Classification of customer churn prediction model for telecommunication industry using analysis of variance

Ronke Babatunde[1], Sulaiman Olaniyi Abdulsalam[2], Olanshile Abdulkabir Abdulsalam[3], Micheal Olaolu Arowolo[4,5]

[1]Department of Computer Science, Kwara State University, Malete, Nigeria
[2]Department of Computer Science, Landmark University, Omu-Aran, Nigeria
[3]Landmark University SDG 9 (Industry, Innovation and Infrastructure Research Group), Omu-Aran, Nigeria
[4]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, United States

Article Info

ABSTRACT

Customer predictive analytics has shown great potential for effective churn models. Thriving in today's telecommunications industry, discerning between consumers who are likely to migrate to a competitor is enormous. Having reliable predictive client behavior in the future is required. Machine learning algorithms are essential to predict customer turnovers, and researchers have proposed various techniques. Churn prediction is a problem due to the unequal dispersal of classes. Most traditional machine learning algorithms are ineffective in classifying data. Client cluster with a higher risk has been discovered. A support vector machine (SVM) is employed as the foundational learner, and a churn prediction model is constructed based on each analysis of variance (ANOVA). The separation of churn data revealed by experimental assessment is recommended for churn prediction analysis. Customer attrition is high, but an instantaneous support can ensure that customer needs are addressed and assess an employee's capacity to achieve customer satisfaction. This study uses an ANOVA with a SVM, classification in analyzing risks in telecom systems It may be determined that SVM provides the most accurate forecast of customer turnover (95%). The projected outcomes will allow other organizations to assess possible client turnover and collect customer feedback.

Corresponding Author:

Micheal Olaolu Arowolo
Department of Computer Science, Landmark University
Omu-Aran, Nigeria
Email: Arowolo.olaolu@gmail.com

1. INTRODUCTION

In the telecommunications industry, data mining techniques have been used to classify and predict client attrition. In the realm of e-commerce, customer churn is a significant issue. Due to cost constraints, conventional churn prediction models are simple and basic, based on available data from an e-commerce platform [1], [2].

Customers who haven't logged in or made any purchases are deemed lost. Customers who are at risk of being lost can be precisely identified with data mining and analysis, also, they can be restored in a timely manner using efficient marketing methods. Although customer acquisition is unpredictable, a customer churn predictions model can help e-commerce enterprises figure out what created a lack and how to eliminate it in the future. The most extensively used customer turnover prediction model is the regularity, repetition, and financial value approach [3], [4].

The model has divided customers into groups groupings characterized by three index values: relevance (once customers can directly make most latest acquisition), recurrence (how often the consumer makes transactions), and financial value (how much the purchaser gets to spend on each purchase), and then gives e-commerce businesses the operational measures they need to keep potential customers and enhance profitability. Way of categorizing, but on the other side, has a significant wait than some other types of client information, especially data on browsing habits [5], [6]. Furthermore, various clients' repurchase cycles are different. As a result, the recency, frequency, monetary (RFM) model's analysis is too straightforward and straightforward.

The term "churn" was coined by combining the English terms "change" and "turn" to represent the phenomenon of a consumer leaving. It refers to the transfer of subscribers from one provider to another in the mobile telecommunications business (also known as customer attrition or subscriber churning). The rate of churn is used to measure it, and it is an essential metric for businesses. Churn usually occurs as a result of better rates or services offered by a competitor, or as a result of various benefits offered by a competitor when signing up [7], [8].

To thrive in today's telecommunications industry, you must be able to discern between consumers who are likely to migrate to a competitor. Customer churn prediction is a method of dealing with the problem, and it has become a significant concern in the telecommunications business. In such a competitive industry, and accurate method of predicting clients' future behavior would be considered invaluable. Telecom's technologies are typically based on statistical pattern-recognition algorithms. These sophisticated solutions dispute the common misconception that client churning is only a data-mining exercise [9], [10]. The machine learning approach, on the other hand, is a rapidly expanding field of study. As a result, this study predicts that the telecommunications industry will catch up with the integration of nonparametric approaches and machine learning models in the next years.

