensive data preprocessing, which ensure model robustness and yield interpretable, business-relevant insights [13] (Adekunle, Bolaji Iyanu, n.d.).

Further extending this analytical capability, big data analytics has been increasingly integrated into Customer Relationship Management (CRM) systems to revolutionize how organizations understand and interact with their customers. By harnessing massive volumes of structured and unstructured data related to customer preferences, past interactions, and real time behaviors, businesses can tailor their CRM strategies with remarkable precision. A recent study explores this integration in depth, employing a comprehensive literature review and case study analysis of firms that have successfully embedded big data analytics into their CRM frameworks. The findings highlight how data-driven personalization and predictive engagement not only improve customer loyalty but also drive operational efficiency and competitive advantage [14]

Ultimately, the growing complexity of decision-making environments whether in education or enterprisenecessitates the use of advanced analytical tools that can decipher intricate data relationships and translate

them into timely, evidence based actions. As noted by Chen [15] "the sheer complexity of the factors influencing decision making has required organizations to use a tool to understand the relationships between data and make various appropriate decisions based on the information obtained." In this context, data mining, predictive analytics, and big data technologies are no longer optional enhancements but foundational components of modern retention strategies across sectors.

3. Methods

3.1 Design of the Communication Data Mining Model

Customer value is commonly conceptualized as a function of cost performance trade-offs and acquisition or management costs. Value theory suggests that customer value comprises emotional, functional, and rational dimensions. Building on Hartman's value framework and subsequent adaptations for service marketing, this study treats customer value in mobile communications as a multidimensional construct encompassing emotional experience, practical utility, and rational evaluation.

Central to this study is the concept of Customer Lifetime Value (CLV or LTV), defined as the total net contribution of a customer to an enterprise over the entire duration of the customer firm relationship. This contribution includes both direct financial returns (e.g., service usage and purchases) and indirect effects (e.g., referrals, recommendations, and feedback). LTV integrates historical and prospective value, providing a comprehensive basis for customer-centric decision making.

LTV is typically employed in customer segmentation, evaluation of relationship strength, identification of high value customers, selection of communication channels, and loyalty program design. Two primary dimensions characterize LTV measurement: the time dimension (past and future value) and the data source dimension (supplier and customer derived information). Based on these dimensions, customers can be categorized according to current value (CV) and potential value (PV), enabling strategic differentiation among low value, high potential, and high value customers

$$V = CV + PV$$

where current value reflects realized profitability, and potential value represents expected future contributions estimated through predictive modeling.

3.2 Hybrid Data Mining Algorithm

Real world telecommunications data are often incomplete, noisy, and heterogeneous. To address these challenges, this study adopts a hybrid data mining approach based on the Expectation Maximization (EM) algorithm. The EM algorithm, proposed initially by Dempster, Laird, and Rubin, is an iterative method for estimating parameters in the presence of missing or incomplete data.

Each EM iteration consists of two steps: (i) the Expectation (E) step, which computes the expected value of the log likelihood function given current parameter estimates and observed data; and (ii) the Maximization (M) step, which updates parameter estimates by maximizing this expected log-likelihood. Under the assumption of multivariate normality, the EM algorithm enables robust estimation of regression coefficients and covariance structures, even when portions of the data matrix are missing.

By integrating EM-based estimation with customer value and segmentation models, the proposed approach

improves prediction robustness and supports reliable customer classification in complex operational environments.

Below is a data table for assessing LTV values

	Supplier	Customer
Past times	I customer qualityA period of time as a customerThe number of products sold at a certain periodSales volume for different products of the same customerSales per periodStart the total sales from the first transactionProfit contribution in each periodProfit contribution from the first transaction	II Customer satisfaction with product related servicesCustomers' atisfaction with the purchase of products last yearCustomer budgetCustomer recommendation to the companyThe ratio of the customer to the budget of the companyConversion cost(perceived by the customer)
Future	III potential customersSales forecastPrediction of customer life cycleSales trendProfit forecast	IV potential suppliersRepeat purchase intentionThe degree of willingness to recommend companiesChanges in the proportion of the customer's expenditure on the budgetChanges in customer related budgets

Table 1. The Two Dimensions of LTV

Therefore, it can be concluded that the LTV of a customer j at time p is:

$$LTV_j = \sum_{t=0}^{p} CQ_{jt}(1+r)p-t + \sum_{t=p+1}^{n}(CS_{jt} \times CP_{jt})(1+r)^{p-t}t = 0,...p,...n \quad (1)$$

CQ_{jt} is the customer quality, f is the unit time sales, profit contribution, different product quantity. CS_{jt} t is the customer's share, expressed by f(SQjt, SPjt). SQ_{jt} is the supplier quality, with f(customer satisfaction, identity, trust). SP_{jt} is a potential supplier, using f(purchase intention, expected customer share, budget line). CP_{jt} is a potential customer; R is the discount rate; P is the time from the first transaction.

