Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8534739, 16 pages
https://doi.org/10.1155/2022/8534739

Research Article

Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study

Nabahirwa Edwine,[1] Wenjuan Wang,[1] Wei Song,[1] and Denis Ssebuggwawo[2]

[1]*College of Information Technology, Shanghai Ocean University, Shanghai 201306, China*
[2]*Department of Computer Science, Kyambogo University, Kampala, Uganda*

Correspondence should be addressed to Wenjuan Wang; wangwj@shou.edu.cn and Wei Song; wsong@shou.edu.cn

Received 30 December 2021; Revised 11 June 2022; Accepted 16 June 2022; Published 18 July 2022

Academic Editor: Melih Yucesan

Churn rate describes the rate at which customers abandon a product or service. Identifying churn-risk customers is essential for telecom sectors to retain old customers and maintain a higher competitive advantage. The purpose of this paper is to explore an effective method for detecting the risk of customer churn in telecom sectors through comparing the advanced machine learning methods and their optimization algorithms. Based on two different telecom datasets, Mutual Information classifier was firstly utilized to select the most critical features relevant to customer churn. Next, the controlled-ratio undersampling strategy was employed to balance both minority and majority classes. Key hyperparameter optimization algorithms of Grid Search, Random Search, and Genetic Algorithms were then combined to fit the three promising machine learning models-Random Forest, Support Vector Machines, and K-nearest neighbors into the customer churn prediction problem. Six evaluation metrics-Accuracy, Recall, Precision, AUC, *F*1-score and Mean Absolute Error, were last used to evaluate the performance of the proposed models. The experimental results have revealed that the RF algorithm optimized by Grid Search based on a low-ratio undersampling strategy (RF-GS-LR) outperformed other models in extracting hidden information and understanding future churning behaviors of customers on both datasets, with the maximum accuracy of 99% and 95% on the applied dataset 1-2 respectively.

1. Introduction

In the telecom sector, situations, where customers withdraw from company services, are common. The customer either voluntarily leaves the services because of reasons like being unsatisfied with company's existing plans and services, changing their jobs and relocating to a different area where they cannot access the services, going out of the country, finding a job mandating them to use a particular service provider, or the service provider involuntarily terminates their services due to some reasons like failure to pay bills or using the services for illegal purposes [1]. Churn is movement of clients from one service provider to another in a specific period with different reasons, reducing revenues for the former service provider [2]. Due to increasing competition in the telecom industry, clients have taken an advantage of available choices for transferring to better and cheaper services. Customers are the main source of revenue for any organization [1]. A precise customer churn prediction can help companies to retain their customers at a large base and thus maintain their revenues at a higher level.

One of the reasons why telecom companies employ customer relationship management is customer retention. Huge volumes of data stored in customer relationship management systems have been currently adapted in predicting customer churn and revealed the factors that drive customers to churn [3]. Acquiring a new client is expensive considering the cost of advertising, promotion, human resource, and time involved [1]. Retaining customers is becoming a major consideration for telecom service providers because it is easier and more economic [4]. Currently, companies have increased the gear in exploiting customer churn prediction method as a data mining task to investigate the high-risk customers who are more likely to quit the company service.

Single classifiers, optimized classification techniques as well as hybrid machine learning models have been widely applied to interpret customer churn [5–7]. However, some of the results are not satisfactory due to a lack of understanding of datasets thoroughly and applying inappropriate methods for data preprocessing, class balancing, and learning models [8, 9].

Random Forest (RF) has been widely employed for solving different classification problems in other industries, such as botnet detection, smart meter data classification, satellite imagery, and employee turnover [10–13]. Although some researchers have applied RF technique to predict customer churn in telecom sector, the predicting results range between 67% and 87% [14–18], which still has a distance from being satisfied. To derive the most powerful method for detecting the risk of customer churn in the telecom sector, an optimized RF model is employed and compared with the other two advanced machine learning models-Support Vector Machines (SVM) and K-nearest neighbors (KNN) in this study. Studies have proved that Machine Learning (ML) algorithms optimized by the optimal hyperparameters have better performance [19, 20]. To boost ML models by hyperparameter optimization (HPO), it is necessary to find out the key hyperparameters that need to be tuned to fit the ML model into specific problems or datasets [21]. This study is trying to optimize the performance of RF, SVM, and KNN by searching appropriate key hyperparameters via comparing Grid Search (GS), Random Search (RS), and Genetic Algorithms (GA) to fit customer churn prediction problems for the telecom sector [21].

Taking dominant determinants as inputs can increase prediction efficiency and lead to higher accurate results [4, 22]. In this study, Mutual Information (MI) classifier has been adopted as feature selection method [23]. Another issue of customer churn prediction is class imbalance, which degrades classification performance due to biased supervision. Class imbalance is a situation where the number of observations belonging to one class is significantly lower than those belonging to the other class. As stated by the study in [24], random sampling methods for class imbalance are not useful in improving the performance of predicting results. Therefore, a controlled-ratio undersampling strategy is employed in this study to favor both minority and majority classes [24]. To evaluate the general performance of the predicted models, six evaluation metrics of Accuracy, Recall, Precision, AUC, $F1$-score and Mean Absolute Error (MAE) were used [25–27]. The major contribution of this work is summarized as follows:

(1) Adopting advanced ML algorithms (Random Forest, Support Vector Machines, K-nearest neighbors) and HPO techniques (GS, RS, GA) to solve customer churn prediction problems in the telecom sector to find out which method works best.

(2) Implementing a controlled-ratio undersampling technique to solve the problem of class imbalance, bringing up a significant improvement in the overall performance of the proposed customer churn predicting models.

(3) Providing a solution using an MI classifier to identify critical factors related to customer churn in the telecom sector which can be used by practitioners for building better customer relationships.

The remainder of this paper is structured as follows. Section 2 presents a literature review of customer churn determinants and prediction models. Section 3 presents the research methodology. Section 4 provides the major research findings and discussion. Section 5 provides both practical and theoretical implications and Section 6 concludes the study and provides future recommendations.

2. Literature Review

2.1. Machine Learning. Machine learning algorithms have been broadly utilized to construct vigorous and proficient classification models. A wide range of research especially in Natural Language Processing has shown that optimized machine learning algorithms can achieve effective results in text classification. Four learning algorithms (Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forest) with five widely utilized ensemble methods (AdaBoost, Bagging, Dagging, Random Subspace, and Majority Voting) were analyzed to evaluate the effectiveness of statistical keyword extraction and language function analysis respectively [28, 29]. Bagging ensemble of the Random Forest algorithm achieved the highest average predictive performance of 93.80% on statistical keyword extraction [28] and 94.43% on language function analysis [29]. An ensemble scheme based on hybrid supervised clustering for text classification was introduced by Onan [30] to partition the data samples of each class into clusters. The ensemble classifier outperforms the conventional classification algorithms. A deep learning architecture that combines a weighted Glove word embedding with CNN-LSTM (Convolutional Neural Network-Long-Short Term Memory) architecture, an approach to sentiment analysis on product reviews obtained from Twitter outperforms the conventional deep learning methods in [31]. Similar results can also be found in the study [32] which constructs an improved word embedding scheme incorporating word vectors obtained by word2vec, POS2vec, word-position2vec, and LDA2vec schemes, and study [33] that builds an effective sarcasm identification framework on social media data by pursuing the paradigms of neural language models and deep neural networks.

Therefore, it is concluded that optimized machine learning algorithms have great superiority in solving classification problems.

2.2. Customer Churn Determinants. To detect customer churn, the first step is to find the most relevant factors which can be done by manual selection or automatic feature selection.

Mahajan et al. [22] conducted a review on factors affecting customer churn in the telecom sector and categorized these factors into two sub-categories: Service Quality and Brand Image. Factors that fall under the Service Quality include Network coverage/signal length, Voice quality, Tariff, Customer service, and Value-added service. Antecedents of customer churn under Brand Image category include Fair Company, Friendly Company, and Innovative Company. Jain et al. [1] found out that Lack of engagement, Lack of new offers/promotions, Lack of customer service support, High call rates/SMS charges, Nonpayment bills, Fraud or Misuse of services by customers, and Change of location or position of customers are the main factors why customers churn. According to various studies conducted on factors that determine customer churn, service quality has been quoted as a vital determinant [1, 34, 35]. Other phenomena of relevant work have shown that the nature of the dataset can largely affect the determinants of customer churn [36].

To determine what factors, contribute more to customer churn, feature selection methods such as principal component analysis [37, 38] have been utilized, but it is difficult to interpret their results. Thus, MI becomes a better feature selection method because its results can be easily interpreted [23].

2.3. Customer Churn Prediction Methods.

Besides the telecom sector, other industries like Banking and Insurance have also conducted research on customer churn prediction [39–41].

Researchers have put a lot of emphasis on applying machine learning techniques, both individual and hybrid classifiers, in predicting customer churn. Some studies have utilized single classification methods to anticipate customer churn in the telecom industry. Sharma and Panigrahi [42] proposed a Neural Network (NN) approach for predicting customer churn in cellular network services. The accuracy of the proposed model was up to 92%. Using a dataset of 5000 instances obtained from an anonymous mobile service provider, Shaaban et al. [43] applied Decision Tree (DT), NN, and SVM to predict customer churn respectively. NN and SVM performed both best with an accuracy of 83.7%. Jamjoom [44] carried out an analysis of customer churn in B2B. Data from a health insurance company were included for providing insights into churn behavior based on a design and application of a prediction model. Three promising data mining techniques were identified for building the prediction models, including logistic regression (LR), NN, and K-means. NN and regression analysis performed both best in their study.