Furthermore, a number of experts suggested existing machine learning methodologies for predicting consumer attrition in the telecom industry. A classification procedure is used in a huge percentage of these techniques [11], [12]. Customer churn, on the other side, is a challenging task due to the inequitable distribution of its categories, with churner clients often being significantly fewer than non-churner individuals. Because of this issue, most individual machine learning approaches are ineffective for recognizing patterns [13], [14]. As a result, several researchers identified artificial intelligence techniques for determining consumer turnover that combine two or more methods, one of which is utilized for information pre-processing prior to conducting the multiclass classification [15], [16]. Some article suggests grouping the information into similar groups and then deleting some of them as a way to filter out misrepresentative information [17], [18]. In this study, the support vector machine (SVM) is presented as a classifier to predict customer churn in a telecommunication sector, and the analysis of variance (ANOVA) feature selection method is developed to extract useful information from the telecommunications data.

2. METHOD

This study was conducted in stages, including data pre-processing, selection of features, and classification. It uses an open-source telecom customer churn dataset from Duke University saved in the Kaggle repository; it consists of 51,047 instances and 53 attributes "Customer ID, Churn, Monthly Revenue, Monthly Minutes, charges, calls, roaming, among others". Data processing (to remove distortion and inconsistency), data transformation (to merge several sources of data), data selections (to extract data relevant to the analytical objective from the data), and data processing were all performed on the dataset (by executing summarize or aggregation processes, data is converted and compacted into forms suitable for mining). In this study, ANOVA is used to select relevant information from the given dataset, SVM is used as a classifier to evaluate the results obtained and evaluated in terms of evaluation measures, Figure 1 shows the proposed model.
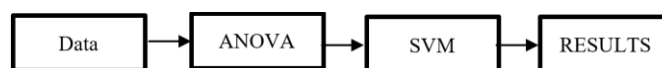

Figure 1. Proposed model

2.1. Equations selection of features using ANOVA

For feature selection, ANOVA was used to gather relevant data that could aid in the classification stage. The ANOVA technique was employed in this study to extract meaningful data from the churning

telecom dataset. A total of 10,149 characteristics were chosen. ANOVA increases its performance and scalability by deleting portions of features at a moment having little impact on the validity of the limited set of features as achievable, using approaches from artificial reinforcement of the procedure [19], [20].

ANOVA improves feature selection by being "greedy," since the features selected of the high-dimensional telecommunications information are picked by identifying the best group of attributes. That is, once a feature has been decided to be removed, it cannot be restored [21]–[23]. The goal was to use a one-way ANOVA F-test to see if all of the various classes of Y have the same means as X [24].

2.2. Classification

Cortes and Vapnik (1995) invented SVM, a prominent widely utilized machine learning approach in a set of real research areas. The creditworthiness ground has been extensively utilized due to the advanced classifier performance and roughly comparable expandability, especially in comparison to its relatively close predecessor artificial neural network (ANN) as well as other classification methods. The purpose of the SVM, dependent on the empirical risk methodology, is to lower the arbitrary limit of misclassifications. It is essential to use training examples to approximate a purpose for assessment in the SVM [25].

Its key premise is to visualize the data input into a high dimensional subspace, then generate a higher dimensional space aided by the support vectors to find the smallest possible margin between the two classes. The support vector's feature can be used to predict the new input sample labels. SVM uses a variety of functions (called kernels) to map input data into high dimension feature space, such as linear, radial basis function (RBF), polynomial, and sigmoid [26].

The discovery of intriguing patterns and information from massive amounts of data is part of the classification stage. The ANOVA features were passed on to SVM, which was used to further analyze the data and identify patterns in the data set, which was divided into churners and non-churners. Classification accuracy, F-measure, sensitivity, specificity, recall, and precision were all used to assess the technique's performance.