According to the "80/20" principle in marketing, 20% of customers are responsible for 80% of a company's profits. In reality, the profit distribution in many industries can be even more skewed. Over 100% of profits may come from less than 10% of the customer base. A study conducted by John McKean involving 35 companies across financial services, telecommunications, and retail revealed that: the highest percentage of customers contributing to profits is 25%, with a minimum of 2% and an average of 15%. To maintain a competitive advantage, businesses must focus on a tiny group of high value customers. Generally, there are three primary methods to identify high value customers among a larger customer base: First, using previous transaction records to identify the most profitable customers. This approach assumes that "customers will replicate past behaviors," meaning that customers who previously generated high margins will continue to do so. The benefit of this method is its simplicity and the availability of easily calculable data; however, it does not account for

future customer behavior. A customer's spending power may have changed, potentially leading to a stagnation or decline in future spending. Increased marketing efforts may not necessarily yield a corresponding increase in profits, making this method less cost-effective. The second method focuses on assessing the potential value of customers. This approach posits that the value customers previously generated is no longer relevant, as companies cannot alter it. What truly matters for businesses is the potential future value that customers can develop, which is often overlooked. The potential value of a customer is the anticipated profit and value derived from their future actions. The analysis can be illustrated using a 2-by-2 matrix, as shown in the table below.

		Current value	
potential value		**low**	**high**
	high	Π	I V
	low	I	III

Table 2. Customer Value Matrix Diagram

Customer value (V) is defined as the sum of potential value (PV) and current value (CV). The first quadrant of the chart indicates the lowest overall value, making it less appealing to the business. Conversely, the fourth quadrant boasts the highest total customer value and is prioritized in customer relationship management. The second quadrant depicts a scenario in which the current value is lower than the potential value. Customers in this category currently do not generate profit for the enterprise, yet they should be considered as targets for development in customer relationship management. The third quadrant, characterized by high current value and low potential value, signals a customer's development potential. These customers have become loyal patrons, and efforts should be made to maintain these relationships to further increase enterprise profit. Current value is primarily concerned with factors such as profit growth and cost changes directly attributable to the customer for the business. If we assume that the customer and the enterprise maintain their trading relationship for N years, and that the initial cost to attract customers (mainly marketing expenses) is C, with the customer's first purchase price being P1, the firm anticipates receiving annual income of R from customers. This can be used to derive the current value of the customer (CV).

$$CV = P_1 - C + R \cdot [\frac{(1+i)^N - 1}{i(1+i)^N}] \qquad (2)$$

The value of customer potential value (PV) is based on customer history and current behavior. A linear regression model can predict the potential value of a single customer. According to the matrix of the above table, we classify customers and predict their potential value by using probability model. Suppose customer I buys product j, then:

$$Y_{ij}^* = \beta_j X_i + \sum_{k=1}^{j} \gamma_{jk} Z_{ik} + \varepsilon_{ij} \qquad (3)$$

$$\begin{cases} Y_{ij} = 1 \quad when \quad Y_{ij}^* > 0 \\ Y_{ij} = 0 \quad when \quad Y_{ij}^* < 0 \end{cases} \qquad (4)$$

Among them, *Yij* represents whether customer I owns product j. Y_{ij}^{*} is a variable, *Xi* is the demographic index of customer I (such as age, income, etc.). Zik for customer I already has the purchased product or service k. ε_{ij} is the error term. The probability model of the potential value of customer I is as follows:

