VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2023, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 11, Number 3, July-September 2023 *pp:73-80*

A Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry

Gul Zaman Khan [1*], **Ikram Ulhaq** [2], **Ihsan Adil** [3], **Sajad Ulhaq** [4], **Inam Ullah** [5]

[1]Department of Computer Software Engineering UET Mardan,Pakistan; [2,3,4]Department of Computer Science Ghazi Umara Khan Degree Collage Samar Bagh, Pakistan.; [5]Department of Computer Science Abdul Wali Khan University Mardan,Pakistan

Keywords: *BAT-ANN, Churn prediction, Telecom industry.*

Journal Info:
Submitted:
September 18, 20223
Accepted:
October 20, 2023
Published:
November 03, 2023

Abstract In recent years, customer churn has been one of the most prominent topics, especially in the telecom industry. The telecommunications industry is producing massive amounts of data every minute. Thus, the telecom industry is seeking more ways to analyze and predict their potential and churn customers. According to telecom analysis, acquiring a new customer is costlier than keeping a current one. To lessen customer churn, it is compulsory for industries to detect an increase in customer churn factors. The number of service suppliers is increasing daily, especially in the telecom industry. Phishing attacks and fraud are crucial points in customer churn. The aim of this study is to predict customer churn with predictive churn models for retention campaigns to satisfy the business requirement of profit maximization. The proposed research used the BAT-ANN classification model with the BigML dataset to predict customer churn in the telecom industry. The proposed model achieved 89.2% accuracy.

***For correspondence:**
gulzamankhan726@gmail.com
DOI: 10.21015/vtse.v11i3.1642

1 Introduction

Customer churn is a major problem in the telecom industry, leading to lost revenue, increased costs, and a decline in customer satisfaction. Churn is the occurrence where customers stop utilizing a certain service completely, or switch from one service provider to another. There are a number of factors that can contribute to customer churn, including price, service quality, customer support, competition, and changes in customer needs. In order to prevent churn, telecom Companies need to be able to predict which customers are most likely to leave [1]. Customer churn prediction can be done using a variety of methods, including statistical analysis, machine learning, and data mining [2]. The most effective method for customer churn prediction will vary depending on the specific data and the goals of the telecom company. However, there are a number of general principles that can be applied to customer churn prediction in the telecom industry

This work is licensed under a Creative Commons Attribution 3.0 License.

[3]. One important principle is to use a variety of data sources. This includes customer demographic data, call detail records, and customer feedback. By using a variety of data sources, telecom companies can get a more complete picture of their customers and their potential for churn [4]. Another important principle is to use a variety of machine learning algorithms. This allows telecom companies to test different algorithms and find the one that best predicts customer churn [5]. In the telecom sector, churn prediction is essential since it aids service providers in understanding and anticipating consumer behavior. Telecom firms must be able to predict churn since it enables them to take preventative action to keep customers and lower attrition rates [6]. Customer retention is critical for sustaining profitability and growth in the highly competitive telecom business. Attrition prediction models use historical data and machine learning approaches to uncover patterns and factors that lead to customer attrition. These algorithms may predict customer turnover by analyzing numerous consumer variables and behaviors [7]. Telecom businesses may use customer demographics, phone records, billing information, service use, complaints, and customer interaction data to develop strong churn prediction algorithms. Telecom businesses can deploy targeted retention efforts once prospective customers are identified. To address consumer problems and increase happiness, these techniques may include personalized offers, enhanced customer service, loyalty programs, or proactive outreach initiatives. Understanding and anticipating churn allows telecom firms to concentrate their efforts and resources on customer retention, enhancing customer loyalty, lowering client acquisition costs, and increasing revenue [8].

In the telecom industry, artificial intelligence (AI) is becoming increasingly important in forecasting churn. AI can be used to evaluate large amounts of customer data to find patterns and trends that may indicate a client is about to churn. This data can then be used to target clients with activities designed to keep them from leaving [9]. Machine learning techniques can be used to analyze historical data to identify elements linked to churn. Based on these criteria, prediction models for locating clients who are likely to leave can be developed [10]. Deep learning is becoming increasingly important in predicting telecom churn due to its capacity to examine massive quantities of customer data to discover patterns and trends that may indicate churn [11].

