University of Hertfordshire

Developing optimization models for customer churn prediction in telecoms using artificial intelligence techniques

Ibrahim AlShourbaji

A thesis submitted to the University of Hertfordshire in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

June 2024

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Na Helian, for her unwavering support and guidance throughout my PhD journey. Her extensive knowledge, expertise, and experience have been instrumental in shaping my research. I am also sincerely thankful to my second supervisors, Dr. Yi Sun and Dr. Stilianos Vidalis, for their valuable contributions to my work.

A special appreciation goes to my father, Hasan AlShourbaji, for his wisdom, motivation, and constant support. I am profoundly grateful to my loving and kind mother, Fatimah AlShourbaji, and to my sister, my irreplaceable treasure, Khitam AlShourbaji, for her unwavering love, patience, and kindness during the most challenging times. I also extend my heartfelt thanks to my beloved daughters, Joud and Alma, who are the light of my life.

I am deeply thankful to my colleagues and friends for their continuous support throughout this PhD journey. Lastly, I would like to express my gratitude to all the students I have encountered, whose contributions and interactions have enriched my knowledge.

Abstract

Customer churn prediction is a critical task in the telecommunication (telecoms) industry, where accurate identification of customers at risk of churning plays a vital role in reducing customer attrition. This research presents a comprehensive study on customer churn prediction using Machine Learning (ML) techniques. Three distinct aspects of churn prediction are investigated: class imbalance handling, feature selection, and model enhancement, each utilizing different AI methodologies. First, to address the issue of highly skewed datasets, we propose an effective oversampling method called HEOMGA. HEOMGA combines the Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) to oversample the minority class. Experimental results on six benchmark datasets from the UCI repository demonstrate the superiority of HEOMGA over popular oversampling methods such as SMOTE, ADASYN, G SMOTE, and Gaussian oversampling, as evaluated by three performance metrics: Recall, G mean, and AUC. The experiment results show the effectiveness of the proposed method compared to some popular oversample methods, such as SMOTE, ADASYN, G SMOTE, and Gaussian oversampling methods. The HEOMGA method significantly outperformed the other oversampling methods in terms of recall, G mean, and AUC when the Wilcoxon signed-rank test is used.

Second, in the preprocessing phase, feature selection plays a crucial role in improving the performance of ML models while reducing computational time. To address this, we introduce an ACO-RSA based-FS approach that combines two metaheuristic algorithms: Ant Colony Optimization (ACO) and Reptile Search Algorithm (RSA). The ACO-RSA approach selects the most salient features for churn prediction. The performance evaluations on six open-source customer churn prediction datasets demonstrate the superiority of ACO-RSA over other competitor algorithms such as PSO, MVO, GWO, standard ACO, and standard RSA.

Third, we focus on enhancing Gradient Boosting Machine's (GBM) learning process for Churn Prediction (CP). In traditional GBM, learning process uses Decision Tree (DT) as a base learner and logistic loss as a loss function. However, using a DT to start the GBM model in the training process could result in poor predictive performance and overfitting. Therefore, a new model, called CP- Enhanced Gradient Boosting Model (CP- EGBM) is proposed. In the CP- EGBM, Support Vector Machine with a Radial Basis Function kernel ($\mathrm{SVM}_{\mathrm{RBF}}$) is employed as a base learner and exponential loss function is utilized as a loss function to enhance learning process of the GBM. In order to effectively tune the hyperparameters of CP-EGBM, Finally, a modified version of Particle Swarm Optimization (PSO) using the consumption operator of the Artificial

Ecosystem Optimization (AEO) method to prevent premature convergence of the PSO in the local optima is developed to tune the hyper-parameters of the CP-EGBM effectively. Six open-source CP datasets are used to evaluate the performance of the developed CP-EGBM model. Comparative analysis reveals the significant superiority of the CP-EGBM over GBM and SVM models, along with promising improvements compared to the recently reported models in the literature. Comparative analysis with state-of-the-art models showcases CP-EGBM's promising improvements, making it a robust and effective solution for churn prediction in the telecom industry.

This research contributes to customer CP by providing effective solutions to address class imbalance, feature selection, and model enhancement challenges. The proposed methods, HEOMGA, ACO-RSA, and CP-EGBM, demonstrate their efficacy in improving CP performance, thereby assisting the telecom industry in understanding customer needs and taking appropriate actions to mitigate churn risks.

Contents

List of Figures

List of Tables

List of abbreviations

ACO	Ant Colony Optimization
ACO-RSA	Ant Colony Optimization- Reptile Search Algorithm
ADASYN	Adaptive Synthetic Sampling Method
AEO	Artificial Ecosystem Optimization
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
CP	Churn Prediction
CP-EGBM	Enhanced Gradient Boosting Machine for Churn Prediction
CRM	Customer Relationship Management
CRM	Customer Relationship Management
CV	Cross Validation
DL	Deep Learning
DT	Decision Tree
EGBM	Enhanced Gradient Boosting Machine
FN	False Negative
FP	False Positive
FS	Feature Selection
GA	Genetic Algorithm
GANs	Generative Adversarial Networks
GBM	Gradient Boosting Machine
GBT	Gradient Boosted Tree
GWO	Gray Wolf Optimizer
HEOM	Heterogeneous Euclidean-Overlap Metric ()
ICT	Information and Communication Technology
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
MA	Metaheuristic Algorithm
MH	MetaHeuristic
ML	Machine Learning
mPSO	Modified Particle Swarm Optimization
MVO	Multi-Verse Optimizer
NB	Naive Bayes
OFS	Optimum Feature Subsets
PSO	Particle Swarm Optimization
RBF	Radial Basis functions

RF	Random Forest
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
$\mathrm{SVM}_{\mathrm{RBF}}$	Support Vector Machine with RBF kernel
TN	True Negative
TP	True Positive
WOA	Whale Optimization Algorithm
XAI	Explainable AI
XGBoost	Extreme Gradient Boosting
XGBoost	Extreme Gradient Boosting

Chapter 1 Introduction

1.1 Introduction

The telecom industry is evolving rapidly over time, and they are facing severe revenue losses because of churning with the growing competition in the telecom market. Many telecom companies are discovering the cause of losing clients by enumerating their loyalty. Here, the word “churn” refers to a customer who tends to leave a company and move to another competitor for some reason [1].

According to several reviews [2, 3, 4, 5, 6], customer churn is divided into Voluntary Churners, Non-voluntary Churners, and Silent Churners. Voluntary churners are those customers who want to quit the contract and move to another service provider. In such a case, the situation can be considered in more detail and is centered on numerous churn motives, such as technology change, regulation, contract expiration, handset change, service quality, competition, etc. Non-voluntary churners are customers who have their service retracted by their service provider. For instance, the service provider can choose to revoke the service with the customer for several reasons, such as abuse of service and not disbursing the bill. Silent churners are those customers who discontinue the contract without any prior knowledge or notifications from either the company or the customer.

In the telecom sector, customers seek good quality service, competitive pricing, and value for money. However, when they do not find what they are expecting, they can easily terminate using the service and switch from one service provider to another without restrictions [7]. This has led telecom companies to offer customers great inducements to encourage them to shift to their services [8].

The phenomenon related to customer abandonment is called customer churn, while the process of calculating the probability of future churning behaviors in the database based on the past prior behavior using a predictive model is usually called customer churn prediction (CCP) [9]. Customer churn is typically calculated as a relative number in percentage (i.e., the churn rate), and in the telecom sector, the churn rates are often reported monthly, which might be misleading since some customers can leave the service at any time, whereas others are locked in more extended contracts that are not due to renew monthly [10].

Churn rate estimation should not be considered for the new customers acquired in the same period. On the other hand, losses of new customers in this period may be considered or not, and it depends on whether the churn rate should be counted for all the losses during the period. When the losses of new customers are included, the churn rate measures the number of customers who have left the company. Otherwise, it measures how many initial customers have left the company. Several ways can be used to compute churn rate and it is usually expressed as follows:

- Fix a conventional period as a month or a year
- Count the number of customers lost in this period
- Split this quantity by the number of customers that the company have at the beginning of this period.

It has been recognized that long-standing consumers are more lucrative in the long term, as new clients are engrossed by persuasive offers and inclined to switch to an alternative competitor in the market at the moment they obtain a better concession. Therefore, companies need to consider churn management as a part of CRM to protect themselves from competitors [11, 12]. Customer satisfaction can be considered one of the main aims of CRM, which plays a role in the success of the telecom market [13]. Companies use CRM to modify their process management to improve their revenues and find new approaches by primarily focusing on customers' needs to avoid losing them rather than a product [14, 15]. These specifics have led competitive companies to capitalize on CRM to up-hold their customers, thus helping to increase customer strength. Figure 1.1 shows the main sections of CRM [16].

- *Collaborative CRM*: It aims to establish customized customer relationships using several ways such as emails, telephone, websites, call centers, face-to-face contact, etc.
- *Operative CRM*: This type offers services for organizations to increase the efficiency of CRM processes.
- *Analytical CRM*: focuses on data collection and analysis to help the management build strategic decisions and plan for the future.

The data of customers are stored in such CRM systems, which can then be transformed into valuable information with the help of ML techniques, which aid telecom companies in formulating new policies, develop campaigns for existing clients and figure out the main reasons behind customer churn. In this way, companies can easily observe their customer's behavior from time to time and manage them effectively. Therefore, ML approaches are needed in telecom sectors which remain the cornerstone of customer churn control and can play a fundamental role in decreasing the probability of churners.

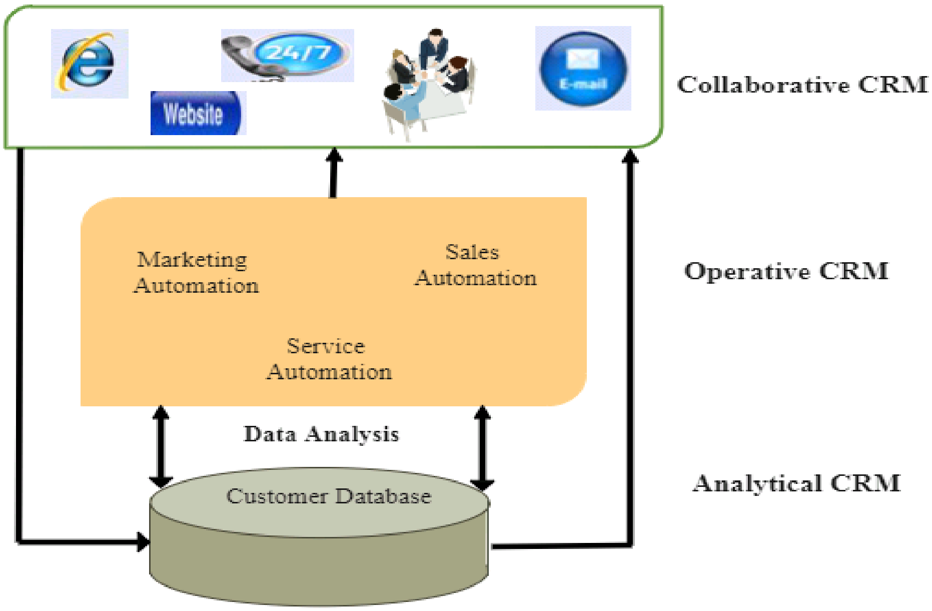


Figure 1.1: CRM areas

1.2 The Importance of CCP in Telecom

ICT has grown and developed rapidly during the last decades, specifically in the mobile industry, which represents the most significant ICT market due to the appearance of the internet and the commercial success in the mobile communication market. Before 1999, the Internet was regarded as a fancy tool that only professionals, computer–savvy users, and “nerds” could play with [17]. During that time, Internet users were less than 5% of the worldwide population. Recently, ITU estimated the number of internet subscribers at the end of 2017 to reach 3.4 billion, over 45% of the young population. By the end of 2023, the number of mobile phone users is expected to reach 5.4 billion, over 67 % of the population globally [18]., as shown in Figure 1.2. However, the ICT market, particularly the telecom industry, has reached market saturation, and the average annual churn rate reaches between 10 - 67% monthly due to the intense competition between service providers to attract new customers [19].

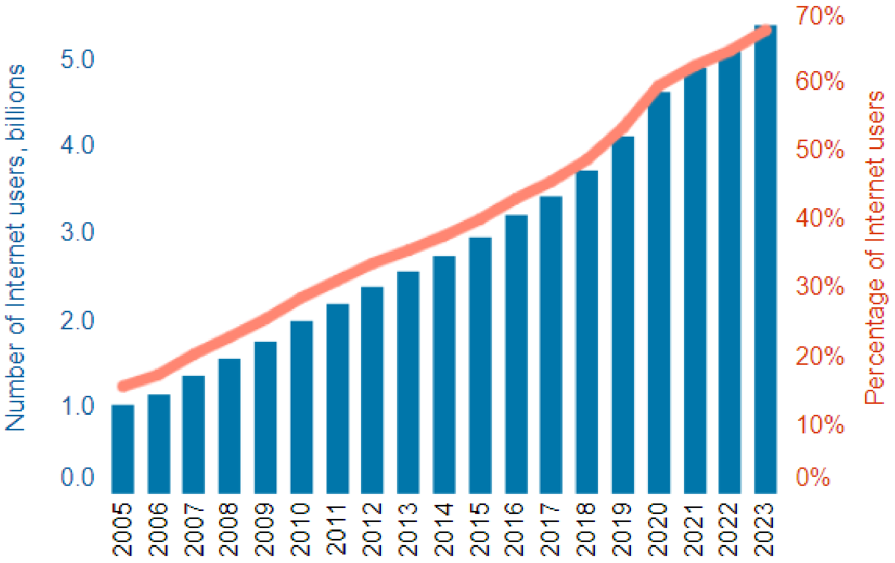


Figure 1.2: Subscribers using the internet 2005- 2023

Commercial companies in general and telecom companies in particular are considered one of the top sectors on the list of industries that suffer from customer churning [20] with annual churn rates ranging from 20% to 46% [21]. This means a company could lose almost half of its customers, resulting in a profit drop. AT&T, a well-known wireless carrier, reported that the total churn rate was 1.01%, consistent in 2023 with the previous year and 0.89% in first quarter and 0.95% in the second quarter in 2024 [22]. Even Twitter considers the effects of churn on their services, and it was reported that 60% of their accounts became idle within one month [23]. Furthermore, several research papers acknowledged the existence of the problem of churn from Telecom companies in different countries such as Nigeria [24], India [25], Kenya [26], Indonesia [27], and Ghana [28].

Churn management is an essential concept in CRM; it manages the most relevant aspects that may change the customers' behavior, such as price, service quality, company reputation, and effective advertising competition. The primary way to reduce churn is to offer retention incentives [29]. Telecom companies use different strategies for churn management and retention: offering all clients incentives without determining which customers to target [30] or utilizing the customers’ transactional data to develop predictive models to specify the customers likely to defect in advance [31]. Once specified, these customers could be targeted with appropriate incentives to encourage them to stay [32]. These incentives can take several forms, such as promotions, discounts, free calls, etc. Price reduction is one of the main market methods for attracting customers willing to churn in the telecom sector [33]. Promotions such as free calls in which the customers only pay an equivalent amount of 12 or $15 monthly and get to take free calls on the same network have become common and result in affordable telecom services. This encourages existing customers to switch their calling plans and join the promotion, as they can easily switch carriers by simply purchasing a new SIM card. As a result, customers may end up with three or four different phone lines, choosing to use the one that offers the best promotion at any time.

Traditional data analysis techniques that focus on mining quantitative and statistical techniques are used for churn detection in the telecom industry [34]. The success of this model depends on observing the customers' behavior with the help of experience and creating rules to classify a client as a churner or not. For example, a telecom company could label the client as a churner when the client makes several calls to customer service. However, these rules are created using only intuition and experience and without a scientific method, so the outcomes may be below expectations. In light of this, a robust method is needed to make reliable predictions and decisions based on experience effectively. Data mining and ML techniques have been widely favored to model customer churn [35, 36]. This is because churn is a rare event in a dataset and making an accurate prediction calls for techniques that emphasize predictive ability [37]. Therefore, to eliminate customer churn probability and develop an effective customer retention program, the utilized model should be accurate [38], or else, these systems would be useless when spending incentives on customers who will not churn. The true classification of churners and non-churners provides the telecom companies enough time to build a specific campaign to decrease customer churn probability and maximize their retention campaign profitability [39].

ML techniques have been broadly employed to model customer churn. This is because a churn is a rare event in a dataset and making accurate decisions requires creating models with high predictive performance. Therefore, to reduce customer churn probability and develop an effective customer retention program, the utilized predictive model should be accurate enough, or else, these systems will be useless when spending incentives on customers who will not churn. The correct classification of churners and non-churners provides telecom companies enough time to build a specific campaign to decrease customer churn possibility and maximize their profitability from the retention campaign [39]. Figure 1.3 shows the main learning model phases:

i. Preparation phase
ii. Learning phase
iii. Performance and evaluation phase
iv. Decision phase

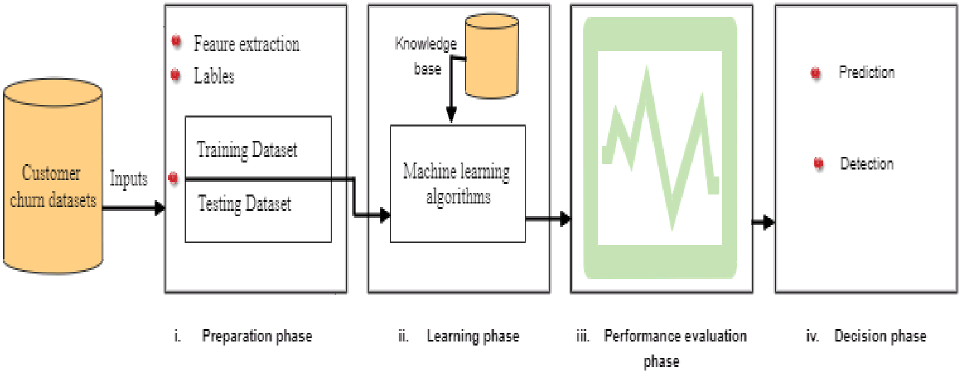


Figure 1.3: Learning system model

Dataset collection for customer churn requires timely monitoring and observation before transformation to conceptual labelling. Study shows it is increasingly becoming difficult to collect required training dataset without labelling simultaneously. While the direct data labelling technique (interactive approach) is ruled out in many practical cases, error prone each time data label manually resulting to data imbalance problem. The data analysis pipeline for prediction focuses on data balancing, feature selection, and modeling. Therefore, in this regard, this study aimed to introduce the HEOMGA method, which combines Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) to minimize data imbalance, ACO-RSA as a feature selection method and improved GBM model for better CP accuracy.

1.3 Research Questions

The fundamental research questions to be investigated throughout this proposal:

- Over-sampling techniques that utilize all available information in the minority class, such as SMOTE or ADASYN, can potentially outperform local neighbor-based methods depending on the context of the CP problem. Techniques like SMOTE generate new synthetic instances by interpolating between existing minority class instances, which can help improve model performance by addressing class imbalance. However, their effectiveness can vary based on the nature of the dataset and the specific problem, and investigations comparing both approaches are necessary to draw definitive conclusions. Therefore, **can over-sampling techniques that utilize all available information in the minority class outperform those that rely on local neighbor information in CP prediction problems?**

- Advanced metaheuristic-based optimization methods can enhance CP predictions by effectively selecting the most relevant features. These methods explore the feature space more thoroughly and can identify feature combinations that traditional methods might overlook. This can improve model accuracy, robustness, and generalization by reducing dimensionality and eliminating irrelevant or redundant features. Therefore, **can features selected via advanced metaheuristic-based optimization methods help improve CP predictions?**
- Gradient-boosting methods, which use tree-based weak learners, can significantly enhance CP predictions. This method builds an ensemble of trees (weak learners), where successive trees correct the errors of the previous ones, leading to a highly accurate final model. Therefore, **can the gradient-boost optimization method be used with tree-based weak learners to improve CP predictions?**

1.4 Research Contributions

The contributions of this project can be summarized as follows:

1. Develop a new method for the class imbalance problem in applying Churn Prediction (CP) in the telecom sector (the details of the proposed method are provided in chapter 4).

For solving the class imbalance problem in CCP, the minority data points are increasing using a hybrid techniques. It comprises two steps: (i) the distance between points within minority class data is calculated, as the square root of summation, to produce the fitness score that indicates the similarity (distance) between the data points. (ii) A genetic algorithm (GA) is applied (crossover and mutation) to produce new data points based on the smallest distance and hence, decreasing the class imbalance problem.

2. Propose a new heuristic searching method-based feature selection method for CP (the details of the proposed method are provided in chapter 5).

Although correlations amongst features and with the class variable are valuable for predicting importance to eliminate weak correlation, the method is good only for linear relation. For exploiting the nonlinear relations amongst the features, a new searching method is needed. The literature on feature selection reported that algorithms like Ant Colony Optimizer (inspired by ants food finding capability) and Reptile search Algorithm (inspired by crocodiles' hunting capability), etc shows superiority in solving several engineering problems and are excellent at solving multiple problems simultaneously. In this work a combination of these two techniques supporting each other is proposed to solve two optimization issues namely feature selection and performance improvement of CCP.

3. Develop a new CP model with high predictive performance that may be used to develop effective strategies and contain customer churn risks in the telecom sector (the details of the proposed method are provided in chapter 6).

The traditional Gradient Boosting Machine (GBM) for CP is improved model by using a stronger starting classifier, known as a base learner. A combination of Particle Swarm Optimization and Artificial Ecosystem-based Optimization for tuning hyper-parameters of the improved model while preventing premature convergence in the local optima.

1.5 Publications

- AlShourbaji, I., Helian, N., Sun, Y., & Alhameed, M. (2021). Customer churn prediction in telecom sector: A survey and way a head. Journal of scientific & technology research. https://uhra.herts.ac.uk/handle/2299/25060
- AlShourbaji, I., Helian, N., Sun, Y., & Alhameed, M. (2021). Anovel HEOMGA approach for class imbalance problem in the application of customer churn prediction. SN Computer Science, 2, 1-12. https://doi.org/10.1007/s42979-021-00850-y
- Al-Shourbaji, I., Helian, N., Sun, Y., Alshathri, S., & Abd Elaziz, M. (2022). Boosting ant colony optimization with reptile search algorithm for churn prediction. Mathematics, 10(7), 1031. https://uhra.herts.ac.uk/handle/2299/25060
- AlShourbaji, I., Helian, N., Sun, Y., Hussien, A. G., Abualigah, L., & Elnaim, B. (2023). An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. Scientific Reports, 13(1), 14441. https://doi.org/10.1038/s41598-023-41093-6

1.6 Thesis Structure

The rest of the thesis is structured as follows:

- Chapter 2 discusses a literature review on customer churn prediction using ML and explores various techniques and their effectiveness in addressing churn challenges.
- Chapter 3 overviews different open-source datasets used in customer churn prediction and their unique characteristics.
- The HEOMGA method combines HEOM and GA to address the class imbalance by oversampling the minority class along with experimental results on benchmark datasets is described in Chapter 4.
- Chapter 5 proposes an ACO-RSA-based-FS approach for feature selection in customer churn prediction and comparative analysis with other competitors like PSO, MVO, and GWO on multiple datasets.
- The CP-EGBM model is introduced in Chapter 6 by incorporating SVM_{rbf} as a base learner and an exponential loss function to enhance GBM's learning process for churn prediction. An mPSO is developed to optimize hyperparameters.
- The report concludes with Chapter 7, which includes a summary of contributions thus far, along with the main achievements of this thesis and also provides plans for future research.

Chapter 2 Literature Review

2.1 Business Domain Knowledge

By looking at the market size in the telecom industry, with the intent of finding the necessary information to understand how significant the churn problem is. The key findings include:

- The number of subscriber milestones in the telecom industry is expected to increase and reach more than 9 billion by the end of 2022 globally [40]
- 62.9% worldwide already owned a mobile phone in 2016 and is expected to reach 67.1% in 2019 [41].
- 35% of the individuals using the Internet are young people aged 15-24, Least developed countries compared with 13% in developed countries and 23% worldwide [42].

The following information is the average across the telecom industry:

- The monthly churn rate reaches between 10 - 67% [20]. This reflects significant differences in customer retention, where some companies lose as few as 1 in 10 customers monthly, while others may lose more than half.
- The average monthly revenue per business customer of Frontier communications reached $ 673.72 in 2016 [43].
- The gross margin in the second quarter of 2017 is 80.38% [44].
- The customer's lifetime is 52 months [45].
- A customer's lifetime value is $1782 in Frontier company [45]
- The acquisition cost of a new customer in the telecom sector is $ 315 [46].

Customer lifetime value can be defined as a measure of how much profit can be generated over the customer's lifetime.

Gross margin is a company’s residual profit after selling a service or product and deducting the cost allied with its production and sale.

Acquisition cost refers to gaining new customers, which involves persuading customers to purchase a company’s services or products. It measures how much value customers bring to telecom companies and their businesses

Several reasons are presented in the work of [47] for when customers decide to stop using the service and why. The authors classified the causes of churn into three groups: controllable churn, uncontrollable churn, and non-pay/abuse. The controllable includes anything under the company's control: Defecting to a competitor, response to poor service, and the service price. Uncontrollable includes all the reasons that are outside the control of the company hands, such as death, illness, and moving to a different country. The last group includes the causes related to nonpayment, abuse, theft of service, or other causes in which the company made the churn decision for the customer and it is unclear whether any of the non-pay/abuse is controllable or not.

A wide variety of factors play a great effect on churn in the telecom industry, such as income level, educational background, marital status, age, gender, geographical location, the effect of family and friends, cultural habits, service quality, and price. In [48], the authors highlighted that account length, international plan, voice mail plan, number of voice mail messages, total day minutes, total evening minutes, total night minutes, total international minutes, and number of calls to customer service are the most important factors for churning. Some other works emphasized that factors such as income level, educational background, marital status, and friend factors [49] and economic patterns: rate plans, tariffs, and the promotion available from different service providers [50] are crucial in determining customer churn.

The author in [51], investigated customer behavioral and demographic characteristics. The behavioral factors include rate plan (i.e., number of rate plan changes made with the carrier), handset changing frequency (i.e., number of handset changes made with the carrier), contract (Customer's service contract), rate plan suitability, customer tenure (i.e., number of months the customer stays with the service provider since service activation) and account status (still active or already churned at the end of the study period). The demographic factors include age, location (Western Canada or Eastern Canada), and language (English or French). The data were extracted from a Canadian wireless carrier, and 4896 residential customers were selected. The final results showed that rate-plan suitability is a key factor in customer churn. It suggests that customers who frequently change their call plans tend to churn less, indicating that offering flexible and suitable rate plans can significantly reduce churn rates.

In another recent work, the authors in [52] demonstrated the lost customer-first behaviors and lifetime experience, the reasons behind defection and the nature of the win-back offer made to lost customers are all related to the likelihood of their reacquisition, lifetime duration, and the lifetime profitability per month in a US telecom products and services company. They include six sets of variables:

- First behaviors and lifetime experience (a member of referrals, number of complaints, and service recovery).
- Defection behavior (price-related reasons, service-related reasons, or price and service-related dummy and time of defection).
- Win back offer nature (price discount, service upgrade, and price discount and service upgrade),

- Interaction between reasons for defection and win-back offers (price-related defection X price discount offer, service-related defection X service upgrade offer).
- First-lifetime marketing contacts (frequency of phone calls, e-mails sent, and direct emails).
- Demographics and first life control variables (age, gender, income, household size, education level, tenure, revenue, level of service plan, and cross-buy).

Their empirical results indicated that referral and complaint are essential pointers for the quality of the first-lifetime experience and how customers with positive first-lifetime experiences were more likely to accept a win-back offer.

Other factors and their complex relationships affect customer churns, such as service quality, a combination of features such as network coverage, signal strength, voice quality, and customer service provided by the service provider. This factor directly influences customers to switch to another service provider, as confirmed in the work of [53, 54].

In [55], the authors surveyed 196 respondents in the Java West area to assess how customer value and service quality affect customer churn. They found that higher service quality positively influences customer value and helps control customer churn

Service usage, switching cost, customer dissatisfaction, and demographic factors play a crucial role in customers switching to another service provider [56] when the authors used a dataset containing 1,000,000 records and 42 factors collected from a telecom company in the US.

Table 2.1: List of literature effort and the various factors used by the researchers in their works.

Works	**Customer churn factors**
Braun & Schweidel, 2011, [47]	Controllable churn, uncontrollable churn, and non-pay /abuse
Antipov & Pokryshevskaya, 2010, [48]	Account length, international plan, voice mail plan, number of voice mail messages, total day minutes, total evening minutes, total night minutes, total international minutes, and number of calls to customer service
Wong, 2011, [49]	Income level, educational background, marital status, and friends
Ranaweera, 2007, [50]	Economic patterns: rate plans, tariffs, and the promotion available from different service providers

Wong, 2011, [51]	Behavioral: Rate plan, handset changing frequency, contract, rate plan suitability, customer tenure and account status, Customer demographic information: Age, location, and language
Kumar et al., 2015, [52]	First behaviors and lifetime experience, defection behavior, win back offer nature, the interaction between reasons for defection and win-back offers, First-lifetime marketing contacts, and customer demographics
Cronin et al., 2000, [53] Al-Rousan et al., 2010, [54]	Service quality
Marwanto & Komaladewi, 2017, [55]	Customer value and service quality
Al-Mashraie et al., 2020, [56]	Customer demographic information: age and gender, number of dripped calls, number of service calls, and geographical area

2.2 Computer Science Domain

Developing a model that accurately predicts customer churn could have several managerial and financial implications for telecom companies. The correct classification of a customer as churner and non-churner can reduce misclassification costs such as cost of incentives and retention rate in real-world decision making, the assumption of equal miss classification, the default operating mode for many classifiers, is most likely violated. Customer churn has significant negative managerial and financial results on the company retention strategies. For instance, a company might lose direct contact with a client, making it impossible to sell additional products to them. Additionally, if an incentive is mistakenly sent to a non-churning customer instead of the actual churner, the intended churner misses out on the incentive meant to retain them. This misallocation could waste the marketing budget and negatively impact the company's financial resources [57, 58].

During the last decade, various studies have applied ML techniques for CCP modeling. Table 2.2 provides an overview of previous works on using ML techniques for modeling CCP in the telecom industry. Some primary reference papers in recent literature, along with their titles, the modeling techniques used, the dataset names and their characteristics, the number of records and variables and whether the datasets are Private (*) or public (#), the applied evaluation measures, the validation method, and research outcomes are summarized in the table.