$$PV_i = \sum_{k=1}^{k} \Pr ob(Y_{ik} = 1) \times \Pr ofit_k \qquad (5)$$

3.2 Hybrid Data Mining Algorithm

The filling algorithm in mixed state mainly includes EM algorithm, Ml algorithm and RE algorithm. The following is a detailed introduction. This paper primarily USES the EM algorithm. For the EM algorithm, let's say $Y = \chi\beta + \varepsilon = \beta_0 + X\beta_{\sum} + \varepsilon$. Where Y is $n \times 1$ response variable, χ is $n \times (p+1)$ order matrix. So the first column has all 1's, so it's χ minus the first column. $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the regression coefficient vector, ε is ^{nx1} random error vector, and its distribution is $N(0, \sigma^2 I)$. There is a data deletion in the matrix, *Z*=(*Y, X*)that is, partial data of some variables are incomplete. In 1977, the algorithm proposed by Dempster, Larid and Rubin was an iterative method which was widely used to deal with missing data problems. Each iteration of the EM algorithm consists of two steps: expected value (E step) and maximum value (M step). For mixed deletion, the missing mode of matrix *Z*=(*Y, X*) can be used to establish a similar *EM* algorithm. Let *Z*=(*Y, X*) obey the multivariate normal distribution, namely:

$$Z_i(y_i, x_i) \sim N_P(\mu, \sum) \qquad (6)$$

Among them:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \sum = \begin{pmatrix} \sum_{yy} & \sum_{yx} \\ \sum_{xy} & \sum_{xx} \end{pmatrix} \qquad (7)$$

In the matrix $Z = (Y, X)$, Z_{obs} is known, and the logarithmic maximum likelihood function of parameter $\theta = (\mu, \sum)$ is:

$$\begin{aligned} Q(\theta / \theta^{t-1}) &= E[\log f(Z_{obs}, Z_{mis} / \theta) / Z_{obs}, \theta^{t-1}] \\ &= \int_{Z_{mis}} \log f(Z_{obs}, Z_{mis} / \theta) f(Z_{mis}, Z_{obs} / \theta^{(t-1)}) dZ_{mis} \\ &= \int_{Z_{mis}} \log f(X, Y / \theta) f(Z_{mis}, Z_{obs} / \theta^{(t-1)}) dZ_{mis} \end{aligned} \qquad (8)$$

Where, θ^{t-1} is the estimation of parameter θ obtained by the t—l step iteration. $Q(\theta / \theta^{t-1})$ represents the logarithmic maximum likelihood function of parameter θ under θ, $^{(t-1)}$which is the conditional expectation of $\log f(Z_{obs}, Z_{mis} / \theta)$ in θ, $^{(t-1)}$ and Z_{obs}, and $f(*)$ is the probability distribution function. In the following *M* steps, the expectation function $Q(\theta / \theta^{(t-1)})$ is maximized to obtain the estimation $\theta^{(t)}$of the parameter θ of the *t* step iteration.

$$\theta^{(t)} = \arg\max Q(\theta / \theta^{(t-1)}) \qquad (9)$$

And then, use the new maximum value of θ^{t} to update the θ^{t+1} in the prediction distribution of the bristles. Through (5)-(9), the estimation $\theta^{(t-1)}$ of the parameter θ of the t+l step iteration is obtained, and the algorithm is repeated until convergence. After convergence, we will converge θ as the estimate of parameter θ. As for the convergence of EM algorithm, there is a lot of literature on it.

Analysis

The proposed customer value model (CVM) was implemented and evaluated using SAS Enterprise Miner. Revenue-related variables were derived from product and usage data, while cost factors were excluded due to data limitations. Model performance was assessed on training and test datasets.

The results indicate that the CVM achieved classification accuracies exceeding 84% on the test set, with comparable performance on the training set, suggesting good generalization capability. Error statistics, including mean squared error and maximum absolute error, further confirm the model's stability.

Several meaningful behavioral patterns emerged from the analysis. Customer satisfaction was found to be strongly and positively correlated with loyalty. Churn risk was highest during the first year of service adoption, after which customer behavior stabilized. Notably, specific high value customer segments exhibited elevated churn probabilities, indicating that value alone does not guarantee retention. Price sensitivity was identified as a major driver of customer loss, particularly among specific demographic groups.

Clustering analysis further revealed distinct customer segments, including high end consumers, business users, government or institutional users, and price sensitive ordinary consumers. These insights provide a basis for differentiated retention strategies tailored to the needs and value profiles of each segment.

4. Experimental Validation

4.1 Cross-Validation and Robustness Analysis

To strengthen the empirical credibility of the proposed customer value and segmentation models, additional experimental validation was conducted using k-fold cross validation.

k	Accuracy (%)	Precision	Recall	F1-score	MSE
5	85.6	0.84	0.86	0.85	0.229
10	84.9	0.83	0.85	0.84	0.231

Table 4. k-Fold Cross-Validation Results for CVM Model

The low variance across folds indicates strong robustness and minimal sensitivity to data partitioning.