3.   RESULTS AND DISCUSSION

Microsoft Excel was used to compile the churn data. The model was created in MATLAB 2015a, a fourth-generation programming language with object-oriented procedures. The dataset was tested, and it included 53 telecom experiments from Duke University, totalling 51,047 instances of expression levels. An ANOVA feature selection algorithm was used in this study to retrieve specific features from the telecom set of data, resulting in a total of 10,149 features. If the null hypothesis is true, the 0.5 p-value employed in the ANOVA feature selection technique is the likelihood of seeing a result (F-critical) as large as the one produced in the experiment (F0). Low p-values suggest that the null hypothesis is likely untrue. From a dataset supplied from Duke University on the Kaggle site, 10149 characteristics were picked using ANOVA. There was a statistically significant difference in group means as a consequence of the analysis. The average time it took each class to complete the spreadsheet activity is 0.05, which is the significant value. The selected features are processed for categorization.

The result of the categorization using the ANOVA-SVM is depicted in Figure 2's Scatter Plot. In addition, Figure 3, and Figure 4 shows the findings as a confusion matrix and the Receiver operating curve respectively. The confusion matrix depicts the true positive (TP), true negative (TN) false positive (FP), and false negative (FN) classes utilized by an SVM classifier to create performance metrics such as accuracy, sensitivity, specificity, precision, recall, and F-Measure. Figure 3 illustrates a confusion matrix graphic with the predicted class (Output class) in the rows and the true class in the columns (Target class). The diagonal cells correspond to observations that have been properly categorised. The observations that were incorrectly classified are represented by the off-diagonal cells. The number of observations as well as the percentage of the total number of observations are displayed in each cell.

In the far-right column of the plot, the percentages of all the occurrences anticipated to belong to each class that are correctly and incorrectly classified are shown. Two often used measures are accuracy (or positive predictive value) and false discovery rate. In the row at the bottom of the picture, the percentages of all the examples belonging to each class that are correctly and incorrectly classified are shown. Two often used measurements are recall (or true positive rate) and false negative rate. In the cell at the bottom right of the plot, the total accuracy is displayed.