This study addressed the key challenges to accurately predicting highly valuable customer turnover based on available research and client attribute features in the telecoms industry. The proposed research work used historical customer data collected in the telecom industry with a BAT artificial neural network to accurately predict customer churn in the telecom sector. The remaining parts of the article are organized as follows: Section 2 discusses the approaches used in existing research work. Section 3 discusses the material and methods used to carry out this experiment, while Section 4 discusses the experimental outcome, Section 5 discusses the comparative analysis, and Section 6 discusses the future work of the proposed experiment.

2 Literature Review

To overcome these challenges, researchers and experts have used various machine learning and deep learning models to predict customer churn in the telecom industry. Previous work done in customer churn prediction in the telecom sector is summarized as follows: Wee How Khoh et al. [1] proposed an effective method for telecom operators to detect prospective churn clients. The system gives accurate predictions and assists organizations in adopting focused marketing strategies to retain clients and achieve significant cost savings by employing an optimal ensemble learning model like KNN, CatBoost, and RF and achieving 84% accuracy. Similarly, Restu Herdian et al. [2] used hybrid models for customer churn prediction and then compared the obtained results with previous approaches such as DT and artificial neural networks. Their proposed hybrid model outperformed the individual decision tree and ANN models with 81% accuracy.

Sudharsan et al., [3] proposed the S-RNN model

to forecast client churn. The system was trained on historical customer data, learning from earlier churn incidents to identify patterns predictive of churn behavior. The CP dataset was used and obtained an accuracy of 95.99%. Likewise, Xiancheng Xiahou et al., [4] combined the k-means clustering technique with the AdaBoost classifier algorithm, allowing customers to be classified into three groups using the Alibaba Cloud Tianchi dataset and achieving 96% prediction accuracy. V. Kavitha et al., [5] conducted an experiment using various machine learning models, including DT, RF, and XGB, to offer telecom firms a useful tool for predicting and reducing client attrition. These algorithms use previous customer data to identify trends and forecast future churn behavior. The proposed work used the Kaggle dataset and achieved 80% accuracy. Similarly, Samah Wael Fujo et al., [12] examined ANN with deep-BP models to address the issue of customer churn and achieved 88.12% accuracy. Pragya Manghnani et al., [6] utilized text analysis and machine learning classes such as LR, DT, RF, and XGBoost to change customer behavior patterns and predict customer churn. The proposed models made use of a simulated dataset from a telecom company, which includes some unbalanced metadata. Their proposed method obtained a 95% accurate result.

Pei Chen et al. [7] used SVC, GBDT, RF, and AdaBoost models to predict European bank customer churn, which may have an impact on retention and service efforts. Their proposed approach used a Kaggle dataset with 10,000 customer records and achieved an accuracy of 83%. Sulaimon Olani Abdulsalam et al., [13] implemented machine learning methods to build a churn prediction model in the telecom industry. The proposed work used a Relief-F function selection algorithm with Random Forest and CNN classifiers with a BIG-ML dataset and achieved 94% accuracy. Ilys Abdullaev et al., [14] utilized AIJOA-CPDE to predict HCI turnover in order to discriminate between churned and non-churned clients. AIJOA-CPDE selects feature subsets using JOA-FS. AIJOA-CPDE used BDLSTM to identify clients who were about to leave.

Further, the BDLSTM hyperparameters are optimized using the CSO approach. AIJOA-CPDE outperformed other techniques in rigorous testing applied to both the TR and TS datasets and achieved 91% accuracy. Chen Zhou [8] proposed the MIPCA-XGBoost approach to predict customer turnover in the telecom business. XGBoost was used to train the dataset and split the sample data into leaf nodes based on classification features. The MIPCA-XGBoost technique achieved a high prediction accuracy of 90.56% utilizing the Kaggle telecom customer dataset. Manal Lousily et al., [9] utilized machine learning algorithms and models to estimate customer churn risk in the telecoms business. The proposed method made use of KNN, LR, RF, and SVM models. The SVM model achieved the highest accuracy of 96.92% compared to the other predictive models with the Kaggle dataset, modified from the original IBM data. Xiancheng Xiahou et al. [10] conducted an experiment in which machine learning models such as SVM and LR were used to anticipate client attrition in B2C e-commerce enterprises. The proposed model uses K-means client segmentation, splits clients into three different categories, and finds the primary categories of customers. The proposed experiment used a dataset collected from Alibaba Cloud Tianchi and achieved 91% accuracy. Amgad Muneer et al., [11] utilized three prominent models, namely RF, AdaBoost, and SVM, to construct a customer churn prediction technique. The initial dataset was collected from the Kaggle online repository and balanced with SMOTE. Their proposed system was 88% accurate.