Table 2.2: Overview of previous works on CP using ML approaches in the telecom industry

works	Title of paper	Used Techniques	Dataset-#rec. - #var. - Private (*) or public (#)	Measures- var. selection - sampling - validation	Outcomes
(Idris et al., 2012, [59])	Genetic programming and adaboosting based churn prediction for telecom	Adaboost style boosting – Artificial Neural Network (ANN) and Random Forest (RF)	Orange telecoms (KDD Cup 2009 small) –50,000 rec. – 230 var. - (#) and Cell2Cell - 70831sub. - 75 var. - (#)	AUC, sensitivity, specificity - Genetic Programming (GP) - uniform numerical format -no validation	GP- AdaBoost based model offers higher accuracy than ANN and RF. The GP- AdaBoost achieved a prediction accuracy of 89% for Cell2Cell and 63% for the other dataset.
(Miguéis et al., 2013, [60])	Customer attrition in retailing: an application of multivariate adaptive regression splines	Logistic Regression (LR)) and Multivariate Adaptive Regression Splines (MARS)	European retail company – 130284 #rec. – 7 var. - (*)	AUC, top-percentile lift - stepwise forward and stepwise backward – no sampling – 10- fold cross validation	MARS achieved better results when the whole set of variables are used. However, the LR outperforms MARS when variable selection procedures are applied.
(Brandusoiu&Toderean, 2013, [61])	Churn prediction in the telecommunications sector using support vector machines (SVM)	Support Vector Machine (SVM) with four kernel functions: Radial Basis Function kernel (RBF), linear kernel (LIN), Polynomial kernel (POLY) and sigmoid kernel (SIG)	UCI Repository, University of California - 3333 rec. – 21 var. - (#)	Recall, Specificity, Precision, Accuracy, Misclassification and F-measure – no var. selection – no sampling – no validation	SVM model using polynomial kernel (SVM-POLY) showed superiority over other SVM models.
(Lemmens & Gupta, 2013, [32])	Managing churn to maximize profits	gain/loss matrix and gradient boosting	Teradata Center at Duke University - -10,000 rec. – 171 var. – (*)	Misclassification - PCA –oversampling – F scores	The results indicated that improvements are achieved by using gain/loss matrix and gradient boosting approach for companies with no additional implementation cost.
(Keramati et al., 2014, [62])	Improved churn prediction in telecommunication industry using data mining techniques	Decision trees, ANN, K-Nearest Neighbors and SVM	Iranian mobile company – 3150 rec. – 11 var. - (*)	Accuracy, precision, recall, F –score – exhaustive recombination of variables - training- test split	ANN shown higher accuracy than Decision trees, K-Nearest Neighbors and SVM.
(Chen et al., 2015, [63])	Predicting customer churn from valuable B2B customers in the logistics industry: a case study	LR, Decision tree (C4.5), ANN (multilayer perceptron, (MLP)) and SVM	Logistics company in Taiwan - 69,170rec. - 18 var. - (*)	Accuracy, precision, recall, F1-measures) - no var. selection - random re-sampling – 10-fold cross validation	C 4.5 outperformed the other models and. the most significant variables for customer churn: Length, recency and monetary.
(Vafeiadis et al., 2015, [64])	A comparison of ML techniques for CCP	SVM-poly, Decision tree, ANN, Naïve Bayes, Regression Analysis,	UCI Repository, University of California - 3333 rec. – 21 var. - (#)	Precision, Accuracy, Recall and F-measure - no var. selection – no sampling – monte carlo based cross validation	SVM-poly using AdaBoost obtained 97% accuracy and F-measure over 84%.

works	Title of paper	Used Techniques	Dataset-#rec. - #var. - Private (*) or public (#)	Measures- var. selection - sampling - validation	Outcomes
(Zhang et al., 2015 [65])	Profit Maximization Analysis Based on Data Mining and the Exponential Retention Model Assumption with Respect to Customer Churn Problems	Decision trees and Regression	Guangxi Mobile Communication Company in China -40,000 rec. – 127 var. – (*)	ROC, normalized profit - no var. selection – no sampling– no validation	The relationship between profit and retention is good when the prediction algorithm sufficiently good, when the capability of retention is good enough, the relationship of profit and retention is convex and when both prediction algorithm and retention capability are not effective enough, operators should not take any actions.
(Hassouna et al, 2016, [66])	Customer Churn in Mobile Markets A Comparison of Techniques	Decision tree (CART, C 5.0, CHAID) and LR	Two UK mobile telecom operator data warehouse - 15,519 and 19, 919 rec. – 017 var.- (*)	AUC, Receiver Operating Characteristic, top decile, accuracy - choosing the most important var. - no validation	C 5.0 model is better than the LR model for CCP.
(Umayaparvathi&Iyakutti, 2016, [67])	Attribute selection and CCP in telecom industry	Gradient Boosting (GB), DT, SVM, RF, K-Nearest Neighbour, Ridge Regression and LR	Cell2Cell - 70831sub. - 75 var. - (#) and CrowdAnalytix – 3333 rec. – 20 var. - (#)	Accuracy, Precession and Recall, F1-measures - Brute force - no sampling - 10-fold cross validation	GB model has a higher performance than other techniques and six attributes: day minutes, voice, mail plan, night charge, international calls, evening calls, and day calls minutes have upmost importance towards churn prediction in the Cell2Cell dataset.
(Abdullaev, Ilyos, et al. 2023 [68])	Artificial intelligence with Jaya optimization algorithm-based churn prediction for data exploration technique	chicken swarm optimization and bidirectional long short-term memory (BDLSTM)	Churn dataset in UCI Repository	Accuracy, Recall, Precision, and F-measure,	The AIJOA-CPDE model has shown a maximum of 91.41%.
(Coussement et al, 2017, [69])	A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry	CART, Bayesian network, J4.8 decision tree, MLP, Naive Bayes, RF, SVM with RBF kernel function and Stochastic gradient boosting	Large European mobile telecom provider - 30, 104 rec. – 956 var. - (*)	AUC, top decile lift (TDL) - correlation-based var.– no sampling – no validation	Data preparation treatment (DPT) improves prediction performance. Logistic regression-DPT approach outperformed the empirical methods remarkably.
(Prashanth et al, 2017, [70])	High Accuracy Predictive Modeling for CCP in Telecoms Industry	Linear: LR, non-linear: RF, Deep Learning: Deep Neural Network, Deep Belief Networks and Recurrent Neural Networks	Asian telecom service provider – 337817 rec. – 36 var. - (*)	Accuracy, Sensitivity, AUC, Specificity- no var. selection - no sampling - 5-fold cross validation	Non-linear techniques performed better than the linear and both RF and deep learning gave comparable performance,
(Amin et al, 2017, [71])	CCP in the telecommunication sector using a rough set approach	Rough Set Theory (RST), Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and LEM2 algorithm (LA)	UCI Repository, University of California - 3333 rec. – 21 var. - (#)	Recall, Specificity, Precision, Accuracy, Misclassification, F-measure - Information Gain Attribute Evaluation - random re-sampling – K-fold cross	RST-GA based model showed satisfactory results for extracting implicit knowledge.

works	Title of paper	Used Techniques	Dataset-#rec. - #var. - Private (*) or public (#)	Measures- var. selection - sampling - validation	Outcomes
(Azeem et al., 2017, [72])	A churn prediction model for prepaid customers in telecoms using fuzzy classifiers	Neural Network, Linear regression, C4.5, SVM, AdaBoost, Gradient Boosting, RF and Fuzzy classifiers: Fuzzy	Telecom company in South Asia – 600,000 rec. – 722 var. – (*)	Recall, Precision, ROC, AUC - domain knowledge – oversampling – training-test split	Fuzzy classifiers shown superior performance compared to other used models
(Zhu et al., 2017, [73])	Benchmarking sampling techniques for imbalance learning in churn prediction	LR, Decision Tree (C4.5), SVM and RF	Chile- 5300 rec. – 41 var. – (*), Duke current– 51, 306 rec. – 41 var. – (*), Duke future – 100, 462 rec. – 173 var. – (*), Korean1 K1 Operator East Asia - 2019 rec. – 10 var. – (*), Korean2 K2 Operator East Asia -	AUC, Maximum Profit (MP), top-decile lift - Fisher score - ADASYN, Borderline-SMOTE, CLUS, MWMOTE, SMOTE, SMOTE–ENN, SMOTE–Tomek – 5 × 2 cross-validation	Sampling approaches power lies in the used evaluation metrics as well as the classifiers.
(Effendy et al., 2014, [74])	Handling Imbalanced Data in CCP Using Combined Sampling and Weighted RF	Weighted RF(WRF)	Telecom company in Indonesia – 48,384 rec. – 24 var. –	Recall, Precision and F-measure – no var. selection - combination of undersampling and SMOTE - 10-fold cross validation	The combined sampling techniques able to help WRF algorithm to achieve better performance and predict the churn effectively
(Prabadevi et al., 2023[75])	Customer churning analysis using machine leaning algorithms	Stochastic gradient booster, RF, LR, and k-nearest neighbors	Kaggle Repository, - 7044 rec. – 21 var. -(#)	ROC and AUC	The stochastic gradient booster model achieved the best performance results with an accuracy of 83.9%
(De Caigny et al., 2018, [76])	A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees	Decision trees and LR	Financial services, 117, 808 rec - 237 var- (*), Retail, 32, 371 rec- 47 var- (*), DIY 3, 827 rec - 16 var- - (*), Newspaper, 427, 833 rec- 165 var- (*), Telecom, 71, 074 rec- 87 var - (*), Financial services, 102, 279 rec- 138 var - (*), Telecom, 47, 761 rec- 43 var- (*),,Telecom, 50,000 rec – 303 var (*), Financial services, 631, 627 rec – 232 var - (*), Financial services, 573, 895 rec – 232 var - (*), Financial services, 398, 087rec – 232 var- (*), Financial services, 316, 578 rec – 232 var - (*), Financial services, 602, 575 rec – 232 var - (*), Energy, 20, 000 rec – 33 var -(*)	AUC, top decile lift	The hybrid model provided more accurate model than using its building blocks; decision trees and LR; as a standalone classification model
(Ullah et al., 2019, [77])	A Churn Prediction Model Using Random Forest: Analysis of ML Techniques for Churn Prediction and Factor Identification in telecom sector	JPK, LR, MLP, Naïve Bayes, AdaBoostM1, attribute selected classifier, decision stump, RF, J48, random tree, and LWL	Call Detail Records company in South Asian– 64,107 rec – 29 var.- (*), UCI Repository, - 3333 rec. – 19 var. - (#)	Accuracy, Recall, Precision, and F-measure	RF performed better in terms of prediction of churners

works	Title of paper	Used Techniques	Dataset-#rec. - #var. - Private (*) or public (#)	Measures- var. selection - sampling - validation	Outcomes
(Jafari-Marandi et al., 2020, [78])	Inferring ML-Based Parameter Estimation for Telecoms Churn Prediction	ANN, Self-organizing map	Telecom company in Iran- –3150 – 12 var.- (*),	Accuracy, Recall, Precision, F-measure, and misclassification error	The proposed profit-driven models proved to be effective for telecom churn prediction
(Jafari-Marandi et al., 2020, [78])	Inferring ML-Based Parameter Estimation for Telecoms Churn Prediction	ANN, Self-organizing map	Telecom company in Iran- –3150 – 12 var.- (*),	Accuracy, Recall, Precision, F-measure, and misclassification error	The proposed profit-driven models proved to be effective for telecom churn prediction
(Khattak at al., 2023, [79])	Customer churn prediction using composite deep learning technique	Deep learning model and bidirectional long/short-term memory	Kaggle Repository, - 7033 rec and 20 var. -(#)	Recall, Precision, and F-measure,	The results showed that proposed model attained a remarkable accuracy of 81%
Faritha Banu2022 [80]	[Artificial Intelligence Based Customer Churn Prediction Model for Business Markets	AI-based CCP model for Telecommunication Business Markets using Chaotic Salp Swarm Optimization-based Feature Selection (CSSO-FS) method for the best feature assortment. In addition, a Fuzzy Rule-based Classifier (FRC)	3 CCP datasets	Sensitivity, Specificity, Accuracy F-Score	AICCP-TBM technique outperformed the other two techniques, with maximum accuracy of 97.25 %, 97.70 % and 94.33 % on the three datasets studied.
Praseeda, C. K., & Shivakumar (2023) [81]	Fuzzy particle swarm optimization (FPSO) based feature selection and hybrid kernel distance based possibilistic fuzzy local information C-means (HKD-PFLICM) clustering for churn prediction in telecom industry	The information gain and fuzzy particle swarm optimization (FPSO) has been executed by the method of feature selection, besides the divergence kernel-based support vector machine (DKSVM) classifier	telecom churn (cell2cell) “ https://www.kaggle.com/jpacse/datasets-for-churn-telecom/version/2” . The second dataset is a publicly accessible churn-bigml dataset “ http://github.com/caroljmcdonald/mapr-sparkmlchurn/tree/master/data” .	Accuracy, Precision, Recall, F-measure, ROC curve	The higher accuracy results of 94.11% and 95.41% for cell2cell dataset and big dataset.

Class imbalance, a common issue in CCP, is often tackled with over-sampling techniques such as SMOTE and ADASYN. Several studies underline the importance of feature selection for improving model accuracy. In [17] and [216] used information gain attribute is evaluation with GA to select key features, leading to improved CCP accuracy. As per literature, performance measures such as AUC, recall, F-measure, and accuracy are consistently used to evaluate model efficacy. Many studies have demonstrated that advanced and hybrid methods outperform baseline models in these key metrics.

Chapter 3 Datasets descriptions

3.1 Benchmark Datasets

Publicly available benchmark datasets for customer churn are used in this research. The characteristics of these datasets are presented in Table 3.1. CP is a binary classification problem (churner or non-churner). Details of each dataset such as data collection, features can be found in the hyperlink given in Table 3.1.

Table 3.1: The characteristics of the open-source customer churn datasets

Dataset	Source	# instances	# features	# classes	# churners	#non-churners
Dataset 1	https://www.kaggle.com/barun2104/telecom-churn	3,333	21	2	483	2,850
Dataset 2	https://www.kaggle.com/datasets/blastchar/telco-customer-churn	7,043	21	2	3,521	3,522
Dataset 3	https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom	71,047	58	2	14,210	56,837
Dataset 4	https://www.kaggle.com/abhinav89/telecom-customer/data	100,000	100	2	49,562	50,438
Dataset 5	https://www.kaggle.com/barun2104/telecom-churn	3,333	11	2	483	2,850
Dataset 6	https://www.kaggle.com//barun2104/telecom-churn	3,150	16	2	495	2,655
Dataset 7	https://www.kaggle.com/code/tusarahmed/internet-service-provider-customer-churn	50,375	10	2	20,331	30,044

To understand the research questions more comprehensively, it is essential to define the types of data utilized in this study, specifically focusing on customer churn datasets. This subsection will provide clarity of some attributes involved and how they relate to the analysis of churn behavior.

1. Customer Demographics: This data includes attributes such as age, gender, income, and geographic location. These features help in understanding customer segments and their likelihood of churning.

2. Churn: This is the target variable that directly answers the primary research question about customer retention. Understanding which factors lead to a "Yes" (churn) or "No" (non-churn) is crucial for predicting customer behavior.
3. Service Usage: This data pertains to how customers interact with the service, such as frequency of use, duration of engagement, and service features utilized. Understanding usage patterns can indicate potential churn.
4. CustomerCareCalls and RetentionCalls: These attributes reflect customer service interactions. A higher number of calls could indicate problems that might lead to churn. Analyzing these interactions can inform the second research question about feature selection and the importance of customer service in predicting churn.
5. Total Day/Night/International Calls and Charges: These features provide insights into customer behavior patterns that can be used to refine the predictive models. They are crucial for exploring how usage patterns correlate with churn and improving model accuracy in the context of the research question.

Shared attributes amongst all datasets are:

- Services signed up by customers – phone, multiple lines, Internet, online security, and streaming TV and movies.
- Customer account information – how long they have been a customer, monthly or total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

3.2 Datasets preprocessing

The pre-processing is performed and the step-by-step procedure for the data pre-processing is as follows:

- The discrete attributes of unique values are ignored since the number of levels does not affect the model training process.
- Convert the textual categorical features such as ‘yes’, ‘no’, ‘true’, and ‘false’ into ‘0’s and ‘1’s.
- Continuous features can be scaled using the Min-Max normalization method and the Linear transformation (L) of the original input range to a newly specified range.

$$L = \frac{max_{old}(i) - x(i)}{max_{old}(i) - min_{old}(i)} \left(max_{new}(i) - min_{new}(i)\right) + min_{new}(i) \qquad (1)$$

where $min_{old}(i)$ and $max_{old}(i)$ are the minimum and maximum values of *i*th feature and $min_{new}(i)$ and $max_{new}(i)$ are the new desired minimum and maximum values of *i*th feature.

Min-Max normalization is applied to transform all features in the dataset into the range of 0–1, which means min_{new} and max_{new} for all features is 0 and 1, respectively.

Chapter 4 A novel HEOMGA Approach for Class Imbalance Problem in the Application of Customer Churn Prediction

4.1 Introduction

In recent years, the problem of imbalance class has been widely studied in ML. Typically, this problem occurs when the classes in a given dataset are unequally distributed between the minority and majority classes. Without consideration of this problem, the effective learning process by classification algorithms will be a challenge since the main goal is detecting minority classes [79]. Addressing this problem has attracted increased attention from the research community due to its importance in different applications; examples include malware detection [80], medical diagnosis domain [81], financial crisis prediction [82], and churn prediction [83]. Several studies compared random sampling techniques to handle the class imbalance problem in the preprocessing phase. The results from these efforts highlighted that these methods were helpful before applying classification algorithms [84, 85]. This is also confirmed by the work of [86] when 26 datasets were used to investigate the influence of class imbalance before and after balancing the datasets. On the other hand, it was reported that random sampling methods for class imbalance were shown not to be useful in improving the performance of prediction results [87, 88].

Balancing class is necessary when learning from highly skewed datasets because an imbalanced dataset could classify all the instances as negative, leading the learner to have a high false-negative rate [89, 90]. For example, in a customer churn scenario where 90% of customers are loyal and only 10% are churners, an imbalanced dataset might cause a model to predict that all customers will stay. This approach would result in a high false-negative rate, missing many of the actual churners who need targeted retention efforts. Balancing the dataset helps the model more accurately identify and address customers who are likely to churn. Therefore, a balancing strategy with betterinterpreting capability is essential in the preprocessing phase to specify churn customers. The cost is usually high, as it means the company misses opportunities to retain valuable customers.

In this work, we propose a novel method based on Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) to generate data points from the existing minority ones rather than use random methods. This work proposes a data-level strategy for addressing the class imbalance problem. The main objective of this work is to investigate the suitability of the proposed method in achieving optimal performance results and facilitating the learning process by learners from imbalanced datasets. A thorough empirical study was carried out, which proves the significant performance gains of the proposed method compared to other popular oversampling algorithms.

The rest of the chapter is organized as follows: The section "Literature review "reviews SMOTE and ADASYN oversampling methods. Section "Proposed Method" presents the proposed method. Section "Experiment Design" describes the imbalance of customer churn datasets used to examine the proposed method, while section. "Results and Discussion" provides the experimental design used in this work. Section "Conclusion and Future Work" presents the results and discussion of this research. The final section summarizes the chapter along with future work.

4.2 Literature Review

Research on synthesizing minority samples has been widely studied to address the problem of class imbalance distribution at the data level. The random sampling method is the simplest way. Its main goal is to improve data quality in the preprocessing phase before training classification algorithms. Random sampling can be divided into two categories: undersampling and oversampling. In the under-sampling technique, the same samples belonging to the same majority of samples are removed from the dataset. For example, 30% under sampling means that 30% of the available majority instances are randomly removed from the dataset. However, by removing significant instances, this method may potentially lose valuable information. The second category attempts to create a superset of the original dataset. This can be achieved by replicating the minority instances from the existing dataset. The replication can be done either randomly or using an intelligent method. For example, 100% oversampling means that the minority instances are replicated once on average. However, a drawback of this method is that creating additional instances could have a significant impact on computational cost and overfitting.

SMOTE is an advanced method of oversampling, and it was developed by Chawla et al. [91]. This approach randomly picks one data point from the k neighbors of a minority class sample and inserts a new synthetic minority class sample on the line that connects the randomly chosen minority class sample and one of its k minorities nearest neighbors belonging to the minority class sample, as illustrated in Figure 4.1.

He et al. [92] proposed ADASYN to overcome the problem of class imbalance. It is an oversampling method developed to reduce generating noise data and ambiguity along the decision boundaries produced by SMOTE. The major difference between SMOTE and ADASYN is in generating synthetic sample points for minority data points. In ADASYN, the data points that are harder to learn are more frequently presented by this method, as shown in Figure 4.2.

Recent developments of SMOTE and ADASYN, Borderline-SMOTE [93], Safe-Level-SMOTE [94], and Local Neighborhood SMOTE [95] are some other extensions to reduce generating noise data and the ambiguity along the decision boundaries that SMOTE produces. These extensions attempt to create data points from the minority class that are close to the borderline between the two classes.

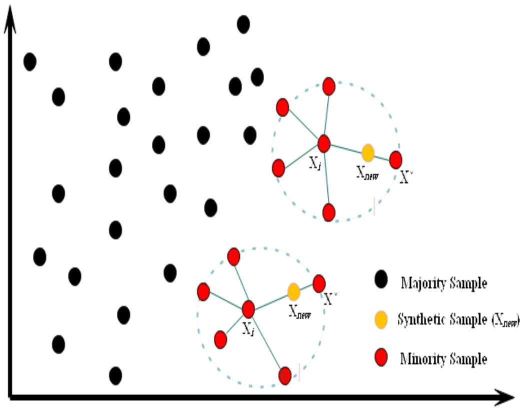


Figure 4.1: Generation of synthetic samples using SMOTE, a randomly selected minority class sample and of its k =5 nearest neighbors

He et al. [92] proposed ADASYN to overcome the problem of class imbalance. It is an oversampling method that was developed as an extension to reduce generating noise data and ambiguity along the decision boundaries that SMOTE produces. The significant difference between SMOTE and ADASYN is in generating synthetic sample points for minority data points. In ADASYN, the data points that are harder to learn are more frequently presented by this method, as shown in Figure 4.2.

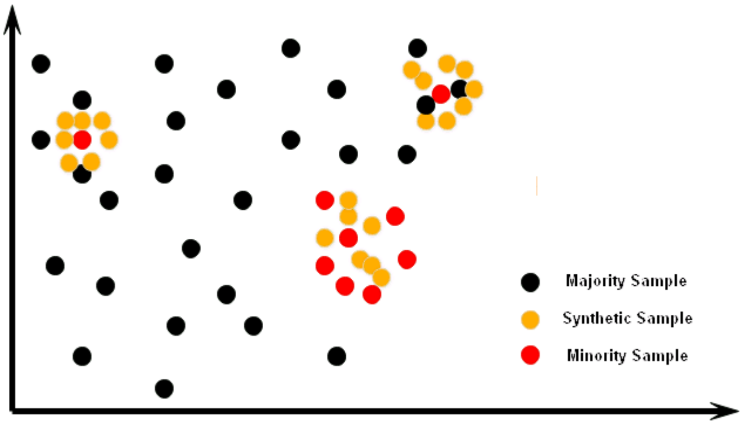


Figure 4.2: Generation of synthetic samples using ADASYN

G-SMOTE, is an extension of the SMOTE algorithm, where it aims to define a safe area around each selected minority data point such that the generated synthetic sample points for minority data points inside this area are not noisy. Also, to increase the variety of generated samples points by expanding the minority class area.

Barua et al. [96] proposed another recent technique for imbalanced data problems: Majority Weighted Minority Oversampling Technique (MWMOTE). This method has several functions, which include a) generating a useful synthetic class sample, b) adding weights to the selected sample based on their importance, and c) clustering to produce suitable synthetic minority class samples.

Zhu et al. [97] assessed the suitability of ADASYN, Borderline SMOTE, Random oversampling, and SMOTE strategies for class imbalance in churn prediction using 11 datasets. The results recommended that suitable sampling strategies need to be selected, and the setting of the class ratio has an impact on the model performance. Another work [98] investigated six sampling techniques and used four customer churn datasets. These methods include Mega-trend Diffusion Function (MTDF), SMOTE, ADASYN, Couples Top-N Reverse k-Nearest Neighbor (TRkNN), MWMOTE, and Immune centroids oversampling technique (ICOTE). Their empirical results demonstrated that MTDF performed better than the other oversampling methods they used in the study. Salunkhe et al. [99], proposed a hybrid data-level approach for handling class imbalance problems. The authors combined SMOTE and under-sampling techniques to achieve better results. The aim was to focus on the majority class's necessary data and avoid removing valuable information when using the under-sampling technique before the model training stage. They achieved results better than the other techniques for class imbalance.

During the last decade, a worldwide range of studies has applied GA for class imbalance problems [100, 101, 102]. In the approach of [103], GA with SMOTE was combined to perform oversampling and they used different sampling rates for different minority examples until reaching the desired oversampling rate. The results showed that the proposed method achieved better performance compared to SMOTE. In another work, GenSample was proposed by [104]. They used the GA method for oversampling minority classes by considering the difficulty in learning an example and the improved performance caused by oversampling it. Their final results showed that the GenSample method achieved better performance than the traditional methods. Distance-based algorithms are widely used for class imbalance problems to provide a numerical description of the similarity between two objects. Several studies confirmed that improving distance metrics performance makes ML algorithms more accurate [105-108]. The aim of the research done by Mahin et al. [109] is to improve the categorization process of the minority class by incorporating the idea of using a dataset-specific distance function and choosing the appropriate distance metric and k nearest neighbor's value among the five used distance metrics for five datasets. They concluded that there is no optimal distance metric for all the datasets.

Modifications can be made at the algorithm level by incorporating the cost of misclassifying minority samples or integrating one class learning algorithm. Bagging and boosting ensemble techniques can be used as cost-sensitive methods, where the classification outcome is some combination of multiple classifiers built on the dataset. Guo et al. [110] applied data boosting to improve the performance of complex, difficult-to-classify examples. The algorithm-level method tries to adapt existing learning algorithms to strengthen their learning capability regarding the majority class. However, this approach requires a deep understanding of the application domain and corresponding classifiers. Hybrid methods are also used to conquer the problem of class imbalance recently. An ensemble of classifiers can be used at the algorithm level and different sampling methods and cost-sensitive learning methods can be hybridized at the data level. Liu et al. [111] incorporated oversampling and undersampling with an ensemble Support Vector Machine (SVM) to improve its prediction performance. Experimental results showed that better performance was achieved by SVM when the problem of class imbalance was contained by the use of oversampling and undersampling methods compared to other classifiers and SVM alone.

Based on the conducted review, the first observation indicates that solving class imbalance at the data level seems the most viable and widely used option in practice to provide the learner with more robust training data [96-111]. The existing techniques conduct oversampling based on the information of subset of the minority data points which only considers local information, However, the idea is to generate artificial minority data points based on all minority data points that considers global information. To achieve this, HEOMGA is utilized as will be discussed in the next section

4.3 Proposed HEOMGA Method

4.3.1 HEOM

A number of distance metrics are designed and used for measuring similarity and dissimilarity among samples within a given dataset. These metrics depend on the nature of the dataset's attributes, whether they are numerical or only contain categorical attributes. For example, Euclidean distance is the most widely used when all the attributes are numerical. Another example, Hamming Distance, can be used when only categorical attributes. However, some other metrics were designed to handle nominal and categorical attributes, i.e., mixed or heterogeneous data such as HEOM. It has become more popular due to its simplicity and efficiency in independently handling continuous and discrete attributes [112 - 115].

Considering input vectors x and y, the HEOM distance can be calculated by

$$\rho(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} d(x_{\mathrm{i}}, y_i)^2} \tag{1}$$

$d(x_i, y_i)$ is the distance between the two cases on its ith attribute, were,

$$d(x_{\mathrm{i}}, y_{\mathrm{i}}) = \begin{cases} 1, & \text{if } x_i \text{ or } y_i \text{ is missing} \\ d_o(x_{\mathrm{i}}, y_{\mathrm{i}}), & \text{if } x_{\mathrm{i}} \text{ and } y_i \text{ are discrete variables} \\ d_e(x_{\mathrm{i}}, y_{\mathrm{i}}), & \text{if } x_{\mathrm{i}} \text{ and } y_i \text{ are continuous variables} \end{cases} \tag{2}$$

HEOM uses the overlap metric $\mathrm{d_o}$, for categorical attributes,

$$d_o(x_{\mathrm{i}}, y_{\mathrm{i}}) = \begin{cases} 0, & \text{if } x_{\mathrm{i}} = y_{\mathrm{i}} \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

The Euclidean distance $d_e(x_{\mathrm{i}}, y_{\mathrm{i}})$, for continuous attributes

$$d_e(x_{\mathrm{i}}, y_{\mathrm{i}}) = |x_i - y_i|^2 \tag{4}$$

4.3.2 GA

A GA uses an iterative method to look for a global solution; a new population at each iteration is created, including the progressions of individuals chosen from the previous iteration. Also, the original population is composed of random resolutions. A data structure called a chromosome codifies people. The chromosomes are reflected by a bit of string in the simple or normal GA. Each bit of a gene reflects the presence (value 1) or absence (value 0) of a particular trait in the individual. In each generation, individuals measure their fitness to resolve the issue. A fitness function that decodes the information stored in each chromosome is considered a measure of its consistency with quality requirements. A chromosome is evaluated to test its "fitness" as a solution. The fitness function plays a vital role in the environment in natural evolution by rating individuals in terms of their fitness. Selecting and formulating an appropriate fitness function is crucial to solve any GA problem efficiently. In our case, selecting the optimal samples (data points) in the initial population, the minority class is set to HEOM.

Random members from the population are chosen for reproduction after assessment, creating offspring, which would form a new population. According to the rules of natural selection, this selection would favor the fittest entity. Their chromosomes are combined in the replication of the chosen individuals to get two descendants. This hybrid method is carried out by adding a crossover operator, a binary operator applied to two individuals. These individuals are called parents, and their genes are mixed to generate two new individuals, called offspring. A typical crossover operator is a one-point crossover for the bit-string representation. The second operator is a mutation, which enforces genetic heterogeneity in the current solutions. The boundary mutation alters genes from the individuals produced in the crossover process. The population generation methods, the assessment of their individuals, and the selection and deployment of genetic operators are iterated and form the basis of the GAs. The GA can generate different solutions to the same problem, depending on the initial population. Therefore, the GA is typically run many times for varying initial populations, and other conditions can be used to stop it. For instance, when the maximum number of generations is reached, the GA can be stopped.

4.3.3 HEOM GA

HEOM measures the distance of the first data point with all other minority data points in the initial population, which is the square root of their summation to produce the final fitness score for that data points. HEOM acts as a fitness function for measuring similarity (distance) between the individuals (data points) in the initial population, which contains all the minority class samples in the training dataset to decide which

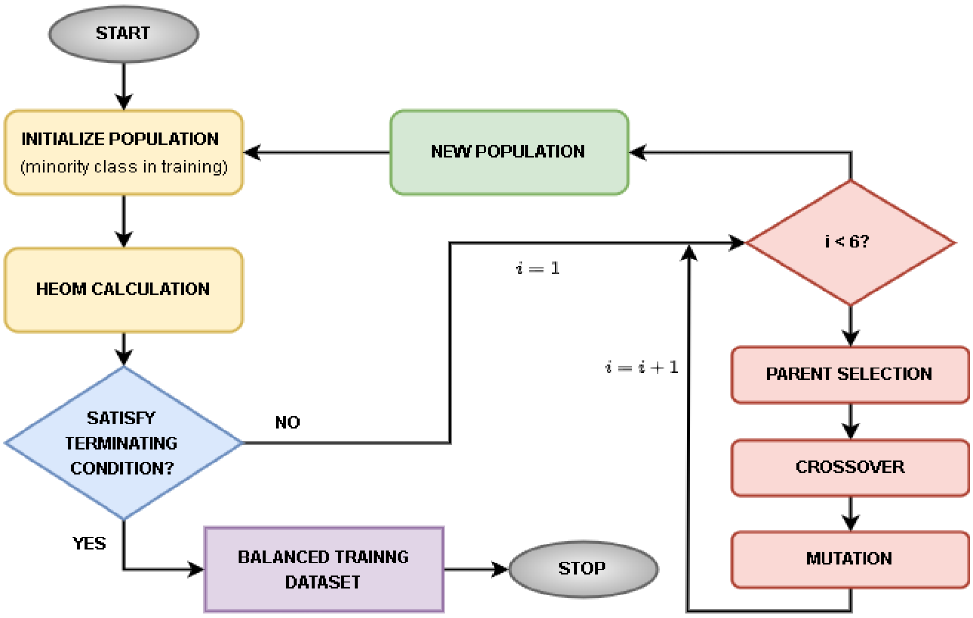


Figure 4.3: Basic structure of the HEOMGA method

data points to use and the data points with small scores which are produced from HEOM are selected as parents for mating and then apply the GA variants (crossover and mutation) to produce offspring within the same iteration. Based on the three genetic operators and the evaluations, the better new populations of a candidate after the specified number of generations (e.g., number of generations =5), the best solution is formed and appended to the initial population. In order to start the next iteration, two data points with the shortest distances are selected as parents by returning the corresponding distance to each data point in the initial population in addition to the appended data points from the distances list produced in the previous iteration and then starts the role of crossover and mutation operators. This procedure will be repeated until the majority data points are equal to the minority data points in the original data set. Finally, to avoid generating newly duplicated data points, the algorithm will check and delete any duplicated ones. Figure 6 depicts the proposed method processes.

SMOTE and ADASYN generate a noise sample that has penetrated the majority class region, increasing by overlapping. These noise samples are less useful because they do not add any new information to the imbalanced datasets and may lead to overfitting. It was confirmed that using the Euclidean distance metric that SMOTE and ADASYN use to measure the distance between two objects introduces issues regarding imbalanced data and performance problems regarding computation or approximation of the square root [108]. Most datasets have both nominal and categorical attributes, and the major weakness of the Euclidean distance is that when some attributes have a large range of values as opposed to the remaining attributes, they may influence a bigger impact on the computed distance. In comparison, attributes with a lower range of values will have a lesser impact on the results.