The following values were acquired for the categorization process. TP=39, TN=18, FP=3, and FN=0. This was used to calculate the following metrics. Figure 4 shows the reciever operating characteristics curve for analysis of varian0ce-SVM. The SVM-RBF was developed using training data that has been reduced to the desired features from each algorithm. In terms of prediction accuracy rate, ANOVA performed well, with 95% accuracy on the test data.
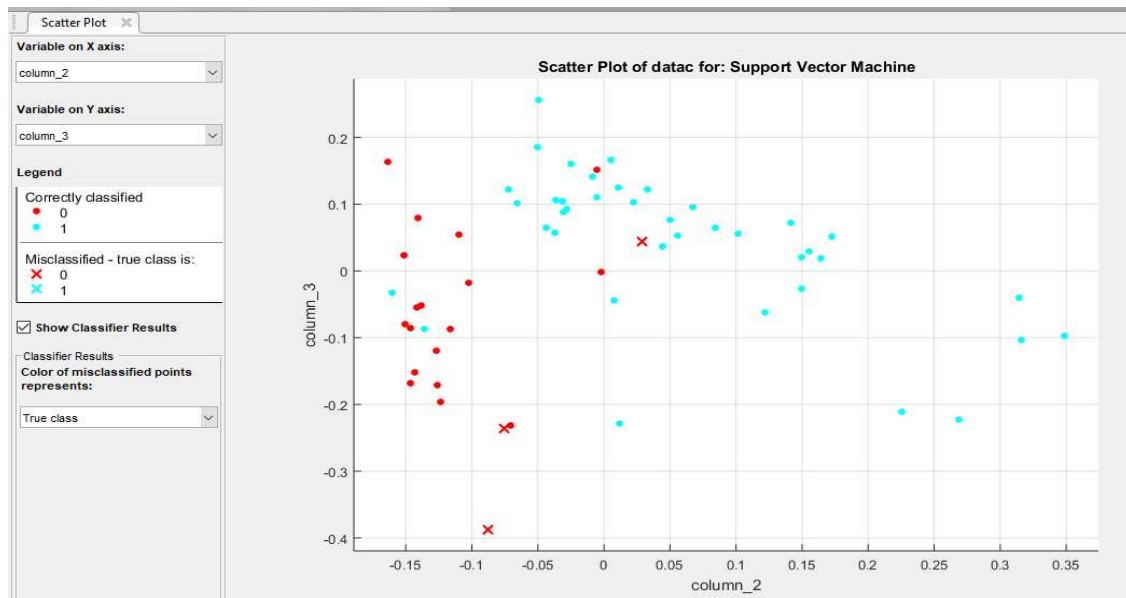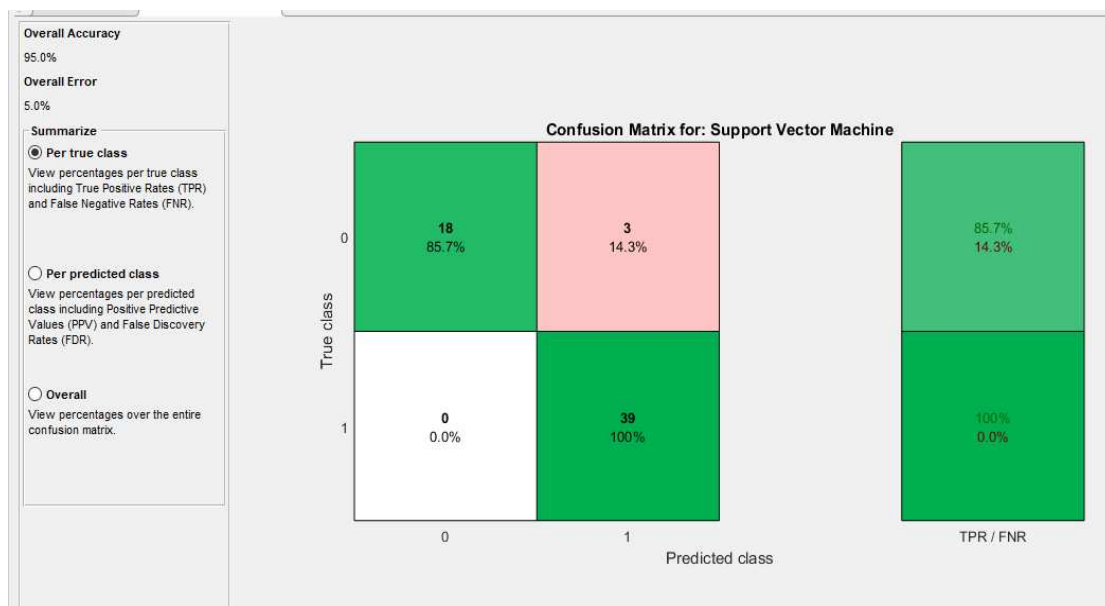
Figure 2. Scattered plot for the ANOVA-SVM

Figure 3. Confusion matrix for ANOVA-SVM

It was discovered that using ANOVA for feature selection increases the efficiency of the classifier algorithm significantly without lowering classification accuracy. In domains with a huge number of features, such as telecom data, the feature selection procedure is therefore made considerably more practical. As the number of samples accessible grows, this enhancement becomes even more critical. Table 1 summarizes the results of the ANOVA-SVM-RBF classification algorithm used in the telecomm unications industry to improve the performance of churned customer data.