Current researchers tend to use ensemble and hybrid ML methods for predicting customer churn in the telecom industry.

A model was suggested by Vijaya and Sivasankar [7] using Rough Set Theory (RST) to identify the efficient features for customer churn prediction in telecommunication sectors. The authors employed ensemble-classification techniques such as Bagging, Boosting, and Random Subspace. The ensemble RST model performed well with an accuracy of 95.13%. Customer churn prediction in the telecom industry using the RST approach was also declared by Amin et al. [45]. In their study, RST was combined with the Exhaustive Algorithm, GA, Covering Algorithm, and the Learning from Examples Module algorithm respectively in developing the prediction model. The authors revealed that RST-GA is the most efficient technique for extracting implicit knowledge of customer churn. Li and Marikannan [19] optimized Naïve Bayes (NB), DT, and Artificial Neural Network (ANN) with GS to predict customer churn. ANN-GS was the best model with 86.6% accuracy. Ahmad et al. [15] employed methods of DT, RF, Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBOOST) to determine the clients that would transfer to a competitor. XGBOOST tree model achieved proficient results in all measurements with an AUC of 93.3%, followed by GBM, RF, and DT successively.

To build the best model that would foretell customer churn for the telecom industry, Ahmed and Maheswari [46] compared various machine learning models, including NB, LR, Multiple Layer Perception (MLP), DT, RF, and Gradient Boost Trees (GBT) and constructed a heuristic model based on ensemble stacking, which performed the best. The predicting results of the model have a higher accuracy of 0.95 but with a lower true positive rate of 0.71. Azeem et al. [47] conducted customer churn prediction research utilizing dataset from a telecom company in South Asia by comparing traditional ML algorithms, ensemble and fuzzy-based models. The results of their study indicated that fuzzy classifier-based models performed better than other models, where Ordered Weighted Average NN is the best method with a true positive rate up to 98%, but with a low AUC score up to 0.68.

There are hybrid methods that have been implemented for customer churn prediction. Zdravevski et al. [48] put forward a cloud-based Extract-Transform-Load (ETL) framework for data fusion and aggregation from a variety of sources. In the churn prediction case, their results showed that over 98% of churners could be detected. A hybrid ANN for customer churn prediction in telecom sector was proposed in [49]. The classification accuracy of ANN-4 hidden layers is 90.34% compared to ANN-2 hidden layers whose maximum accuracy attained is 88.14%.

Research has revealed that class imbalance has a negative impact on prediction results. Yahaya et al. [9] performed an enhanced churn prediction model using GA, K-means clustering, and ANN. The results have shown that the training performance will be improved as the noises in the data are reduced, nevertheless, the testing results were not improved with filtering. The accuracy of their hybrid model (GA-K-means-ANN) dropped from 100% on training to 76.6% on testing. This is because the imbalance between the positive and negative classes can also largely affect the testing results. Similarly, Adhikary and Gupta [50] compared the performance of 114 classifiers for churn prediction in telecom companies by determining various performance parameters. They have found that class imbalance affects the performance of the classifiers, and the ensemble family of classifiers yields better results than SVM, with an accuracy ranging between 70%–75%.

It is concluded that when conducting research towards an effective method for predicting customer churn, examination of different datasets, use of better class imbalance and feature selection techniques, comparison between different ML models and their optimization strategy will be a feasible research methodology.

3. Methodology

The overall research framework is displayed in Figure 1. The methods of data preprocessing and feature selection, controlled-ratio undersampling, ML algorithms, and HPO, and model performance evaluation are depicted in step 1, step 2, step 3 and step 4 respectively. Through step 1, the original data will be preprocessed and critical factors closely related to customer churn will be identified. Then in step 2, different undersampling strategies will be examined to balance the data classes to improve the overall classification performance of the predicted models. In step 3, three ML methods of RF, SVM, and KNN are optimized by three HPO techniques of GS, RS and GA respectively to develop the customer churn prediction models. Finally, the performance of the optimized prediction modes are evaluated and compared with six evaluation metrics.

3.1. Data Preprocessing and Feature Selection

3.1.1. Data Preprocessing. Python tool was utilized in preprocessing the datasets. A uniform numerical format in the training set was established. Redundant variables like Phone, Customer ID, State, and Area Code were deleted because they are insignificant in predicting customer churn. Null values were filled up with zeros (0s). For features with three categories of Yes, No, and No Internet service/No phone service, the "No internet service/No phone service" category was converted into "No" category. Moreover, variables that had more than two categories were converted into dummy variables to obtain numerical values for data analysis.

3.1.2. Feature Selection with Mutual Information Classifier. Feature selection is a critical step in most classification problems to select an optimal subset of features, increasing the classification accuracy and efficiency [51].

Mutual information underlies the concept of measuring the mutual dependence between two random variables by picking out how much information one of the variables can be obtained from the other one. Unlike techniques like the Pearson correlation coefficient, the MI can capture nonlinear dependencies and is constant under distinguishable variations of the random variables. The MI identifies a suitable subset of features by measuring how much the value of one variable reduces the uncertainty on the other [52]. With MI, irrelevant features futile for foreseeing the target can be disposed of precisely and effectively. This presents the idea of pertinence, repetition, and complementarity [53]. Hence, MI is favored as a reasonable strategy for selecting significant features for predicting customer churn in the telecom sector. The fundamental aim of feature selection is to identify m most important attributes out of the d original attributes, where $m < d$. Mutual Information classifier is a filter technique, which chooses the most relevant features without using a learning algorithm. Mutual information $I\,(X;Y)$ refers to the amount of uncertainty in X due to the knowledge of Y according to Information theory. Mathematically, mutual information can be defined as follows:

$$I\,(X;Y) = \sum_{x,y} p\,(x,y)\log\frac{p\,(x,y)}{p\,(x)p\,(y)}, \tag{1}$$

where $p(x,\ y)$ represents the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ represent the marginal probability distribution functions of X and Y, respectively.

3.2. Controlled-Ratio Undersampling. Class imbalance is a common problem in the real-world datasets [54] because it degrades the classification performance of ML algorithms due to biased supervision [55]. In situations where the majority class outweighs the minority class, undersampling is a potent approach. Sampling ratio is defined as the ratio of the sampled majority size over the minority size. Because different undersampling ratios have different effects toward the two classes [54], the minority class is regarded as a positive class while the majority class is regarded as negative class in this study. A 1.0 sampling ratio means that the majority examples are randomly undersampled up to the same size as minority examples. A ratio which is below 1.0 means that the number of sampled majority examples is smaller than that of the minority examples. In this paper, we endorse low-ratio (0.3, 0.4) and high-ratio (0.6, 0.7) undersampling strategies.

In the study conducted by Komamizu et al. [54], it is noted that classifiers trained with excessively undersampled datasets (the 1.0 strategy) favor the minority class while those trained with moderately undersampled datasets favor the majority class, therefore, 1.0 strategy may not be best balancing ratio. To construct a model which favors both majority and minority classes, controlled-ratio strategy is adopted in this study.

To automatically determine the ratios, one approach is to simplify the parameter to a ratio interval. The user assigns an interval value for splitting the imbalance ratio (IR) into a sequence of sampling ratios. The IR differences are determined by the nature of dataset; the larger the dataset, the larger the sampling ratio value [48].

Let p and $\mathcal{N}$ denote sets of minority examples and majority examples, respectively. At the sampling step c, $\mathcal{N}$ is undersampled to $\mathcal{N}'$ so that $(|\mathcal{N}'|/|p|) = R\,(c, np, n\mathcal{N})$, where $R\,(.)$, np and n$\mathcal{N}$ are sampling ratio and the numbers of sampling ratios in the splitting strategy.

Given np and n$\mathcal{N}$ are the sampling ratio determination function R calculates the sampling ratio as a cumulative ratio of blocks as follows:

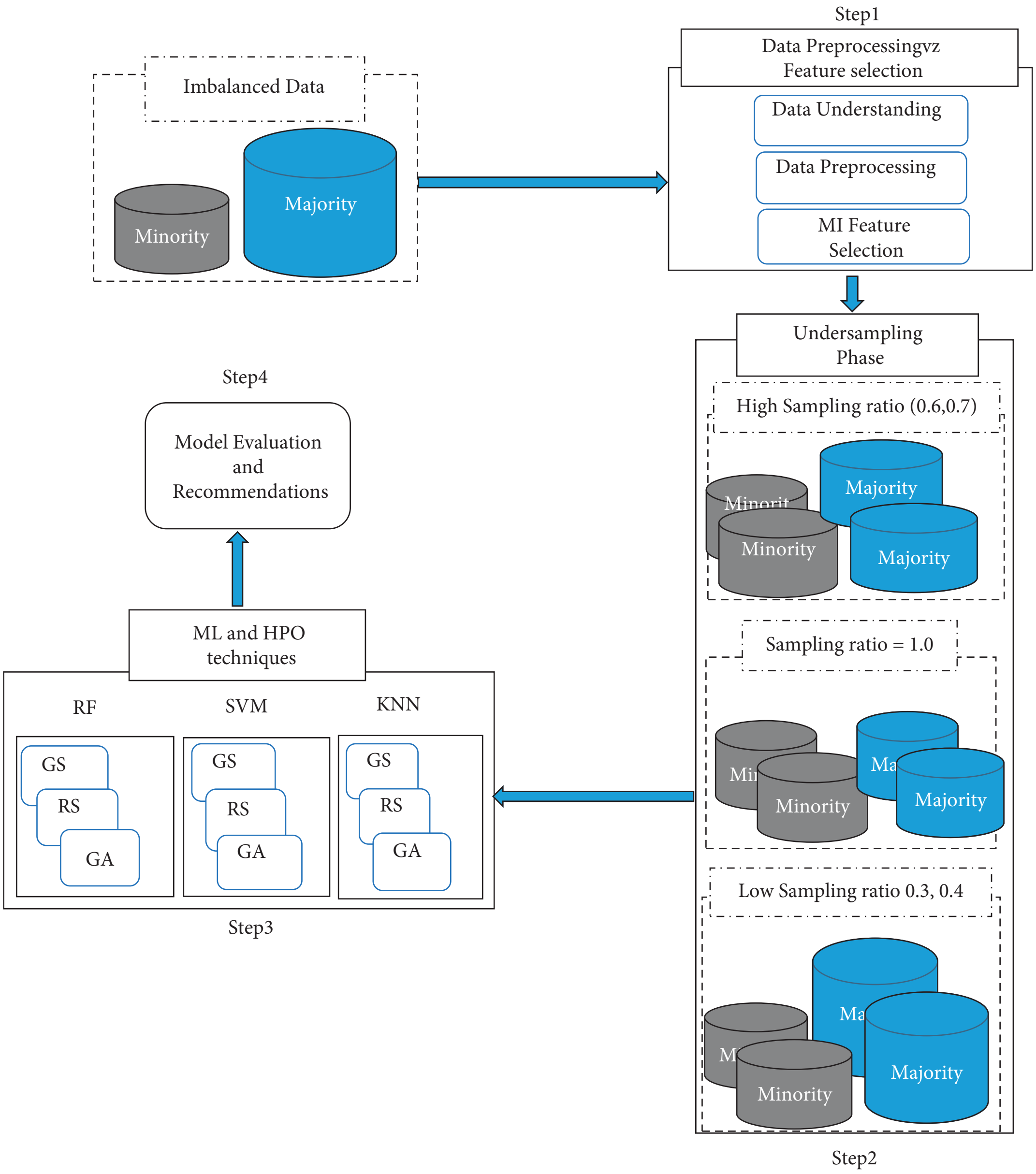


Figure 1: Overview of research design.

$$R(c,np,n\mathcal{N}) = \begin{cases} \dfrac{c}{np+1}, & \text{if}c \le np, \\ 1, & \text{if}c = np+1, \\ 1 + \dfrac{IR}{n\mathcal{N}}(c-np), & \text{if}c > np+1, \end{cases} \qquad (2)$$

where $IR = (|\mathcal{N}|/|p|)$ is the imbalance ratio, the number of majority examples over that of minority examples.

3.3. Prediction Models. Three ML classification algorithms - RF, SVM, and KNN are implemented to address the customer churn prediction problem in telecom sector. To optimize the performance of the proposed prediction models, HPO techniques (Grid search, Random search, Genetic algorithm) are employed and compared, following the methods proposed by [21].

3.3.1. Random Forest. RF combines multiple decision trees to improve model performance [56]. The schematic diagram of RF is displayed in Figure 2.

RF has a lot of hyperparameters to be tuned to build an effective model, which are depicted as follows [21]:

(1) Measuring function denoted by "criterion" in sklearn which has two main types-gini impurity and information gain;

(2) The number of important features for generating the best split (max_features);

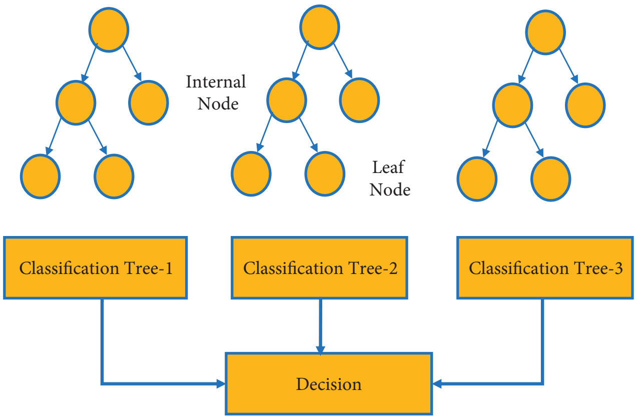


Figure 2: Schematic diagram of RF.

(3) Splitter as the split selection method that can be set to choose the best split (best) or to select a random split (random);

(4) The minimum number of data points to split a decision node or to obtain a leaf node, denoted by “min_samples_ split” and “min_samples_leaf,” respectively;

(5) The “max_leaf_nodes,” indicating the maximum number of leaf nodes;

(6) The “min_weight_fraction_leaf” that means the minimum weighted fraction of the total weights.

3.3.2. SVM. A support vector machine is a supervised learning algorithm that can be used for both classification and regression problems. SVM algorithms are based on the concept of mapping data points from low-dimensional into high-dimensional space [21]. Support vector machines are sensitive to outliers and noise [57, 58]. The schematic diagram of SVM is drawn in Figure 3.

The kernel type is the main hyperparameter to be tuned in SVM. Common SVM kernel types include linear kernels, radial basis function (RBF), polynomial kernels, and sigmoid kernels. The coefficient y, denoted by “gamma” in sklearn, is the conditional hyperparameter of the “kernel type” hyperparameter when it is set to polynomial, RBF, or sigmoid; r, specified by “coef0” in sklearn, is the conditional hyperparameter of kernels polynomial and sigmoid. Moreover, the polynomial kernel has an additional conditional hyperparameter d representing the “degree” of the polynomial kernel function [21]. This study examines all the four types of kernels and the best is automatically selected for predicting customer churn in the telecom sector.

3.3.3. KNN. K-nearest neighbor calculates distances between data points. In KNN, the predicted class of each test sample is set to the class to which most of its k-nearest neighbors in the training set belong. The most important hyperparameter to be tuned is the number of considered nearest neighbors-k [21]. The schematic diagram of KNN is described in Figure 4.

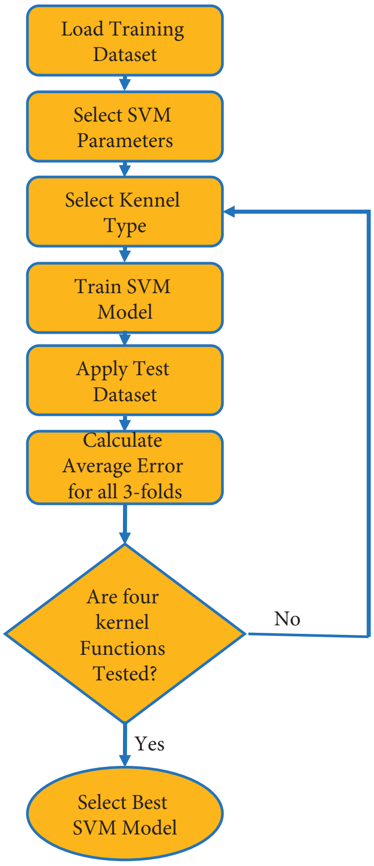


Figure 3: Schematic diagram of SVM.

3.4. HPO Techniques. To boost ML model performance by HPO, finding out the key hyperparameters that need to be tuned to fit the ML model into specific problems or datasets is crucial. The hyperparameter optimization process consists of four main components: an estimator (a classifier) with its objective function, a search space (configuration space), search or optimization method used to find hyperparameter combinations, and an evaluation function to compare the performance of different hyperparameter configurations. Specifically, the main process of HPO is shown below [21]:

(1) Choose the objective function and the performance metrics;

(2) Selecting the hyperparameters that need to be tuned, note their types, and choose the suitable optimization technique;

(3) Using the default hyperparameter configuration, train the ML model as the baseline model;

(4) Begin the optimization process with a large search space as the hyperparameter achievable domain;

(5) Reduce the search space based on the regions of currently tested well-performing hyperparameter values, or if necessary, explore new search spaces;

(6) Return the standout hyperparameter configuration.

Three HPO techniques of Grid search, Random Search, and Genetic Algorithms were used in this study.

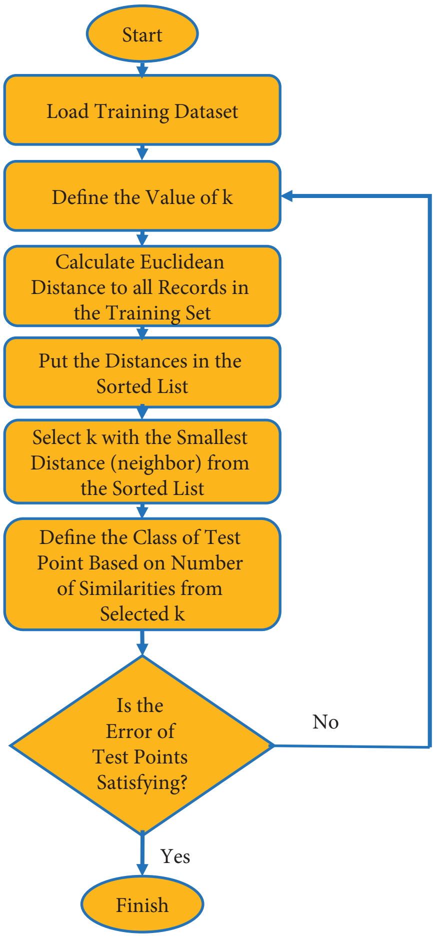

Start
Load Training Dataset
Define the Value of k
Calculate Euclidean Distance to all Records in the Training Set
Put the Distances in the Sorted List
Select k with the Smallest Distance (neighbor) from the Sorted List
Define the Class of Test Point Based on Number of Similarities from Selected k
Is the Error of Test Points Satisfying?
No
Yes
Finish

Figure 4: Schematic diagram of KNN.