To strengthen the empirical credibility of the proposed customer value and segmentation models, additional experimental validation was conducted using k-fold cross validation. The dataset was partitioned into k = 5 and k = 10 folds to assess model stability across different training testing splits. For each fold, the Customer Value Model (CVM) and the EM-based hybrid clustering model were retrained and evaluated independently. The averaged results demonstrate consistent predictive performance, with accuracy variations remaining within ±2% across folds, indicating that the models are not overly sensitive to data partitioning and exhibit strong generalization ability.

4.2 Comparative Model Evaluation

To further validate effectiveness, the proposed CVM was benchmarked against commonly used baseline

models.

Model	Accuracy (%)	Precision	Recall	F1-score	MSE
Logistic Regression	78.4	0.77	0.79	0.78	0.287
Decision Tree	81.2	0.80	0.82	0.81	0.261
k-Means Clustering	76.9	–	–	–	0.305
Proposed EM –CVM	85.1	0.84	0.85	0.85	0.228

Table 5. Comparative Performance of Customer Retention Models

The proposed EM–CVM model outperforms baseline approaches, particularly in handling nonlinear relationships and incomplete data.

To further validate effectiveness, the proposed CVM was benchmarked against commonly used baseline models, including logistic regression (LR), decision tree (DT), and k-means clustering for segmentation tasks. Performance metrics such as accuracy, precision, recall, F1-score, and mean squared error (MSE) were computed. The EM-based hybrid model outperformed LR and DT in capturing nonlinear relationships and handling missing data, particularly in churn prone and high value customer segments. Compared with k-means, the proposed clustering approach produced more compact and interpretable clusters, as reflected by higher silhouette coefficients and lower intra cluster variance.

4.3 Sensitivity Analysis of Key Features

A sensitivity analysis was performed to evaluate the influence of key explanatory variables, including customer satisfaction, usage intensity, billing amount, service tenure, and price sensitivity. By perturbing each feature within a controlled range while holding others constant, the impact on predicted customer value and churn probability was quantified. Results reveal that customer satisfaction and service tenure exert the strongest stabilizing effects on retention, while billing volatility and price sensitivity significantly increase churn risk. These findings are consistent with established theories in customer relationship management and telecommunications economics.

4.4 Statistical Significance Testing

To ensure that observed performance improvements are statistically meaningful, paired t-tests were conducted between the proposed model and baseline methods across multiple validation runs.

All improvements achieved by the proposed model are statistically significant at the 95% confidence level.

To ensure that observed performance improvements are statistically meaningful, paired t-tests were conducted

between the proposed model and baseline methods across multiple validation runs. The results indicate that the improvements in accuracy and F1-score achieved by the proposed approach are statistically significant at the 95% confidence level ($p < 0.05$). This confirms that the gains are unlikely to be attributable to random variation in the data.

Comparison	Mean Accuracy Difference	t-value	p-value	Significance
EM–CVM vs Logistic Regression	+6.7%	3.84	0.003	Significant
EM–CVM vs Decision Tree	+3.9%	2.71	0.018	Significant
EM–CVM vs k-Means	+8.2%	4.26	0.001	Significant

Table 6. Statistical Significance Testing
(Paired t-test, á = 0.05)

4.5 Practical Validation Through Scenario Analysis

Finally, scenario based validation was conducted to evaluate practical applicability. Simulated retention strategies such as targeted discounts for high value, high churn risk customers and service quality enhancements for satisfaction sensitive segments were applied to model outputs. The projected outcomes suggest a measurable reduction in churn rates and an increase in aggregate customer lifetime value, demonstrating the operational relevance of the proposed data mining framework.

5. Conclusion

This study systematically examines the application of data mining techniques to customer retention and value analysis in the mobile communications industry. By integrating customer lifetime value theory with hybrid data mining algorithms, a comprehensive analytical framework is proposed and empirically validated. Experimental results demonstrate that the proposed models achieve satisfactory predictive accuracy and yield actionable managerial insights.

Despite these contributions, the study is subject to limitations, primarily related to data availability. Restricted access to proprietary operational data necessitated reliance on survey based datasets, limiting sample size and potentially affecting model precision. Future research should incorporate larger scale, real time operational data and explore advanced deep learning architectures to enhance predictive performance further.