In the proposed method, all the minority data points are selected as the initial population, and the HEOM finds the distance between them by calculating the square root of their summation to produce the final fitness scores. In HEOM, normalized Euclidean distance is used for numeric features, and the overlap distance for categorical features is employed to find the distance between two instances, x1 and x2, as provided in the above equations (3 and 4). Applying the HEOM distance metric allows better handling of nominal and categorical attributes following the dataset nature. In addition, HEOM will help obtain better representation capability for minority data points and will enable us to appropriately select the data points that will be used as input for mating in the GA.

Crossover and mutation operators in GA realize the search exploration and exploitation, respectively. Exploration is the ability to create diversity in the population by exploring the search space, while exploitation reduces diversity by focusing on individuals of higher fitness. Therefore, the newly generated synthetic data points will be produced in a safe region within the boundaries of the minority data points that the HEOM selects. As shown in Figure 4.4, overlapping and overfitting problems will be somehow eliminated by causing the distance (d) between the generation area (the pink dotted oval) and the decision boundary to be larger and spread the newly generated data points far from the majority space.

Crossover and mutation operators improve the learning process by providing rich information about the newly generated data points since they are inherited from the

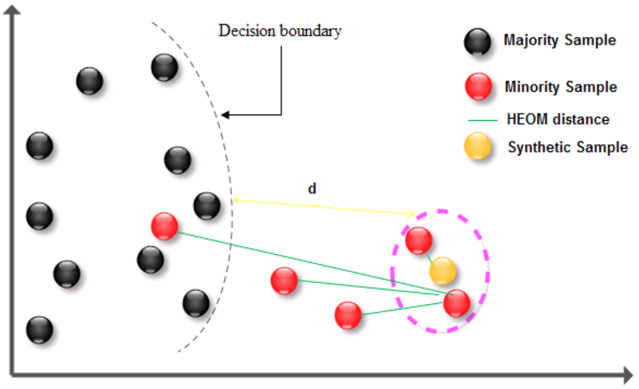


Figure 4.4: An example of how can HEOMGA avoid overlapping

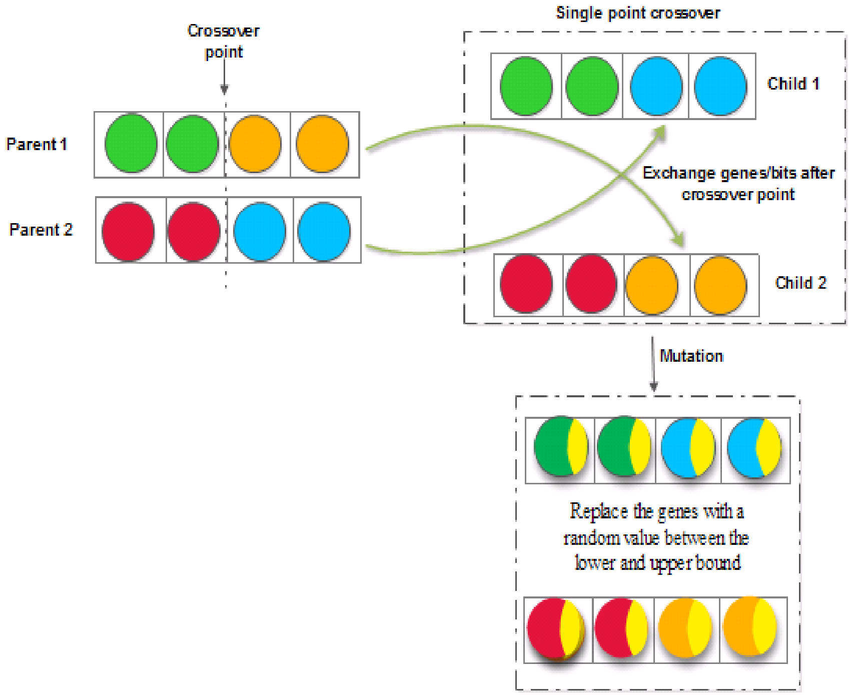

Crossover point
Single point crossover
Parent 1
Parent 2
Child 1
Exchange genes/bits after crossover point
Child 2
Mutation
Replace the genes with a random value between the lower and upper bound

Figure 4.5: GA operators’ processes

original data points, as shown in Figure 4.5. This will make the learning process by a given learner easier. Finally, the HEOMGA will check and delete any duplicated data points during the generation process to avoid the generation of newly repeated samples,

4.4 Experimental Design

A set of publicly available datasets for customer churn prediction are used in this work. Table 3.1 gives the details for each dataset. Evaluation of data mining and ML methods on publicly available datasets offers different advantages [116]. These benefits include:

- In terms of comparability of results, ranking methods, and evaluation of existing methods with new techniques
- Study the impact of the data and their characteristics on the performance of a technique
- Using available datasets provides insight into the effect of each phase of the followed methodology

Please note that the datasets provided in Table 3.1-chapter 3, excepting Dataset 2 because it is a balanced dataset, are used to evaluate and validate the results of the proposed HEOMGA method.

4.4.1 Baseline approaches and Learners

Three different learners are used to examine the capability of the method: Decision Trees (DTs, i.e., C4.5 algorithm), Bagging, and SVM with radial basis function kernel (SVM_{rbf}). These learners are selected due to their popularity with classification problems and sensitivity to imbalanced datasets [117, 118].

The DTs are one of the most popular and interpretable machine learning algorithms, commonly used for both classification and regression tasks. Their simplicity, ability to handle both categorical and numerical data, and ease of visualization make them a fundamental tool in data science. DTs work by splitting the dataset into smaller subsets based on certain criteria or questions at each node, ultimately leading to a decision or prediction at the leaf node. The DTs rely on greedy-search heuristics that checks one variable at a time [119], and therefore, it can attain a high level of accuracy by predicting the majority class, mainly if the majority class constitutes most of the dataset.

SVMs are a powerful class of supervised learning algorithms primarily used for classification, though they can also be adapted for regression tasks. SVMs are designed to find the optimal hyperplane that separates data points of different classes with the maximum margin. This geometric approach makes SVMs highly effective for both linear and non-linear classification problems. An SVM learner tries to find the hyperplane by splitting instances of two classes based on the largest distance between them. It is useful mainly due to its capability to work in high-feature space since the learner can map complex nonlinear relationships between input and output with relatively high-accuracy [120].

Bagging is a powerful ensemble technique that enhances the performance of machine learning models by reducing variance and improving accuracy. By leveraging the diversity of multiple models trained on different subsets of data, bagging creates a robust and reliable predictive system. Despite its computational cost and potential loss of interpretability, bagging remains a fundamental approach in machine learning, widely applied across various fields to achieve better and more stable predictions. Bagging is an ensemble learning learner who can effectively handle class imbalance problems [118]. The nearest neighbors (K) parameter number in both SMOTE and ADASYN was set to 5 [121].

10-fold cross-validation is used to avoid picking particular parts for training and testing. The number of k was adjusted to 10; the resulting data was split into 10 parts; the procedure started by splitting the dataset into 90% for training and 10% for testing.

In order to finalize the process, the procedure was repeated 10 times to allow each part of the data to be tested; finally, the average results were considered for the used datasets on the 10 partitions.

4.4.2 Experimental setup

All the experiments are implemented using Python scikit-learn and he DTs SVM$_{rbf}$ and bagging learners are constructed based on the default parameters on Windows 7 with 2 Duo CPU running on 3.13 GHz PC with 44.25 GB RAM. The settings of these methods are defined based on their implementations in original works and they are listed in Table 4.1.

Table 4.1. Parameters settings

Algorithm	Parameters
DTs	Max-depth=10; min_samples_split=10, max-feature= number of features in each dataset, criterion= entropy
SVM	Regularization=10; kernel= radial basic function; gamma= 0.01.
LR	Regression type = Lasso (L1); Regression coefficient = 1.
SMOTE	Nearest neighbors (K) =5
ADASYN	Nearest neighbors (K) =5

4.4.3 Evaluation metrics

In order to assess learners' results, a confusion matrix was employed to count: True Positive (TP) and True Negative (TN) denote number of positive and negative examples that are classified correctly, while False Negative (FN) and False Positive (FP) represents number of misclassified positive and negative examples respectively. Table 4.2 shows a confusion matrix of a two-class problem. The table's first column is the examples' actual class label, and the first row presents their predicted class label.

Table 4.2: Confusion matrix for a two-class problem

	Actual	
	Churn Customers	Non-churn Customers

Predicted	Churn Customers	TP	FP
	Non-churn Customers	FN	TN

Recall: It is the True Positive rate, which refers to the percentage of positive instances correctly predicted as positive class instances [40].

$$\mathrm{Re}\,call = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

Geometric mean (G mean): It is a good indicator that can be used to assess the overall performance of a given learner because it combines the learner's accuracy on the positive class and negative class samples [44]. Therefore, a large value of this measure indicates that the learner performs well on both classes' samples.

$$Gmean = \sqrt{\frac{TP}{\mathrm{TP} + \mathrm{FN}} \times \frac{TN}{\mathrm{T}N + \mathrm{FP}}} \tag{6}$$

Area Under Curve (AUC): Receiver Operating Curve (ROC), usually known as AUC. The ROC graph plots true-positive rates versus false-positive rates. Learners can be selected based on their trade-off between true and false positives. Rather than visually comparing curves, the ROC metric aggregates the performance of classification methods into a single number, which makes it easier to compare the overall performance of different learners. This metric can also be applied to evaluate learning from imbalanced data [45]. The higher the AUC indicates, the better the generalization of the methods. The AUC can be determined as follows:

$$AUC = \frac{\left(1 + \frac{TP}{\mathrm{TP} + \mathrm{FN}} - \frac{FP}{\mathrm{FP} + \mathrm{TN}}\right)}{2} \tag{7}$$

The above evaluation metrics can reasonably evaluate the learning process from imbalanced datasets since their formulae are relative to the rare class, which is the churn class. These measurements are used to evaluate the proposed method and its effectiveness in overcoming the class imbalance.

4.5 Results and Discussion

The results without using any balancing method (i.e., 0% balancing) and the results of the proposed method against SMOTE, ADASYN, G SMOTE [120] and Gaussian method [121] were applied over three customer churn datasets to study the impact of different balancing technique on the evaluation measures used in this work. The results are summarized in Tables 4.3–4.5.

Table 4.3: DTs results based on the evaluation metrics for all the datasets

Dataset	Method	Recall	G-mean	AUC
Dataset 1	0% balancing	0.780	0.852	0.856
	SMOTE	0.741	0.798	0.833
	ADASYN	0.725	0.802	0.812
	Proposed	**0.926**	**0.944**	**0.944**
	G-SMOTE	0.758	0.841	0.847
	Gaussian	0. 852	0.921	0.924
Dataset 2	0% balancing	0.581	0.620	0.665
	SMOTE	0.602	0.645	0.695
	ADASYN	0.582	0.632	0.692
	Proposed	**0.823**	0.**825**	**0.827**
	G-SMOTE	0.635	0.705	0.793
	Gaussian	0.723	0.768	0.820
Dataset 3	0% balancing	0.522	0.523	0.523
	SMOTE	0.466	0.544	0.551
	ADASYN	0.464	0.540	0.546
	Proposed	**0.523**	**0.552**	**0.554**
	G-SMOTE	0.473	0.536	0.541
	Gaussian	0.478	0.539	0.543
Dataset 4	0% balancing	0.806	0.835	0.868
	SMOTE	0.772	0.807	0.846
	ADASYN	0.767	0.800	0.835
	Proposed	**0.925**	**0.932**	**0.940**
	G-SMOTE	0.789	0.822	0.858
	Gaussian	0.876	0.895	0.914
Dataset 5	0% balancing	0.831	0.855	0.880
	SMOTE	0.804	0.830	0.859
	ADASYN	0.810	0.833	0.858
	Proposed	**0.925**	**0.930**	**0.936**
	G-SMOTE	0.821	0.844	0.869
	Gaussian	0.876	0.890	0.904
Dataset 6	0% balancing	0.882	0.892	0.903
	SMOTE	0.866	0.875	0.885
	ADASYN	0.894	0.898	0.903
	Proposed	**0.923**	**0.925**	**0.927**
	G-SMOTE	0.883	0.887	0.891

	Gaussian	0.876	0.880	0.884

Bold values indicate best results

Table 4.4: SVM results based on the evaluation metrics for all the datasets

Dataset	Method	Recall	G mean	AUC
Dataset 1	0% balancing	0.719	0.706	0.724
	SMOTE	0.814	0.816	0.817
	ADASYN	0.676	0.739	0.739
	Proposed	**0.845**	**0.919**	**0.919**
	G SMOTE	**0.845**	**0.919**	0.918
	Gaussian	0.837	0.913	0.914
Dataset 2	0% balancing	0.581	0.607	0.636
	SMOTE	0.605	0.640	0.681
	ADASYN	0.539	0.586	0.642
	Proposed	**0.674**	**0.705**	**0.740**
	G SMOTE	0.673	0.697	0.724
	Gaussian	0.668	0.699	0.734
Dataset 3	0% balancing	0.443	0.537	0.547
	SMOTE	0.395	0.524	0.545
	ADASYN	0.402	0.535	0.544
	Proposed	**0.502**	**0.557**	**0.560**
	G SMOTE	0.500	0.529	0.530
	Gaussian	0.498	0.551	0.553
Dataset 4	0% balancing	0.763	0.765	0.767
	SMOTE	0.836	0.838	0.839
	ADASYN	0.730	0.755	0.781
	Proposed	**0.864**	**0.889**	**0.916**
	G SMOTE	**0.864**	0.885	0.912
	Gaussian	0.852	0.879	0.907
Dataset 5	0% balancing	0.806	0.808	0.811
	SMOTE	0.859	0.860	0.861
	ADASYN	0.785	0.803	0.823
	Proposed	**0.884**	**0.898**	**0.913**
	G SMOTE	**0.884**	**0.889**	0.905
	Gaussian	0.868	0.883	0.900
Dataset 6	0% balancing	0.893	0.895	0.897
	SMOTE	0.903	0.904	0.905
	ADASYN	0.893	0.900	0.907
	Proposed	**0.922**	**0.914**	**0.906**
	G SMOTE	0.903	0.897	0.892
	Gaussian	0.898	0.891	0.885

Bold values indicate best results

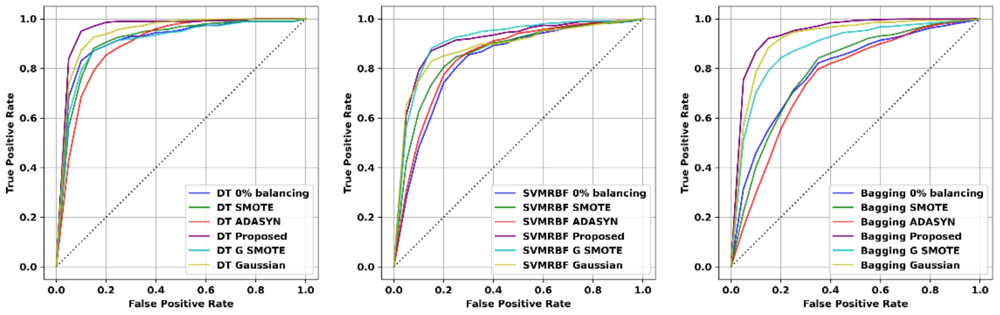
Table 4.5: Bagging results based on the evaluation metrics for all the datasets

Dataset	Method	Recall	G mean	AUC
Dataset 1	0% balancing	0.737	0.656	0.591
	SMOTE	0.631	0.605	0.581
	ADASYN	0.698	0.612	0.545
	Proposed	**0.875**	**0.904**	**0.934**
	G SMOTE	0.762	0.815	0.876
	Gaussian	**0.875**	0.896	0.928
Dataset 2	0% balancing	0.455	0.646	0.794
	SMOTE	0.724	0.745	0.786
	ADASYN	0.534	0.680	0.733
	Proposed	**0.776**	**0.849**	**0.853**
	G SMOTE	0.554	0.678	0.796
	Gaussian	0.651	0.801	0.802
Dataset 3	0% balancing	0.418	0.521	0.533
	SMOTE	0.416	0.513	0.524
	ADASYN	0.413	0.512	0.523
	Proposed	**0.480**	**0.537**	**0.540**
	G SMOTE	**0.480**	0.529	0.538
	Gaussian	0.437	0.523	0.539
Dataset 4	0% balancing	0.779	0.717	0.665
	SMOTE	0.700	0.679	0.660
	ADASYN	0.749	0.687	0.635
	Proposed	**0.887**	**0.905**	**0.925**
	G SMOTE	0.795	0.836	0.880
	Gaussian	0.872	0.894	0.917
Dataset 5	0% balancing	0.820	0.777	0.739
	SMOTE	0.768	0.753	0.738
	ADASYN	0.800	0.760	0.725
	Proposed	**0.900**	**0.907**	**0.915**
	G SMOTE	0.829	0.856	0.885
	Gaussian	0.877	0.891	0.906
Dataset 6	0% balancing	0.903	0.895	0.887
	SMOTE	0.905	0.900	0.895
	ADASYN	0.901	0.902	0.904
	Proposed	**0.924**	**0.910**	**0.896**
	G SMOTE	0.895	0.894	0.893
	Gaussian	0.886	0.884	0.883

Bold values indicate best results

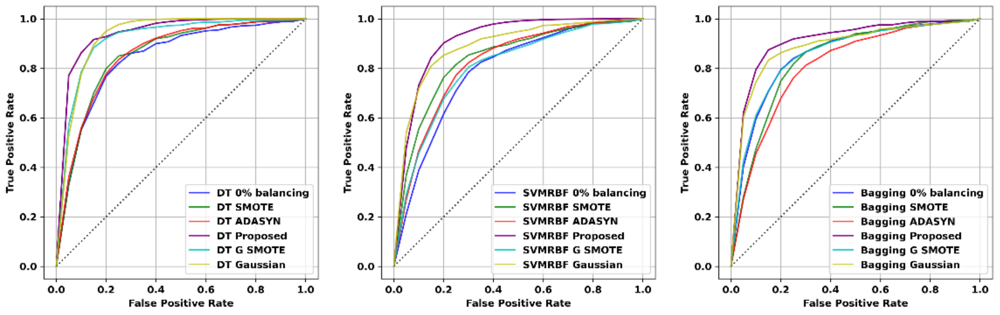
Tables 4.3–4.5 show that HEOMGA performs better than 0% balancing, SMOTE, ANDSYN, G SMOTE and Gaussian method in terms of Recall in all the used datasets. Therefore, an improvement in the churn rate is achieved by the proposed methods among the other used oversampling methods.

The bigger the AUC and G mean indicate, the better the methods. Empirical results indicated that the proposed method outperformed the tested oversampling methods in terms of G mean and AUC in the datasets. The proposed method for the three datasets obtained the best G mean and AUC values compared to other methods. This can be explained by the fact that the proposed method provided rich information to the learners, which improved prediction results and the learning process. The ROC graph computes the learner performance by changing the DTs' confidence level, SVM$_{rbf}$, and Bagging scores to get distinct values of TP$_{rate}$ and FP$_{rate}$ as shown in Figures. 4.6–4.11.



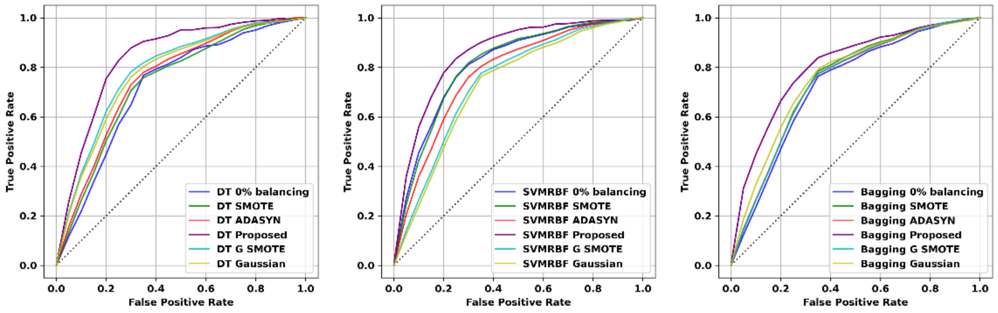
(a) (b) (c)

Figure 4.6: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 1 using **a** DTs, **b** SVM$_{rbf}$, and **c** Bagging



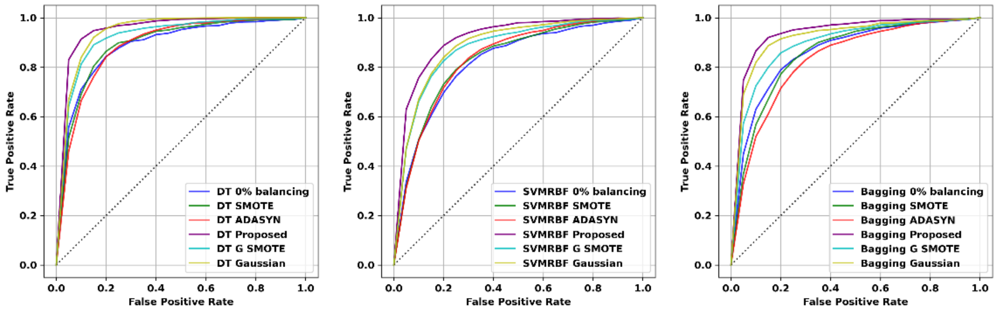
(a) (b) (c)

Figure. 4.7: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 2 using **a** DTs, **b** SVM$_{rbf}$, and **c** Bagging



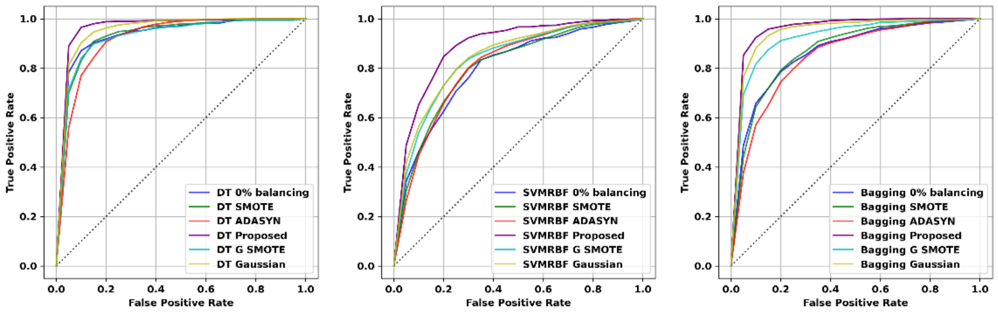
(a) (b) (c)

Figure. 4.8: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 3 using **a** DTs, **b** SVM$_{rbf}$, and **c** Bagging



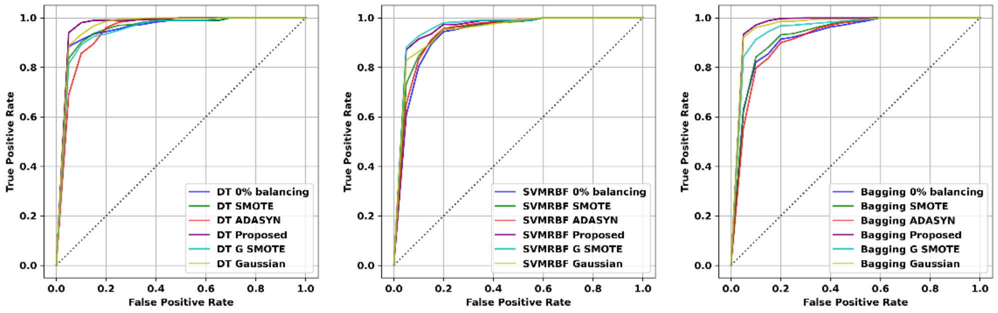
(a) (b) (c)

Figure. 4.9: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 4 using **a** DTs, **b** SVM$_{rbf}$, and **c** Bagging



(a) (b) (c)

Figure. 4.10: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 5 using **a** DTs, **b** SVM$_{\mathrm{rbf}}$, and **c** Bagging



(a) (b) (c)

Figure. 4.11: ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed, G SMOTE, and Gaussian for dataset 6 using **a** DTs, **b** SVM$_{\mathrm{rbf}}$, and **c** Bagging

To further check the statistical significance of the proposed method and whether it significantly outperforms the other used oversampling algorithms in terms of Recall, G mean, and AUC, the Wilcoxon signed-rank test [122] is performed. The results of the test are provided in Tables 4.6 – 4.8. The test's confidence level is set at 0.05, given the null hypothesis that the learners' performance varies significantly across the various algorithms and evaluation metrics with the proposed method as a control algorithm.

Table 4.6: Wilcoxon signed-rank test evaluation results based on Recall

Comparison	p value	W value	Mean diff.	R^+	R^-	Z-value	Mean (W)	Std (W)	Significance
Proposed vs. 0% balancing	0.05	10	0.50	455	10	− 4.5765	232.5	48.62	+
Proposed vs. SMOTE	0.05	1	− 0.10	464	1	4.7616	232.5	48.62	+
Proposed vs. ADASYN	0.05	0	0.04	465	0	− 4.7821	232.5	48.62	+
Proposed vs. G-SMOTE	0.05	0	− 0.13	435	0	− 4.7030	217.5	46.25	+
Proposed vs. Gaussian	0.05	0	− 0.13	406	0	− 4.6226	203.0	43.91	+

R^+ is the sum of ranks for the datasets in which the first method outperforms the second and R^- is the sum of ranks of the opposite, Std is the standard deviation (W), and + refers to significance at 0.05 level

Table 4.7: Wilcoxon signed-rank test evaluation results based on G mean

Comparison	p value	W value	Mean diff.	R^+	R^-	Z-value	Mean (W)	Std (W)	Significance
Proposed vs. 0% balancing	0.05	0	0.3	465	0	-4.7821	232.5	48.62	+
Proposed vs. SMOTE	0.05	1	-0.05	464	1	-4.7616	232.5	48.62	+
Proposed vs. ADASYN	0.05	1	0.03	464	1	-4.7616	232.5	48.62	+
Proposed vs. G-SMOTE	0.05	8.5	-0.15	426.5	8.5	-4.5192	217.5	46.25	+
Proposed vs. Gaussian	0.05	20	-0.14	445	20	-4.3708	232.5	48.62	+

Table 4.8: Wilcoxon signed-rank test evaluation results based on AUC

Comparison	p value	W value	Mean diff.	R^+	R^-	Z-value	Mean (W)	Std (W)	Significance
Proposed vs. 0% balancing	0.05	0	0.05	465	0	-4.7821	232.5	48.62	+
Proposed vs. SMOTE	0.05	0	-0.04	465	0	-4.7821	232.5	48.62	+
Proposed vs. ADASYN	0.05	0	0.04	465	0	-4.7821	232.5	48.62	+
Proposed vs. G-SMOTE	0.05	0	-0.15	435	0	-4.703	217.5	46.25	+
Proposed vs. Gaussian	0.05	0	-0.14	435	0	-4.703	217.5	46.25	+

The Recall, G mean, and AUC test results are given in Tables 4.6–4.8, validating that the proposed method significantly outperforms 0% balancing, SMOTE, ADASYN, G SMOTE, and Gaussian methods.

CCP significantly impacts business's revenue and long-term sustainability. Predicting churn accurately allows businesses to proactively engage at-risk customers, reducing customer loss and the costs associated with acquiring new customers, which can be up to five times more expensive than retaining existing ones. In highly competitive telecom industries, even a small improvement in retention can result in substantial financial gains. With class imbalance, common in churn datasets where churners are a minority, using proper metrics such as AUC-ROC, F1-score, and precision-recall curves is crucial to avoid misleading results from accuracy alone. Properly addressing class imbalance ensures that the model identifies churners effectively, maximizing the business's ability to intervene and retain valuable customers, ultimately driving profitability.

4.6 Summary

This work proposes a practical preprocessing approach, HEOMGA, to overcome class imbalance issues and assist the learners in improving their performance. This work has conducted experiments on publicly available customer churn prediction datasets to assess the proposed method's efficiency. Experimental results showed the efficiency of the proposed method than the other tested oversampling methods. In addition, the HEOMGA method also significantly outperformed the other oversampling methods in terms of Recall, G mean, and AUC by the Wilcoxon signed-rank test analysis.

Chapter 5 Boosting Ant Colony Optimization with Reptile Search Algorithm

5.1 Introduction

The rapid evolution in the telecommunication industry over time has increased the competition between the service providers in the market, resulting in severe revenue losses because of churning [123]. Churner customers refer to those who leave a service provider and develop a new relationship with another operator in the market. It was confirmed that attracting a new customer cost about five to six folds the cost of retaining an existing one [124]. For this reason, Telecommunication companies employ CRM as an integrated approach in their strategic plan to understand their customers' needs and reduce customer churn [125]. The customers' historical data are stored in such CRM systems and can be transformed into valuable information with the help of ML methods. The results from these techniques can assist these companies in formulating new policies, detecting customers who have a high tendency to end their relationship with the company, and developing retention strategies for the existing customers [126].

In ML techniques, data preprocessing is vitally essential, and Feature Selection (FS) is generally considered a foremost preprocessing step. FS techniques aim to determine the OFS by removing redundant and irrelevant features from high dimensional data without changing the original data representation. It has been proven that using FS in the ML learning process has several benefits [127, 128], such as reducing the amount of required data to achieve a good learning process. It improves prediction performance and minimizes CT.

FS techniques have been successfully applied in different applications and deliver promising results. Among these techniques, MAs have shown significant success in several applications such as vehicle routing [129], energy consumption [130] and fuzzy control design [131], e-commerce [132], medical diagnosis [133] and other, mainly because of their capability to provide high-quality OFS [134]. MAs utilize two search principles: exploration, where the algorithm investigates different candidate regions in the search space and exploitation, the algorithm searches around the obtained promising solutions to improve the existing ones.

According to [135], MAs can be grouped based on their behavior as (i) single solution-based algorithms and (ii) population-based algorithms. The first group exploits prior search knowledge to expand the search space in some promising environments; Tabu search [136], greedy randomized adaptive search procedure [137], and vector neighborhood search [138] are examples belonging to this group of algorithms. Population-based algorithms generate optimal solutions by exploring a new region in the search space via an iterative process for generating a new population through nature-inspired selection; GWO [139], cuckoo search algorithm [140], PSO [141], Firefly Algorithm (FFA) [152], crow search algorithm [143], dragonfly optimization algorithm [144], ACO [145], MVO [146], and RSA [147] are examples of the well-known MAs in this group.

In recent years, various researchers explored MAs for customer churn prediction. For example, in [148], a Customer Churn Prediction Business Intelligence using Text Analytics with Metaheuristic Optimization (CCPBI-TAMO) model is reported. It used Pigeon Inspired Optimization (PIO) to select OFS from a customer churn dataset collected from a business sector and used as inputs to Long Short-Term Memory (LSTM) with Stacked Auto Encoder (LSTM-SAE) model. Their proposed model outperformed other existing models. In [149], applied FFA for classification and FS using a huge publicly available dataset of churn prediction and the authors reported that FFA performed well for this application in both classification and FS. The potential of ACO to predict customer churn is discussed in [150]. The results reported that the ACO attained an effective performance compared to other MAs. In [151], combined Multi-Objective Cost-Sensitive ACO (MOCS-ACO) with Genetic Algorithm (GA) is reported to enhance the classification results. The GA was employed to select OFS, while MOCS-ACO was used as a classification model. Experimental results reported that the model performed well when validated on a customer churn prediction dataset from a company in Turkey. In [152], ACO identified OFS and fed identified features to the Gradient Boosting Tree model that predicts churners. The results reported that the proposed PSO-GBT models achieved good results.

In [153], using a public dataset, PSO is used to choose the OFS for ELM for churn prediction. In [154], fuzzy PSO is employed to determine the OFS from two publicly available churn prediction datasets and then use the selected features from the fuzzy PSO in their model. In [155], a hybrid model based on PSO and feed-forward neural networks for churn prediction is reported to select OFS from one public and private datasets. In [156], three variants of PSO are reported for churn prediction using a public dataset. These variants comprised PSO is incorporated with FS as a preprocessing, PSO is embedded with Simulated Annealing (SA), and PSO is combined with FS and SA.