3.4.1. Grid Search. GS is an exhaustive search or a brute-force method that evaluates all the hyperparameter combinations given to the grid of configurations. It is one of the most commonly-used methods to explore hyperparameter configuration space [59].

3.4.2. Random Search. RS and GS work almost the same way. The only difference is that RS randomly selects a predefined number of samples between the upper and lower bounds of candidate hyperparameter values, and then trains these candidates until the defined budget is exhausted instead of testing all values in the search space as GS [59].

3.4.3. Genetic Algorithm. GA is a metaheuristic algorithm based on evolutionary theory. With GA, the individuals that will survive and continue to the next generations are those with high capability and adaptability of survival to the environment [21].

The next generation will also inherit their parents' characteristics and may involve better and worse individuals. Better individuals will be more likely to survive and have more capable offspring; The worse individuals will vanish slowly. The individual with the best adaptability will be identified as the global optimum after many generations have passed. When implementing GA to solve HPO problems, each chromosome or individual represents a hyperparameter. Moreover, its decimal value is the actual input value of the hyperparameter in each evaluation. Every chromosome has several genes, which are binary digits; and then crossover and mutation operations are performed on the genes of this chromosome. The population involves all possible values within the initialized chromosome/parameter ranges, while the fitness function characterizes the evaluation metrics of the parameters [21].

3.4.4. Configuration Space of HPO Methods. Three HPO methods of GS, RS and GA are compared using the same hyperparameter configuration space (see Table 1). For KNN, "n_neighbors," is set to be in the same range of 1 to 20 for each optimization method evaluation. The maximum number of iterations for all HPO methods is set to 50 for RF and SVM model optimizations, and then 10 for KNN model optimization. All the experiments are repeated ten times with different random seeds to avoid problems that may arise due to randomness. Finally, the HPO experiments are implemented based on the main processes stated in Section 3.4.

Many machine learning models have a range of hyperparameters and they may interact in nonlinear ways. ML performance is highly sensitive to these hyperparameters. Usually, to study the parameter sensitivity, we need to iterate over all the possible values of hyperparameters to find the set of hyperparameters that result in the best performance of a model on a dataset. It is time-consuming and difficult. Therefore, we used HPO techniques to get an alternative solution. HPO aims to achieve optimal or near-optimal model performance by tuning hyperparameters within the given budgets and returning the majority vote for classification problems [21].

3.5. Evaluation Metrics. Wang et al. [26] denoted that accuracy is inappropriate for performance evaluation in class imbalance learning, because it can be overwhelmed by the high accuracy of the majority class and hide the poor performance of the minority class. Therefore, to prove the general strength of the proposed prediction models, this study adopts six performance measures, including Accuracy, Recall, Precision, AUC, $F1$-score, and MAE.

Accuracy determines the overall performance of the model in predicting churners as churners and nonchurners as nonchurners correctly. Precision refers to how many customers were correctly predicted as churners out of all positive churner predictions. Recall means the rate of customers that were correctly identified as churners out of all true churners. A higher recall value demonstrates a higher performance in predicting churn customers. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average of the absolute differences between prediction and actual observation where all individual differences have equal weight. $F1$-measure is the harmonic mean of the model's precision and recall used as a measure for predicting the performance

Table 1: Configuration space for the hyperparameters of tested ML models.

ML model	Hyperparameter	Type	Search space
RF classifier	n-estimator	Discrete	[10, 100]
	max-depth	Discrete	[5, 50]
	min_samples_split	Discrete	[2, 11]
	min_samples_leaf	Discrete	[1, 11]
	criterion	Categorical	[“gini,” “entropy”]
	max_features	Discrete	[1, 20]
	C	Continuous	[0.1, 50]
SVM classifier	kernel	Categorical	[“linear,” “poly,” “rbf,” “sigmoid”]
KNN classifier	n_neighbors	Discrete	[1, 20]

of the classifier [1, 26, 58]. The equations for calculating Accuracy, Precision, Recall, $F1$-score, and MAE are as follows:

$$\text{Acuracy} = \frac{TP + TN}{TP + FP + TN + FN},$$

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

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 - \text{Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|,$$

where TP (True Positive) means the number of churn customers that the ML learning technique correctly identifies as churn customers, FP (False Positive) means the number of nonchurn customers that the ML technique incorrectly predicts as churn customers, FN (False Negative) means the number of churn customers that ML method incorrectly classifies as nonchurn, and TN (True Negative) means the number of nonchurn customers that ML method correctly identifies as nonchurn.

AUC is the aggregate measurement of the entire ROC Curve which reflects the model’s ability to correctly classify customers who are churned and those who are not. A higher AUC value means that the model is better at correctly classifying outcomes. The AUC value is an effective tool for demonstrating balanced accuracies in both positive (minority) and negative (majority) classes [60] calculated using the following equation:

$$AUC = \frac{1 + TPR - FPR}{2}, \tag{4}$$

where TPR means True Positive Rate and FPR means False Positive Rate.

4. Results and Discussion

4.1. Datasets and Preprocessing. Two datasets [61, 62] from the telecommunication industry are used in the experiments. The details of the two datasets are listed in Table 2.

Table 2: Dataset details.

Description	Dataset 1	Dataset 2
No. of instances	3333	7043
No. of features	21	21
No. of classes	2	2
Percentage of positive samples (churn)	14.5%	27%
Percentage of negative samples (nonchurn)	85.5%	73%

Dataset 1 consists of 3333 records along with 21 variables. Dataset 1 consists of 20 independent variables and one dependent variable (churn). Dataset 1 is heavily imbalanced with the number of churners significantly lesser than the number of nonchurners. Out of the 3333 instances, 2850 (85.5%) are nonchurners while only 483 (14.5%) are churners.

Dataset 2 consists of 7043 rows and 21 columns with 1869 (27%) customers categorized as churners and 5174 (73%) customers as nonchurners. Dataset 2 has 3 numerical variables (tenure, monthly charges, and total charges) and 18 categorical variables relevant to demographic and account service information about the customer.

Open-source Python libraries and frameworks including sklearn, DEAP (Distributed Evolutionary Algorithm in Python) and TPOT (Tree-based Pipeline Optimization Tool) were utilized in preprocessing and building the models. A uniform numerical format in the training set was established. For Dataset 1, the variable of Phone which stores the phone numbers of customers was deleted from the dataset as the values of this variable are unique for each customer, thus, this variable is insignificant in predicting customer churn. In addition to that, variables of State and Area Code were also removed from Dataset 1. For Dataset 2, the column of Total Charges contains 11 null values which were filled up with zeros (0s). For features with three categories of Yes, No, and No internet service/No phone service, the “No internet service/No phone service” category was converted into “No” category and six relevant attributes (Online Security, Online Backup, Device Protection, Tech Support, Streaming TV’, Streaming Movies) were affected. Redundant variables (such as Customer ID) were eliminated from analysis, since it is only description information and irrelevant to customer churn. Variables that had more than two categories were converted into dummy variables to obtain numerical values for data analysis; Payment method, Contract, and Internet Service were affected.

4.2. Critical Factors Related to Customer Churn. The threshold value of feature selection was determined by calculating the average MI of the determining factors and thus 0.3 for Dataset 1 and 0.4 for Dataset 2 were confirmed. After feature selection by MI technique, four factors of Customer service calls, Account length, Day charge, and Day minutes were screened in Dataset 1, as displayed in Figure 5(a); seven factors of Contract month to month, Tenure, Contract two-year, Internet service fibre optic, Total charges, Payment method electronic check, and Monthly charges were filtered in Dataset 2, as shown in Figure 5(b).

Based on the depiction in Figure 5, the critical factors related to customer churn can largely depend on the nature of the datasets. This is consistent with the findings in [36]. However, the four critical factors in Dataset 1 can be categorized into: the length of the service contract, means of charge, and customer service quality, so do the seven critical factors in Dataset 2. If further integrated, they can both be included by a higher-level category of Service Quality, which is also highlighted by the studies [1, 22, 34, 35]. However, there exist some differences. For example, the significance of “Brand Image” highlighted in the study [22] has not been proved in our study.

Those critical factors play a critical role in explaining the predicted results of customer churn in telecom sector and will provide useful guidelines for practical decision makers. For the two telecom companies which are related to Dataset 1 and Dataset 2, the companies can reduce the cost of fibre optic internet connection, persuade customers who use month-to-month and two-year contracts to change to one-year contract, check the problem with an electronic payment method, pay attention to customers who call for help and provide incentives to attract customers with day charges.

4.3. Performance of the Customer Churn Prediction Models. The experimental results of applying the three different HPO methods (GS, RS, GA) to the three ML models (RF, SVM, KNN) are summarized in Table 3 (RF), Table 4 (SVM), and Table 5 (KNN) respectively. Their graphical results on the two datasets are compared and displayed in Figures 6–11. In the first step, each ML model with its default hyperparameter configuration was trained and evaluated as baseline model. After that, each HPO algorithm is implemented on the ML models to evaluate and compare their performance based on different undersampling ratios.