The performance assessment of classifier using SVM-RBF on the Telecom dataset uncovers that the ANOVA feature selection approach achieved the considered necessary greater value in the sets of data on performance parameters such as accuracy, timing, sensitivity, specificity, and prediction, according to the findings of this study. The feature selection algorithm using ANOVA is very useful when the dataset has a large number of dimensions. That matter improves the performance of feature selection methods as well as the classifier algorithm "SVM" in terms of accuracy, sensitivity, specificity, and precision.
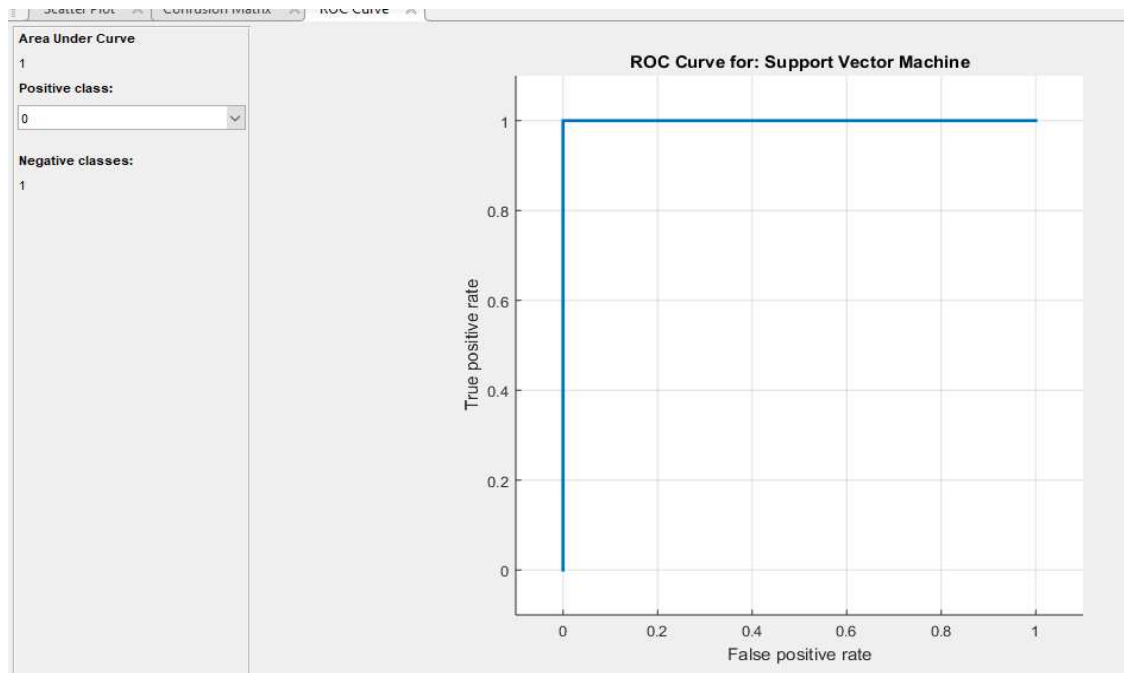
Figure 4. ROC curve for ANOVA -SVM-RBF classifier

Table 1. Analysis of the classification of ANOVA-SVM-RBF for telecommunication industry

Performance Metrics (%)	ANOVA-SVM-RBF
Accuracy	95.0
Sensitivity	100
Specificity	92.86
Precision	85.71
F1 Score	92.31
Matthews Correlation Coefficient	89.21
Misclassification	5
Time (Sec)	91.3532

4.   CONCLUSION

The greatest threat to the planet is complexity in the telecommunications realm. Churners and non-churners alike now have new options thanks to the expansion of telecom data and the development of statistical methods. The basic technologies of client retention are feature selection and classification. They're both important for recognition and prediction. Limited to telecom data characteristics, many common solutions in this field still require further attention to overcome their drawbacks. The key characteristics of telecom data, as well as the main problems for researchers conducting telecom data analysis, are small sample size, high dimensionality, and class imbalance. Researchers in this sector rarely look on class imbalance when pre-processing datasets. This problem is solved in this study by using the ANOVA resampling method. For classification, SVM-RBF was used, which worked effectively by decreasing unnecessary processing costs for large-scale linear separable data like telecom data. This research can be expanded in the future to include more feature extraction algorithms and classifiers.