All these studies have reported promising results for using MAs to select the most informative features in churn prediction. Although most of these efforts have used MAs to select OFS, a quantitative analysis of methods' capabilities in accuracy, number of features in OFS, fitness values, and CT in this application is not reported. Thus, further work is needed to propose new MAs for FS in this application. Also, most of these works are limited to using individual MA and combination MAs to produce a hybrid FS method for this application, which needs to be investigated. Whereas selecting OFS in this application is very important for reliable and safe predictions to the customers who will end the relationship and develop a new one with another competitor. Motivated by these limitations, we proposed a new metaheuristic-based approach called ACO-RSA that combines standard ACO and RSA in a serial collaborative manner to find the most appropriate features for churn prediction. The comparison with five MAs, including PSO, MVO, GWO, ACO, and RSA, validates the effectiveness of the ACO-RSA approach. The contributions can be summarized as follows:

- A new metaheuristic-based approach, namely ACO-RSA, is proposed for churn prediction.
- The standard ACO and RSA are combined in a serial collaborative mechanism to achieve an exploration-exploitation balance in the proposed ACO-RSA and avoid being stuck in local optima.
- Six publicly available benchmark customer churn datasets with different records and features are investigated to check the stability of ACO-RSA performance.
- We also investigate convergence behavior, statistical significance, and the exploration-exploitation balance of proposed ACO-RSA against competitor MAs.

A brief overview of ACO and RSA is provided in the next section, followed by detailed explanations of the suggested ACO-RSA approach. The experimental results are discussed in section 5.4. Eventually, conclusions are noted in section 5.5.

5.2 Ant Colony Optimization (ACO)

ACO is a nature-inspired MA that mimics the ants' searching process for food sources [145]. The characteristics of ACO make the model more sensible than other MAs as it supports parallel processing while avoiding process dependency, and it gives feedback on the ants' behaviors in the search space [157]. Ants are not blind when searching for food; they find the shortest route between their nest and the food source. While moving, ants deposit a chemical material, a pheromone, along their trail. The pheromone is a medium for communication between the ants and represents the shortest path to collect food. The ants move towards the food by sensing the pheromone deposition by the ants that have previously traveled the path, subsequently increasing the probability of other ants traversing via the same path as shown in Figure 5.1.

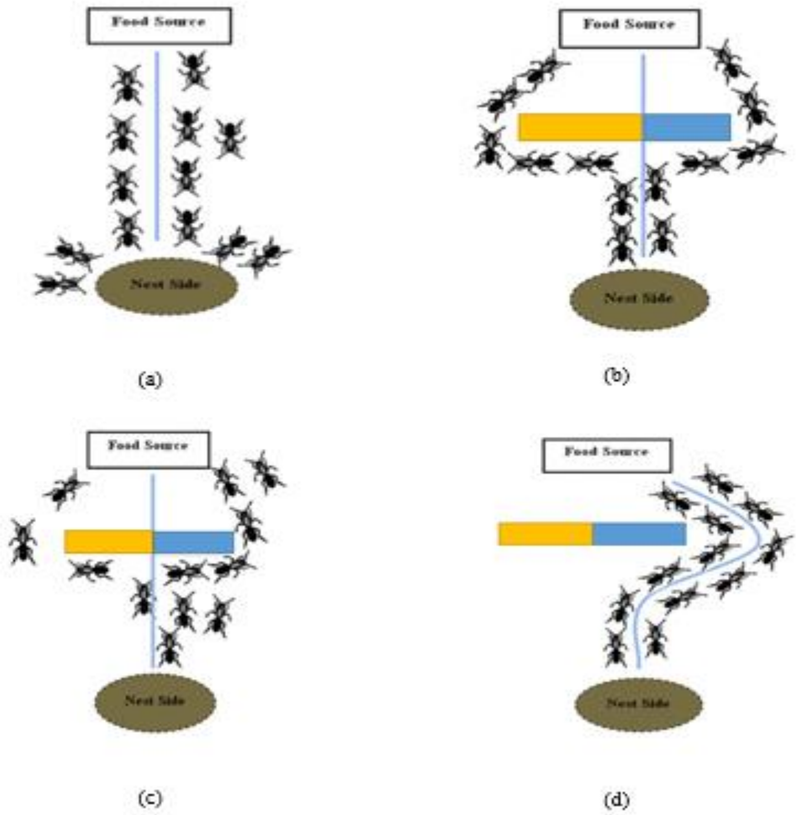


Figure 5.1. Process of path finding for ants: (a) original path between the nest and the food source, (b) path after introducing an obstacle having a larger yellow side than blue, (c) when pheromone deposition on the blue side increases cumulatively, and (d) converged to the shortest path.

ACO uses pheromone trails and heuristic information to make the probabilistic decision. The ants update the pheromone level at any feature as they traverse a path. The more ants traverse a feature, the more pheromone deposition at that feature, resulting in a higher probability of the feature being part of the shortest path. The maximum number of ants will follow the path with the higher pheromone level and be the shortest. The pheromone value $\tau_0 = 1$ is initialized at all M features, and ants are positioned randomly on a set of features with a predefined maximum number of generations T. At every generation g, the transition probability $TP_i^k(g)$ of kth ant at ith feature is shown below [158]:

$$TP_i^k(g) = \begin{cases} \dfrac{[\tau_i(g)]^\alpha [\eta_i]^\beta}{\sum_{j \in j_i^k} [\tau_j(g)]^\alpha [\eta_j]^\beta} & \text{if } j \in j_i^k \\ 0, & otherwise \end{cases} \tag{1}$$

where j_i^k is set of possible neighbors of ith features that the kth ant does not visit. The relative importance of pheromone level and heuristic information for ants' movements are specified by non-negative parameters α and β, respectively.

After choosing the next feature in the ant's path, a Fitness Function (FF) is employed to quantify the new set of selected features. The movement of kth ant is stopped if the improvement in the fitness value is not attained after adding any new feature [145]. If stopping criteria is not reached, the amount of pheromone level at next generation (g+1) at ith feature is updated as [159]:

$$\tau_i(g+1) = (1-p)\tau_i(g) + \sum_{k=1}^{N} \Delta\tau_i^k(g) \tag{2}$$

where,

$$\Delta\tau_i^k(g) = \begin{cases} FF\left(S^k(g)\right) / \left|S^k(g)\right|, & if\ i \in S^k(g) \\ 0, & otherwise \end{cases} \tag{3}$$

where p is the pheromone decay rate ($0 \leq p \leq 1$), N is the number of ants, $\left|S^k(g)\right|$ presents a number of the selected features, and $\Delta\tau_i^k$ represents the pheromone deposited by kth ant if ith feature is in the shortest path of the ant; otherwise, it is 0. The stopping criteria are achieved when g reaches the predefined maximum T. The features with the highest pheromone level and smallest fitness value will be selected as an OFS. Figure 5.2 shows the overall process of the ACO.

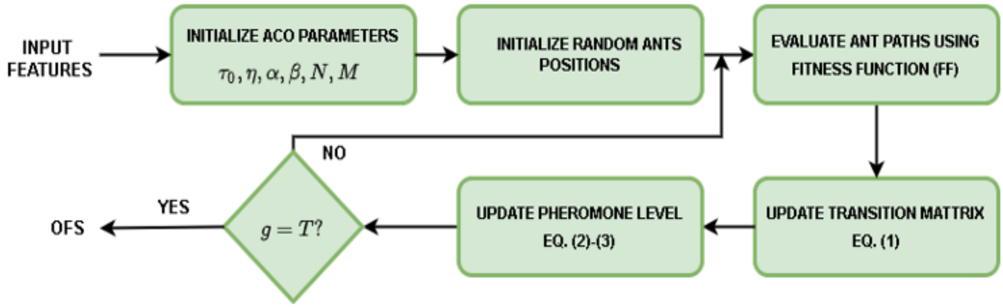

INPUT FEATURES
INITIALIZE ACO PARAMETERS $\tau_0, \eta, \alpha, \beta, N, M$
INITIALIZE RANDOM ANTS POSITIONS
EVALUATE ANT PATHS USING FITNESS FUNCTION (FF)
UPDATE TRANSITION MATTRIX EQ. (1)
UPDATE PHEROMONE LEVEL EQ. (2)-(3)
$g = T?$
NO
YES
OFS

Figure 5.2: The flow diagram of the ACO

5.3 Reptile Search Algorithm *(RSA)*

Reptile Search Algorithm (RSA) is another nature-inspired MA proposed by [147] in 2021 to simulate Crocodiles' encircling and hunting behavior. It is a gradient-free algorithm that starts by generating random solutions as follows:

$$x_{i,j} = rand_{\in[0,1]} * \left(UB_j - LB_j\right) + LB_j \qquad for\ \ i \in \{1, \dots, N\}\ and\ j \in \{1, \dots, M\} \tag{4}$$

where, $x_{i,j}$ is the *i*th solution for *j*th input feature for total *N* solutions comprising *M* features, $rand_{\in[0,1]}$ is a random number distributed uniformly in the range [0, 1], and the *j*th feature has upper UB_j and lower LB_j boundaries.

Like the other nature-inspired MAs, RSA can be understood in two principles: exploration and exploitation. These principles are facilitated by Crocodile movement while encircling the target prey. Total iterations of RSA are divided into four stages to take advantage of the natural behavior of Crocodiles. In the first two stages, RSA achieves the exploration based on the encircling behavior comprising the high and the belly walking movements. Crocodiles begin their encircling to search the region, facilitating a more exhaustive search of the solution space. This behavior can be mathematically modeled as follows:

$$x_{i,j}(g+1) = \begin{cases} \left[-n_{i,j}(g)\,.\,\gamma\,.\,Best_j(g)\right] - \left[rand_{\in[1,N]}\,.\,R_{i,j}(g)\right], & g \leq \dfrac{T}{4} \\ ES(g)\,.\,Best_j(g)\,.\,x_{(rand_{\in[1,N]},j)}, & g \leq \dfrac{2T}{4}\ \ and\ g > \dfrac{T}{4} \end{cases} \tag{5}$$

where, $Best_j(g)$ is the best solution for *j*th feature, $n_{i,j}$ refers to the hunting operator for the *j*th feature in the *i*th solution (calculated as in Eq. (6)), parameter γ controls the exploration accuracy throughout the length of iterations and is set as 0.1. The reduce function $R_{i,j}$ is used to reduce the search region and is computed as in Eq. (9), $rand_{\in[1,N]}$ is a number between 1 to *N* used to select one of the possible candidate solutions randomly, and Evolutionary Sense $ES(g)$ stands for the probability ratio reducing from 2 to −2 over iterations, calculated as in Eq. (10).

$$n_{i,j} = Best_j(g)\ \times P_{i,j} \tag{6}$$

where, $P_{i,j}$ indicates the percentage difference between the *j*th value of the best solution to its corresponding value in the current solution and is calculated as:

$$P_{i,j} = \theta + \frac{x_{i,j} - M(x_i)}{Best_j(g) \times \left(UB_j - LB_j\right) + \epsilon} \tag{7}$$

where θ denotes a sensitive parameter that controls the exploration performance, ϵ is a small floor value, and $M(x_i)$ refers to the average solutions and is defined as:

$$M(x_i) = \frac{1}{n}\sum\nolimits_{j=1}^{n} x_{i,j} \tag{8}$$

$$R_{i,j} = \frac{Best_j(g) - x_{(rand_{\in[1,N]},j)}}{Best_j(g) + \epsilon} \tag{9}$$

$$ES(g) = 2 \times rand_{\in[-1,1]} \times (1 - \frac{1}{T}) \tag{10}$$

where the value 2 acts as a multiplier to provide correlation values in the range [0, 2], and $rand_{\in[-1,1]}$ is a random integer number between [−1, 1].

RSA implements the exploitation (hunting) in the last two stages to search feature space for optimal solutions using hunting coordination and cooperation. The solution can update its value during the exploitation using the following equation:

$$x_{i,j}(g+1) = \begin{cases} rand_{\in[-1,1]} \,.\, Best_j(g) \,.\, P_{i,j}(g), & g \leq \frac{3T}{4} \; and \; g > \frac{2T}{4} \\ [\epsilon \,.\, Best_j(g) \,.\, n_{i,j}(g)] - [rand_{\in[-1,1]} \,.\, R_{i,j}(g)], & g \leq T \;\; and \; g > \frac{3T}{4} \end{cases} \tag{11}$$

The quality of candidate solutions at each iteration is measured using the predefined FF, the algorithm stops after T iteration, and a candidate solution with the least fitness value is selected as OFS. The process of the RSA is shown in Figure 5.3.

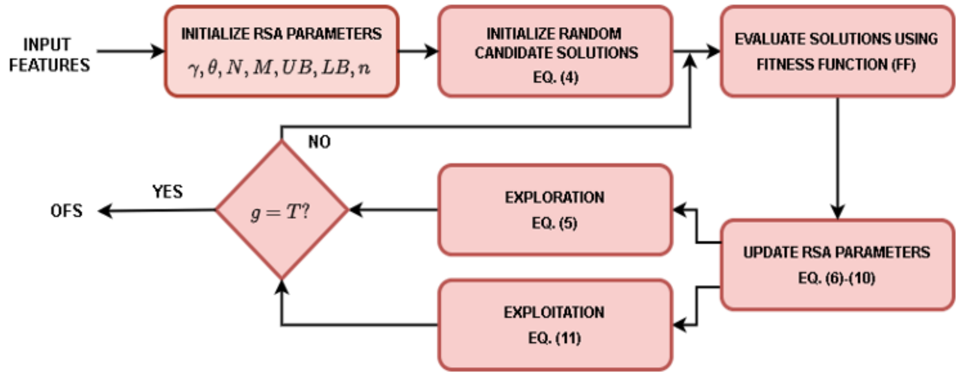

INPUT FEATURES
INITIALIZE RSA PARAMETERS
$\gamma, \theta, N, M, UB, LB, n$
INITIALIZE RANDOM CANDIDATE SOLUTIONS
EQ. (4)
EVALUATE SOLUTIONS USING FITNESS FUNCTION (FF)
NO
OFS
YES
$g = T?$
EXPLORATION
EQ. (5)
UPDATE RSA PARAMETERS
EQ. (6)-(10)
EXPLOITATION
EQ. (11)

Figure 5.3: The flow diagram of the RSA

5.4 ACO-RSA based FS

In ACO, a path with the highest pheromone level is the shortest path to transport the food from the source to the nest. Most ants will follow this path unless there is some obstruction and might limit ACO from exploring the quality of existing solutions by searching only within the current search space [160]. High exploration in MA reduces the quality of the optimum solutions, and fast exploitation prevents the algorithm from finding global optimum solutions. RSA is the most recent MA, which shows superiority in solving several engineering problems and has an excellent exploration capability. It has an inbuilt exploration-exploitation balance significantly enhances its performance [161]. Different MAs can be combined effectively to use the algorithm's merits while maintaining exploration-exploitation balance and avoiding premature convergence in local optima.

According to [162], hybrid MAs have several ways; the High-level Relay Hybrid (HRH) strategy is one of these methods. In HRH, two MAs can be executed in homogenous (i.e., same algorithms) or heterogeneous (i.e., different algorithms) sequences. The proposed ACO-RSA method uses the heterogeneous HRH strategy to achieve exploitation-exploration balance as in RSA, with high exploitation as in ACO. Figure 5,4 illustrates the overall process of the ACO-RSA approach. At first, ACO, RSA, and shared parameters are initialized. A random number uniformly distributed in the range [-1, 1] initializes N candidate solutions $\left\{\mathrm{x}_{i,j} \in \mathbf{X}(0) \mid 1 \geq i \geq N \text{ and } 1 \geq j \geq M\right\}$ each for M-dimensional feature vectors. Then FF evaluates candidate solutions to judge the enhancement by comparing current solutions with those obtained in the previous iteration. If the current solution is better than the previous solution, it will be accepted; else rejected.

The threshold used to convert MAs candidate solutions during the searching process for the informative features into binary vectors is set to 0.5, as recommended by [163], to produce a small number of features. K-Nearest Neighbor (KNN) is a widely used classifier due to its simple, fast, and flexible working even in the presence of noisy data [164, 165]. KNN with a Euclidean distance measure (k = 5) is employed as the classifier. Hence, the FF is considered to achieve dimensionality reduction (minimizing the number of the selected OFS) and maximum accuracy (reducing classification error). Therefore, it is defined using the following equation:

$$FF = \gamma \times \left(1 - \frac{N_c}{N}\right) + \beta \times \frac{d_i}{M} \tag{12}$$

where, γ and β are weighted factors that vary in the range of [0, 1] (subject to $\gamma + \beta = 1$) to balance the number of features in OFS d_i out of M features in the original dataset and the is the number of correctly classified instances $N_c$out of total N instances in the original dataset by the KNN classifier. Each feature in the OFS follows:

$$d_i = \begin{cases} 1 & if\ x_i > 0.5 \\ 0 & otherwise \end{cases} \tag{13}$$

Then the best solution is determined and the current solution $\mathbf{X}(0)$ are assigned to the candidate solutions of ACO. In addition, the ACO starts with assigning each candidate solution $\mathrm{x}_{i,j}$ as an initial path for an ant in the colony. An ith ant initially traverses a subset of features initialized with a pheromone value $\mathrm{x}_{\mathrm{i,j}}$ greater than 0.5. ACO updated candidate solutions $\mathbf{X}^{\mathrm{new}}$ according to Eq. (1)– (3). The FF evaluates the enhancement in the candidate solutions. A candidate solution is updated only if the fitness value for the solution has decreased after the update according to the following equation:

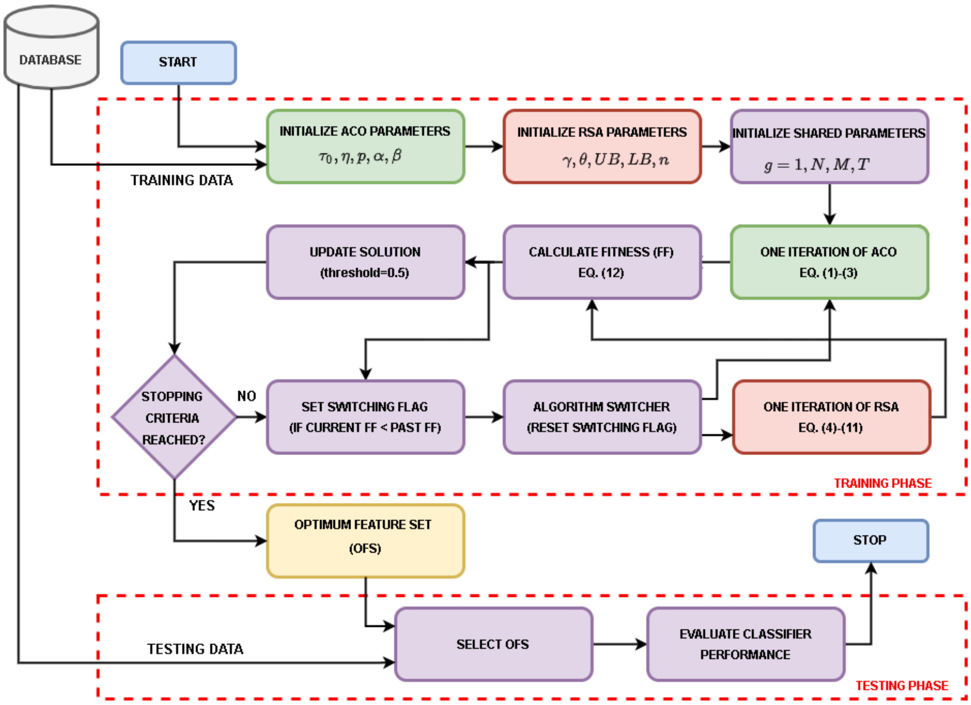


Figure 5.4: The flow diagram of the ACO-RSA FS approach.

$$\mathrm{x}_i(g+1) = \begin{cases} \mathrm{x}_\mathrm{i}^{\mathrm{new}}(g)\,, & if\ \mathrm{FF}\big(\mathrm{x}_i(g)\big) > \mathrm{FF}\big(\mathrm{x}_i(g+1)\big) \\ \mathrm{x}_i(g)\,, & else \end{cases} \tag{14}$$

In the next iteration, the set of candidate solutions $\mathbf{X}(g+1)$ are given as initial candidate solutions (after thresholding) to either ACO or RSA to extend the searching process in other promising regions in the feature space. If the least FF value in the current iteration is smaller than the smallest FF value in the previous iteration ($min(\mathrm{FF}(x_i)|\mathrm{x}_i \in \boldsymbol{X}(g)) < min(\mathrm{FF}(\mathrm{x_i})|\mathrm{x}_i \in \boldsymbol{X}(g-1))$), the same algorithm continues in the next iteration; otherwise, an algorithm switching flag is set to switch between the two algorithms. The main goal behind the switching between the two algorithms is that if the ACO cannot improve the candidate solutions, it might get stuck in local optima. At this point, RSA moves the candidate solutions into another search region using Eq. (4)– (11) to find some better solutions. This process repeats until the maximum iteration T is reached. A candidate solution with the smallest FF value $\{min(FF(\mathrm{x}_i) \mid \mathrm{x}_i \in \mathbf{X}(T)\}$ is used to extract OFS. During the testing phase, a reduced feature set is obtained by filtering only the selected features present in OFS. This OFS is used to evaluate classifier performance metrics, as discussed later in section 4.3. The steps of ACO-RSA are shown in Algorithm 5.1.

Algorithm 5.1: Proposed ACO-RSA approach

```
Form mutually exclusive and exhaustive training and testing subsets.
    Training Phase
Load training dataset
Initialize ACO parameters $\tau_0, \eta, p, \alpha, \beta$
Initialize RSA parameters $\gamma, \theta, UB, LB, n$
Initialize shared parameters $N, M, T$
for g = 1 to T do
    if first iteration
        Perform one iteration of ACO using Eq. (1)– (3)
    else
        if switch flag = 1
            Perform one iteration of an alternate algorithm that was not executed in the previous iteration
                ACO: Eq. (1)– (3) or RSA: Eq. (4)– (11)
            switch flag = 0
        else
            Continue the same algorithm as in the previous iteration
                ACO: Eq. (1)– (3) or RSA: Eq. (4)– (11)
        end if
    end if
    Evaluate fitness function (FF) using Eq. (13) for updated candidate solutions
    Update candidate solutions using Eq. (12) and a threshold of 0.5
    if min($FF_{new}$) < min($FF_{old}$)
        switch flag = 1
    end if
end for
Extract OFS by applying a threshold of 0.5 to a candidate solution with the smallest FF.
    Testing Phase
Load testing dataset
Select only optimum features as described in OFS
Evaluate performance using KNN classifier
```

5.5 Experiments results

This section provides the experiments performed to assess the ACO-RSA and compare its performance with PSO [141], MVO [146], GWO [139], ACO [145], and RSA [147], for FS on seven datasets.

5.5.1 Experimental setup

All the experiments are implemented using Python and executed on a 3.13 GHz PC with 16 GB RAM and Windows 10 operating system. The performance of the proposed ACO-RSA is validated by conducting experiments on publicly available benchmark datasets for customer churn. The characteristics of these datasets are provided in Chapter 3 in Table 3.1 above. It shows the number of classes, the number of features, the number of instances, and the dataset source. Each dataset is divided randomly into the ratio of 50 % as a training set and the remaining as a test set.

The ACO-RSA approach is examined with several well-known MAs, and these algorithms include PSO, MVO, GWO, ACO, and RSA. Parameter settings play a critical role in enhancing the performance of MAs. For all MAs, the population of 20 and the maximum iterations of 50 are selected empirically. The number of independent runs is 20 to calculate statically significant inference. In addition, the default parameter settings of each comparative MA are defined according to its implementations, and they are presented in Table 5.1.

Table 5.1: Parameters settings

Algorithm	Parameters
PSO	Individual acceleration factor (c_1) increases, global acceleration factor (c_2) decreases linearly in range [0.5–2.5], and inertia weight linearly decreases in range [0.9–0.4]
MVO	WEP_{min}(Wormhole Existence Probability) = 0.2, WEP_{max} = 1, p = 6, variable (α) linearly decreases in range [2–0]
GWO	Variable (α) linearly decreases in range [2–0], variable (C) is a random value in the range [0,2], variable (A) decreases linearly from 1 to -1
ACO	$\tau_0 = 1, p = 0.95, \alpha = 1.2, \beta = 0.5$
RSA	UB and LB vary according to feature in the dataset, $\gamma = 0.9, \theta = 0.5$

5.5.2 Evaluation measures

In order to assess the reliability and performance of the ACO-RSA approach against the other comparative MAs, a set of evaluation measures including accuracy, fitness function, number of selected OFS, and computational time are used [149]–[151].

Average accuracy (Avg_{ACC}): It calculates the average accuracy for all runs. The proposed ACO-RSA and the other MAs are executed 20 times ($N_r = 20$):

$$\text{Avg}_{ACC} = \frac{1}{N_r}\sum_{k=1}^{N_r} ACC_{best}^{k} \tag{15}$$

Average Fitness Function (Avg_{FitF}): This metric quantifies the performance of the proposed ACO-RSA and the other MAs, which puts the relationship between maximizing classification accuracy and minimizing the number of the selected OFS, its average is computed by using the following:

$$AvgFitF = \frac{1}{N_r}\sum_{k=1}^{N_r} FitF_{best}^{k} \tag{16}$$

Average OFS (Avg_{ofs}): It represents the average number of the selected OFS to the total number of features in each dataset (D) at run number *i*:

$$\text{Avg}_{ofs} = \frac{1}{N_r}\sum_{k=1}^{N_r} \frac{d_i}{D} \tag{17}$$

Average Computational Time (Avg_{CT}): It measures the average CPU time in seconds for the proposed ACO-RSA and the other MAs at the run number i:

$$\text{Avg}_{CT} = \frac{1}{N_r}\sum_{k=1}^{N_r} CT_i \tag{18}$$

5.6 Results and analysis

In this subsection, the performance results of ACO-RSA and the comparative MAs are demonstrated not only using the performance measurements in subsection 4.3, but also based on the convergence behavior, boxplot graphs, statistical analysis and exploration and exploitation effects.

5.6.1 Performance results

The performance of the ACO-RSA and the other MAs on the seven open-source customer churn datasets are given in Tables 5.2 – 5.5. Each MA is executed 20 times independently to obtain statistically reliable analysis and conclusions. Table 5.2 compares all the algorithms regarding the average (Avg) testing accuracy and the number of OFS. Table 5.3 reports the best and worst fitness values obtained by the ACO-RSA and other MAs, while the Avg and standard deviation (Std) of the fitness values are summarized in Table 5.4. The average CT in seconds for the ACO-RSA and other MAs on all seven datasets are provided in Table 5.5.

The testing accuracy varies in the range 0-1; 0 means a total misdetection, while 1 means a perfect detection. The number of features in OFS varies from 1 to the total number of features in the respective datasets. A good MA should maximize classification accuracy and minimize the number of selected OFS. In Table 3, the ACO-RSA gained better accuracy than other MAs on five out of seven datasets. Comparing OFS for each dataset, ACO-RSA required fewer informative features than the other MAs. This proves the capability of ACO-RSA in reducing the selected OFS while obtaining a higher accuracy result.

Table 5.2: The average results obtained by different algorithms in terms of the accuracy and OFS

Dataset	Metric	PSO	MVO	GWO	ACO	RSA	ACO-RSA
Dataset 1	ACC	0.8963	0.8836	0.8842	0.8434	0.8989	**0.9036**
	OFS	12	9	10	8	8	**4**
Dataset 2	ACC	0.8312	0.8127	0.8312	0.8084	0.8319	**0.8330**
	OFS	6	6	5	**4**	5	**4**
Dataset 3	ACC	0.6910	0.6906	0.6907	0.6893	0.6922	**0.6923**
	OFS	35	32	28	27	26	**25**
Dataset 4	ACC	**0.5586**	0.5534	0.5468	0.5291	0.5519	0.5538
	OFS	45	42	38	38	**15**	**15**
Dataset 5	ACC	0.9008	0.8724	0.8948	0.8484	**0.9052**	0.9047
	OFS	5	5	6	5	4	**3**
Dataset 6	ACC	0.9361	0.8646	0.9185	0.8437	0.9382	**0.9471**
	OFS	10	9	8	7	7	**4**
Dataset 7	ACC	0.9385	0.8654	0.9248	0.8563	0.9342	**0.9390**
	OFS	4	4	4	3	3	**2**

The fitness value is a singular measure that varies from 0 to 1, with a preference for a value closer to 0 (an ideal value that cannot be achieved), indicating better detection with fewer features. As seen in Table 5.3, the best fitness value for the ACO-RSA arrived at the minimum value in five out of seven datasets, while the worst fitness value for ACO-RSA is the smallest in six datasets. Although RSA scored the smallest fitness value in datasets 2 and 6, the ACO-RSA has better testing accuracy than the standard RSA. Similarly, the PSO achieved the smallest worst-case fitness for dataset 7, but Table 5.3 confirms a slightly superior performance of ACO-RSA than the PSO. The

ACO-RSA and standard RSA obtained the first and second rank in the best and worst fitness value range, respectively.

Table 5.3: The best and worst fitness values of the ACO-RSA and the other MAs

Dataset	Metric	PSO	MVO	GWO	ACO	RSA	ACO-RSA
Dataset 1	Best	0.0827	0.0797	0.0906	0.1773	0.0788	**0.0746**
	Worst	0.1008	0.1006	0.1175	0.1952	0.1050	**0.0958**
Dataset 2	Best	0.1426	0.1435	0.1462	0.1446	**0.1399**	0.1420
	Worst	0.1495	0.1534	0.1544	0.1572	0.1532	**0.1489**
Dataset 3	Best	0.2803	0.2785	0.2796	0.3187	0.2411	**0.2386**
	Worst	0.2894	0.2889	0.2918	0.3287	0.2877	**0.2837**
Dataset 4	Best	0.4239	0.4208	0.4284	0.4286	0.4190	**0.4179**
	Worst	0.4356	0.4337	0.4392	0.4468	0.4295	**0.4326**
Dataset 5	Best	0.0755	0.0804	0.0818	0. 0915	0.0745	**0.0728**
	Worst	0.0904	0.0927	0.1046	0.1737	0.0966	**0.0894**
Dataset 6	Best	0.0347	0.0382	0.0438	0.1002	**0.0291**	0.0368
	Worst	0.0908	0.0599	0.0465	0.1866	0.0657	**0.0501**
Dataset 7	Best	0.0609	0.0636	0.0669	0.1219	0.0615	**0.0605**
	Worst	**0.0647**	0.0704	0.0805	0.0997	0.0737	0.0651

Table 5.4 provides the Avg and Std of the fitness values for all the MAs and datasets over 20 independent runs. A good MA should have a smaller Avg and Std of fitness values to signify the stability and consistency of the MAs. As shown in Table 5.4, the ACO-RSA has the smallest Avg fitness value in six out of seven datasets and the smallest Std in five. The PSO and MVO have the least Avg and Std for datasets 5 and 7, respectively. It is evident from Table 5.4 that the ACO-RSA approach is better than the other comparative algorithms. Although the Std values of the PSO and the standard RSA are smaller than those corresponding to ACO-RSA for datasets 3 and 4, the Avg fitness values of ACO-RSA are slightly smaller in both cases.