Abhinav Sudhir Thorat et al., [15] used various machine learning algorithms such as XGBoost, RF, and LR to predict customer churn in the telecoms business. Their proposed approach achieved 88% accuracy. A lot of researchers have worked on customer churn prediction, but there are still some limitations. Some techniques used an imbalanced and small dataset and achieved satisfactory accuracy to some extent, while others achieved less accuracy. Still, a lot of research work is needed in customer churn prediction.

3 Proposed Methodology

The main objective of the suggested approach is to develop a churn prediction model for the telecom indus-

try to identify customers who are likely to churn and devise effective strategies to retain them. This study uses BAT-ANN models with a churn prediction dataset to predict churn efficiently and accurately. The purpose of combining a bat method and an artificial neural network is to provide better performance estimates and predictions. Data pre-processing, data cleaning, filtering, and feature selection are important steps before feeding data into a model. Applying data cleaning techniques (imputation, controlling outliers, resolving inconsistencies) and selecting relevant features for churn prediction (demographics, usage behavior, service history) to improve the model prediction accuracy The proposed experiment consists of six phases, as shown in Figure 1.

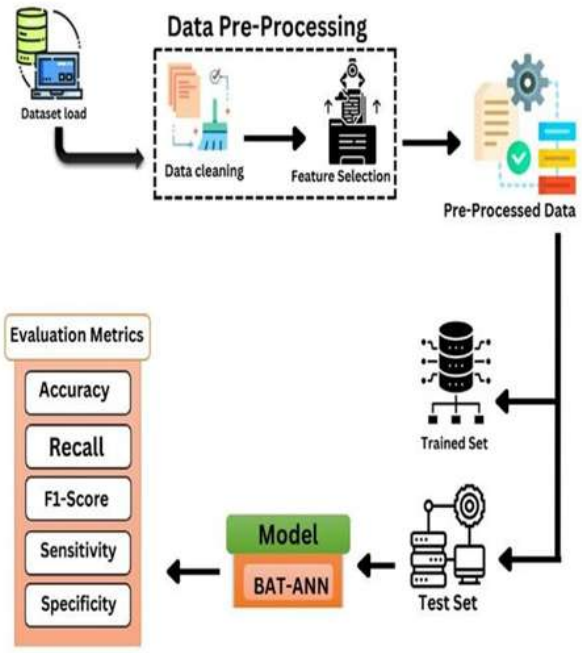

Figure 1. Proposed Methodology

3.1 Data Acquisition

The dataset used in the proposed research experiment, obtained from the BigML website (https://bigml.com/gallery/datasets), contains information about customers' usage patterns, interactions, and relevant features of churn prediction for a telecommunications company, such as demographics, usage patterns, and whether or not they churned (i.e., stopped being customers). The dataset contains 3333 instances, 21 features, and the churn variables (0 for no, 1 for yes).

3.2 Data Pre-processing

Data cleansing involves identifying and rectifying incompleteness, anomalies, and discrepancies in a dataset. Various techniques are employed to address these issues, including imputation methods such as mean, median, or mode for handling missing values. Outliers can be managed through removal, transformation, or acceptance as valid data points. Inconsistencies are resolved by cross-referencing information, error correction, or making reasonable assumptions. These techniques guarantee data accuracy and enhance the reliability of the churn prediction model. Additional methods like standard scaling ensure uniform coefficients for all attributes and scale features between 0 and 1 using min-max scaling. Finally, lines containing confused data are eliminated to preserve data integrity.

3.3 Feature Selection

A critical factor that enhances classification performance and reduces calculation time for classification algorithms is the thorough and correct selection of features by selecting a subset of relevant and informative features. Eliminating unnecessary or noisy features that can enhance machine learning model accuracy and minimize model training time by lowering the data set size Additionally, the practice of standardizing numerical data through the normalization process of converting written material to numerical data Irrelevant data is eliminated, and unstructured data is turned into a structured format that machines can interpret. The min-max scaling technique converts numerical features into a comparable range. This is done to avoid bias induced by differing feature scales and to ensure that machine learning models treat all features equally.