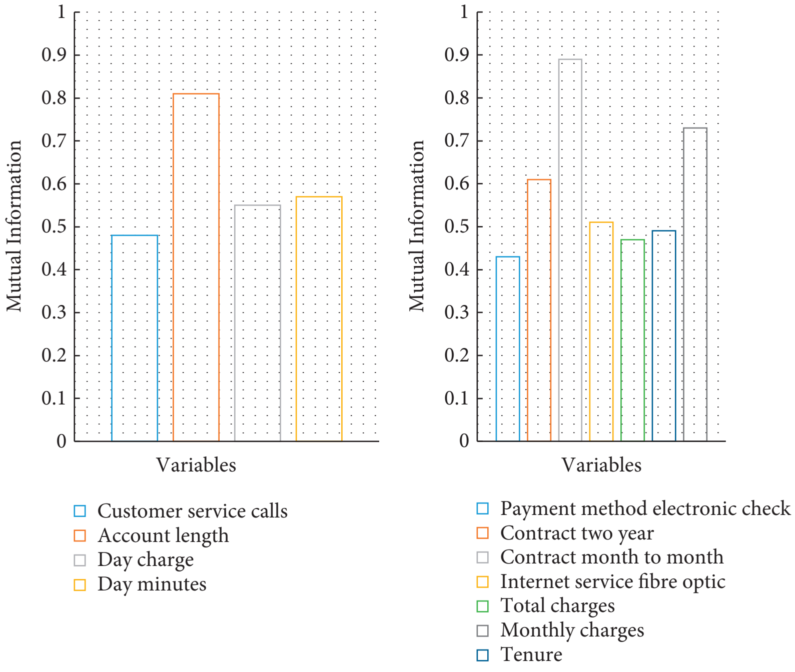
It is observed from Table 3, Figures 6, and 7 that the RF prediction model with default HPs performs worst on both Dataset 1 and Dataset 2. The RF with default HPs (84%) is only slightly better than the RF with RS (83%) and GA-DEAP (82%) in terms of the measure of precision on Dataset 1. For all the other measures, it gets the lowest values. This indicates the default parameters for an RF model are not effective. On both datasets, the RF with GS performs best with superiority of all six measures, and the RF with GA-TPOT is the second best model. More specifically, the RF optimized by GS achieves 99% and 95% accuracy, 100% and 94% recall, 91% and 89% precision, 99% and 95% AUC, 97% and 94% *F*1-score, and 0.014 and 0.049 MAE on Dataset 1 and Dataset 2, respectively. The performance measures of RF with GA-TPOT are much closer to those of RF with GS on Dataset 1 than on Dataset 2. It could be inferred that when the dataset is small (such as Dataset 1), the RF model with GA-TPOT can also be acceptable with 96% accuracy, 87% recall, 84% precision, 91% AUC, 91% *F*1-score, and 0.043 MAE.

Through comparing the results of Table 4, Figures 8, and 9, it is found that the SVM model with default HPs still performs worst on Dataset 1 and Dataset 2. It is surprising that for Dataset 1 there are almost no differences in the performance measures between the SVM models with GS, RS, GA-DEAP, GA-TPOT, and for Dataset 2, the SVM optimized by GS, RS, GA-TPOT also have similar performance. This may be explained by the small number of SVM parameters, and any of the HPO strategies can obtain the parameter settings with similar optimal performance. However, on Dataset 2 with a larger amount of data, the SVM with GA-DEAP is superior to the other optimization approaches (83% accuracy, 63% recall, 72% precision 76% AUC, 78% *F*1-score, and 0.173 MAE). It is shown that as the data grows, the GA-DEAP may have superiority over other HPO strategies for the SVM prediction model. Nevertheless, there is still a large gap between its performance and RF models with different HPs.

By analyzing the data and graphs in Table 5, Figures 10, and 11, it is revealed that on both Dataset 1 and Dataset 2, the KNN model with default HPs performs slightly poorer than the KNN with GS, RS, GA-DEAP, GA-TPOT (except close precision values on Dataset 1). Moreover, when changing the HPs on KNN model, their predicting results are quite adjacent on both Dataset 1 and Dataset 2. Therefore, it is deduced that regardless of big or small datasets, the HPO methods to the KNN classifier can only increase the customer churn prediction performance to a less extent, and an HPO strategy with GS, RS, and GA on KNN classifier has no prominent improvements and differences. The bet performance model of KNN is the prediction model optimized by RS, with 91% accuracy, 59% recall, 77% precision, 78% AUC, 81% *F*1-score, and 0.085 MAE on Dataset 1, which is not competitive with the optimized RF and SVM classifiers.

By comparing Tables 3–5, it is observed that using the default HP configurations does not yield the best model performance. This proves the importance of utilizing HPO methods. As a whole, it is shown that GS is much better than the other two optimization methods of RS and GA. With the same search space size, GS together with RF (RF-GS) model performs best in both datasets, considering all the six performance measures, closely followed by GA together with RF (RF-GA). RS cannot guarantee to detect the near-optimal hyperparameter configurations of ML models, especially for RF and SVM models, which have a larger search space than KNN.

Generally, the performance of RF model is much better than the other two models of SVM and KNN through comparison between Tables 3–5 (also see Figures 12 and 13). This differs in scenarios like handwritten character recognition showing that KNN and SVM are better than RF [21].



(a) (b)

Figure 5: Critical factors related to customer churn. (a) Dataset 1. (b) Dataset 2.

Table 3: Performance evaluation of applying HPO methods to the RF classifier based on low-ratio undersampling strategy (0.3 and 0.4).

Dataset	Optimization algorithm	Accuracy	Recall	Precision	AUC	$F1$-score	MAE
Dataset 1	Default HPs	0.92	0.58	0.84	0.78	0.82	0.078
	GS	**0.99**	**1.00**	**0.91**	**0.99**	**0.97**	**0.014**
	RS	0.94	0.74	0.83	0.85	0.88	0.059
	GA-DEAP	0.93	0.68	0.82	0.82	0.85	0.068
	GA-TPOT	0.96	0.87	0.84	0.91	0.91	0.043
Dataset 2	Default HPs	0.77	0.47	0.61	0.67	0.69	0.233
	GS	**0.95**	**0.94**	**0.89**	**0.95**	**0.94**	**0.049**
	RS	0.83	0.61	0.73	0.76	0.77	0.174
	GA-DEAP	0.83	0.63	0.72	0.76	0.78	0.173
	GA-TPOT	0.83	0.65	0.73	0.77	0.79	0.167

Table 4: Performance evaluation of applying HPO methods to the SVM classifier based on low-ratio undersampling strategy (0.3 and 0.4).

Dataset	Optimization algorithm	Accuracy	Recall	Precision	AUC	$F1$-score	MAE
Dataset 1	Default HPs	0.89	0.36	0.73	0.66	0.71	0.113
	GS	**0.92**	0.65	0.76	0.80	0.83	0.081
	RS	**0.92**	0.67	0.75	0.81	0.83	0.080
	GA-DEAP	**0.92**	0.66	0.74	0.81	0.83	0.083
	GA-TPOT	**0.92**	0.65	0.75	0.80	0.83	0.082
Dataset 2	Default HPs	0.77	0.46	0.64	0.67	0.69	0.226
	GS	0.78	0.54	0.64	0.71	0.72	0.215
	RS	0.78	0.54	0.64	0.71	0.72	0.215
	GA-DEAP	0.83	0.63	0.72	0.76	0.78	0.173
	GA-TPOT	0.77	0.50	0.62	0.68	0.70	0.229

Therefore, it is inferred that ML models have different performances in solving different practical problems. In the area of customer churn predictions, more datasets from diverse industries could be further tested on distinct ML models to explore the best one.

The low-ratio undersampling ratio (0.3 and 0.4) yields better results than high-ratio undersampling (0.6 and 0.7) as displayed in Table 6. The undersampling ratio strategy of 0.3 and 0.6 was applied on Dataset 1. The same strategy was supposed to be applied on Dataset 2. But since the

Table 5: Performance evaluation of applying HPO methods to the KNN classifier based on low-ratio undersampling strategy (0.3 and 0.4).

Dataset	Optimization algorithm	Accuracy	Recall	Precision	AUC	*F*1-score	MAE
Dataset 1	Default HPs	0.89	0.45	0.71	0.70	0.74	0.107
	GS	0.90	0.57	0.70	0.76	0.79	0.098
	RS	**0.91**	0.59	0.77	0.78	0.81	0.085
	GA-DEAP	**0.91**	0.58	0.73	0.77	0.80	0.092
	GA-TPOT	0.90	0.63	0.69	0.79	0.80	0.091
Dataset 2	Default HPs	0.75	0.46	0.56	0.66	0.67	0.253
	GS	0.80	0.60	0.66	0.73	0.74	0.202
	RS	0.80	0.64	0.65	0.75	0.75	0.199
	GA-DEAP	0.80	0.57	0.67	0.73	0.74	0.199
	GA-TPOT	0.80	0.57	0.67	0.73	0.74	0.199

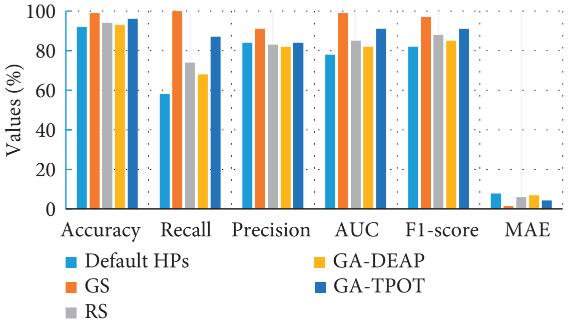


Figure 6: Result analysis of HPO methods on RF using dataset 1.