Table 5.4: The Avg and Std of fitness values of the ACO-RSA and the other MAs

Dataset	Metric	PSO	MVO	GWO	ACO	RSA	ACO-RSA
Dataset 1	Avg	0.0928	0.0896	0.1032	0.1140	0.0914	**0.0877**
	Std	0.0057	0.0064	0.0054	**0.0041**	0.0081	**0.0041**
Dataset 2	Avg	0.1452	0.1491	0.1454	0.1478	0.1475	**0.1436**
	Std	0.0021	0.0030	0.0022	0.0032	0.0035	**0.0016**
Dataset 3	Avg	0.2853	0.2849	0.2872	0.3241	0.2684	**0.2496**
	Std	**0.0023**	0.0024	0.0033	0.0029	0.0155	0.0125

Dataset 4	Avg	0.4287	0.4299	0.4344	0.4380	0.4257	**0.4248**
	Std	0.0029	0.0027	0.0035	0.0055	**0.0024**	0.0040
Dataset 5	Avg	**0.0812**	0.0876	0.0935	0.1186	0.0831	0.0814
	Std	0.0048	0.0043	0.0062	0.0086	0.0069	**0.0038**
Dataset 6	Avg	0.0429	0.0525	0.0722	0.1067	0.0460	**0.0409**
	Std	0.0043	0.0064	0.0140	0.0126	0.0082	**0.0038**
Dataset 7	Avg	0.0659	**0.0631**	0.0771	0.1040	0.0670	0.0635
	Std	0.0011	0.0017	0.0073	0.0186	0.0032	**0.0010**

The number of features in each dataset and its size (i.e., samples) affecting the CT. For instance, with more features and size, as in datasets 3, 4, and 7, the algorithms take more CT to find OFS. Table 6 provides the Avg results in terms of CT. As per Table 5.5, the standard RSA gets the smallest CT, followed by the ACO-RSA, to finish a job compared with other MAs. For small datasets, the difference in CT between the standard RSA and the proposed ACO-RSA is not significant, while the CT increases to a significant value for large datasets. It should be noted that CT is essential only during the training phase for practical applications and is independent of the FS algorithm in the testing phase. Hence, ACO-RSA would still be suitable for most real-time implementations of the application of churn prediction.

Table 5.5: The Avg CT of the ACO-RSA and the other MAs

Dataset	**PSO**	**MVO**	**GW**	**ACO**	**RSA**	**ACO-RSA**
Dataset 1	26.6507	26.6087	28.061	25.8793	16.0972	16.4592
Dataset 2	97.3320	94.7404	94.791	93.7622	45.1266	46.4270
Dataset 3	472.8719	469.9067	473.323	474.1874	143.6752	175.6912
Dataset 4	1062.6525	1080.8391	1060.753	1060.2647	316.8873	351.1655
Dataset 5	22.6211	23.6389	22.635	22.6300	11.7968	14.4268
Dataset 6	22.3199	22.9185	22.085	21.7553	15.2821	16.0060
Dataset 7	355.7554	441.0723	424.725	1498.0671	202.5721	268.2194

Figure 5.5 presents the switching behavior of the proposed ACO-RSA during fifty exploitation-exploration iterations for all seven datasets. The total number of iterations for ACO and RSA are displayed in each switching behavior in the last column in Figure 5.4. In datasets 3 and 4, ACO uses slightly more iterations than RSA to exploit many features. The iterative design of ACO requires more than one iteration to build confidence in the estimated shortest path. Hence, Table 5.6 shows a significantly higher

CT for these datasets. On the other hand, datasets 1, 2, 5, and 6 have comparatively fewer features, and therefore, more iterations are used by the RSA, resulting in a very close CT to the one obtained using the proposed ACO-RSA. For dataset 7, a very high number of training examples causes a larger delay for each iteration of ACO. This results in a significant impact on the CT of the proposed ACO-RSA than the fastest RSA algorithm. The most informative features that are selected by the ACO-RSA are provided in Table 5.6.

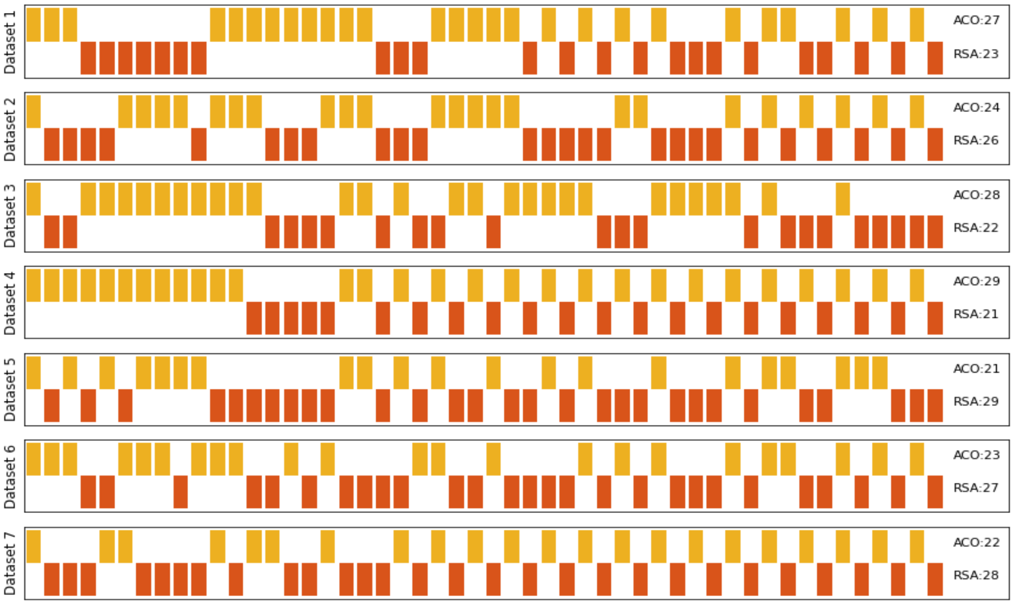


Figure 5.5: Switching behavior of proposed ACO-RSA for sample runs using all seven datasets

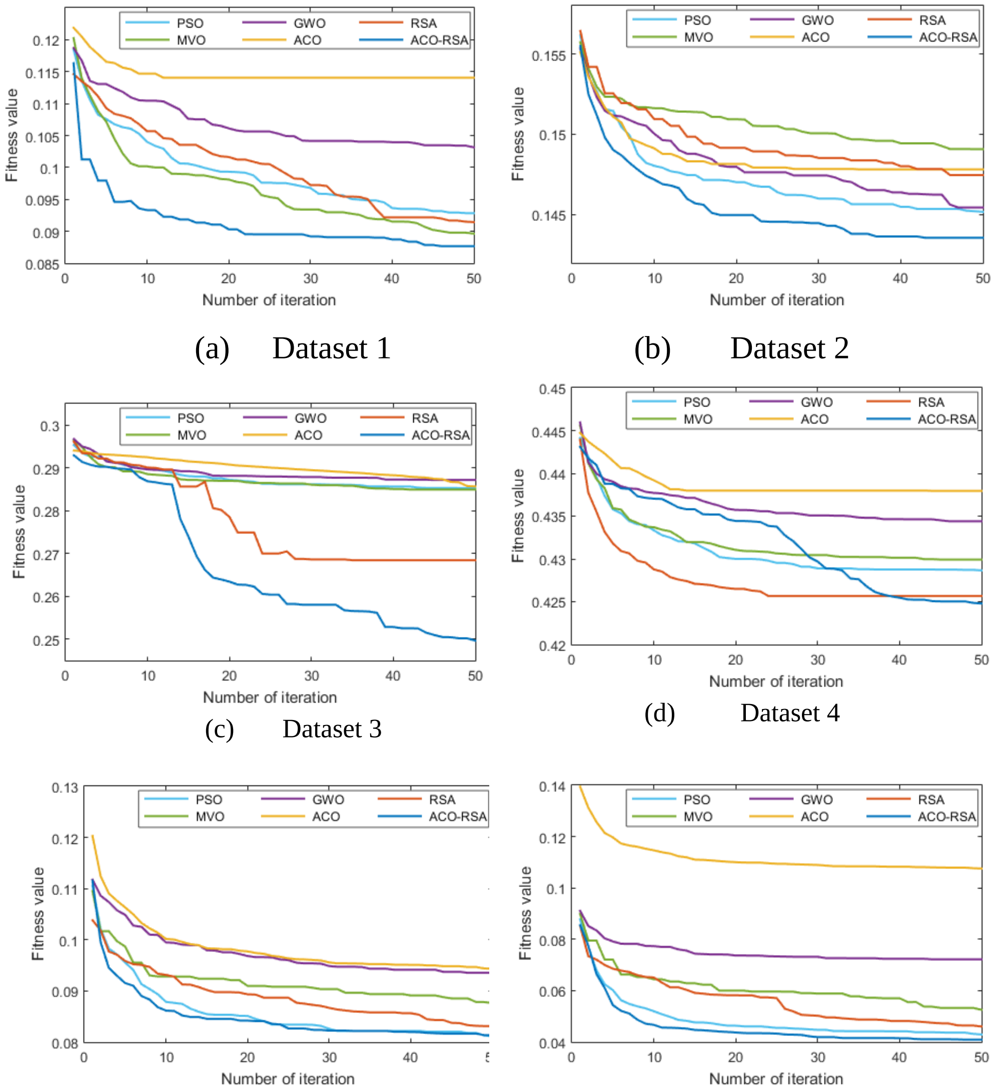
Table 5.6: Selected features by the ACO-RSA

Dataset	Selected features
Dataset 1	Total day calls, Number customer service calls, Total intl calls, Total night calls
Dataset 2	Tenure, PhoneService, InternetService, TotalCharges
Dataset 3	mou_Mean, blck_dat_Mean, mouowylisv_Mean, mou_peav_Mean, opk_vce_Mean, mou_opkv_Mean, mou_opkd_Mean, dualband, phones, ownrent, dwlltype, marital, forgntvl, ethnic, kidSelected
Dataset 4	Call Failure, Subscription Length, DayCalls
Dataset 5	Call Failure, Subscription Length, Frequency of SMS, Foreign Phone
Dataset 6	subscription_age, service_failure_count
Dataset 7	subscription_age, service_failure_count

5.6.2 Convergence behavior

Figure 5.6 demonstrates the convergence behavior of the ACO-RSA and the other comparative MAs for all datasets over the defined number of iterations on the x-axis and the fitness values on the y-axis. It presents the average convergence behavior obtained by executing each algorithm 20 times. In these convergence curves, the rapid convergence method is the best.

Although the standard RSA converges faster than ACO-RSA in dataset 4, the final fitness value of the ACO-RSA is slightly smaller than the RSA. It is observed in Figure 5.5 that ACO-RSA shows a faster convergence rate and finds OFS in the least iterations for datasets 1, 2, 3, 5, and 6. This proves that the proposed ACO-RSA is suitable for churn prediction compared to other comparative methods.



(a) Dataset 1

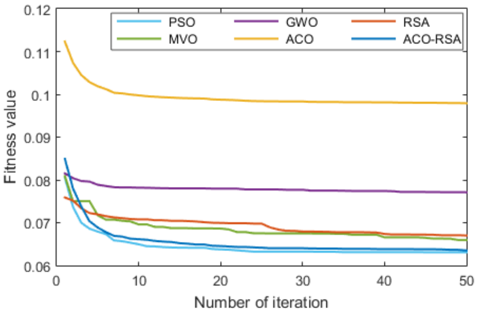
(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

(f) Dataset 6



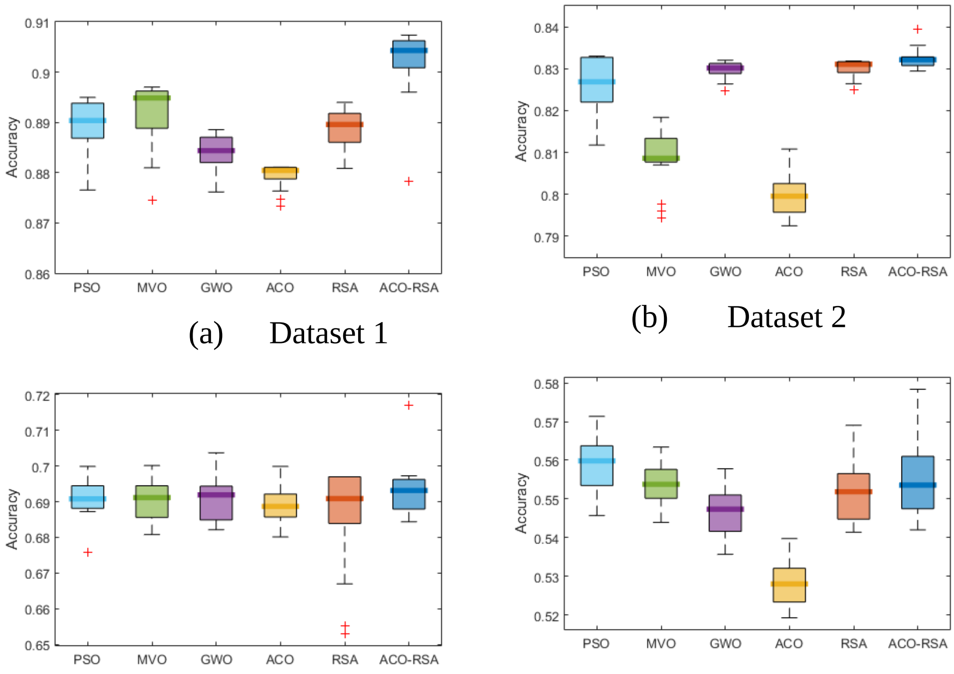
(g) Dataset 7

Figure 5.6: The convergence curves of the ACO-RSA and the other MAs

5.6.3 Boxplots

Figure 5.7 demonstrates the boxplots used to visualize the distribution of classification accuracy for the ACO-RSA and the comparative MAs. In this figure, the x-axis represents the MAs and the y-axis represents the average accuracy.

In boxplots, small degree of dispersion (the gap between the best, the median, and the worse) refers to the algorithm's robustness that achieves the same results in the experiment. It can be seen from Figure 5.7 that the ACO-RSA is more robust than the other comparative MAs on most of the datasets. This indicates the efficacy and robustness of the ACO-RSA approach compared to the PSO, MVO, GWO, standard ACO, and standard RSA methods.

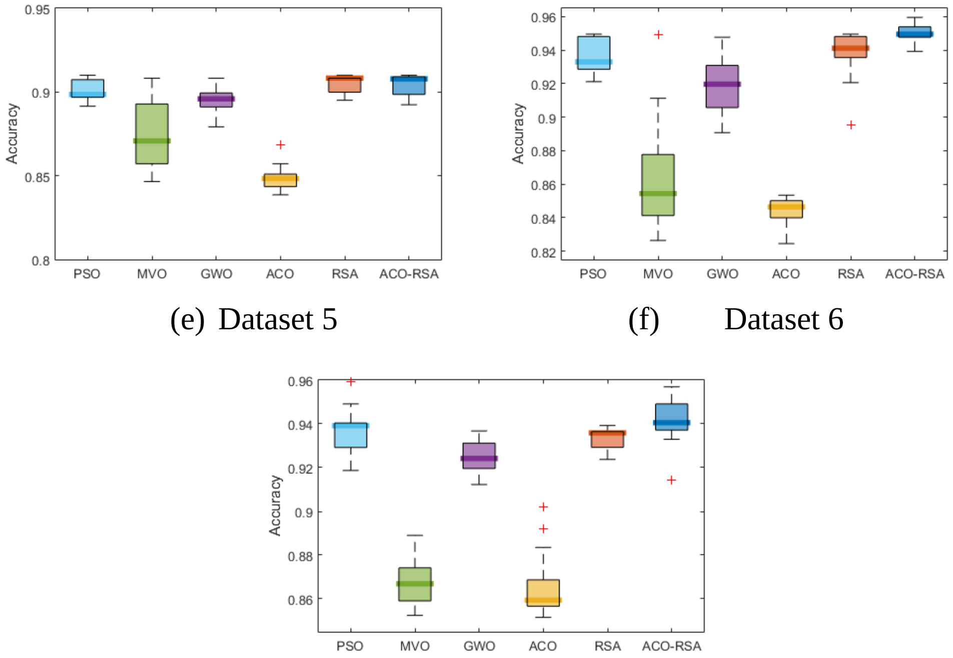


(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4



(e) Dataset 5

(f) Dataset 6

(g) Dataset 7

Figure 5.7: The Boxplot graphs of each MAs for each dataset

5.6.4 Statistical analysis

A widely used non-parametric two-way analysis of variances by ranks [166], to show the significance of the ACO-RSA results Friedman test is performed on seven datasets for 20 independent runs. In this test, the null hypothesis H_0 affirms the equal behavior of the comparative methods, while the alternative hypothesis H_1 indicates the difference in behaviors of the comparative methods. In the Friedman test, the higher (lower) rank refers to the best measure algorithm assuming the larger (smaller) value is preferred. In the current scenario, H_0 points out that all the MAs have the same behaviors, while H_1 points out that there is a significant difference in the MAs behaviors.

Table 5.7 provides the Avg ranking for each algorithm in terms of accuracy, the number of features in OFS, and fitness value. The significance level (α = 0.05) is employed to reveal the statistically reliable results. The highest p-value calculated using Friedman's test for all seven datasets is 0.0026, less than α. The lower the p-value, the greater the statistically significant difference; therefore, the results are statistically significant. For the classification accuracy metric, the higher value is better, indicating that the method with the highest rank performs better, while for the OFS and fitness value metrics, the method with the lower rank is preferred. In Table 5.7, the proposed ACO-RSA gained the best accuracy, OFS, and fitness value metrics results than the

PSO, MVO, GWO, standard ACO, and standard RSA in five out of seven datasets. However, in the case of OFS, the RSA achieved slightly better results than the proposed ACO-RSA for datasets 5 and 7.

Table 5.7: Friedman ranking results for the ACO-RSA and the other MAs across all metrics

Dataset	Metric	PSO	MVO	GWO	ACO	RSA	ACO-RSA
Dataset 1	ACC	4.00	4.75	2.05	1.00	3.25	**5.95**
	OFS	5.45	4.75	3.90	3.10	2.65	**1.15**
	Fitness	3.24	2.14	4.98	6.00	3.62	**1.02**
Dataset 2	ACC	4.10	1.90	3.80	1.15	4.45	**5.60**
	OFS	4.05	4.50	3.15	2.35	3.95	**3.00**
	Fitness	2.18	5.86	3.20	3.70	5.02	**1.04**
Dataset 3	ACC	1.55	2.80	4.05	4.55	2.40	**5.65**
	OFS	5.40	4.60	3.95	3.95	1.60	**1.50**
	Fitness	2.54	3.40	4.96	5.96	2.88	**1.26**
Dataset 4	ACC	3.60	3.50	3.35	2.95	3.65	**3.95**
	OFS	5.05	4.05	2.85	2.35	4.60	**2.10**
	Fitness	3.76	2.92	4.88	5.76	2.62	**1.06**
Dataset 5	ACC	3.90	3.40	1.10	5.00	2.45	**5.15**
	OFS	3.80	3.35	4.70	4.55	**2.15**	2.45
	Fitness	3.85	1.9	5.00	6.00	2.60	**1.65**
Dataset 6	ACC	5.05	1.85	3.10	1.35	4.40	**5.25**
	OFS	4.75	4.95	4.15	3.00	2.80	**1.35**
	Fitness	3.95	2.15	5.00	6.00	2.85	**1.05**
Dataset 7	ACC	4.70	2.65	1.05	3.85	2.95	**5.80**
	OFS	3.30	4.20	4.40	3.60	**2.30**	3.20
	Fitness	3.20	5.00	6.00	3.80	1.95	**1.05**

Highlight (bold) denotes the best performance of the corresponding metric.

Holm's procedure is used as a post hoc method to statistically confirm the differences in the behavior between the controlled algorithm and the other methods. In Holm's test, p-values are adjusted to control the probability of false positives. The controlled and alternate hypotheses are evaluated using a pairwise comparison of p-values. The alternate hypothesis is rejected if the adjusted p-value is smaller than the original p-value. A hypothesis is rejected if there is a significant difference between the controlled method and comparative methods; otherwise, it is not rejected.

In the current work, ACO-RSA is employed as the controlled algorithm. The results of Holm's procedure regarding fitness values for the controlled method and other comparative algorithms are given in Table 5.8. This table shows a significant difference between the controlled method and other MAs in most cases. However, the controlled method shows no significant results than the standard ACO and standard RSA in dataset 3 and the MVO in dataset 5. The overall performance results of the ACO-RSA approach are significantly different from the rest of the MAs. These results prove the superiority of the ACO-RSA approach as an FS method for customer churn prediction.

Table 5.8: Significant tests of the controlled method (ACO-RSA) and other MAs using Holm's test

Dataset	**Algorithm**	***p*-Value**	**Adjusted *p*-Value**	**Hypothesis**
Dataset 1	PSO	1.7845×10^{-28}	1.7845×10^{-27}	**Rejected**
	MVO	1.3164×10^{-19}	5.2655×10^{-19}	**Rejected**
	GWO	1.9346×10^{-44}	2.9019×10^{-43}	**Rejected**
	ACO	2.7827×10^{-43}	3.8958×10^{-42}	**Rejected**
	RSA	5.4055×10^{-21}	3.2433×10^{-20}	**Rejected**
Dataset 2	PSO	2.5520×10^{-33}	3.5728×10^{-32}	**Rejected**
	MVO	7.7200×10^{-32}	8.4920×10^{-31}	**Rejected**
	GWO	2.0184×10^{-32}	2.6239×10^{-31}	**Rejected**
	ACO	9.4165×10^{-26}	7.5332×10^{-25}	**Rejected**
	RSA	5.0025×10^{-40}	7.5038×10^{-39}	**Rejected**
Dataset 3	PSO	7.5939×10^{-2}	1.5188×10^{-1}	**Rejected**
	MVO	8.1198×10^{-1}	8.1198×10^{-1}	**Rejected**
	GWO	1.3254×10^{-10}	3.976×10^{-8}	**Rejected**
	ACO	1.9803×10^{-14}	9.9012×10^{-15}	Not rejected
	RSA	5.5713×10^{-11}	2.2286×10^{-11}	Not rejected
Dataset 4	PSO	3.2151×10^{-16}	2.5721×10^{-15}	**Rejected**
	MVO	1.2848×10^{-15}	8.9933×10^{-15}	**Rejected**
	GWO	1.9558×10^{-16}	1.7602×10^{-15}	**Rejected**
	ACO	1.1534×10^{-17}	1.2687×10^{-16}	**Rejected**
	RSA	7.9148×10^{-18}	9.4978×10^{-17}	**Rejected**
Dataset 5	PSO	1.2760×10^{-13}	7.656×10^{-13}	**Rejected**
	MVO	5.1020×10^{-1}	5.1020×10^{-1}	Not rejected
	GWO	3.5554×10^{-18}	3.5554×10^{-17}	**Rejected**
	ACO	1.0817×10^{-27}	1.5144×10^{-26}	**Rejected**
	RSA	1.7354×10^{-3}	5.2063×10^{-3}	**Rejected**
Dataset 6	PSO	5.2644×10^{-12}	4.7380×10^{-11}	**Rejected**
	MVO	6.0216×10^{-10}	3.0108×10^{-9}	**Rejected**
	GWO	8.7912×10^{-11}	6.1539×10^{-10}	**Rejected**
	ACO	1.9293×10^{-25}	2.3152×10^{-24}	**Rejected**

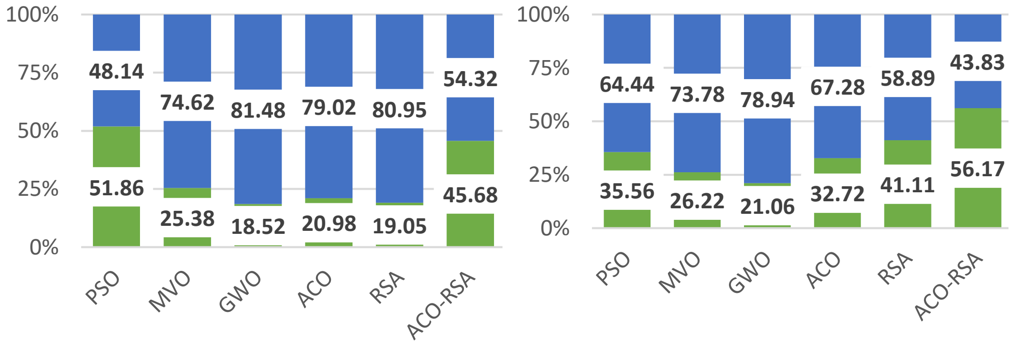
	RSA	2.8887×10^{-4}	5.7774×10^{-4}	**Rejected**
Dataset 7	PSO	2.0999×10^{-12}	2.0999×10^{-11}	**Rejected**
	MVO	5.34510×10^{-9}	3.7416×10^{-8}	**Rejected**
	GWO	6.3569×10^{-32}	7.0121×10^{-31}	**Rejected**
	ACO	1.8961×10^{-7}	7.5843×10^{-7}	**Rejected**
	RSA	2.9728×10^{-7}	8.9185×10^{-7}	**Rejected**

Highlight (bold) denotes that there is a significant difference

5.6.5 Exploration and exploitation effects

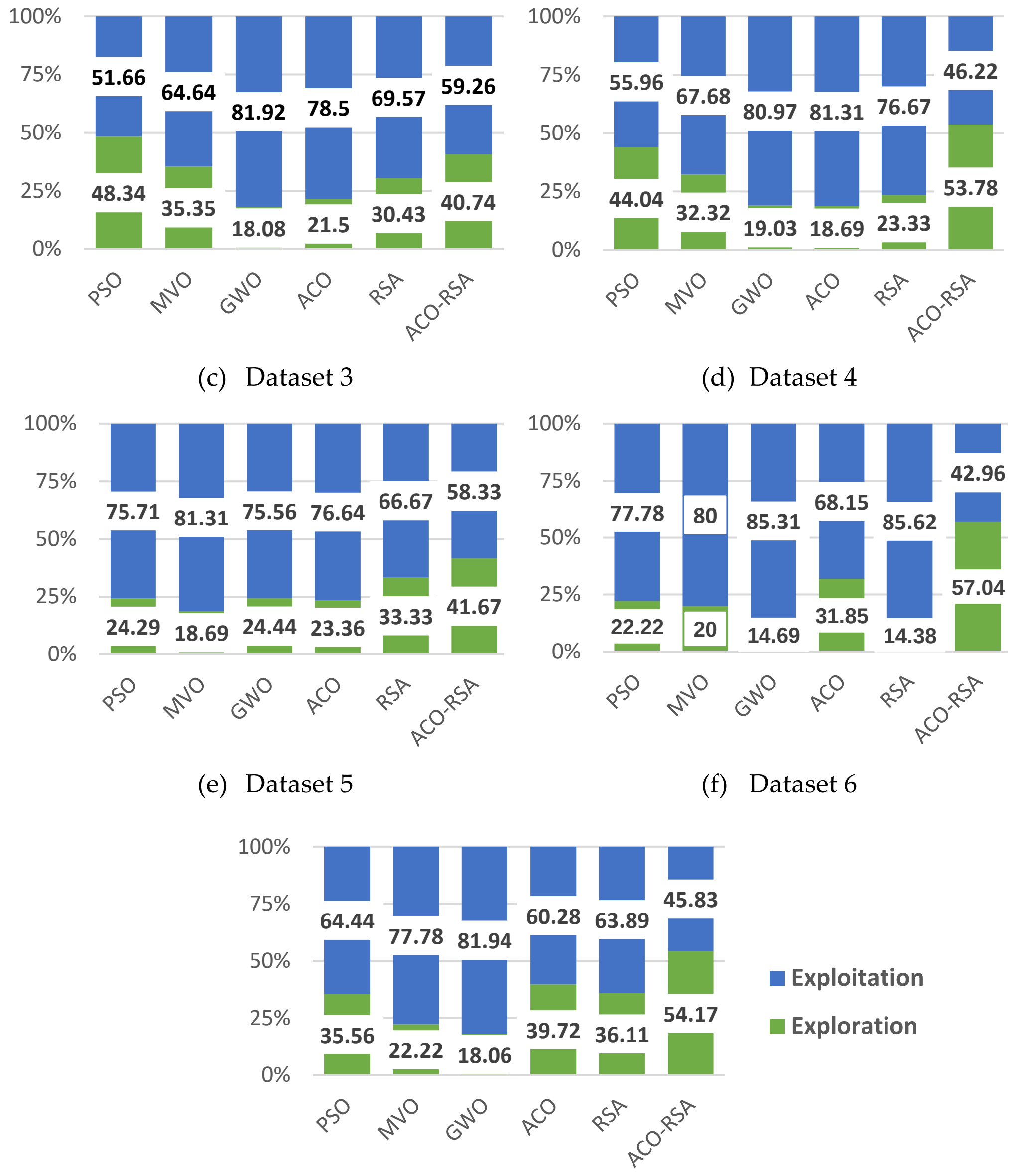
As mentioned earlier, exploration and exploitation are the two main principles in any search algorithm. These phases are obtained using the dimension-wise diversity measurement presented in [167]. In this approach, the exploration can be measured during the search process by the increased mean value of distance within dimensions of the population and exploitation phase by the reduced mean value, where search agents are in a concentrated region.

Figure 5.8 provides exploration-exploitation ratios for all MAs on each dataset for 50 iterations during the search process. From the bar charts in Figure 5.8, ACO-RSA maintains a better balance between exploration and exploitation for all the seven datasets. Although PSO balanced exploration-exploitation for the first four datasets, it fails to maintain the balance (has high exploitation) for the remaining three datasets. Most other algorithms have shown high exploitation, which can be confirmed through the literature or by analyzing the algorithm design. In standard ACO, ants travel the path iteratively to find the best solution, representing higher exploitation than exploration. In standard RSA possessed this balance by splitting the total iteration into four stages, but it failed for four out of seven datasets.



(a) Dataset 1

(b) Dataset 2



(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

(f) Dataset 6

(g) Dataset 7

Figure 5.8: Exploration and exploitation ratio maintained by MAs on each dataset

5.6.6 CEC 2019 test functions

To show the capability of the ACO-RSA compared to standard ACO and standard RSA, ten standard well-known test functions from CEC 2019 test functions with dimension 50 and search range as in the work of [147], which have been widely used in recent years, are chosen. Table 5.9 provides a summary of these functions.

Table 5.9. CEC 2014 Test Functions

Nu.	Functions	$F_i^* = F_i(X^*)$
F1	Storn’s Chebyshev Polynomial Fitting Problem	1

F2	Inverse Hilbert Matrix Problem	1
F3	Lennard-Jones Minimum Energy Cluster	1
F4	Rastrigin’s Function	1
F5	Griewangk’s Function	1
F6	Weierstrass Function	1
F7	Modified Schwefel’s Function	1
F8	Expanded Schaffer’s F6 Function	1
F9	Happy Cat Function	1
F10	Ackley Function	1

To achieve the simulation criteria, i.e., the Avg and Std values, the algorithm for each function of CEC 2019 has been performed by each algorithm 20 independent runs and the results are given in Table 5.10.

Table 5.10. Avg and Std results using CEC 2019 test functions

Function	Metric	ACO	RSA	ACO-RSA
F1	Avg	2.6843×10^{-15}	**0**	**0**
	Std	6.4655×10^{-8}	**0**	**0**
F2	Avg	9.8746×10^{-23}	**0**	**0**
	Std	7.4512×10^{-6}	**0**	**0**
F3	Avg	1.5476×10^{-21}	**0**	6.8764×10^{-28}
	Std	6.5241×10^{-9}	**0**	3.6481×10^{-9}
F4	Avg	**0**	**0**	**0**
	Std	**0**	**0**	**0**
F5	Avg	1.6784×10^{2}	4.9000×10^{2}	2.6818×10^{2}
	Std	4.7864×10^{-3}	6.5260×10^{-3}	$\mathbf{3.4510 \times 10^{-3}}$
F6	Avg	2.5420×10^{1}	1.2382×10^{1}	$\mathbf{1.0307 \times 10^{-1}}$
	Std	3.6857×10^{-2}	7.1253×10^{-2}	$\mathbf{1.0541 \times 10^{-2}}$
F7	Avg	3.5407×10^{-3}	1.9745×10^{-3}	$\mathbf{2.5438 \times 10^{-4}}$
	Std	2.6743×10^{-4}	6.6287×10^{-3}	$\mathbf{7.6842 \times 10^{-5}}$
F8	Avg	-6.8741×10^{3}	$\mathbf{-7.1505 \times 10^{3}}$	-4.8366×10^{3}
	Std	5.6874×10^{3}	6.8674×10^{2}	$\mathbf{6.8133 \times 10^{2}}$
F9	Avg	2.6845×10^{-38}	**0**	**0**
	Std	8.6451×10^{-39}	**0**	**0**
F10	Avg	9.6872×10^{-12}	$\mathbf{8.8818 \times 10^{-16}}$	$\mathbf{6.8766 \times 10^{-16}}$
	Std	2.6851×10^{-12}	**0**	**0**

It can be observed that ACO-RSA achieved better performance in three out of ten functions than standard ACO and standard RSA. For functions F1, F2, F4, and F9, both

ACO-RSA, and RSA achieved the best Avg and Std results. For functions, F5 and F8, ACO and RSA reported the best average performance, respectively.