3.4 Prediction model

The proposed research work used a single BAT-ANN algorithm on a BigML dataset.

3.4.1 Bat Algorithm

Yang first presented the BAT algorithm, motivated by how natural bats find food within their native environment. It is an evolutionary algorithm used to solve several problems. This method optimizes a variety of activities by utilizing echolocation, flying patterns, and bats' capacity to adjust their behavior in response to environmental changes. Because of the BAT algorithm's effectiveness in resolving challenging optimization issues, it has attracted considerable interest in the scientific community. It uses a population-centered approach, where each bat in the population represents a potential solution, and it iteratively updates the positions and velocities of the bats using both local and global search techniques. The BAT algorithm exhibits potential abilities in resolving practical optimization challenges, such as feature selection, clustering, and neural network training. It does this by emulating the characteristics of bat behavior, such as the emission of high-frequency calls and the exploration-exploitation trade-off. It is an appealing option for academics because of its versatility, convergence speed, and resilience.

3.4.2 Artificial neural networks(ANN)

Artificial neural networks are a well-known artificial intelligence technique for simulating the operation of a human brain mechanism. It is a way of handling data coming from various points in the structure, termed neurons. Such nodes are integrated into various layers that will cooperate to address a challenging problem. The input layer, hidden layer, and output layer are all components of the ANN structure. NNs are made up of multiple layers of neurons, which are nodes that are linked together. Beginning at the input layer and moving through hidden levels, data moves via the network until it reaches the output layer. A weighted total of each neuron's inputs is applied, subsequently followed by an activation function that adds nonlinearity. In order to reduce the discrepancy between the expected output and the actual result, training techniques such as reverse propagation are used to alter the weights as well as the biases of the neurons. ANNs are able to recognize complex patterns in the data and generate precise predictions or classifications through this continuous learning process. Numerous fields, including image identification, natural language processing, and recommendation systems, have found extensive use for artificial neural networks (ANN). Figure 2 shows the three-layered architecture of ANN.

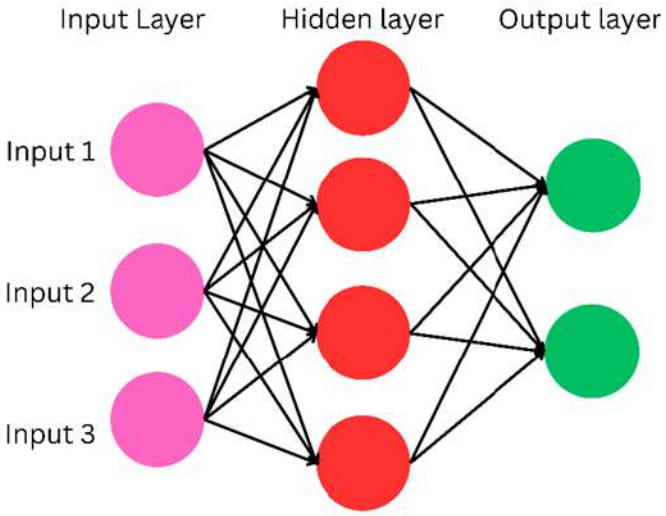

Figure 2. ANN

3.5 Evaluation matrics

Evaluation metrics are used to measure the efficiency of a model in the context of artificial intelligence and data science. These metrics measure the extent to which predictions the model makes match with actual facts. Following are some examples of common evaluation measures together with their formulas:

Accuracy, Precision, Recall, F1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

4 Results and Discussion

This section of the study presents the experimental outcome together with the corresponding statistical analysis. Initially, the performance of the BAT-ANN

model is evaluated and then compared with previous studies. Table 2 shows the overall performance of the BAT-ANN model.

Table 1. Performance evaluation matrics

Model	Accuracy	Precision	Recall	F1- Score
BAT-ANN	0.89	0.90	0.89	0.89

The proposed model achieved an accuracy of 89%, precision of 90%, recall of 89%, and f1-score of 89%. Figure 3 shows the graphical representation of the overall performance of the BAT-ANN model.