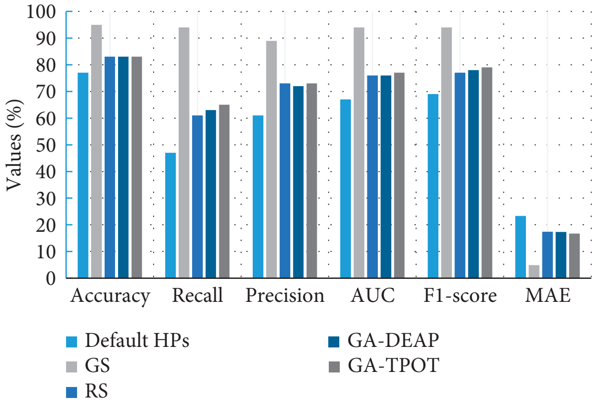


Figure 7: Result analysis of HPO methods on RF using dataset 2.

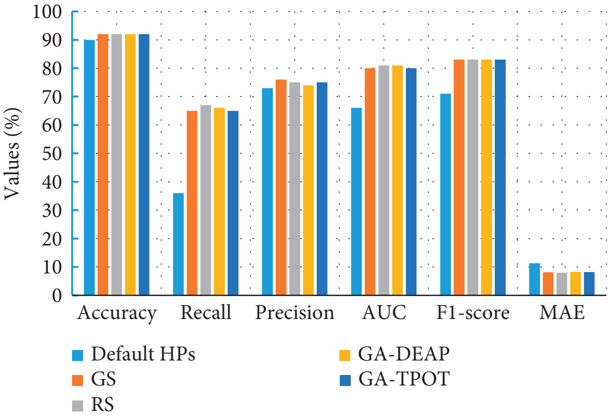


Figure 8: Result analysis of HPO methods on SVM using dataset 1.

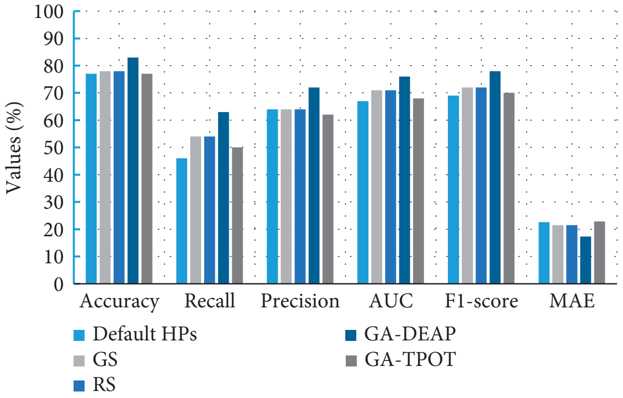


Figure 9: Result analysis of HPO methods on SVM using dataset 2.

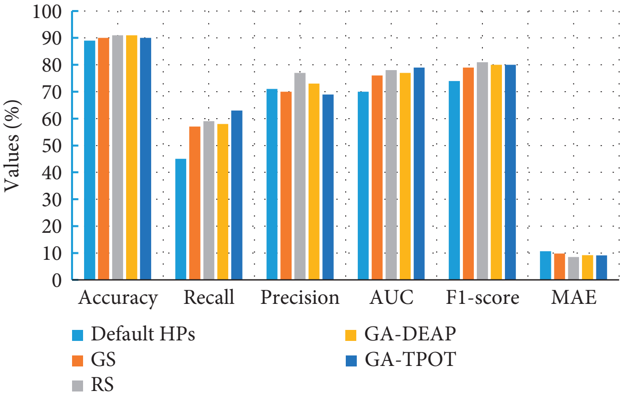


Figure 10: Result analysis of HPO methods on KNN using dataset 1.

percentage of negative samples in the total sample size is different between Dataset 1 and Dataset 2, the threshold for undersample ratio cannot be the same for the two datasets. The threshold for undersample ratio is supposed to be a little more than the target class. For example, Dataset 2 has got target class of 0.27 (27%) of the total sample size. It is impossible to set a threshold that is below or closer to 0.27 for this dataset. In this case, 0.4 suits this situation. Dataset 1 has got 0.145 (14.5%) churn, and the threshold ratio could be set as 0.2. But to keep a small range of threshold ratios between datasets, we selected 0.3 for Dataset 1 and 0.4 for Dataset 2.

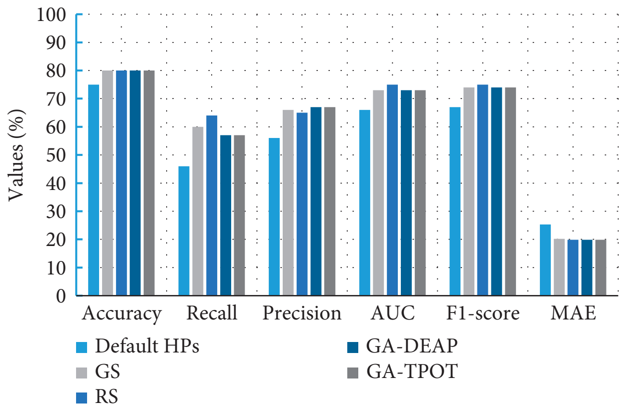


Figure 11: Result analysis of HPO methods on KNN using dataset 2.

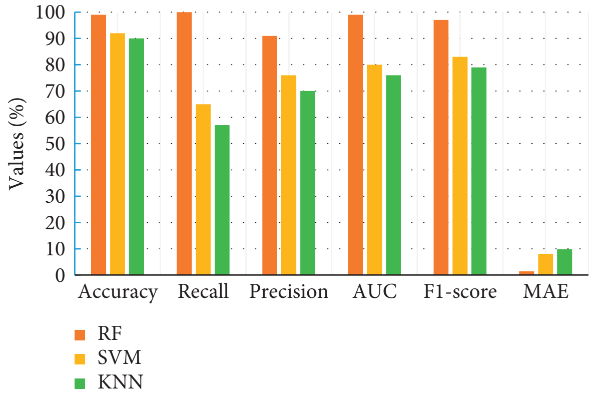


Figure 12: Result analysis of ML models on dataset 1 based on GS.

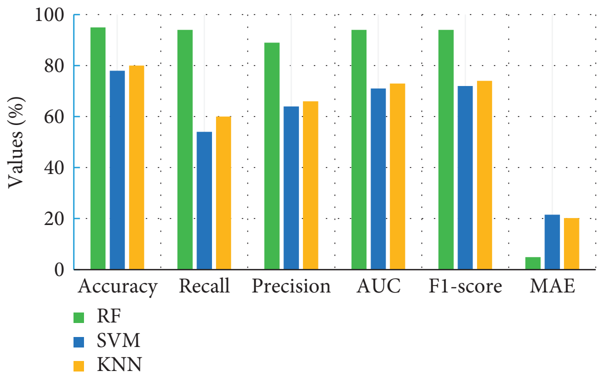


Figure 13: Result analysis of ML models on dataset 2 based on GS.

It is important to know that low-ratio undersampling secures important information for training. Moreover, after training, both minority and majority classes are favored with high-performance results. Even when the majority class is much more than the minority class after low-ratio undersampling, the performance of RF still stands out. Therefore, RF, if fed with well-processed datasets, is an extraordinary model for highly imbalanced datasets such as the scenario of detecting the churn-risk customers in the telecom sector.

The proposed RF-GS model with a low-ratio undersampling strategy (0.3) (RF-GS-LR) based on the Mutual Information feature selection technique has been able to capture all churn customers with 99% accuracy, 100% recall, 91% precision, 99% AUC, 97% $F1$-score and 0.014 MAE on Dataset 1, which is better than the performance of the proposed RF-GS model with high-ratio undersampling strategy (0.6) (RF-GS-HR) with 96% accuracy, 96% recall, 79% precision, 95% AUC, 92% $F1$-score, and 0.043 MAE.

Although the recall value of the RF-GS-HR model with a high-ratio undersampling strategy (0.7) is slightly higher than that of the RF-GS-LR model with a low-ratio undersampling strategy (0.4), with 96% and 93% respectively, its other five performance measures are much lower than those of the RF-GS-LR model with low-ratio undersampling strategy (0.4). Thus, considering the overall performance, the proposed RF-GS-LR model with a low-ratio undersampling strategy (0.4) still performs best on Dataset 2, with 95% accuracy, 93% recall, 95% AUC, 94% F1-score, 90% precision, and 0.048 MAE, a little lower than that of Dataset 1. This could be attributed to the differences in the nature (dimensionality, velocity, and veracity) of the datasets. However, the performance of the RF-GS-LR model on the larger Dataset 2 are still better than the performance of other ML models in predicting customer churn as stated in studies of Vijaya and Sivasankar [7], Pamina et al. [14], Ahmad et al. [15], Li and Marikannan [19] (see Table 7), which shows its superiority.