5.7 Summary

In the telecommunication sector, churn prediction models are broadly employed to analyze and discover patterns in massive data using ML so that past customers' behavior can be used to predict the ones likely to join other operators. FS is a typical preprocessing problem in ML concerning the discrimination of salient and redundant features from each dataset's complete set of features. A new FS approach is presented by combing the standard ACO and standard RSA for customer churn prediction. The combined method, ACO-RSA, utilized a serial mechanism to balance exploration and exploitation while eliminating trapped in local optima. The efficiency of the proposed ACO-RSA is evaluated using six public benchmark datasets from the churn prediction application and ten CEC 2019 test functions. The reliability and performance of the ACO-RSA are compared with the standard ACO, the standard RSA, and three other MAs: PSO, MVO, and GWO. The results showed that the ACO-RSA approach has higher accuracy with the minimum number of features over the other comparative methods. Statistical analysis also confirmed the superiority of the ACO-RSA in terms of various measures. Therefore, the proposed ACO-RSA provides a high-reliability FS approach for applying churn prediction. The main limitation of the proposed approach is the slightly high CT requirement during the training phase to specify the best combination of the tested element.

Chapter 6 An Improved Churn Prediction Model

6.1 Introduction

Nowadays, ML techniques are used to predict future patterns and behaviors of customers [169], so marketing strategies can be improved according to the produced results from these models. ML approaches can play a critical part in the success of different applications, such as oil price prediction [170], sentiment analysis [171], energy consumption [172], medical diagnosis [173], and CP [174]. These applications use one type of ML family of algorithms, called ensemble methods, which are inspirited by the human cognitive system. These methods have the powerful capability to deal with high-dimensional data and generate several diverse solutions for a given task [175].

Ensemble methods build many base models and then merge them into one to achieve better prediction results than using a single base model. Bagging and boosting are the most popular ensemble methods [176]. The bagging method, also known as "bootstrap aggregation," is based on averaging the base models, while the boosting methods are built upon a constructive iterative mechanism. In boosting algorithms, several weak learners are combined stage-wise to obtain a strong learner with improved prediction accuracy [177]. The family of boosting methods depends on different constructive strategies of ensemble formation. A gradient-descent-based formulation of boosting methods, called Gradient Boosting Machine (GBM), is derived by [178]. The GBM can be considered an optimization model aiming to train a series of weak-learner models, which sequentially minimizes a pre-defined loss function.

According to [179], several essential choices of differentiable weak-learner models and loss functions can be customized to a given task in the GBM model, making this model highly flexible to be applied in several ML applications based on the task requirements [180, 181, 182]. The aim is to develop a new model by improving GBM's structure to effectively predict customer churn in the telecom sector. The main contributions can be summarized as follows:

- CP-EGBM is a new model with high predictive performance that may be used to develop effective strategies and contains customer churn risks in the telecom sector. It can enhance the learning ability of the GBM model structure by using SVM as a base learner and exponential loss as a loss function.

- Boosting the capability of the PSO in the exploration phase using the consumption operator of the AEO method could effectively find the most suitable values of the CP-EGBM's hyper-parameters.
- The performance of the proposed CP-EGBM is assessed using six datasets in several evaluation metrics.
- The CP-EGBM model outperformed either GBM or SVM alone, and it is superior to several earlier reported models in the literature, making it more suitable for CP.

6.2 Literature review

Many works applied ensemble ML models to predict customer churn [174, 183, 184]. Wang *et al.* [185] investigated the capability of the GBM model for CP. They used a large customer dataset obtained from the Bing-Ads platform company to identify whether the customers would leave or stay based on the analysis of their historical data records. The results showed that GBM was an effective and efficient model for predicting churner customers in the near future.

Several comparative analyses are conducted for CP using ML models. Ahmad *et al.* [186] compared four ML models, including Decision Trees (DTs), Random Forest (RF), GBM, and XGBoost, for customer churn prediction. The results showed that the XGboost method outperformed other models when they evaluated the models using big data provided by a telecom company in Syria. Jain *et al.* [187] used four models for CP in the banking, telecom, and IT sectors, where they used Logistic Regression (LR), RF, SVM, and XGBoost. The results showed that XGBoost performed better than others in the telecom sector. In another work, Dhini *et al.* [188] compared RF and XGBoost to find the best model for CP. They used a private dataset collected from different companies in Indonesia to evaluate the models. The results showed that the predictive performance of the XGBoost was better than that of the RF model. In Sabbeh [189], the author compared a set of ML models using a publicly available dataset for CP. The results showed that RF attained the best results compared to other models used in their work.

Sandhya *et al.* [190] applied LR, KNN, SVM, and RF models to a publicly available dataset for CP. The authors first preprocessed the dataset and overcame the class imbalance problem using Synthetic Minority Oversampling Technique (SMOTE). The obtained results showed that RF performed better than the other models. Kimura [191] used six ML models: LR, RF, SVM, CatBoost, XGBoost and LightGBM. For data preprocessing, the authors used SMOTE Tomek Link and SMOTE-ENN sampling methods to balance class distribution in a publicly available dataset for CP. The results showed that CatBoost with SMOTE is the best model. Zhu & Liu [192] conducted a comparative study between ten ML models for churn prediction using a publicly available dataset; the results indicated that XGBoost obtained the best accuracy compared to the other models.

Kanwal *et al.* [193] employed a hybrid CP model using PSO to select the most informative features in a publicly available dataset for CP. Then, the selected features are used as inputs to DTs, KNN, Gradient Boosted Tree (GBT), and NB models. The findings indicate that the PSO with the GBT model obtained successful accuracy outcomes compared to the other models. Bilal *et al.* [194] introduced a CP model based on hybrid clustering and classification methods to predict customer churn from two publicly available datasets. The results showed that this model is more robust than the other existing models in the literature.

The stacking model technique (i.e., a mechanism that aims to leverage the benefits of a set of base models while ignoring their disadvantages) is also used for CP. Karuppaiah & Gopalan [195] presented a stacked Customer Lifetime Value-based heuristic incorporated ensemble model to predict customer churn. The authors used a publicly available dataset to evaluate the proposed model, and the obtained accuracy results showed good performance compared to the other existing models in the literature. Rabbah *et al.* [196] proposed a new CP model using deep learning and stacked models. They used a publicly available dataset to validate their model; the dataset is first preprocessed and balanced by the SMOTE method and then used a pre-trained Convolutional Neural Network (CNN) to select the essential features from the dataset. They employed the stacking model technique (i.e., a mechanism that aims to leverage the benefits of a set of base models while ignoring their disadvantages) to predict customer churn. The results demonstrated high efficacy of the developed model than the DTs, LR, RF, XGBoost, and Naive Bayes (NB) models.

Karamollaoglu *et al.* [197] used to separate datasets for CP in telecommunication industry. Eight ML models comprising LR, KNN, DT, RF, SVM, AdaBoost, NB, and multi-layer perceptron are explored. Although all models reported good performance, ensemble-based RF models showed highest performance. Akinrotimi *et al* [198] used oversampling techniques for class imbalance problems and applied the dimensionality reduction technique to pick out optimal features with strong predictive ability. They used LR and the NB models as classification strategies for CP. The results showed that NB provided more efficient results than LR. Akbar and Apriono [199] used XGBoost, Bernoulli NB, and DT models for CP and showed that XGBoost attained the best performance compared to other models.

Based on the provided research works on customer CP, the following research gaps can be identified:

- Limited exploration of ensemble models: While some studies have applied ensemble models for CP, such as stacking models, there is still a need for further exploration and evaluation of different ensemble techniques and their effectiveness in improving prediction accuracy.
- Limited investigation of hybrid models: Hybrid models that combine different ML algorithms or feature selection techniques have shown promising results in CP. However, there is still a lack of comprehensive studies comparing various hybrid models and evaluating their performance on different datasets.
- Lack of focus on industry-specific CP: Many studies have evaluated CP models on publicly available datasets, but there is a need for more research focusing on specific industries, such as banking, telecom, and IT. Different industries may have unique characteristics and churn patterns, requiring customized CP approaches.
- Preliminary analysis of feature selection techniques: Feature selection plays a crucial role in CP, as it helps identify the most informative features for accurate prediction. However, the existing literature lacks comprehensive analyses and comparisons of different feature selection techniques and their impact on CP performance.
- Lack of comparison across multiple performance metrics: Many studies focus on a single performance metric, such as accuracy or F1-measures, for evaluating CP models. However, a comprehensive comparison across multiple metrics, including, recall, and area under the receiver operating characteristic curve (AUC-ROC), is essential to understand different models' overall performance and effectiveness.

Addressing these research gaps would contribute to advancing the field of customer churn prediction by providing insights into the effectiveness of different models, techniques, and approaches in various industry contexts and facilitating more accurate and proactive customer retention strategies.

Although existing models based on ensemble methods achieved tremendous success in the application of CP, there is still a need for more efforts to provide this sector with an efficient and accurate model which can identify churner and non-churner customers accurately and can assess decision-makers in this sector to develop more effective strategies in order to reduce customer churn rate. The GBM model shows excellent potential in classification problems. It typically uses a DT as a base learner to initialize the model, which is sub-optimum [179]. SVM is a powerful mathematical model that proves its ability to solve CP problems [190, 197]. Choosing an effective base- learner as a starting point for the GBM learning process could produce an effective GBM model. Hence, in this work, the base- learner in the GBM is replaced with the SVM. In addition, the hyper-parameters for the modified GBM are optimized using a modified version of the PSO method. To the authors' best knowledge, optimizing GBM has never been applied in CP so far. A new FS model is presented that relies on improving the GBM structure and optimizing its hyper-parameters. This paper presents a new model that relies on improving the GBM structure and optimizing its hyper-parameters to predict customer churn effectively. The proposed model can assess improving CP's efficiency and designing optimal decisions and policies in this sector.

6.3 Proposed CP-EGBM

The overall process flow of our CP is depicted in Figure 6.1, with the proposed CP-EGBM classification model in red. The following sub-sections provide the details of the model.

6.3.1 Data preprocessing and feature selection

Let the dataset consist of N examples of M-dimension feature vectors $\{x_{n,m}, 1 \leq n \leq N \text{ and } 1 \leq m \leq M\}$ and target label $\{y, 1 \leq y \leq C\}$ where C is the number of classes. Each feature in the dataset is normalized in the range [0, 1] as per Eq. (1) to improve classification capability.

$$\hat{x}_{n,m} = \frac{x_{n,j} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{1}$$

where, x_j^{min} and x_j^{max} are minimum and maximum values for the jth feature dimension and $\hat{x}_{i,j}$ is the normalized value of jth feature for ith example.

The performance of most ML models degrades for a class-imbalanced dataset. A dataset balance can be checked by comparing the number of examples for each class label y. For balancing the dataset, the minority class examples are oversampled to match the number of examples using the Heterogeneous Euclidean-Overlap Metric Genetic Algorithm (HEOMGA) approach [200].

Another critical factor affecting the performance of ML models is the input feature dimensional space. The significant features for classification are selected from the normalized-balanced dataset using Ant Colony Optimization- Reptile Search Algorithm (ACO-RSA) approach [201]. The ACO-RSA is a recent Meta-Heuristic (MH) approach published as a feature selection method for CP. The optimal feature set comprises only the most significant features for classification. Finally, the datasets are split into two exclusive and exhaustive sets for training and testing the proposed CP-EGBM model.

6.3.2 Classification using CP-EGBM

An overview of the GBM, a description of the developed CP-EGBM, and Hyper-parameter optimization for the CP-EGBM are given in this section.

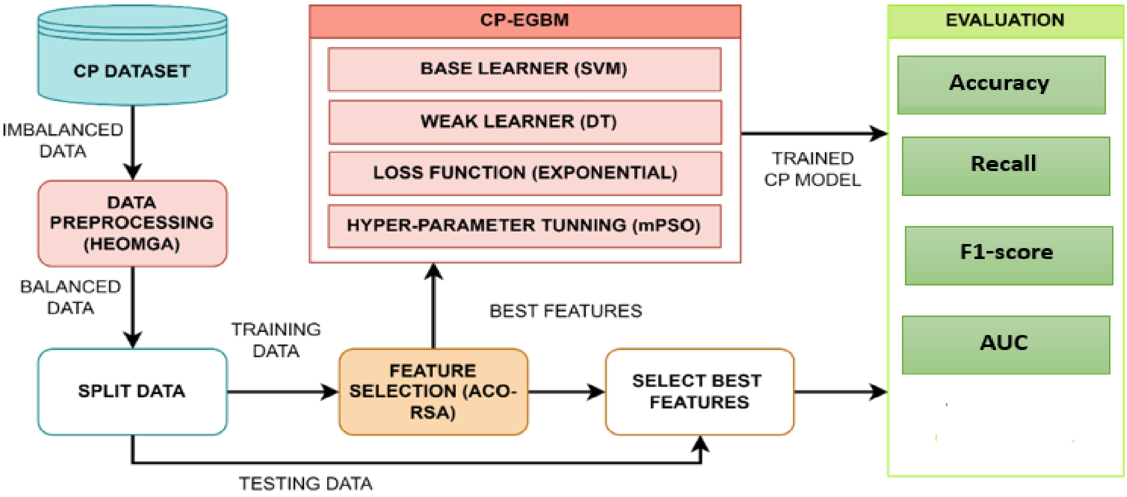


Figure 6.1: Block diagram of CP-EGBM model

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) [178] combines a set of weak learners by focusing on the resulting error at each iteration until a strong learner is obtained as a sum of the successive weak ones.

Let $D = \{x_n, y_n\}_{n=1}^{N}$ denote training examples where the goal of gradient boosting is to find an optimal estimate $F(x)$ of an approximation function $F^{*}(\text{x})$, which maps the instances x_n to y_n to minimize the expected value of a given loss function $L\ (y, F(x))$ over the distribution of all training examples.

$$F^{*}(\mathrm{x}) = \mathrm{argmin}_{F(x)} L_{x,y}\big(y, F(x)\big) \tag{2}$$

GBM uses a logistic loss function for classification tasks to estimate approximation function $L\big(y, F(x)\big) = \big(y - F(x)\big)^2$ [202]. GBM starts with a weak learner $F(x)$ that is usually a constant value, and then it fits each weak learner to correct the errors made by the previous weak learner to strengthen prediction performance by minimizing loss function over each boosting stage [203]. At each stage, the local minimum proportional takes steps to the loss function's negative gradient to find the local minimum. The gradient direction of the loss function at ith boosting stage can be calculated as

$$r_{i,n} = -\left[\frac{\partial L\big(y_n, F(x_n)\big)}{\partial F(x_n)}\right]_{F(x)=F_{i-1}(x)} \tag{3}$$

GBM generalizes the gradient's calculation range when regression trees are is used with parameter a as weak-learners, usually a parameterized function of the input variables x, characterized by the parameters a. The tree can be obtained by solving the following:

$$a_i = \mathrm{argmin}_{a,\beta} \sum_{n=1}^{N} \left[r_{i,n} - \beta h(x_n, a)\right]^2 \tag{4}$$

where, a_i is a parameter that is obtained at iteration i, and β is the weight value (i.e., the expansion coefficient of the weak learner). Then the optimal length p_i is determined, and the model $F_i(x)$ is updated at each iteration i, with t = 1 to the number of iterations T, as in steps 5 and 6 below in the GBM algorithm. GBM is detailed in Algorithm 1 [178].

Algorithm 1: GBM training

```
Input: Training dataset $D = \{x_n, y_n\}_{n=1}^{N}$, the maximum number of boosting stages $B$
Output: GBM $F_i(x)$
1. $F_0(x) = \mathrm{argmin}_p \sum_{n=1}^{N} L\,(y_n\,, p)$
2. For $m$ =1 to $B$ do
3.     $r_{i,n} = -\left[\frac{\partial L(y_n, F(x_n))}{\partial F(x_n)}\right]_{F(x)=F_{i-1}(x)}$
4.     $a_i = \mathrm{argmin}_{a,\beta} \sum_{n=1}^{N} \left[r_{i,n} - \beta h(x_n, a)\right]^2$
5.     $p_i = \mathrm{argmin}_p \sum_{n=1}^{N} L\,(y_n\,, F_{i-1}(x_n) + p\, h(x_n, a_i))$
6.     $F_i(x) = F_{i-1}(x) + p_i\, h(x, a_i)$
7. End for
```

The choices of base learners and loss functions derived from the GBM model facilitate the capacity to design and further development in this model by researchers based on the task requirements [178, 179]. This work aims to develop a new classification model for the application of CP by enhancing the structure of the GBM and its hyper-parameters, as will be discussed in the following subsections.

Developing CP-EGBM

As mentioned earlier, the GBM model typically uses a DT as the base learner. At each boosting stage, a new DT (weak learner) is fitted to the current residual and concatenated to the previous model to update the residual. This process continues until the maximum number of boosting stages is reached [179]. However, using DT as a base learner might not optimally approximate a smooth function since DT extrapolates the relationship between the input/output data points with a constant value [204]. Thus, using a DT to start the GBM model training process could result in poor predictive performance and overfitting.

In the GBM model, various base-learners are derived, divided into linear, smooth, and DTs models [179], and optimized the GBM using different manners [205, 206, 207]. However, no previous works focused on changing the base learner of the GBM to improve its structure using the SVM in CP. The SVM model introduced in [208] proves its ability to solve various classification problems [209]. As for most classifiers, SVM depends on the training data to build its model by finding the best decision hyperplane that separates the class labels (i.e., response variables). The main goal of the SVM is to find the optimum hyperplane by maximizing the margin and minimizing classification error between each class. In addition, using kernel functions strategy and its applicability to the linearly non-separable data can be extended to map input data into a higher dimensional space. The hyperplane can be described as follows [210]:

$$\mathrm{w}.x_i + \mathrm{b} = 0, \tag{5}$$

where, w is an average vector, and b is the position of the relative area to the coordinate center.

The optimization of the margin to its support vectors can be converted into a constrained programming problem as:

$$\min \frac{1}{2} [\![w]\!]^2 + C \sum_{i=1}^{N} \zeta_i \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \zeta_i \quad \text{and} \quad \zeta_i \geq 0 \tag{6}$$

where, ζ_i represents the misclassified samples to the corresponding margin hyperplane, and C is the cost of the penalty.

The SVM model's most widely used kernel functions are Linear, Polynomial, Sigmoid, and Radial Basis functions (RBF). Among them, RBF is preferable due to its reliability in implementation, adaptability to handle very complex parameters and simplicity [210]. In this research, the SVM with RBF kernel (SVM_{RBF}) is integrated as a base learner in the GBM's structure to boost its learning capability and provide a more accurate approximation of the target label. The RBF kernel function can be given as:

$$k(\mathrm{x_n}\,,\mathrm{x_i}) = \exp(-\gamma\ \|\mathrm{x_n} - \mathrm{x_i}\|^2 + C) \tag{7}$$

where $\mathrm{x_n}\,,\mathrm{x_i}$ are vectors of features computing from training or test data points, γ determines the influence of each training example, and C is the cost or penalty.

The GBM learning performance for a given task depends greatly on the loss function [179, 203]. Therefore, it is essential to carefully select the loss function and the function to calculate the corresponding negative gradients in the GBM model's structure. Several loss functions are reported in the literature for classification, including logistic regression (i.e., deviance), Bernoulli, and exponential. A comparison between them is in the next section. The pseudo-code of CP-EGBM is given in Algorithm 2.

Algorithm 2: Pseudo-code of the developed CP-EGBM model.

```
Data preprocessing and feature selection
    1. Normalize the features in the dataset, Eq. (1).
    2. Balance the dataset for all classes using HEOMGA [200].
    3. Calculate the optimum feature set using the ACO-RSA approach [201].
    4. Split the dataset into training and testing.
CP-EGBM training phase
    5. Load training dataset.
    6. Initialize the CP-EGBM model with SVM as the base learner, DT as weak
    learners, logistic/ Bernoulli/exponential as a loss function.
    7. Tune hyper-parameters of CP-EGBM using the mPSO.
    8. Train SVM as a base learner using optimum hyper-parameters, Eq. (5)– (7).
    9. Train the GBM model using optimum hyper-parameters per Algorithm 6.1.
CP-EGBM testing phase
    10. Load testing dataset.
    11. Select only optimum features as calculated training phase.
    12. Evaluate performance metrics using the trained CP-EGBM model.
```

6.3.3 Hyper-parameter optimization

Parameter setting is essential in enhancing the models' efficacy and performance. Traditionally, hyper-parameters can be selected using a trial-and-error. However, manually tuning the parameters is often time-consuming, yielding unsatisfactory results without deep expertise. MH method can tune the model's hyper-parameters for solving this problem. Two MH methods, PSO and AEO, are presented in the following subsections, and the modified PSO (mPSO) method is introduced.

Particle Swarm Optimization (PSO)

PSO is an MH method inspired to simulate the social and group behaviors of animals, humans, and insects [211]. This method uses a set of particles (initial population) to traverse a given search space randomly. In each iteration, the position of each particle x and the velocity υ of this particle is updated using the best position in the current population.

Let there be P particles in the *K*-dimensional search space. The position $x(t)$ and velocity $\upsilon(t)$ at the time of *t* are expressed as:

$$\begin{aligned} x_i(t) &= [x_{i1}(t), x_{i2}(t) \cdots x_{iK}(t)]^T \\ \upsilon_i(t) &= [\upsilon_{i1}(t), \upsilon_{i2}(t) \cdots \upsilon_{iK}(t)]^T \end{aligned} \qquad \text{for } 1 \leq i \leq \ P \tag{8}$$

The fitness, the local best position P_{best} and global best position G_{best} at time *t* are represented as:

$$\begin{aligned} P_{best}(t) &= [\mathrm{P}_1(t), \mathrm{P}_2(t) \cdots \mathrm{P}_K(t)]^T \\ G_{best}(t) &= [\mathrm{G}_1(t), \mathrm{G}_2(t) \cdots \mathrm{G}_K(t)]^T \end{aligned} \tag{9}$$

At time $t + 1$, the velocity $\upsilon(t + 1)$ of the particle is updated as,

$$\begin{aligned} \upsilon_i(t + 1) = w\upsilon_i(t) + c_1 r_1\big(\mathrm{P}_{besti}(t) - \mathrm{x}_1(t)\big) \\ + c_2 r_2\big(G_{best}(t) - x_i(t)\big) \end{aligned} \tag{10}$$

where w is an inertia weight factor that controls the velocity and allows the swarm to converge, c_1 is the cognitive factor and c_2 is the social factor that controls the randomness added to the velocity $\upsilon(t + 1)$ for the next position $x_i(t + 1)$, r_1 and r_2 are two random vectors in the range [0,1].

$$x_i(t + 1) = x_i(t) + v_i(t + 1) \tag{11}$$

where the next position $x_i(t + 1)$ of *i*th particle is computed using the current position $x_i(t)$ and updated velocity $v_i(t + 1)$ as generated in Eq (10). Finally, x_i vectors present solutions while υ_i presents the momentum of particles.

Artificial Ecosystem-based Optimization (AEO)

AEO is another MH method motivated by the energy flow in the natural ecosystem, introduced by [212]. AEO uses three operators to achieve optimal solutions, as described below.

1. Production

In this operator, the producer represents the worst individual in the population. Thus, it must be updated concerning the best individual by considering the upper and lower boundaries of the given search space so that it can guide other individuals to search other regions. The operator generates a new individual between the best individual x_{best} (based on fitness) and the randomly produced position of individuals in the search space x_{rand} by replacing the previous one. This operator can be given as,

$$x_i(t+1) = (1-\alpha)x_{best}(t) + \alpha x_{rand}(t) \tag{12}$$

$$\alpha = (1 - t/\mathrm{T})r_1 \tag{13}$$

$$x_{rand} = \quad {}_2(UB - LB) + LB \tag{14}$$

where $x_{rand}(t)$ guides the other individuals to broadly explore search space in the subsequent iterations, $x_i(t+1)$ leads the other individuals to exploitation in a region around $x_{best}(t)$ intensively, α is a linear weight coefficient to move the individual linearly from a random position to the position of the best individual $x_{best}(t)$ through the pre-defined maximum number of iterations T, r_1 and r_2 are random numbers in the interval [0, 1], and UB and LB represent the upper and lower boundaries of the search space.

2. Consumption

This operator starts after the production operator is completed. It may eat a randomly chosen low-energy consumer, a producer, or both to obtain food energy. A Levy flight-like random walk, called Consumption Factor (CF), is employed to enhance exploration capability, and it is defined as follows:

$$\mathrm{CF} = \frac{1}{2}\frac{v_1}{|v_2|}, \qquad v_1, v_2 \in N(0,1) \tag{15}$$

where, $N(0,1)$ is a normal distribution with zero mean and unity standard deviation

Different types of consumers adopt different consumption behaviors to update their positions. These strategies include:

- Herbivore behavior: A herbivore consumer would eat only the producer and can be formulated as:

$$x_i(t+1) = x_i(t) + CF.\left(x_i(t) - x_1(t)\right), \qquad i \in [2, \ldots P] \tag{16}$$

- Carnivore behavior: A carnivore consumer would only eat another consumer with higher energy, and it can be modeled as:

$$x_i(t+1) = x_i(t) + CF.\left(x_i(t) - x_{\mathrm{rand}\in(0,\ 2i-1)}(t)\right), \qquad i \in\ [3, \ldots P] \tag{17}$$

- Omnivore behavior: An omnivore consumer can eat a random producer or a producer with higher energy, and this behavior can be formulated as:

$$x_i(t+1) = x_i(t) + CF(r_2(x_i(t) - x_1(t))) + (1 - r_2)(x_i(t) - x_{\mathrm{rand}\in(0,\ 2i-1)}(t)), i \in [3, \ldots P] \tag{18}$$

3. Decomposition

In this final phase, the ecosystem agent dissolves. The decomposer breaks down the remains of dead individuals to provide the required growth nutrients for producers. The decomposition operator can be expressed as:

$$x_i(t+1) = x_P(t) + De(e\ .x_P(t) - h. x_{\mathrm{rand}\in(0,\ 2i-1)}(t)), \qquad i \in [1, \ldots P] \tag{19}$$

where $De = 3u \quad u \in\ N(0,1)$, $e = r_3\ .randi([1,2]) -\ 1$, and $h = 2r_3 - 1$and e, h, and De, are weight coefficients designed to model decomposition behavior.

Modified PSO (mPSO) method

The exploration phase is integral to MH algorithms, aiming to find better solutions by investigating search space. PSO suffers from premature convergence to a local minimum, which makes it spend most of the time on locally optimal solutions. Hence, it is weak in exploring new areas in the search space [213, 214].

A modified PSO (mPSO) method aims to avoid premature convergence in the local optima and, thus, enhance its capability to tune optimum hyper-parameters for the CP-EGBM model. The mPSO method integrates the consumption operator of the AEO into the PSO method's structure. As discussed in the previous subsection, the consumption phase in the AEO method is responsible for exploration, and it has three leading operators: Herbivore, Carnivore, and Omnivore. Both herbivores and omnivores are based on the producer solution (i.e., equals to the best solution in the swarm); the last operator depends on two randomly selected solutions, which helps explore new regions in the search space. The mPSO method utilizes the strength of the AEO in exploration (Eq. (15)) and the strength of the PSO in exploitation (Eq. (10)) to select optimum hyper-parameters for the CP-EGBM model. The mPSO can be presented as (Eq. (20)): The pseudo-code of the mPSO is described in Algorithm 3.

$$v_i(t+1) = wv_i(t) + c_1r_1\big(CF - x_1(t)\big) + c_2r_2\big(G_{best}(t) - x_i(t)\big) \tag{20}$$

Algorithm 3: Pseudo-code of the mPSO approach.

```
1. Initialize particles' positions and velocity, Eq. (8).
2. For $t$ = 1 to $T$ do
3.     Calculate local and global best positions w.r.t. minimum fitness, Eq. (9).
5.     Calculate CF, as in AEO Eq. (15).
6.     Update the velocity of particles, Eq. (20).
7.     Update the positions of particles, Eq. (11).
8. End for
```

6.4 Evaluation measures

In this study, the CP-EGBM model is assessed using a set of evaluation measures, including, Accuracy, Recall, F1-score, and Area Under the ROC Curve (AUC), and they are computed as follows:

$$\text{Accuracy } (AC) = \frac{TP + TN}{TP + TN + FN + FP} \tag{21}$$

$$\text{Recall } (R) = \frac{TP}{TP + FN} \tag{22}$$

$$\text{F1-score } (F) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{23}$$

$$\text{AUC} = \frac{1}{2}\left(1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN}\right) \tag{24}$$

where True Positive and (TP) and True Negative (TN) denote the correctly detected samples as positive and negative, respectively; similarly, False Negative (FN) and False Positive (FP) represent the number of misclassified positive and negative examples.

6.5 Experimental results

The experiments performed to assess the CP-EGBM model, comparing its performance with the GBM and SVM$_{RBF}$ models, are described.

6.5.1 Experimental setup

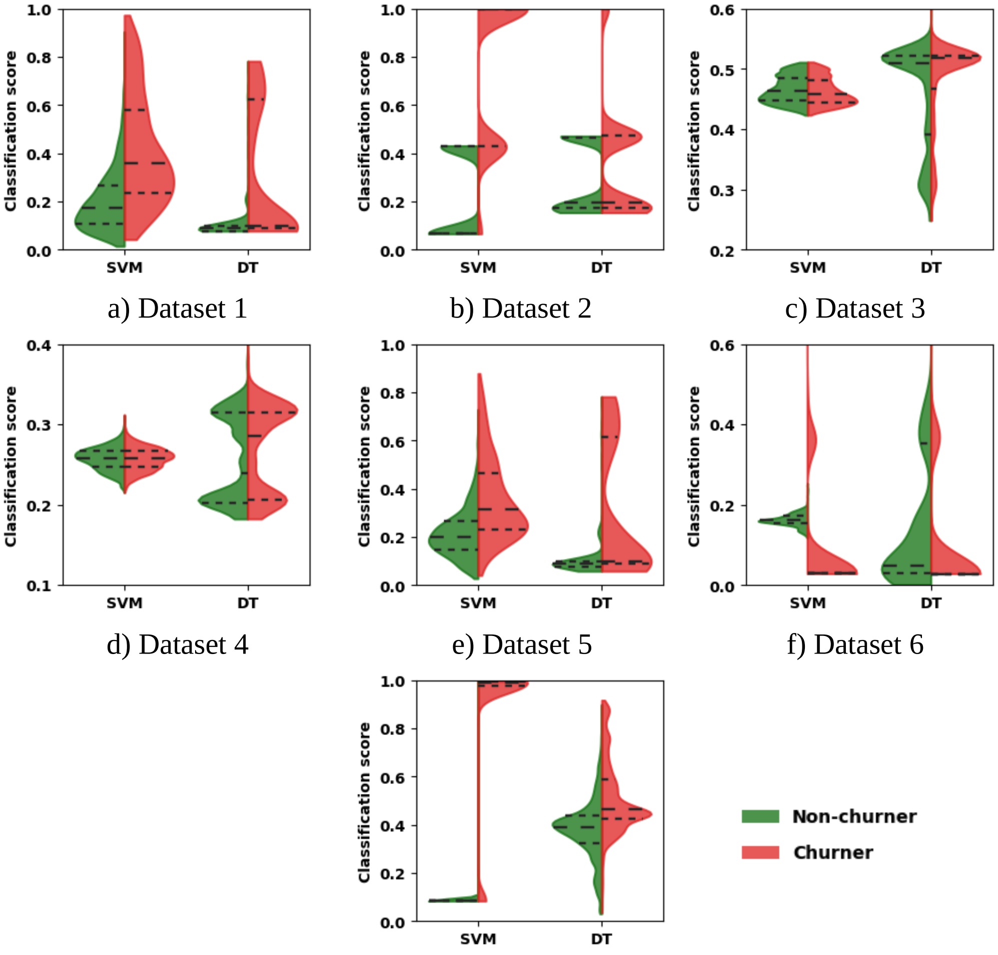
The performance of the CP-EGBM is validated by conducting experiments on the datasets that are given in Chapter 3, Table 3.1 above. The HEOMGA [200] is used for data balancing, and ACO-RSA [201] is employed for FS on all the datasets. Possible bias in selecting the training and testing datasets is avoided using the 10-fold cross-validation (CV) technique is employed. All the experiments are implemented using Python and executed on a 3.13 GHz PC with 16 GB RAM and Windows 10 operating system.