REFERENCES

[1]   A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0191-6.
[2]   W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012, doi: 10.1016/j.ejor.2011.09.031.

[3]   A. Rehman and A. Raza Ali, "Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry," *ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference*, no. July, pp. 1–9, 2014, [Online]. Available: https://www.researchgate.net/publication/304822719.

[4]   X. Li and Z. Li, "A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm," *Ingenierie des Systemes d'Information*, vol. 24, no. 5, pp. 525–530, 2019, doi: 10.18280/isi.240510.

[5]   R. H. Lin, W. W. Chuang, C. L. Chuang, and W. S. Chang, "Applied big data analysis to build customer product recommendation model," *Sustainability (Switzerland)*, vol. 13, no. 9, 2021, doi: 10.3390/su13094985.

[6]   R. de Villiers, P. Tipgomut, and D. Franklin, "International Market Segmentation across Consumption and Communication Categories: Identity, Demographics, and Consumer Decisions and Online Habits," *Promotion and Marketing Communications*, 2020, doi: 10.5772/intechopen.89988.

[7]   J. Lappeman, M. Franco, V. Warner, and L. Sierra-Rubia, "What social media sentiment tells us about why customers churn," *Journal of Consumer Marketing*, vol. 39, no. 5, pp. 385–403, 2022, doi: 10.1108/JCM-12-2019-3540.

[8]   I. Brandusoiu and G. Toderean, "Churn Prediction Modeling in Mobile Telecommunications Industry Using Decision Trees," *Journal of Computer Science and Control Systems*, vol. 6, no. 1, pp. 14–19, 2013.

[9]   L. Sook Ling, N. Mustafa, and S. F. Abdul Razak, "Customer churn prediction for telecommunication industry: A Malaysian Case Study," *F1000Research*, vol. 10, 2021, doi: 10.12688/f1000research.73597.1.

[10]   R. Sudharsan and E. N. Ganesh, "A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy," *Connection Science*, vol. 34, no. 1, pp. 1855–1876, 2022, doi: 10.1080/09540091.2022.2083584.

[11]   I. Antonopoulos *et al.*, "Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review," *Renewable and Sustainable Energy Reviews*, vol. 130, 2020, doi: 10.1016/j.rser.2020.109899.

[12]   A. Amin, S. Shehzad, C. Khan, I. Ali, and S. Anwar, "Churn prediction in telecommunication industry using rough set approach," *Studies in Computational Intelligence*, vol. 572, pp. 83–95, 2015, doi: 10.1007/978-3-319-10774-5_8.

[13]   F. E. Usman-Hamza *et al.*, "Intelligent Decision Forest Models for Customer Churn Prediction," *Applied Sciences (Switzerland)*, vol. 12, no. 16, 2022, doi: 10.3390/app12168270.

[14]   F. Devriendt, J. Berrevoets, and W. Verbeke, "Why you should stop predicting customer churn and start using uplift models," *Information Sciences*, vol. 548, pp. 497–515, 2021, doi: 10.1016/j.ins.2019.12.075.

[15]   M. E. Sánchez-Gutiérrez and P. P. González-Pérez, "Multi-Class Classification of Medical Data Based on Neural Network Pruning and Information-Entropy Measures," *Entropy*, vol. 24, no. 2, 2022, doi: 10.3390/e24020196.

[16]   I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Computer Science*, vol. 2, no. 5, 2021, doi: 10.1007/s42979-021-00765-8.