4.4. *Validity of Feature Selection.*

The increase in dimensionality and complexity of predictor variables may result in a decrease in classification accuracy, which can be ascribed to a deficiently number of training data to characterize the expanded complexity related to the bigger dimensionality of the feature space. High classification accuracies have been reported when using machine learning with no feature reduction such as Random Forest and SVM, despite the drawbacks of high dimension [63]. Even if the incorporation of enormous variables does not reduce the classification accuracy, a limited number of predictor variables may simplify the model for reproducibility and speed [64]. A great example is given by research [65], where a reduction in model complexity with little loss in the overall classification accuracy was recorded, with a reduction of 60% features for mapping land cover with RF. In addition, reducing the number of features in a model can address the multicollinearity problem and the curse of dimensionality, remove redundancy, and avoid overfitting of the models, which provides more advantages than training high-dimensional data [66]. Therefore, feature selection is an important step to take when building customer churn prediction models.

To objectively assess the influence of feature selection on classification performance, we compared the results of the RF-GS-LR model with feature selection by MI and without feature selection, as shown in Table 8.

According to Table 8, although the RF-GS-LR model performs slightly better without feature selection in other performance measures, the value of Recall is less than the RF-GS-LR model with feature selection by MI. In other

Table 6: Performance of RF on different undersampling strategies.

Dataset	Optimization algorithm	Accuracy	Recall	Precision	AUC	$F1$-score	MAE
Dataset 1			0.3 Undersampling				
	GS	**0.99**	**1.00**	**0.91**	**0.99**	**0.97**	**0.014**
	RS	0.94	0.74	0.83	0.85	0.88	0.059
	GA-DEAP	0.93	0.68	0.82	0.82	0.85	0.068
	GA-TPOT	0.96	0.87	0.84	0.91	0.91	0.043
			0.6 Undersampling				
	GS	0.96	0.96	0.79	0.95	0.92	0.043
	RS	0.92	0.89	0.67	0.90	0.86	0.079
	GA-DEAP	0.93	0.88	0.71	0.91	0.87	0.074
	GA-TPOT	0.92	0.88	0.67	0.90	0.85	0.082
Dataset 2			0.4 Undersampling				
	GS	**0.95**	0.93	**0.90**	**0.95**	**0.94**	**0.048**
	RS	0.83	0.61	0.73	0.76	0.77	0.174
	GA-DEAP	0.83	0.63	0.72	0.76	0.78	0.173
	GA-TPOT	0.83	0.65	0.73	0.77	0.79	0.167
			0.7 Undersampling				
	GS	0.89	**0.96**	0.73	0.91	0.88	0.108
	RS	0.78	0.73	0.59	0.76	0.75	0.218
	GA-DEAP	0.80	0.78	0.62	0.79	0.77	0.196
	GA-DEAP	0.80	0.78	0.62	0.79	0.77	0.196
	GA-TPOT	0.80	0.75	0.62	0.78	0.77	0.201

Table 7: Performance comparison with existing studies.

Methods		Accuracy (%)	AUC (%)
RSFS-boost [7]		95.1	None
XGBOOST [14]		79.8	None
XGBOOST [15]		None	93.3
Grid search ANN [19]		86.8	None
RF-GS-LR	Dataset 1	**99.0**	**99.1**
	Dataset 2	**95.2**	**94.6**

words, feature selection by MI has improved the chance of picking out positive churn customers, which is the most important goal and the reason why customer churn prediction is carried out. Therefore, under the scenario of customer churn prediction for telecom sector in this study, although with a selected subset of features filtered by MI, the accuracy and efficiency of the proposed ML prediction models have still increased and been satisfactory.

5. Implications

5.1. Theoretical Implications. First of all, we contribute to the literature on customer churn prediction in the telecom sector by improving the predicting accuracy through optimizing ML methods with HPO techniques. We are among the first to attempt to examine the key HPO techniques of GS, RS, GA in improving the prediction performance of Random Forest, Support Vector Machines, and K-nearest neighbors, and thus to reveal that Random Forest optimized with GS performs superior on both small and bigger telecom datasets.

Second, we emphasize the influence of the class imbalance problem on the predicting results of customer churn in the telecom sector. Most of the existing literature on customer churn prediction has focused on feature selection techniques, neglecting the class imbalance of datasets. The present study has tried low and high controlled-ratio undersampling strategies on two different datasets and thus has brought up a significant improvement in the overall performance of the proposed customer churn predicting models.

Third, we enrich the literature of feature section techniques by employing MI classifier to identify critical factors related to customer churn in the telecom sector, rather than PCA and other widely used feature selection techniques in customer churn prediction, to better interpret the dominant determinants of customer churn for practitioners.

5.2. Practical Implications. Our study can facilitate the construction of an effective prediction tool for managers in the telecom sector to discover the underlying churn-risk customers which have a high probability of transferring to their competitors. Under such circumstances, the managers can take favorable measures to re-attract those underlying churners by promoting the dominant determinants revealed by this study through MI feature selection. Specifically, the decision makers can emphasize improving of the length of the service contract, means of charge, and customer service quality based on customer preferences, such as reducing the cost of fibre optic Internet connection, persuading customers who use month-to-month and two-year contracts to change into a one-year contract, checking the problem with the electronic payment method, paying attention to customers who call for help and providing incentives to attract customers with day charges. Thus, our study can assist managers in the telecom sector to build a good relationship with their customers and maintain a competitive advantage.

Table 8: Comparison of RF-GS-LR with feature selection by MI (w/FS) and without feature selection (wo/FS).

Datasets	GS	Accuracy	Recall	Precision	AUC	$F1$-score	MAE
Dataset 1	w/FS	0.99	**1.00**	0.91	0.99	0.97	0.014
	wo/FS	0.99	0.94	1.00	0.97	0.98	0.008
Dataset 2	w/FS	0.95	**0.94**	0.89	0.95	0.94	0.049
	wo/FS	0.96	0.92	0.93	0.95	0.95	0.046

6. Conclusions and Future Research

Optimized ML procedures can be successful in extracting hidden information and reveal customer's information, which could be quite useful in assisting decision makers. To apply ML models for practical problems of detecting the churn-risk customers in the telecom sector, their hyperparameters need to be tuned to fit specific datasets. In this study, RF optimized by GS HPO techniques has shown its superior abilities in predicting customer churn in telecom sectors, regardless of being tested on small or bigger datasets. RF is an extraordinary model to be optimized for solving class imbalance problems. Low-ratio undersampling works better with an RF classifier.

The most critical factors captured by this study are: "Day charges," "Account length," "Customer service calls," "Day minutes" (Dataset 1), "Contract month to month," "Tenure," "Contract two year," "Internet connection fibre optic," "Monthly charges," "Total charges" and "Payment with electronic check" (Dataset 2). This gives insights to telecom company directors to understand areas that require customer retention campaigns. The telecom managers can prevent churn by considering some strategies like providing the facilities required, improving the quality of services and products, identifying the different needs of the consumers, and increasing customer responsiveness.

Future research could examine the RF-GS-LR model or other ML models for customer churn prediction on larger telecom datasets and other industries. Because the nature of the datasets will affect the performance of prediction models, a variety of telecom customer datasets from different countries with different volumes will be examined and compared. At the same time, more advanced optimization techniques such as different methods for data preprocessing, feature selection, and class balancing will be encouraged to test their adaptability and progressiveness in predicting customer churn in the telecom sector. In addition, researchers can design and conduct experiments with more novel and advanced ensemble and hybrid ML methods to achieve a higher prediction performance of customer churn in the telecom sector or other industries. More transparent and interpretable intelligent optimization algorithms and models can be experimented with to assist decision makers to better understand their customers' needs and maintain a stronger customer relationship. It could also be a future research direction that combines ML methods with social science research methods to better serve the decision makers and practitioners in the telecom sector.

Data Availability

The data used to support the findings of this study are available from the corresponding authors upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Program for the Capacity Development of Shanghai Local Colleges (Grant no. 20050501900).

References

[1] H. Jain, A. Khunteta, and S. Srivastava, "Telecom churn prediction and used techniques, datasets and performance measures: a review," *Telecommunication Systems*, vol. 76, no. 4, pp. 613–630, 2020.

[2] M. Al-Mashraie, S. H. Chung, and H. W. Jeon, "Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: a machine learning approach," *Computers & Industrial Engineering*, vol. 144, Article ID 106476, 2020.

[3] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 1, no. No. 10, pp. 2–10, 2016.

[4] I. V. Pustokhina, D. A. Pustokhin, R. Aswathy et al., "Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms," *Information Processing & Management*, vol. 58, no. 6, Article ID 102706, 2021.

[5] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 2, pp. 165–172, 2013.

[6] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[7] J. Vijaya and E. Sivasankar, "Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector," *Computing*, vol. 100, no. 8, pp. 839–860, 2018.

[8] A. R. Safitri and M. A. Muslim, "Improved accuracy of naive bayes classifier for determination of customer churn uses smote and genetic algorithms," *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 70–75, 2020.

[9] R. Yahaya, O. A. Abisoye, and S. A. Bashir, "An enhanced bank customers churn prediction model using a hybrid genetic algorithm and k-means filter and artificial neural network," in *Proceedings of the Paper presented at the IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*, Abuja, Nigeria, February 2021.

[10] A. Moubayed, M. Injadat, and A. Shami, “Optimized random forest model for botnet detection based on DNS queries,” in *Proceedings of the Paper presented at the 32nd International Conference on Microelectronics (ICM)*, Aqaba, Jordan, December 2020.