6.5.2 Base learner and its behavior in the GBM model

To examine the effect of changing the base learner from DT to SVM_{RBF} in the GBM model, Probability Density Distribution is used, and the test dataset classification score (which is a number between '0' and '1', indicating the degree how much a testing example belongs to Churner/Non-churner class) generated by both base learners are visualized using the Violin plot method [48], as shown in Figure 6.2. A classification score is a raw continuous-valued probabilistic output of the ML model. For binary classification, one class (assume Churner) has a classification score then another class will have a score $1 - p$.

The Violin plot is a method similar to the box plot with an additional characteristic called probability density, typically smoothed by a kernel density estimator. An interquartile range is calculated for each distribution to compare base learners' dispersion of non-churner and churner classes. The horizontal dotted lines in each class group indicate the first (25th percentile of the data), the second (50th percentile of the data or median), and the third (75th percentile of the data) quartiles to the corresponding distribution. The similarity/closeness of the two distributions is directly proportional to the closeness of these quartiles.

Figure 6.2 shows probability density distribution of testing dataset classification scores for all the datasets used. The visualization in this figure shows that the quartiles of classification score using SVM_{RBF} as a base learner in Dataset 1, Dataset 2, Dataset 5, Dataset 6, and Dataset 7 well-separate churners (in red) and non-churners (in green) than the quartiles using the DT. Using SVM_{RBF} as a base learner better classifies the Churner and Non-churner than DT. In Dataset 3 and Dataset 4, distributions for churners and non-churners are similar for both base learners, also indicated by closer quartiles for both classes, resulting in poor classification for both base learners. These results confirm and prove the suitability of the SVM_{RBF} to be used as a base learner in the developed CP-EGBM model.



a) Dataset 1

b) Dataset 2

c) Dataset 3

d) Dataset 4

e) Dataset 5

f) Dataset 6

g) Dataset 7

Figure 6.2: Probability density distribution of testing dataset classification scores generated by SVM$_{\mathrm{RBF}}$ and DT base learners for non-churner and Churner samples in all the datasets

6.5.3 Loss function selection

The loss function gives a general picture of how well the model is performed in predictions. If the predicted results are much closer to the actual values, the loss will be minimum, while if the results are far away from the original values, then the loss value will be the maximum.

In this section, an experiment is conducted using three loss functions to figure out the most suitable one for the application of CP, and they include:

- Logistic, deviance, or cross-entropy loss is the negative log-likelihood of the Bernoulli model. It is the default loss function in the GBM, and it is defined as [216]:

$$\mathrm{L}_{Logi}(\mathrm{y}, \widehat{y}) = -y\, log\ (\widehat{y}) + (1 - y)\, log\ (1 - \widehat{y}) \tag{26}$$

- Bernoulli, it can be formulated as follows [47].

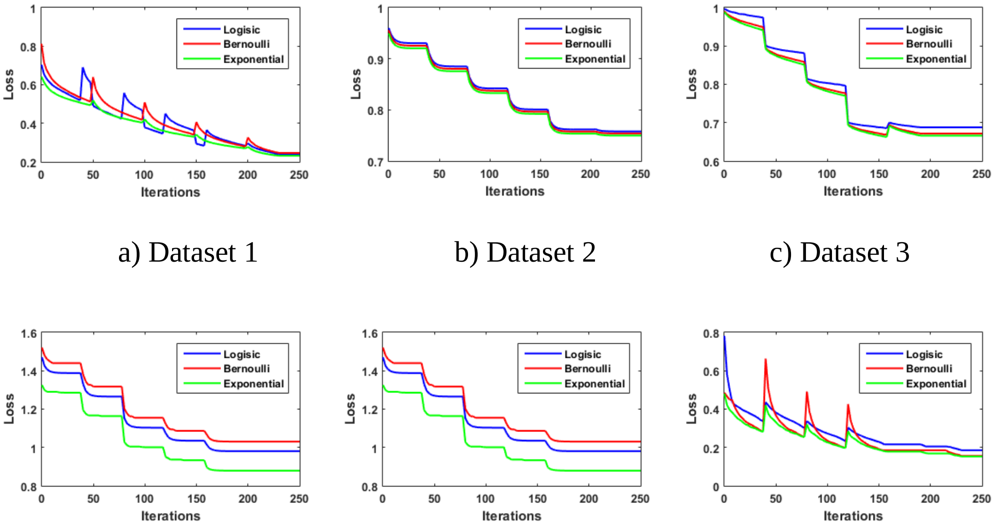
$$\mathrm{L}_{Bern}(\mathrm{y}, \widehat{y}) = \log(1 + \exp(-2y.\, \widehat{y})), \tag{27}$$

- Exponential is also used in the Adaboost algorithm, and it can be defined as [216]:

$$\mathrm{L}_{Ada}(\mathrm{y}, \widehat{y}) = \exp(-y.\, \widehat{y}), \tag{28}$$

where y is a binary class indicator, either 0 or 1, and $\widehat{y}$ is the probability of class 1, while $1 - \widehat{y}$ is the probability of class 0

Figure 6.3 plots the behavior of the loss functions over the defined number of iterations on all the DSs using the developed CP-EGBM. It can be seen in Figure 6.3 that the exponential loss function obtains a smaller loss value on all the DSs. This can be explained by exponentially effectively contrasting misclassified data points much more, enabling the CP-EGBM to capture outlying data points much earlier than the logistic and Bernoulli functions. The results from this experiment confirm that the exponential loss function is more suitable than the other two competitor loss functions for the application of CP.



a) Dataset 1

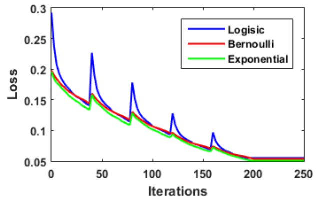
b) Dataset 2

c) Dataset 3

d) Dataset 4

e) Dataset 5

f) Dataset 6

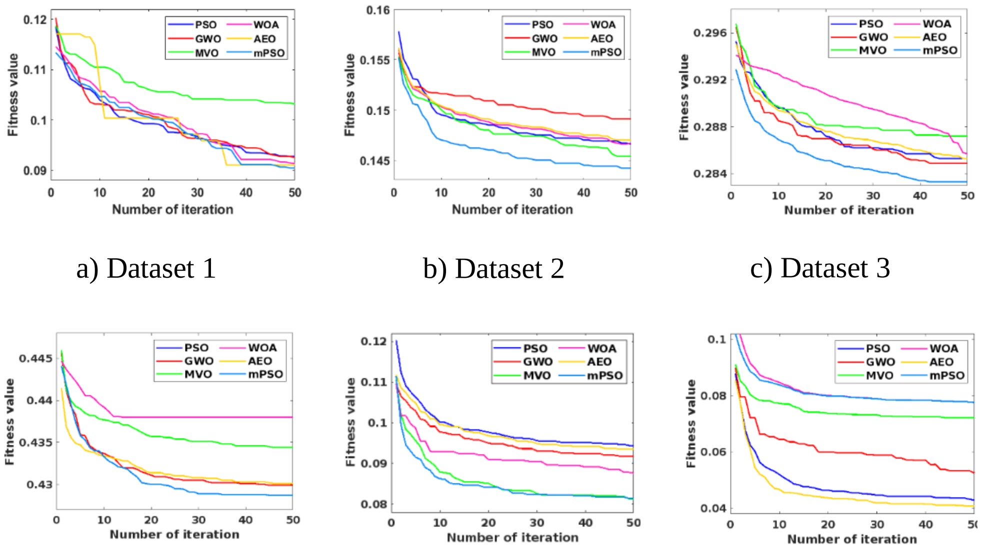


g) Dataset 7

Figure 6.3: Loss functions behavior on all the used datasets

6.5.4 Hyper-parameter setting

To better understand the behavior of the introduced mPSO, convergence curves are generated over 50 iterations on the x-axis and fitness values on the y-axis, as shown in Figure 6.4. A wide range of MH methods introduced in the literature can be used for hyper-parameters tuning. However, the mPSO is compared with Multi-Verse Optimizer (MVO) [217], Whale Optimization Algorithm (WOA) [218], Gray Wolf Optimizer (GWO) [219], PSO [220], and AEO [221]. For all the methods, the population size is set to 20 and the maximum iterations are equal to 50. Each is run 20 times, and these settings are selected after empirically studying them. From Figure 6.4, the convergence speed of the mPSO is faster than the other MH methods in five out of seven datasets, as it stabilizes to shallow fitness values in fewer iterations. Overall, the suggested improvement in the PSO leads to better convergence attributes and less computation time, making mPSO more suitable for tuning the CP-EGBM model's hyper-parameters.



a) Dataset 1

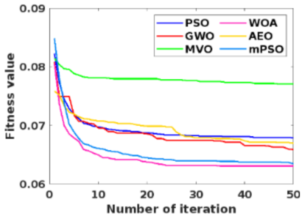
b) Dataset 2

c) Dataset 3

d) Dataset 4

e) Dataset 5

f) Dataset 6



g) Dataset 7

Figure 6.4: mPSO convergence behavior on all the datasets

Several hyper-parameters need to be initialized in the developed CP-EGBM. The mPSO method is used to optimize them. The hyper-parameter settings and the optimized information for each dataset are listed in Table 6.1 and Table 6.2, respectively.

Table 6.1: Hyper-parameters settings of the developed CP-EGBM model

Model	Function	Default value	Search space
SVM_{RBF}	C	1	LB: 1E-1, UB: 1E
	Feature space map (γ)	1 / (#features)	LB: 1E-4, UB:1E4
GBM	Number of estimators	100	LB: 100, UB:3000
	Learning rate	0.1	LB: 1E-3, UB:1
	Maximum depth of DTs	3	LB: 1, UB: 10
	Minimum samples for split	2	LB: 2, UB: 10
	Maximum features	sqrt(#features)	LB: 1, UB: #features
	Sub-sample	1	LB: 0.5, UB:1

LB: Lower Boundary and UB: Upper Boundary

Table 6.2: Optimization results by MPs for all the datasets

Model	Function	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7
SVM_{RBF}	Regularization (C)	100	156	50	65	25	120	87

	Kernel coeff. (γ)	0.213	0.302	0.030	0.001	0.003	0.203	0.137
GBM	No. of estimators	315	503	223	418	438	250	305
	Learning rate	0.093	0.103	0.132	0.312	0.034	0.001	0.003
	Max. depth	5	5	4	6	6	7	6
	Min. sample split	5	8	6	10	7	8	9
	Max. features	8	12	25	40	8	8	6
	Sub-sample	1	0.82	0.90	0.95	0.83	0.97	0.83

6.6 Experimental results and discussion

The results of the GBM, SVM$_{RBF,}$ and the developed CP-EGBM models using evaluation metrics, Receiver Operating Characteristic (ROC), Statistical test, and model stability are discussed in this section. Also, a comparison between the CP-EGBM and other used models in recent works is provided.

6.6.1 Performance results

The performance assessment of the GBM alone, SVM$_{RBF}$ alone, and the developed CP-EGBM models on the datasets is carried out in this section. After applying 10- fold-CV and fine-tuning the model's hyper-parameters using the mPSO, the average results are computed and recorded in Tables 6.3, 6.4, and 6.5, respectively.

Table 6.3: Performance evaluation of the GBM alone on all the datasets

Dataset	***AC***	***R***	***F***	***AUC***
Dataset 1	0.9401	0.7931	0.8439	0.8246
Dataset 2	0.8677	0.8514	0.8200	0.8062
Dataset 3	0.6737	0.6528	0.6813	0.7062
Dataset 4	0.5631	0.6063	0.5902	0.6160
Dataset 5	0.9352	0.7825	0.8413	0.8187
Dataset 6	0.9520	0.8747	0.8672	0.8774
Dataset 7	0.9520	0.7747	0.8150	0.8274

Table 6.4: Performance evaluation of SVM$_{RBF}$ alone on all the datasets

Dataset	***AC***	***R***	***F***	***AUC***

Dataset 1	0.8799	0.8050	0.8410	0.8462
Dataset 2	0.8376	0.6743	0.7394	0.7971
Dataset 3	0.6836	0.6889	0.7009	0.7070
Dataset 4	0.6157	0.6157	0.6146	0.6261
Dataset 5	0.8821	0.8050	0.8407	0.8462
Dataset 6	0.8711	0.8749	0.9004	0.8875
Dataset 7	0.8711	0.7549	0.8202	0.8275

Table 6.5: Performance evaluation of CP-EGBM on all the datasets

Dataset	***AC***	***R***	***F***	***AUC***
Dataset 1	0.9623	0.9121	0.8698	0.8579
Dataset 2	0.8649	0.8456	0.8211	0.8991
Dataset 3	0.6949	0.7138	0.7044	0.7091
Dataset 4	0.6250	0.6298	0.6287	0.6329
Dataset 5	0.9482	0.9175	0.8727	0.8599
Dataset 6	0.9779	0.9033	0.9152	0.9273
Dataset 7	0.9520	0.9275	0.8609	0.8473

The results in Tables 6.3– 6.5 show that the developed CP-EGBM performs better than the other models on all the datasets for individual evaluation metrics. Figures 6.5–6.8 show the models' performance on all the datasets. These figures reveal that the CP-EGBM has accomplished effective outcomes compared to GBM and SVM$_{\mathrm{RBF}}$. For instance, in dataset 6, the CP-EGBM obtained an accuracy of 97.79%, a recall of 90.33%, F1-measure of 91.52%, and AUC of 92.73%. The results in Tables 6.6– 6.8 and Figures 6.5–6.8 confirm the superiority of CP-EGBM compared to other models.

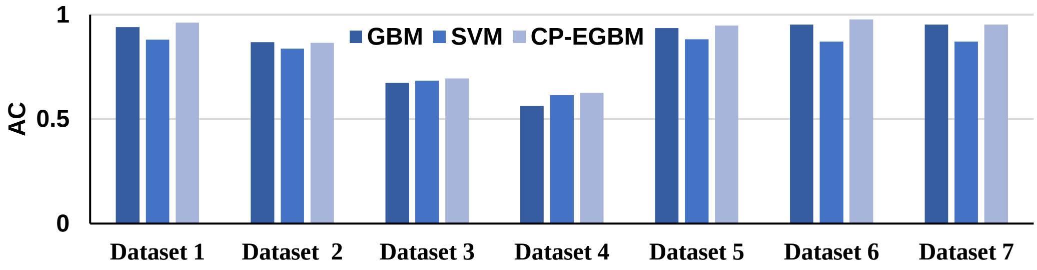


Figure 6.5: Accuracy analysis of GBM, SVM$_{RBF}$, and CP-EGBM on all the datasets

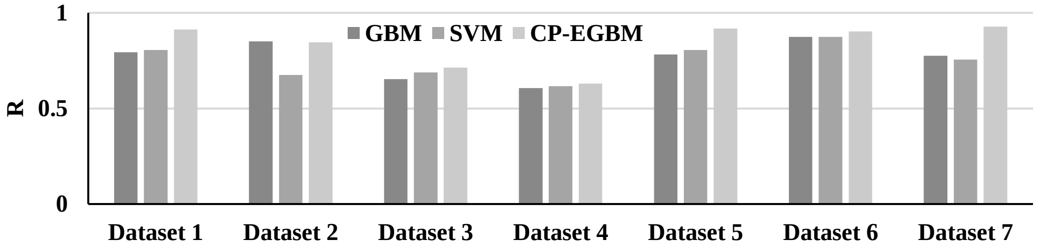


Figure 6.6: Recall analysis of GBM, SBM$_{RBF}$, and CP-EGBM on all the datasets

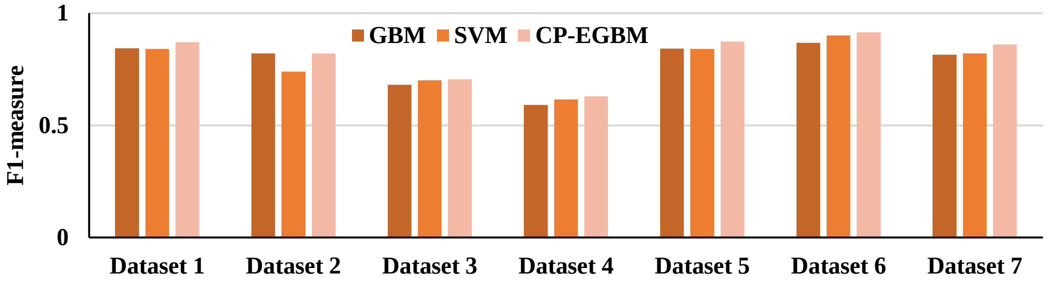


Figure 6.7: F1-measure of GBM, SVM$_{RBF}$, and CP-EGBM on all the datasets

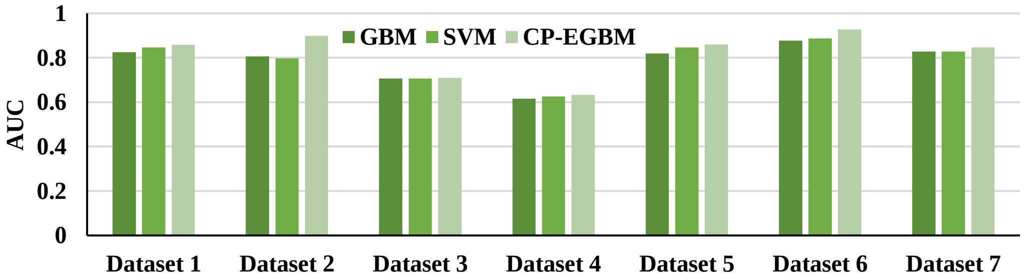


Figure 6.8. AUC analysis of the GBM, SVM$_{RBF}$, and CP-EGBM on all the datasets

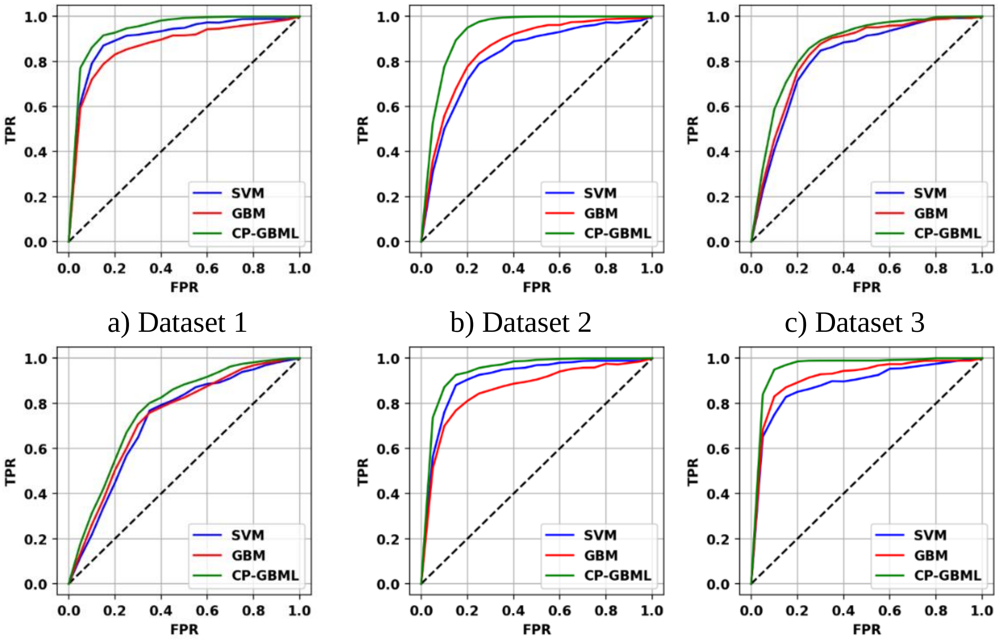
For Dataset 1, SVM has relatively good accuracy (0.8799) and F1-measure (0.8410), while GBM has high better accuracy (0.9401) and F1-measure (0.8439). The developed CP-EGBM outperforms both with the highest accuracy (0.9623) and F1-measure (0.8698). For Dataset 2, SVM alone has moderate performance with accuracy (0.8376) and F1-measure (0.7394), GBM provides accuracy (0.8677) and F1-measure (0.8200), while CP-EGBM provides relatively high accuracy (0.8649) and F1-measure (0.8211). In Dataset 3, GBM shows the worst performance in accuracy (0.6737) and F1-measure (0.6813). While CP-EGBM has best accuracy (0.6949) and F1-measure (0.7044). Similarly, for Dataset 4 GBM has the lowest accuracy (0.5631) and F1-measure (0.5902) and CP-EGBM shows high performance with accuracy (0.6250) and F1-measure (0.6287). On the other hand, GBM performs better than SVM for Dataset 5. CP-EGBM outperforms both with high accuracy (0.9482) and F1-measure (0.8727).

Similar observations can be made for Dataset 6 with an outstanding performance of CP-EGBM by providing very high accuracy (0.9779) and F1-measure (0.9152). For Dataset 7, both GBM and CP-EGBM provide the same accuracy but later have higher F1-measure than earlier.

Overall, CP-EGBM consistently outperforms both GBM and SVM across most of the datasets in terms of accuracy, F1-measure, and AUC. However, GBM and SVM show competitive performance, achieving high accuracy and F1-measure on some datasets but lower performance on others.

6.6.2 ROC curve

The ROC curve computes model performance by changing the confidence level of the model score to get distinct values of the True-Positive Rate (TPR) and False Positive Rate (FPR), as illustrated in Figure 6.9. As this figure shows, the CP-EGBM curves dominate the GBM and SVM$_{\mathrm{RBF}}$ models in all points on all the considered datasets, which indicates the suitability of the developed CP-EGBM.



a) Dataset 1

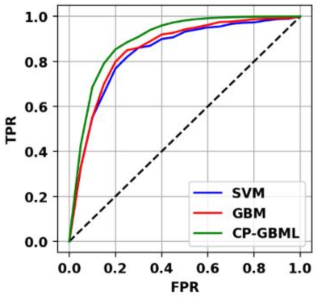
b) Dataset 2

c) Dataset 3

d) Dataset 4

e) Dataset 5

f) Dataset 6



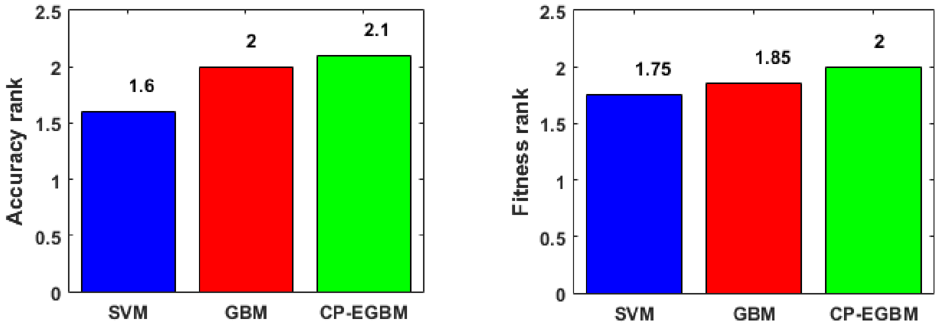
g) Dataset 7

Figure 6.9: ROC graph of GBM. $\mathrm{SVM}_{\mathrm{RBF}}$ and CP-EGBM for all the datasets

6.6.3 Statistical test and model's stability

The developed CP-EGBM is selected as the control model in the Friedman ranks test, as shown in Figure 6.10. In this figure, CP-EGBM gets the highest accuracy (Figure 6.10 a) and fitness values ranks (Figure 6.10 b), followed by GBM as the second and the $\mathrm{SVM}_{\mathrm{RBF}}$ ranked last. Therefore, this work concludes that the CP-EGBM is significantly better than the other models for CP.

The relative stability results associated with the standard deviation (Std) of the developed CP-EGBM and the other models are also calculated and provided in Table 6.6. According to the results in Table 6.9, the developed CP-EGBM model achieved the smallest Std values compared to the GBM and $\mathrm{SVM}_{\mathrm{RBF}}$ models on all the datasets. This reflects the stability and robustness of the developed CP-EGBM for applying CP.



(a) (b)

Figure 6.10: Friedman ranks test for a different model, a) accuracy, b) fitness values

Table 6.6: Std values for the models on all the datasets

Dataset	Measure	Model		
		GBM	$\mathrm{SVM}_{\mathrm{RBF}}$	CP-EGBM
Dataset 1	Std	0.0137	0.0111	**0.0091**
Dataset 2	Std	0.0678	0.0401	**0.0204**
Dataset 3	Std	0.1025	0.0913	**0.0467**
Dataset 4	Std	0.1008	0.0925	**0.0381**
Dataset 5	Std	0.0123	0.0102	**0.0086**

Dataset 6	Std	0.0106	0.0097	**0.0055**
Dataset 7	Std	0.0107	0.0094	**0.0078**

6.6.4 Performance comparison with existing models

Several studies have recently used ML models to predict customer churn in the telecom sector. A comparison between the developed CP-EGBM and other studies for CP is given in Table 6.10. We can see in Table 6.7 that the studies utilized Dataset 1, Dataset 2, and Dataset 5 to evaluate ML models used in their works. Therefore, we can use the same datasets to compare the performance of the CP-EGBM with them. As per the results in Table 6.8, the developed CP-EGBM model has great potential to predict customer churn in terms of accuracy and F1-measure with better prediction results than the existing models.

Table 6.7: Comparison between the existing models and the proposed CP-EGBM model in terms of accuracy and F1-measure on DS 1, DS 2, and DS 5

Author (s)	**Method**	**Dataset 1**		**Dataset 2**		**Dataset 5**	
		AC	*F*	*AC*	*F*	*AC*	*F*
Sabbeh, (2018)	RF	0.9600	-	-	-	-	-
Sandhya *et al.*, (2021)	RF	0.9550	0.8210	-	-	-	-
Kimura. (2022)	CatBoost	-	-	0.7710	0.613	-	-
Zhu & Liu, (2021)	XGBoost	-	-	0.7998	-	-	-
Kanwal *et al.*, (2021)	PSO- GBT	0.9300	0.8110	-	-	-	-
Bilal *et al.*, (2022)	K-means	0.9243	0.7181	-	-	0.9470	0.8063
Karuppaiah & Gopalan, (2021)	Stacked Customer Lifetime Value-based heuristic incorporated ensemble model	0.8900	-	-	-	-	
Rabbah *et al.*, (2022)	DL and stacking	-	-	0.8350	0.8190	-	-
Karamollaoglu *et al.*,(2021)	RF	0.9540	0.9440	-	-	0.7900	0.8630

Akbar & Apriono, (2023)	XGBoost	-	-	0.8156	0.7476	-	-
This work	**CP-EGBM**	**0.9623**	**0.8698**	**0.8649**	**0.8211**	**0.9482**	**0.8727**

The proposed framework uses MH algorithms for feature selection and hyper-parameter tuning. Although MH algorithms have shown effectiveness in many domains, they also have certain limitations. MH algorithms may converge prematurely to get stuck in a local optimum or fail to explore the search space adequately. MH algorithms require a large number of iterations and evaluations of objective functions, which can be computationally expensive for complex problems. Several control parameters need to be set appropriately to achieve good performance. The search process becomes more challenging in high-dimensional spaces, and MH algorithms may struggle to explore and exploit the search space effectively. Despite these limitations, MH algorithms remain valuable tools for solving complex optimization problems.

6.7 Summary

The telecom sector has accumulated a massive amount of customer information during its development, and on the other hand, the widespread data warehouses technology and application make it possible to gain insight into historical customer data. Therefore, it has become clear to managers in this sector that customer information can be used to create prediction models to contain customer churn and risk. A CP-EGBM model is developed to provide a prediction model for the application of CP. The CP-EGBM model uses SVM as a base learner and DTs as weak learners in the GBM's structure. Moreover, a modified version of PSO, mPSO, is introduced to optimize the CP-EGBM model's hyper-parameters by injecting the AEO consumption operator into the PSO's structure. The CP-EGBM is assessed using six CP datasets. The experimental results and statistical test analysis show higher efficacy of the CP-EGBM model than the other tested and reported models in the literature.

Chapter 7 Conclusions and future research

In this dissertation, a range of new algorithms have been developed and applied in real-life case studies. As such, this doctoral thesis contributes from both a theoretical and an application point of view. The main findings and conclusions will be recapitulated in in Section 7.1. The innovative character of the presented approaches opens new perspectives toward future research is discussed in Section 7.2.

7.1 Conclusions

In conclusion, this research uses ML techniques to focus on the important problem of customer churn prediction in the telecommunications sector. Three distinct approaches have been proposed and evaluated: HEOMGA for addressing the class imbalance, ACO-RSA for feature selection, and CP-EGBM for churn prediction. The results obtained from these approaches demonstrate their effectiveness in improving the performance of churn prediction models, thereby providing valuable insights for the telecom industry.

The first approach, HEOMGA, addresses the issue of class imbalance in the datasets by combining the Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) for oversampling the minority class. The HEOM defines a fitness function for the GA, allowing the algorithm to generate synthetic samples that balance the class distribution. The performance evaluation on benchmark datasets from the UCI repository in the domain of customer churn prediction showcases the effectiveness of the HEOMGA method compared to popular oversampling techniques such as SMOTE, ADASYN, G SMOTE, and Gaussian oversampling methods. The results demonstrate that the HEOMGA significantly outperforms these methods.

The second approach, ACO-RSA, focuses on the crucial preprocessing step of feature selection (FS) to enhance the performance of churn prediction models. ACO-RSA combines two metaheuristic algorithms, Ant Colony Optimization (ACO) and Reptile Search Algorithm (RSA), to select the most salient features from the complete feature set. The integration of ACO and RSA enables a balanced exploration-exploitation trade-off, mitigating the risk of getting trapped in local optima. The proposed ACO-RSA approach is evaluated on six open-source customer churn prediction datasets and compared with other optimization algorithms, including Particle Swarm Optimization (PSO), Multi-Verse Optimizer (MVO), Grey Wolf Optimizer (GWO), standard ACO, and standard RSA. The experimental results demonstrate that ACO-RSA outperforms the comparative methods in terms of accuracy and feature dimensionality reduction, highlighting its efficiency in feature selection for churn prediction. The list of the most OFS selected by the developed ACO-RSA is as follows:

The third approach, CP-EGBM, introduces an Enhanced Gradient Boosting Machine (GBM) model for churn prediction. The CP-EGBM model leverages a Radial Basis Function-based Support Vector Machine (SVMRBF) as the base learner and an exponential loss function to enhance the learning process of the GBM. Additionally, a modified version of Particle Swarm Optimization (PSO), called mPSO, is developed to effectively tune the CP-EGBM model's hyperparameters. The proposed model is evaluated on six open-source churn prediction datasets, and its performance is compared with GBM, SVM, and other reported models in the literature. The comparative analysis demonstrates that CP-EGBM outperforms the other models, emphasizing its superior predictive capabilities.

The findings of this research have practical implications for the telecom industry. Customer churn is a significant concern, and accurate prediction of churners can help companies take proactive measures to retain their customers. The proposed approaches contribute to improving churn prediction accuracy and providing valuable insights into the factors influencing churn behavior. The selected features obtained from ACO-RSA highlight the most important variables affecting customer churn, allowing telecom companies to focus on improving those areas.

While the proposed approach offers improved prediction accuracy and optimization, there are certain limitations to consider. One key limitation is the limited availability and access to relevant datasets for CP, which are necessary to test and validate the effectiveness of the CP-EGBM model. Additionally, the reliance on iterative processes makes the method computationally expensive, particularly when dealing with large datasets. The need for a high number of search agents further increases resource consumption and computational time. Moreover, the sequential nature of the optimization techniques restricts the potential to leverage parallel computing, such as GPU optimization, which could otherwise help reduce computation time.

In summary, this research offers valuable contributions to the field of CP in the telecom sector. The proposed approaches, HEOMGA for addressing the class imbalance, ACO-RSA for feature selection, and CP-EGBM for churn prediction, demonstrate their effectiveness in improving the performance of churn prediction models. The findings provide insights and practical implications for the telecom industry, enabling companies to develop robust strategies for retaining customers and mitigating churn risks.

7.2 Future Research

The future research can explore the integration of feature selection methods with the HEOMGA approach to enhance its effectiveness further. Additionally, investigating alternative distance metrics for addressing class imbalance and addressing class overlap situations would be valuable. Also, the proposed HEOMGA method can be compared with other synthetic data generation techniques such as GANs and diffusion models. These methods can generate new samples for the minority class by learning the distribution of the minority class and creating realistic data points that can be added to the training set.