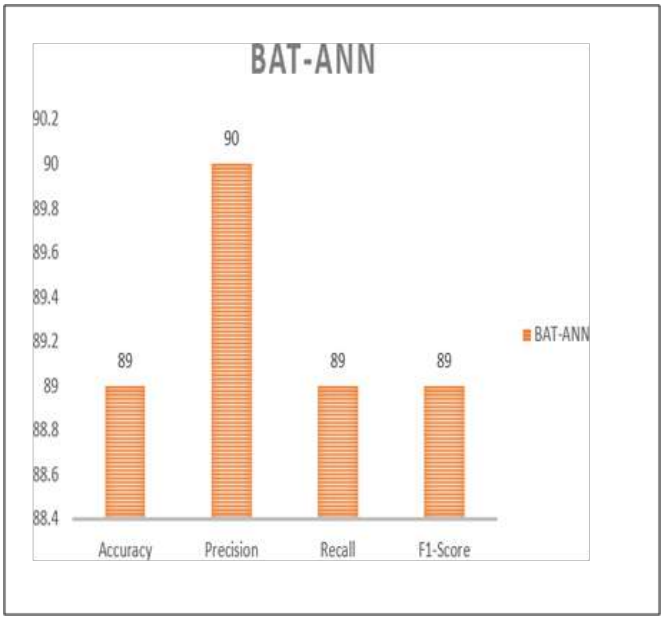

Figure 3. Performance evaluation of BAT-AN

The confusion matrix of the proposed model is depicted in Figure 4.

In Figure 4, we see that the BAT-ANN predicted 568 records correctly and 3 records incorrectly. Similarly, out of 96 records, the BAT-ANN successfully predicted 27 of them.

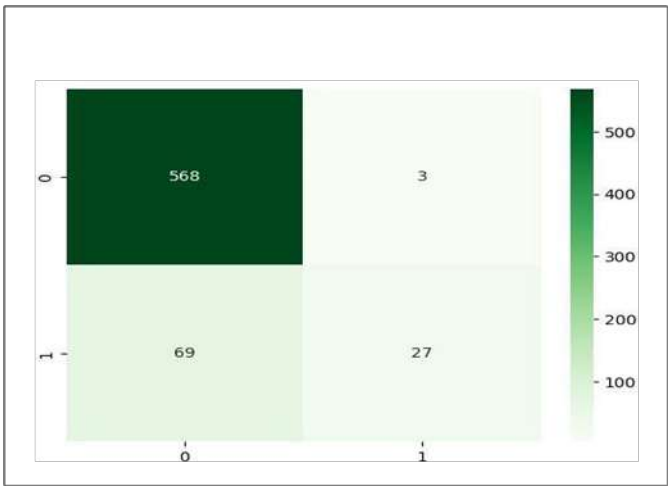

Figure 4. Confusion matrix of BAT-ANN.

The ROC curve is created by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis. Figure 5 depicts the ROC curve of the proposed classifier.

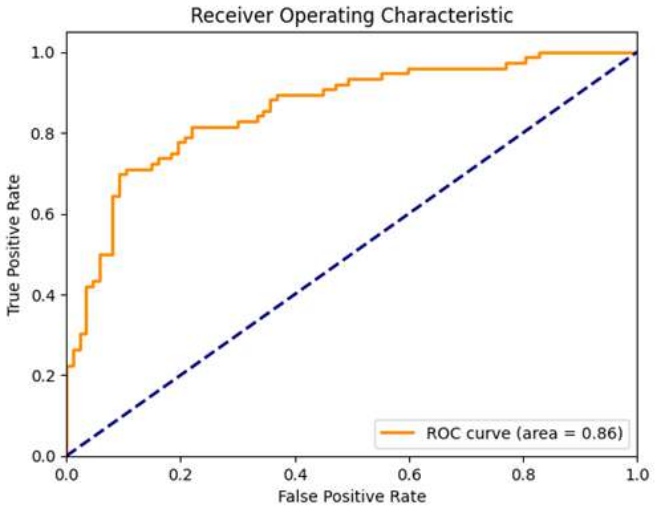

Figure 5. ROC curve of BAT-ANN

4.1 COMAPRATIVE ANALYSIS

This section of the paper presents a comparison of our proposed work with previous studies. The proposed research work used the BAT-ANN model for real-time

telecom industry data and achieved 89.2% accuracy, which is quite good. In contrast, the below-mentioned studies achieved less accuracy. Table II shows the comparison of the proposed approach with previous studies.