[11] A. Zakariazadeh, “Smart Meter Data Classification Using Optimized Random forest Algorithm,” *ISA Transactions*, vol. S0019-0578, pp. 1–9, 2021.

[12] T. Zhang, J. Su, Z. Xu, Y. Luo, and J. Li, “Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier,” *Applied Sciences*, vol. 11, no. 543, pp. 1–17, 2021.

[13] X. Gao, J. Wen, and C. Zhang, “An improved random forest algorithm for predicting employee turnover,” *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 2019.

[14] J. Pamina, B. Raja, S. SathyaBama, M. Sruthi, and A. Vj, “An effective classifier for predicting churn in telecommunication,” *Jour of Adv Research in Dynamical and Control Systems*, vol. 11, pp. 221–229, 2019.

[15] 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. 28, pp. 6–28, 2019.

[16] V. Kavitha, G. H. Kumar, S. Kumar, and M. Harish, “Churn prediction of customer in telecom industry using machine learning algorithms,” *International Journal of Engineering Research and Technology*, vol. 9, no. 5, pp. 181–184, 2020.

[17] G. Jiao and H. Xu, “Analysis and comparison of forecasting algorithms for telecom customer churn,” *Journal of Physics: Conference Series*, vol. 1881, no. 3, Article ID 032061, 2021.

[18] Q. Zhu, X. Yu, Y. Zhao, and D. Li, “Customer churn prediction based on LASSO and random forest models,” *IOP conference series materials science and engineering*, vol. 631, no. 5, Article ID 052008, 2019.

[19] K. G. Li and B. P. Marikannan, “Hyperparameters tuning and model comparison for telecommunication customer churn predictive models,” in *Proceedings of the Paper presented at the 3rd Global Conference on Computing and Media Technology*, Kuala Lumpur, Malaysia, July 2020.

[20] T. Wang, X. Wang, R. Ma, X. Li, and J. Ruan, “Random forest-bayesian optimization for product quality prediction with large-scale dimensions in process industrial cyber–physical systems,” *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8641–8653, 2020.

[21] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[22] V. Mahajan, R. Misra, and R. Mahajan, “Review on factors affecting customer churn in telecom sector,” *International Journal of Data Analysis Techniques and Strategies*, vol. 9, no. 2, pp. 122–144, 2017.

[23] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, “A mutual information-based feature selection method,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[24] I. AlShourbaji, N. Helian, Y. Sun, and M. Alhameed, “Anovel HEOMGA approach for class imbalance problem in the application of customer churn prediction,” *SN Computer Science*, vol. 2, no. 6, pp. 1–12, 2021.

[25] H. Jain, A. Khunteta, and S. Srivastava, “Churn prediction in telecommunication using logistic regression and logit boost,” *Procedia Computer Science*, vol. 167, pp. 101–112, 2020.

[26] S. Wang, L. L. Minku, and X. Yao, “Online class imbalance learning and its applications in fault detection,” *International Journal of Computational Intelligence and Applications*, vol. 12, no. 4, Article ID 1340001, 2013.

[27] F. Zhao and Z. Yao, “Predicting the voluntary donation to online content creators,” *Industrial Management & Data Systems*, vol. 120, no. 10, pp. 1941–1957, 2020.

[28] A. Onan, S. Korukoğlu, and H. Bulut, “Ensemble of keyword extraction methods and classifiers in text classification,” *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[29] A. Onan, “An ensemble scheme based on language function analysis and feature engineering for text genre classification,” *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[30] A. Onan, “Hybrid supervised clustering based ensemble scheme for text classification,” *Kybernetes*, vol. 46, no. 2, pp. 330–348, 2017.

[31] A. Onan, “Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks,” *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, Article ID e5909, 2021.

[32] A. Onan, “Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering,” *IEEE Access*, vol. 7, Article ID 145614, 2019.

[33] A. Onan and M. A. Tocoglu, “A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification,” *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[34] N. Jacob and K. Christian, “Determinants of customer loyalty and subscriber churn of mobile phone services in Ghana,” *International Journal of Research in Commerce, IT, Management*, vol. 2, no. 12, pp. 38–41, 2012.

[35] P. Rajeswari and P. Ravilochanan, “Churn analytics on Indian prepaid mobile services,” *Asian Social Science*, vol. 10, no. 13, pp. 169–183, 2014.

[36] V. Umayaparvathi and K. Iyakutti, “A survey on customer churn prediction in telecom industry: datasets, methods and metrics,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 4, pp. 2395–0072, 2016.

[37] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, “A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA,” *Journal of Big Data*, vol. 7, no. 9, pp. 2–23, 2020.

[38] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[39] E. Domingos, B. Ojeme, and O. Daramola, “Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector,” *Computation*, vol. 9, no. 34, pp. 1–19, 2021.

[40] M. Rahman and V. Kumar, “Machine learning based customer churn prediction In banking,” in *Proceedings of the Paper presented at the 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, November 2020.

[41] C. C. Günther, I. F. Tvete, K. Aas, G. I. Sandnes, and R. Borgan, “Modelling and predicting customer churn from an insurance company,” *Scandinavian Actuarial Journal*, vol. 2014, no. 1, pp. 58–71, 2014.

[42] A. Sharma and P. K. Panigrahi, “A neural network-based approach for predicting customer churn in cellular network services,” *International Journal of Computer Application*, vol. 27, no. 11, pp. 26–31, 2013.

[43] E. Shaaban, Y. Helmy, A. Khedr, and M. Nasr, “A proposed churn prediction model,” *International Journal of Engineering Research in Africa*, vol. 2, no. 4, pp. 693–697, 2012.

[44] A. A. Jamjoom, “The use of knowledge extraction in predicting customer churn in B2B,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–14, 2021.

[45] A. Amin, S. Anwar, A. Adnan et al., "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, 2017.

[46] A. A. Ahmed and D. Maheswari, "An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling," *International Journal of Information Technology*, vol. 11, no. 2, pp. 381–391, 2019.

[47] M. Azeem, M. Usman, and A. C. M. Fong, "A churn prediction model for prepaid customers in telecom using fuzzy classifiers," *Telecommunication Systems*, vol. 66, no. 4, pp. 603–614, 2017.

[48] E. Zdravevski, P. Lameski, A. Dimitrievski, M. Grzegorowski, and C. Apanowicz, "Cluster-size optimization within a cl1oud-based ETL framework for Big Data," in *Proceedings of the Paper presented at the 2019 IEEE International Conference on Big Data*, Los Angeles, USA, December 2019.

[49] P. Ramesh, J. J. Emilyn, and V. Vijayakumar, "Hybrid artificial neural networks using customer churn prediction," *Wireless Personal Communications*, vol. 124, no. 2, pp. 1695–1709, 2022.

[50] D. D. Adhikary and D. Gupta, "Applying over 100 classifiers for churn prediction in telecom companies," *Multimedia Tools and Applications*, vol. 80, no. 28, Article ID 35123, 2021.

[51] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, no. No.1, pp. 25–38, 2017.

[52] N. Barraza, S. Moro, M. Ferreyra, and A. de la Peña, "Mutual information and sensitivity analysis for feature selection in customer targeting: a comparative study," *Journal of Information Science*, vol. 45, no. 1, pp. 53–67, 2019.

[53] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, "Feature selection via mutual information: new theoretical insights," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, IEEE, Budapest, Hungary, 2019.

[54] A. A. Toor and M. Usman, "Adaptive telecom churn prediction for concept-sensitive imbalance data streams," *The Journal of Supercomputing*, vol. 78, no. 3, pp. 3746–3774, 2022.

[55] A. Onan, "Consensus clustering-based undersampling approach to imbalanced learning," *Scientific Programming*, vol. 2019, Article ID 5901087, 2019.

[56] T. Komamizu, Y. Ogawa, and K. Toyama, "An ensemble framework of multi-ratio Undersampling-based imbalanced classification," *Journal of data intelligence*, vol. 2, no. 1, pp. 030–046, 2021.

[57] D. Gupta, P. Borah, U. M. Sharma, and M. Prasad, "Data-driven Mechanism Based on Fuzzy Lagrangian Twin Parametric-Margin Support Vector Machine for Biomedical Data Analysis," *Neural Computing And Applications*, vol. 33, pp. 1–11, 2021.

[58] K. Shung, "Accuracy, precision, recall or F1 score," 2018, https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

[59] P. Borah and D. Gupta, "Affinity and Transformed Class Probability-Based Fuzzy Least Squares Support Vector Machines," *Fuzzy Sets And Systems*, vol. 443, pp. 1–33, 2022.

[60] J. S. Lee, "AUC4.5: AUC-based C4.5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, Article ID 106034, 2019.

[61] Guo, "DataFountain Telecom customer churn," 2019, https://www.datafountain.cn/datasets/60.

[62] J. Seshapanpu, "Telecom data, kaggle datasets," 2019, https://www.kaggle.com/spscientist/telecom-data.

[63] G. P. Petropoulos, C. Kalaitzidis, and K. P. Prasad Vadrevu, "Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery," *Computers & Geosciences*, vol. 41, pp. 99–107, 2012.

[64] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: an applied review," *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784–2817, 2018.

[65] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sensing of Environment*, vol. 118, pp. 259–272, 2012.

[66] Y. Wang, B. Lei, A. Elazab et al., "Breast cancer image classification via multi-network features and dual-network orthogonal low-rank learning," *IEEE Access*, vol. 8, Article ID 27779, 2020.