[17]   M. K. Manoj, S. Atalla, N. Almuraqab, and I. A. Moonesar, "Detection of COVID-19 Using Deep Learning Techniques and Cost Effectiveness Evaluation: A Survey," *Frontiers in Artificial Intelligence*, vol. 5, 2022, doi: 10.3389/frai.2022.912022.

[18]   A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data*, vol. 2, 2019, doi: 10.3389/fdata.2019.00013.

[19]   R. Boutaba *et al.*, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," *Journal of Internet Services and Applications*, vol. 9, no. 1, 2018, doi: 10.1186/s13174-018-0087-2.

[20]   G. V. Miltiadis D. Lytras, Ernesto Damiani, John M. Carroll, Robert D. Tennyson, David Avison, Ambjörn Naeve, Adrian Dale, Paul Lefrere, Felix Tan, Janice Sipior, "Visioning and Engineering the Knowledge Society. A Web Science Perspective," *Springer Berlin Heidelberg*, vol. 5736, 2009, doi: 10.1007/978-3-642-04754-1.

[21]   M. Ahsan, R. Gomes, M. M. Chowdhury, and K. E. Nygard, "Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector," *Journal of Cybersecurity and Privacy*, vol. 1, no. 1, pp. 199–218, 2021, doi: 10.3390/jcp1010011.

[22]   M. Kuhn and K. Johnson, "An Introduction to Feature Selection," *Applied Predictive Modeling*, pp. 487–519, 2013, doi: 10.1007/978-1-4614-6849-3_19.

[23]   K. O. Alabi, S. O. Abdulsalam, R. O. Ogundokun, and M. O. Arowolo, "Credit Risk Prediction in Commercial Bank Using Chi-Square with SVM-RBF," *Communications in Computer and Information Science*, vol. 1350, pp. 158–169, 2021, doi: 10.1007/978-3-030-69143-1_13.

[24]   A. A. Jamjoom, "The use of knowledge extraction in predicting customer churn in B2B," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00500-3.

[25]   Y. J. Son, H. G. Kim, E. H. Kim, S. Choi, and S. K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients," *Healthcare Informatics Research*, vol. 16, no. 4, pp. 253–259, 2010, doi: 10.4258/hir.2010.16.4.253.

[26]   J. F. P. da Costa and M. Cabral, "Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works," *Mathematics*, vol. 10, no. 6, 2022, doi: 10.3390/math10060993.

BIOGRAPHIES OF AUTHORS

Ronke Babatunde (Ph.D.) 🆔 🔍 sc ▶ is a faculty at the Department of Computer Science, Kwara State University, Malete. She received her Ph.D. in Computer Science, with her area of research including Data Science, Computer Vision, Pattern Recognition, Computational Intelligence, and Medical Image Processing. She has authored and presented several works in prestigious journals and conferences international and locally. She can be contacted at email: ronke.babatunde@kwasu.edu.ng.

Sulaiman Olaniyi Abdulsalam (Ph.D.) is a Lecturer at the Department of Computer Science, Kwara State University, Malete, with his research interests in Artificial Intelligence and Datamining. He has authored in respected journals and conferences globally. He can be contacted at email: sulaiman.abdulsalam@kwasu.edu.ng.

Olanshile Abdulkabir Abdulsalam recently graduated from his B.Sc. from the Department of Computer Science, Kwara State University, he works in the research area of Machine Learning. He can be contacted at email: abdulsalamkabeer@gmail.com.

Micheal Olaolu Arowolo (Ph.D.) is presently a Research Scholar at the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia. Prior to joining Missouri, he was a member of Staff at the Department of Computer Science, Landmark University, Omu-Aran, where he worked and also earned his Ph.D., in Computer Science, 2021. His research focuses on; Bioinformatics, Machine Learning, Big Data Analytics, Artificial Intelligence and Data Science. He has published in impressive journals and member of several computing bodies. He can be contacted at email: arowolo.olaolu@gmail.com.