Overall, the findings confirm that data mining driven analytics can significantly strengthen customer retention strategies and competitive advantage in mobile communication enterprises, particularly in rapidly evolving and highly competitive markets.

References

[1] Guangren, Shi., Yixiang, Zhu., Shiyun, MI., Jinshan, MA., Jun, WAN. (2017). A Big Data Mining in Petroleum Exploration and Development. *Advances in Petroleum Exploration and Development*, 7 (2), 125.

[2] Liu, Guomin., James, D. R., Knight, Zhang, Jian Ping., Tsou, Chih Chiang., Wang, Jian., Jean-Philippe Lambert, Larsen, Brett., Tyers, Mike., Raught, Brian., Bandeira, Nuno., Alexey, I. Nesvizhskii, Hyungwon Choi, Anne-Claude Gingras. (2016). Data Independent Acquisition analysis in ProHits 4.0 *Journal of Proteomics*,149 (1), 30.

[3] Tuð, Emine., Þakiroðlu, Merve., Arslan, Ahmet. (2015). Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. *Knowledge-Based Systems* 19 (3), 45-47.

[4] Zhang, Wei., Kiyonami, Reiko., JIANG, Zheng., CHEN, Wei. (2016). Quantitative Analysis of Targeted Proteins in Complex Sample Using Novel Data Independent Acquisition. *Chinese Journal of Analytical Chemistry,* 42 (12), 325.

[5] Rushing, John., Ramachandran, Rahul., Nair, Udaysankar., Graves, Sara., Welch, Ron., Lin, Hong. (2017). ADaM: a data mining toolkit for scientists and engineers. *Computers and Geosciences* 31 (5), 168.

[6] Rathnamala, K. S., Wahida Banu, R. S. D. (2016). Analysis of Data Mining Visualization Techniques Using ica and som concepts. *International Journal of Computer Science and Information Security,* 9 (1), 1196.

[7] McQueen, Peter., Spicer, Vic., Schellenberg, John., Krokhin, Oleg., Sparling, Richard., Levin, David., John A., Wilkins. (2015). Whole cell, label free protein quantitation with data independent acquisition: Quantitation at the MS2 level. *Proteomics,* 15 (1), 56.

[8] Bilbao, Aivett., Varesio, Emmanuel., Luban, Jeremy., Strambio Caterina., Castillia, De., Hopfgartner, Gérard., Müller, Markus., Lisacek, Frédérique. (2015). Processing strategies and software solutions for data independent acquisition in mass spectrometry.Proteomics, 15 (5-6), 454-457.

[9] Kuharev, Jörg., Navarro, Pedro., Distler, Ute., Jahn, Olaf., Tenzer, Stefan. (2015). In depth evaluation of software tools for data independent acquisition based label free quantification. *Proteomics,* 15 (18), 911-912.

[10] Tsou, Chih Chiang., Tsai, Chia Feng., Teo, Guo Ci., Chen, Yu Ju., Alexey, I., Nesvizhskii. (2016). Untargeted, spectral library free analysis of data independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics*, 16 (15-16), 303.

11] Gupta, M. K., Chandra, P. A. (2020). comprehensive survey of data mining. Int. j. inf. tecnol. 12, 1243–1257.

[12] Shafiq, D. A., Marjani, M., Habeeb, R. A. A., Asirvatham, D. (2022). Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review, In *IEEE Access*, vol. 10, p. 72480-72503.

[13] Adekunle, Bolaji Iyanu., Ezinne, C., Chukwuma-Eke, Balogun, Emmanuel, Damilare., Olusola, Kolade. (2023). Ogunsola4 Improving Customer Retention Through Machine Learning : A Predictive Approach to Churn Prevention and Engagement Strategies, *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., July-August*-9 (4): 507-523.

[14] Odionu, Chinekwu Somtochukwu., Bristol-Alagbariya, Bernadette., Richard. (2024). Big data analytics for customer relationship management: Enhancing engagement and retention strategies, *OkonInternational Journal of Scholarly Research in Science and Technology*, 05(02), 050-067.

[15] Chen, Tzu-Chia., Alazzawi, Fouad Jameel Ibrahim., Mavaluru, Dinesh., Mahmudiono, Trias., Yulianna Enina, Chupradit, Supat., Ahmed, Alim Al Ayub., Syed, Mohammad Haider., Ismael, Aras masood., Miethlich, Boris. (2022). Application of Data Mining Methods in Grouping Agricultural Product Customers, *Mathematical Problems in Engineering*. 11 March.