Table 2. Comparison of proposed work with previous studies

Reference	Model Used	Accuracy
1	KNN,CatBoost, RF	84%
2	DT-ANN, DT,ANN	81%
5	RR, LR, AdaBoost.	80%
6	NB, LR, XG-Boost, KNN, DEEP BP-ANN	88%
8	LR, SVC, GBDT, RF,AdaBoost	83%
Proposed	BAT-ANN	89.2%

5 Conclusion and future work

The proposed research study focused on resolving the essential issue of customer turnover in rapidly growing telecom companies. The rising amount of data collected by communications companies demands innovative analytical methods to identify future churn customers as well as maintain current clients. The study recommended using the BAT-ANN classifier using BigML data for accurately predicting customer attrition. The outcome of the study showed an outstanding validation accuracy of 89.2%, showing the model's efficacy in detecting possible churn customers. Telecommunications companies may enhance their company's needs for maximizing revenue by using such effective churn systems for customer churn prediction. In the future, we will use hybrid machine learning and deep learning models with the hybrid datasets of BigMl and IBM and then compare their results with approaches that used these datasets separately with simple machine learning and deep learning models.

Author Contributions

Gul Zaman Khan: supervision, conceptualization, methodology, implementation **Ikram Ulhaq**: Proof reading, literature collection. **Ihsan Adil**: visualization, investigation. **Sajad Ulhaq**: Result creation,writting original draft **Inam Ullah**: Writing,reviewing and editing.

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

Funding Information

This article doesn't receive any external funding.

References

[1] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang, and Q. W. Poh, "Predictive churn modeling for sustainable business in the telecommunication industry: Optimized weighted ensemble machine learning," *Sustainability*, vol. 15, no. 11, p. 8631, 2023.

[2] R. H. Herdian and A. S. Girsang, "The implementation of hybrid methods in data mining for predicting customer churn in the telecommunications sector," *IEEE*, vol. 7, no. 1, pp. 216–228, 2023.

[3] 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.

[4] X. Xiahou and Y. Harada, "Customer churn prediction using adaboost classifier and bp neural network techniques in the e-commerce industry," *American Journal of Industrial and Business Management*, vol. 12, no. 03, pp. 277–293, 2022.

[5] V. Kavitha, G. H. Kumar, S. V. M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms," *International Journal of Engineering Research and*, vol. V9, no. 05, 2020.

[6] P. Manghnani and U. Kumari, "Customer churn prediction," *IEEE*, vol. 8, pp. 259–292, 2023.

[7] P. Chen, N. Liu, and B. Wang, "Evaluation of customer behaviour with machine learning for churn prediction: The case of bank customer churn in europe," in *Proceedings of the International Conference on Financial Innovation, FinTech and Information Technology, FFIT 2022*, Shenzhen, China, October 28-30 2022.

[8] C. Zhuo, "Prediction of telecom customer churn based on mipca-xgboost method," *Frontiers in Computing and Intelligent Systems*, vol. 3, no. 1, pp. 1–5, 2023.

[9] M. Loukili, F. Messaoudi, and M. E. Ghazi, "Supervised learning algorithms for predicting customer churn with hyperparameter optimization," *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 3, pp. 50–63, 2022.

[10] X. Xiahou and Y. Harada, "B2c e-commerce customer churn prediction based on k-means and svm," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, 2022.

[11] A. Muneer, "Predicting customers churning in banking industry: A machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 539–549, 2022.

[12] S. W. Fujo1, S. Subramanian, and M. A. Khder, "Customer churn prediction in telecommunication industry using deep learning," *Information Sciences Letters*, vol. 11, no. 1, pp. 185–198, Jan 2022.

[13] S. O. Abdulsalam, J. F. Ajao, B. F. Balogun, and M. O. Arowolo, "A churn prediction system for telecommunication company using random forest and convolution neural network algorithms," *ICST Transactions on Mobile Communications and Applications*, vol. 6, no. 21, 2022.

[14] I. A. et al., "Leveraging metaheuristics with artificial intelligence for customer churn prediction in telecom industries," *Electronic Research Archive*, vol. 31, no. 8, pp. 4443–4458, 2023.

[15] M. A. S. Thorat and V. R. Sonawane, "Customer churn prediction in telecommunication industry using deep learning," *IEEE*, vol. 38, no. 3, p. 1417, 2023.