Furthermore, applying the ACO-RSA approach to various other applications, such as renewable energy, IoT, and signal processing, would expand its scope and utility. Also, future research could investigate the use of ACO-RSA in real-time for churn prediction in dynamic environments. ACO-RSA could adapt to changing customer behavior patterns in real-time, ensuring that churn prediction models remain accurate and relevant. This would be particularly useful for industries where customer behavior changes rapidly, such as telecommunications and e-commerce.

Integrating additional ML techniques to further enhance the CP-EGBM model and testing its efficacy on larger and more diverse datasets would ensure its robustness and generalizability. The proposed CP-EGBM model could be further extended and compared with other advanced deep learning techniques to enhance its effectiveness and applicability. One promising avenue for extension is the integration of DL-based data augmentation methods, which can help improve model generalization, especially in cases where labeled data is limited. Techniques such as GANs be employed to synthetically enhance the dataset, ensuring better performance in diverse scenarios.

Additionally, the development of XAI techniques for CP can be a crucial area of exploration. While current ML models, such as DL, ensemble methods, and decision trees, often achieve high accuracy in predicting customer churn, they tend to operate as "black boxes," making it challenging for decision-makers to understand how predictions are derived. Future research could focus on integrating explainability into the CP-EGBM model by employing XAI methods like SHAP or LIME. This would provide interpretable insights, offering transparency in the decision-making process. Ensuring that AI model explanations are meaningful and actionable is essential to building trust and enabling informed decisions. By making these models interpretable, organizations can not only enhance model performance but also ensure the reliability and ethical application of AI-driven CP systems.

References

[1] Sharma, A., Panigrahi, D., & Kumar, P. (2013). A neural network based approach for predicting customer churn in cellular network services. arXiv preprint arXiv:1309.3945.

[2] Yu-Teng, C. (2015) ‘Measuring the impact of datamining on churn management’, Internet Research: Electronic Networking Applications and Policy, Vol. 11, No. 5, pp. 375–387.

[3] Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer Churn Prediction in Telecommunication A Decade Review and Classification. International Journal of Computer Science, 10(5), 271-282.

[4] Saradhi, V. V., &Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999-2006.

[5] AlOmari, D., & Hassan, M. M. (2016, September). Predicting Telecommunication Customer Churn Using Data Mining Techniques.In International Conference on Internet and Distributed Computing Systems (pp. 167-178).Springer International Publishing.

[6] Coussement, K., & Van den Poel, D. (2013). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, 36, 6127-6134.

[7] Owczarczuk, M. (2010), Churn models for prepaid customers in the cellular telecommunication industry using large data marts, Expert Systems with Applications, 37(6) pp.4710-4712.

[8] Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. Technology in society, 24(4), 483-502.

[9] Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629-1636.

[10] Statista (2016), Average Monthly Churn Rate for Wireless Carriers in the United States from 1st Quarter 2013 to 1st Quarter 2016,available at: https://www.statista.com/statistics/283511/average-monthly-churn-rate-top-wireless-carriers-us/(Accessed 14, October 2017).

[11] Risselada, H., Verhoef, P. C., &Bijmolt, T. H. (2010). Staying power of churn prediction models. Journal of Interactive Marketing, 24(3), 198-208.

[12] Coltman, T. (2007). Why build a customer relationship management capability?. The Journal of Strategic Information Systems, 16(3), 301-320.

[13] Ling, R., & Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. Journal of computer information systems, 41(3), 82-97.

[14] Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. Expert Systems with Applications, 38(1), 8-15.

[15] Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., ...&Sriram, S. (2006). Modeling customer lifetime value. Journal of service research, 9(2), 139-155.

[16] Buttle, F. (2009). Customer relationship management: concepts and technologies. Routledge.

[17] ITU (2017), ICT Facts and Figures 2017, available at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf (accessed 19, September 2017).

[18] Statista, mobile users worldwide from 2020 to 2025, available at: https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/ (Accessed 30, July 2024).

[19] Database Market institute (2017), Churn reduction in the telecomsindustry, available at: http://www.dbmarketing.com/telecom/churnreduction.html (Accessed 19, September 2017).

[20] TamaddoniJahromi, A., Sepehri, M. M., Teimourpour, B., &Choobdar, S. (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. Journal of Strategic Marketing, 18(7), 587-598

[21] AT&T INC. FINANCIAL REVIEW available at: https://about.att.com/story/2024/q1-earnings.html (Accessed 30, July 2024).

[22] AT&T, sustainable grouth strategy pays off with strong 2Q Results, available at: https://about.att.com/story/2023/q2-earnings.html (Accessed 30, July 2024).

[23] Gürsoy, U. T. Ş. (2010). Customer churn analysis in telecommunication sector. Istanbul University Journal of the School of Business, 39(1), 35-49.

[24] Oghojafor, B., Mesike, G., Bakarea, R., Omoera, C., &Adeleke, I. (2012). Discriminant analysis of factors affecting telecoms customer churn. International Journal of Business Administration, 3(2), 59.

[25] Tenhunen, S. (2008). Mobile technology in the village: ICTs, culture, and social logistics in India. Journal of the Royal Anthropological Institute, 14(3), 515-534.

[26] Halim, J &Vucetic J (2015). Causes of Churn in the Wireless Telecommunication Industry in Kenya. DOI: 10.14355/ijmser.2016.0301.01

[27] Saefuddin, A., Setiabudi, N. A., &Achsani, N. A. (2011). The effect of overdispersion on regression based decision with application to churn analysis on Indonesian mobile phone industry.

[28] Sey, A. (2009). Exploring mobile phone-sharing practices in Ghana. info, 11(2), 66-78.

[29] Shaffer, Greg, and Z. John Zhang. "Competitive one-to-one promotions." Management Science 48, no. 9 (2002): 1143-1160.

[30] Hadden, J., Tiwari, A., Roy, R., &Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902-2917.

[31] Coltman, T. (2007). Why build a customer relationship management capability?. The Journal of Strategic Information Systems, 16(3), 301-320.

[32] Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason. "Defection detection: Measuring and understanding the predictive accuracy of customer churn models." Journal of marketing research 43, no. 2 (2006): 204-211.

[33] Chen, Y., Zhang, G., Hu, D., & Fu, C. (2007). Customer segmentation based on survival character. Journal of intelligent manufacturing, 18(4), 513-517.

[34] Jiang, W., Au, T., &Tsui, K. L. (2007). A statistical process control approach to business activity monitoring. Iie Transactions, 39(3), 235-249.

[35] Wierenga, B. (2010). Marketing and artificial intelligence: Great opportunities, reluctant partners. In Marketing intelligent systems using soft computing (pp. 1-8). Springer Berlin Heidelberg.

[36] Risselada, H., Verhoef, P. C., &Bijmolt, T. H. (2010). Staying power of churn prediction models. Journal of Interactive Marketing, 24(3), 198-208.

[37] Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., ...& Wedel, M. (2005). Choice models and customer relationship management. Marketing Letters, 16(3), 279-291.

[38] Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management, 45(3), 164-174.

[39] Gordini, N., &Veglio, V. (2017). Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry. Industrial Marketing Management, 62, 100-107.

[40] Statista, a valuable at: https://www.statista.com/topics/1147/mobile-communications/, (Accessed 5, October 2020).

[41] Statista, a valuable at:https://www.statista.com/statistics/470018/mobile-phone-user-penetration-worldwide/, (Accessed 5, October 2020)

[42] ICT Facts and Figures, a valuable at: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf, (Accessed 5, October 2020)

[43] Statista, a valuable at:https://www.statista.com/statistics/692056/frontier-communications-average-monthly-revenue-per-customer/, (Accessed 5, October 2020)

[44] Seeking Alpha, Chart: TelecomsCompanies - Gross Profit Margins, a valuable at:https://seekingalpha.com/article/10168-chart-telecom-companies-gross-profit-margins, (Accessed 5, October 2020)

[45] Churn in the telecomsindustry – identifying customers likely to churn and how to retain them.Aditya Kapoorm a valuable at:https://wp.nyu.edu/adityakapoor/2017/02/17/churn-in-the-telecom-industry-identifying-customers-likely-to-churn-and-how-to-retain-them/, (Accessed 5, October 2020).

[46] Propeller a valuable at:https://www.propellercrm.com/blog/customer-acquisition-cost,, (Accessed 5, October 2020)

[47] Braun, M., &Schweidel, D. A. (2011). Modeling customer lifetimes with multiple causes of churn. Marketing Science, 30(5), 881-902.

[48] Antipov, E., & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement. Journal of Targeting, Measurement and Analysis for Marketing, 18(2), 109-117.

[49] Wong, K. K. K. (2011). Getting what you paid for: Fighting wireless customer churn with rate plan optimization. Journal of Database Marketing & Customer Strategy Management, 18(2), 73-82.

[50] Ranaweera, C. (2007). Are satisfied long-term customers more profitable? Evidence from the telecommunication sector. Journal of Targeting, Measurement and Analysis for Marketing, 15(2), 113-120.

[51] Wong, K. K. K. (2011). Using Cox regression to model customer time to churn in the wireless telecommunications industry. Journal of Targeting, Measurement and Analysis for Marketing, 19(1), 37-43.

[52] Kumar, V., Bhagwat, Y., & Zhang, X. A. (2015, May). Regaining “Lost” Customers: The Predictive Power of First-Lifetime Behavior, the Reason for Defection, and the Nature of the Win-Back Offer. American Marketing Association.

[53] Lemmens, A., &Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286.

[54] Foster P, and Fawcett T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data Analytic Thinking, O'Reilly Media, 2013.

[55] Marwanto, S. T., & Komaladewi, R. (2017). How to restrain customer churn in telecommunication providers: study in west java Indonesia. Review of Integrative Business and Economics Research, 6, 51.

[56] Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: a machine learning approach. Computers & Industrial Engineering, 106476.

[57] Coussement, K. (2014). Improving customer retention management through cost-sensitive learning. European Journal of Marketing, 48(3/4), 477-495.

[58] Gordini, N., &Veglio, V. (2017). Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry. Industrial Marketing Management, 62, 100-107.

[59] Idris, A., Khan, A., & Lee, Y. S. (2012, October). Genetic programming and adaboosting based churn prediction for telecom. In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on (pp. 1328-1332).

[60] Miguéis, V. L., Camanho, A., & e Cunha, J. F. (2013). Customer attrition in retailing: an application of multivariate adaptive regression splines. Expert Systems with Applications, 40(16), 6225-6232.

[61] Brandusoiu, I., &Toderean, G. (2013). Churn prediction in the telecommunications sector using support vector machines. Margin, 1, x1.

[62] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., &Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing, 24, 994-1012.

[63] hen, K., Hu, Y. H., & Hsieh, Y. C. (2015). Predicting customer churn from valuable B2B customers in the logistics industry: a case study. Information Systems and e-Business Management, 13(3), 475-494.

[64] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., &Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

[65] Zhang, Z., Wang, R., Zheng, W., Lan, S., Liang, D., &Jin, H. (2015, November). Profit Maximization Analysis Based on Data Mining and the Exponential Retention Model Assumption with Respect to Customer Churn Problems.

In Data Mining Workshop (ICDMW), 2015 IEEE International Conference on(pp. 1093-1097).

[66] Hassouna, M., Tarhini, A., Elyas, T., &AbouTrab, M. S. (2016). Customer Churn in Mobile Markets A Comparison of Techniques. arXiv preprint arXiv:1607.07792.

[67] Umayaparvathi, V., &Iyakutti, K. (2016, March). Attribute selection and Customer Churn Prediction in telecomsindustry. In Data Mining and Advanced Computing (SAPIENCE), International Conference on (pp. 84-90).IEEE.

[68] Abdullaev, I., Prodanova, N., Ahmed, M. A., Lydia, E. L., Shrestha, B., Joshi, G. P., & Cho, W. (2023). Leveraging metaheuristics with artificial intelligence for customer churn prediction in telecom industries. Electronic Research Archive, 31(8), 4443-4458.

[69] Coussement, K., Lessmann, S., &Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

[70] Prashanth, R., Deepak, K., &Meher, A. K. (2017, July). High Accuracy Predictive Modelling for Customer Churn Prediction in TelecomsIndustry.In International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 391-402).Springer, Cham.

[71] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254.

[72] Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecomsusing fuzzy classifiers. Telecommunication Systems, 1-12.

[73] Zhu, B., Baesens, B., &Backiel, A. (2017). Benchmarking sampling techniques for imbalance learning in churn prediction. Journal of the Operational Research Society.

[74] Effendy, V., & Baizal, Z. A. (2014, May). Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In Information and Communication Technology (ICoICT), 2014 2nd International Conference on (pp. 325-330)..

[75] Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. Artificial Intelligence Research, 6(2), 93.

[76] De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772.

[77] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecoms sector. IEEE Access, 7, 60134-60149.

[78] Jafari-Marandi, R., Denton, J., Idris, A., Smith, B. K., & Keramati, A. (2020). Optimum profit-driven churn decision making: innovative artificial neural networks in telecomsindustry. Neural Computing and Applications, 32(18), 14929-14962

[79] Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Int J Pattern Recogn Artif Intell. 2009;23(04):687–719.

[80] Faritha Banu, J., Neelakandan, S., Geetha, B. T., Selvalakshmi, V., Umadevi, A., & Martinson, E. O. (2022). Artificial intelligence based customer churn prediction model for business markets. Computational Intelligence and Neuroscience, 2022(1), 1703696.

[81] Praseeda, C. K., & Shivakumar, B. L. (2023). Fuzzy particle swarm optimization (FPSO) based feature selection and hybrid kernel distance based possibilistic fuzzy local information C-means (HKD-PFLICM) clustering for churn prediction in telecom industry. SN Applied Sciences, 3, 1-18.

[82] Chen Z, Yan Q, Han H, Wang S, Peng L, Wang L, Yang B. Machine learning based mobile malware detection using highly imbalanced network traffic. Inf Sci. 2018;433:346–64.

[83] Jain A, Ratnoo S, Kumar D (2017) Addressing class imbalance problem in medical diagnosis: a genetic algorithm approach. In: 2017 international conference on information, communication, instrumentation and control (ICICIC) (pp. 1–8), IEEE

[84] Ramli NA, Ismail MT, Wooi HC. Measuring the accuracy of currency crisis prediction with combined classifiers in designing early warning system. Mach Learn. 2015;101(1–3):85–103.

[85] Dwiyanti E, Ardiyanti A (2016) Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt. telekomunikasiindonesia regional 7). In: International conference on soft computing and data mining (pp 376–385). Springer, Cham

[86] He B, Shi Y, Wan Q, Zhao X. Prediction of customer attrition of commercial banks based on SVM model. Procedia Comput Sci. 2014;31:423–30.

[87] Huang PJ (2015) Classication of imbalanced data using synthetic over-sampling techniques, Doctoral dissertation, University of California

[88] Chawla NV (2009) Data mining for imbalanced datasets: an overview. In: Data mining and knowledge discovery handbook (pp 875–886). Springer, Boston

[89] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.

[90] Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. J Bus Res. 2019;94:290–301.

[91] Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. ACM SIGKDD Explor Newsl. 2004;6(1):1–6.

[92] Liu XY, Wu J, Zhou ZH (2009) Exploratory under sampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B Cybern 39(2):539–550

[93] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

[94] He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Neural networks, 2008. IJCNN 2008 (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on (pp 1322–1328), IEEE

[95] Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing (pp 878–887). Springer, Berlin, Heidelberg

[96] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-levelsmote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Adv Knowl Discov Data Min. 2009;2009:475–82.

[97] Maciejewski T, Stefanowski J (2011) Local neighbourhood extension of SMOTE for mining imbalanced data. In: Computational intelligence and data mining (CIDM), 2011 IEEE symposium on (pp 104–111), IEEE

[98] Barua S, Islam MM, Yao X, Murase K. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2014;26(2):405–25.

[99] Zhu B, Broucke S, Baesens B, Maldonado S (2017) improving resampling-based ensemble in churn prediction. In: First international workshop on learning with imbalanced domains: theory and applications, pp 79–91

[100] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hussain A, et al. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access. 2016;4:7940–57.

[101] Salunkhe UR, Mali SN. A hybrid approach for class imbalance problem in customer churn prediction: a novel extension to under sampling. Int J Intell Syst Appl. 2018;10(5):71.

[102] Zou S, Huang Y, Wang Y, Wang J, Zhou C (2008) SVM learning from imbalanced data by GA sampling for protein domain prediction. In: 2008 the 9th international conference for young computer scientists (pp 982–987), IEEE

[103] Haque MN, Noman N, Berretta R, Moscato P. Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification. PLoS ONE. 2016;11:1.

[104] Cervantes J, Li X, Yu W (2013) Using genetic algorithm to improve classification accuracy on imbalanced data. In: 2013 IEEE international conference on systems, man, and cybernetics (pp 2659–2664), IEEE

[105] Jiang K, Lu J, Xia K. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. Arab J Sci Eng. 2016;41(8):3255–66.

[106] Karia V, Zhang W, Naeim A, Ramezani R (2019) GenSample: a genetic algorithm for oversampling in imbalanced datasets. arXiv: 1910.10806

[107] Mahin M, Islam MJ, Khatun A, Debnath BC (2018) A comparative study of distance metric learning to find sub-categories of minority class from imbalance data. In: 2018 international conference on innovation in engineering and technology (ICIET) (pp 1–6), IEEE

[108] El Hindi K. Specific-class distance measures for nominal attributes. AI Commun. 2013;26(3):261–79.

[109] Li C, Li H. A survey of distance metrics for nominal attributes. J Softw. 2010;5(11):1262–9.

[110] Wilson DR, Martinez TR. Improved heterogeneous distance functions. J Artif Intell Res. 1997;6:1–34.

[111] Mahin M, Islam MJ, Debnath BC, Khatun A (2019) Tuning distance metrics and K to find sub-categories of minority class from imbalance data using K nearest neighbours. In: 2019 international conference on electrical, computer and communication engineering (ECCE) (pp 1–6), IEEE

[112] Guo H, Viktor HL. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM SIGKDD Explor Newsl. 2004;6(1):30–9.

[113] Liu Y, Yu X, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manage. 2011;47(4):617–31.

[114] Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform. 2015;58:49–59.

[115] Kagie M, van Wezel M, Groenen PJ (2009) An empirical comparison of dissimilarity measures for recommender systems

[116] Tsymbal A, Pechenizkiy M, Cunningham P (2006) Dynamic integration with random forests. In: European conference on machine learning, pp 801–808. Springer, Berlin, Heidelberg

[117] El-Sappagh S, Elmogy M, Ali F, Abuhmed T, Islam SM, Kwak KS. A comprehensive medical decision-support framework based on a heterogeneous ensemble classifier for diabetes prediction. Electronics. 2019;8(6):635.

[118] Vandecruys O, Martens D, Baesens B, Mues C, De Backer M, Haesen R. Mining software repositories for comprehensible software fault prediction models. J Syst Softw. 2008;81(5):823–39.

[119] Rokach L, Maimon OZ (2008) Data mining with decision trees: theory and applications (vol 69). World scientific

[120] Das B, Krishnan NC, Cook DJ (2013) Handling class overlap and imbalance to detect prompt situations in smart homes. In: 2013 IEEE 13th international conference on data mining workshops, pp 266–273, IEEE

[121] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2008;9:1263–84.

[122] Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Inf Sci. 2019;501:118–35.

[123] Zhang H, Wang Z (2011) A normal distribution-based over-sampling approach to imbalanced data classification. In: International conference on advanced data mining and applications, pp 83–96. Springer, Berlin, Heidelberg

[124] García S, Molina D, Lozano M, Herrera F. A study on the use of non-parametric tests for analyzing the evolutionary algorithms’ behaviour: a case study on the CEC’2005 special session on real parameter optimization. J Heuristics. 2009;15 (6):617.

[125] Li, Y., Hou, B., Wu, Y., Zhao, D., Xie, A., & Zou, P. (2021). Giant fight: Customer churn prediction in traditional broadcast industry. Journal of Business Research, 131, 630-639

[126] Kim, S., Chang, Y., Wong, S. F., & Park, M. C. (2020). Customer resistance to churn in a mature mobile telecommunications market. International Journal of Mobile Communications, 18(1), 41-66..

[127] Ascarza, E., & Hardie, B. G. (2013). A joint model of usage and churn in contractual settings. Marketing Science, 32(4), 570-590.

[128] Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. Journal of Marketing Research, 53(1), 46-60.

[129] Hassani, Z., Hajihashemi, V., Borna, K., &SahraeiDehmajnoonie, I. (2020). A Classification Method for E-mail Spam Using a Hybrid Approach for Feature Selection Optimization. Journal of Sciences, Islamic Republic of Iran, 31(2), 165-173.

[130] Manochandar, S., &Punniyamoorthy, M. (2018). Scaling feature selection method for enhancing the classification performance of Support Vector Machines in text mining. Computers & Industrial Engineering, 124, 139-156.

[131] Rajput, U., &Kumari, M. (2017). Mobile robot path planning with modified ant colony optimization. International Journal of Bio-Inspired Computation, 9(2), 106-113.

[132] Manjhi, Y., &Dhar, J. (2016, May). Forecasting energy consumption using particle swarm optimization and gravitational search algorithm. In 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) (pp. 417-420). IEEE.

[133] Chrouta, J., Chakchouk, W., Zaafouri, A., &Jemli, M. (2018). Modeling and control of an irrigation station process using heterogeneous cuckoo search algorithm and fuzzy logic controller. IEEE Transactions on Industry Applications, 55(1), 976-990.

[134] Al-Shourbaji, I., & Zogaan, W. (2021). A new method for human resource allocation in cloud-based e-commerce using a meta-heuristic algorithm. Kybernetes.

[135] Oladele, T. O., Olorunsola, B. J., Aro, T. O., Akande, H. B., & Olukiran, O. A. (2021). Nature-Inspired Meta-heuristic Optimization Algorithms for Breast Cancer Diagnostic Model: A Comparative Study. FUOYE Journal of Engineering and Technology, 6(1).

[136] Hussain, K., Salleh, M. N. M., Cheng, S., & Shi, Y. (2019). Metaheuristic research: a comprehensive survey. Artificial Intelligence Review, 52(4), 2191-2233.

[137] Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019). IEEE Access, 9, 26766-26791.

[138] Fred Glover (1989). "Tabu Search – Part 1". ORSA Journal on Computing. 1 (2): 190–206. doi:10.1287/ijoc.1.3.190

[139] Feo, Thomas A.; Resende, Mauricio G. C. (April 1989). "A probabilistic heuristic for a computationally difficult set covering problem". Operations Research Letters. 8 (2): 67–71. doi:10.1016/0167-6377(89)90002-3

[140] Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. Computers & operations research, 24(11), 1097-1100.

[141] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

[142] Gandomi, A. H., Yang, X. S., & Alavi, A. H. (2013). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. Engineering with computers, 29(1), 17-35.

[143] Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). IEEE.

[144] Yang, X. S. (2008). Nature-Inspired Metaheuristic Algorithms. Luniver Press. ISBN 978-1-905986-10-1.

[145] De Souza, R. C. T., dos Santos Coelho, L., De Macedo, C. A., & Pierezan, J. (2018, July). A V-shaped binary crow search algorithm for feature selection. In 2018 IEEE congress on evolutionary computation (CEC) (pp. 1-8). IEEE..

[146] Mirjalili, S. (2016). Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Computing and Applications, 27(4), 1053-1073.

[147] Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. IEEE computational intelligence magazine, 1(4), 28-39.

[148] Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-verse optimizer: a nature-inspired algorithm for global optimization. Neural Computing and Applications, 27(2), 495-513.

[149] Abualigah, L., Abd Elaziz, M., Sumari, P., Geem, Z. W., & Gandomi, A. H. (2021). Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer. Expert Systems with Applications, 116158.

[150] Pustokhina, I. V., Pustokhin, D. A., Aswathy, R. H., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706.

[151] Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecomsdata using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), 215-220.

[152] Sivasankar, K. (2016). Effective Customer Churn Prediction on Large Scale Data using Metaheuristic Approach. Indian Journal of Science and Technology, 9, 33.

[153] Özmen, M., Aydoğan, E. K., Delice, Y., & Toksarı, M. D. (2020). Churn prediction in Turkey's telecommunications sector: A proposed multiobjective–cost-sensitive ant colony optimization. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1338.

[154] Venkatesh, S., and Jeyakarthic, M. (2020), Metaheuristic based Optimal Feature Subset Selection with Gradient Boosting Tree Model for IoT Assisted Customer Churn Prediction. Journal of Seybold Report ISSN NO, 1533, 9211.

[155] Li, K. G., & Marikannan, B. P. (2019). Hybrid particle swarm optimization-extreme learning machine algorithm for customer churn prediction. Journal of Computational and Theoretical Nanoscience, 16(8), 3432-3436.

[156] Praseeda, C. K., & Shivakumar, B. L. (2021). Fuzzy particle swarm optimization (FPSO) based feature selection and hybrid kernel distance based possibilistic fuzzy local information C-means (HKD-PFLICM) clustering for churn prediction in telecomsindustry. SN Applied Sciences, 3(6), 1-18.

[157] Faris, H. (2018). A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, 9(11), 288.

[158] Vijaya, J., & Sivasankar, E. (2019). An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. Cluster Computing, 22(5), 10757-10768.

[159] Wu, Y., Gong, M., Ma, W., & Wang, S. (2019). High-order graph matching based on ant colony optimization. Neurocomputing, 328, 97-104.

[160] Dorigo, M., &Stützle, T. (2019). Ant colony optimization: overview and recent advances. Handbook of metaheuristics, 311-351.

[161] Kanan, H. R., Faez, K., & Taheri, S. M. (2007, July). Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In Industrial conference on data mining (pp. 63-76). Springer, Berlin, Heidelberg.

[162] Beer, C., Hendtlass, T., & Montgomery, J. (2012, June). Improving exploration in ant colony zoptimization with antennation. In 2012 IEEE Congress on Evolutionary Computation (pp. 1-8). IEEE.

[163] Ibrahim, R. A., Abd Elaziz, M., Ewees, A. A., El-Abd, M., & Lu, S. (2021). New feature selection paradigm based on hyper-heuristic technique. Applied Mathematical Modelling, 98, 14-37.

[164] Talbi, E. G. (2002). A taxonomy of hybrid metaheuristics. Journal of heuristics, 8(5), 541-564..

[165] Wang, A., An, N., Chen, G., Li, L., &Alterovitz, G. (2015). Accelerating wrapper-based feature selection with K-nearest-neighbor. Knowledge-Based Systems, 83, 81-91.

[166] AlShourbaji, I., Helian, N., Sun, Y., & Alhameed, M. (2021). Anovel HEOMGA Approach for Class Imbalance Problem in the Application of Customer Churn Prediction. SN Computer Science, 2(6), 1-12.

[167] Pustokhina, I. V., Pustokhin, D. A., Aswathy, R. H., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706.

[168] Martin, L., Leblanc, R., & Toan, N. K. (1993). Tables for the Friedman rank test. Canadian journal of statistics, 21(1), 39-43.

[169] Hussain, K., Salleh, M. N. M., Cheng, S., & Shi, Y. (2019). On the exploration and exploitation in popular swarm-based metaheuristic algorithms. Neural Computing and Applications, 31, 7665-7683.

[170] Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications, 40(14), 5635-5647.

[171] De Bock, K. W., Coussement, K., & Van den Poel, D. (2010). Ensemble classification based on generalized additive models. Computational Statistics & Data Analysis, 54(6), 1535-1546.

[172] Zhou, Y., Li, T., Shi, J., & Qian, Z. (2019). A CEEMDAN and XGBOOST-based approach to forecast crude oil prices. Complexity, 2019.

[173] Athanasiou, V., & Maragoudakis, M. (2017). A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern Greek. Algorithms, 10(1), 34.

[174] Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543.

[175] Bibault, J. E., Chang, D. T., & Xing, L. (2021). Development and validation of a model to predict survival in colorectal cancer using a gradient-boosted machine. Gut, 70(5), 884-889.

[176] Sharma, T., Gupta, P., Nigam, V., & Goel, M. (2020). Customer Churn Prediction in Telecommunications Using Gradient Boosted Trees.

In International Conference on Innovative Computing and Communications (pp. 235-246). Springer, Singapore.

[177] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[178] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

[179] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

[180] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[181] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[182] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural water management, 225, 105758.

[183] Martınez-Munoz, G., & Superior, E. P. (2019). Sequential training of neural networks with gradient boosting. arXiv preprint arXiv:1909.12098.

[184] Feng, J., Yu, Y., & Zhou, Z. H. (2018). Multi-layered gradient boosting decision trees. Advances in neural information processing systems, 31.

[185] Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. arXiv preprint arXiv:1802.03396.

[186] Jaisakthi, S. M., Gayathri, N., Uma, K., & Vijayarajan, V. (2018). Customer Churn Prediction Using Stochastic Gradient Boosting Technique. Journal of Computational and Theoretical Nanoscience, 15(6-7), 2410-2414.

[187] Wang, Q. F., Xu, M., & Hussain, A. (2019). Large-scale ensemble model for customer churn prediction in search ads. Cognitive Computation, 11(2), 262-270.

[188] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecomsusing machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[189] Jain, H., Yadav, G., & Manoov, R. (2021). Churn Prediction and Retention in Banking, Telecomsand IT Sectors Using Machine Learning Techniques. In Advances in Machine Learning and Computational Intelligence (pp. 137-156). Springer, Singapore.

[190] Dhini, A., & Fauzan, M. (2021). Predicting Customer Churn using ensemble learning: Case Study of a Fixed Broadband Company. International Journal of Technology, 12(5), 1030-1037.

[191] Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of advanced computer Science and applications, 9(2).

[192] Sandhya, G., Samarpana, S., & Sangeetha Vani, R. (2021). A Hybrid Learning System for TelecomsChurn Prediction Using Ensemble Learning. In Computer Networks and Inventive Communication Technologies (pp. 927-934). Springer, Singapore

[193] Kimura, T. (2022). CUSTOMER CHURN PREDICTION WITH HYBRID RESAMPLING AND ENSEMBLE LEARNING. Journal of Management Information & Decision Sciences, 25(1).

[194] Zhu, M., & Liu, J. (2021, December). TelecomsCustomer Churn Prediction Based on Classification Algorithm. In 2021 International Conference on Aviation Safety and Information Technology (pp. 268-273).

[195] Kanwal, S., Rashid, J., Kim, J., Nisar, M. W., Hussain, A., Batool, S., & Kanwal, R. (2021, November). An Attribute Weight Estimation Using Particle Swarm Optimization and Machine Learning Approaches for Customer Churn Prediction. In 2021 International Conference on Innovative Computing (ICIC) (pp. 1-6). IEEE.

[196] Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecomsindustry. PeerJ Computer Science, 8, e854.

[197] Karuppaiah, S., & Gopalan, N. P. (2021). Enhanced Churn Prediction Using Stacked Heuristic Incorporated Ensemble Model. Journal of Information Technology Research (JITR), 14(2), 174-186.

[198] Rabbah, J., Ridouani, M., & Hassouni, L. (2022). A New Churn Prediction Model Based on Deep Insight Features Transformation for Convolution Neural Network Architecture and Stacknet. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 17(1), 1-18.

[199] Rodan, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2014). A support vector machine approach for churn prediction in telecomsindustry. International journal on information, 17(8), 3961-3970.

[200] Al-Shourbaji, I., Helian, N., Sun, Y., Alshathri, S., & Abd Elaziz, M. (2022). Boosting Ant Colony Optimization with Reptile Search Algorithm for Churn Prediction. Mathematics, 10(7), 1031.

[201] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.

[202] Badirli, S., Liu, X., Xing, Z., Bhowmik, A., Doan, K., & Keerthi, S. S. (2020). Gradient boosting neural networks: Grownet. arXiv preprint arXiv:2002.07971.

[203] Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543.

[204] Martınez-Mufioz, G., & Superior, E. P. (2019). Sequential training of neural networks with gradient boosting. arXiv preprint arXiv:1909.12098.

[205] Feng, J., Xu, Y. X., Jiang, Y., & Zhou, Z. H. (2020). Soft gradient boosting machine. arXiv preprint arXiv:2006.04059.

[206] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. Artificial intelligence, 137(1-2), 239-263.

[207] Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

[208] Patle, A., & Chouhan, D. S. (2013, January). SVM kernel functions for classification. In 2013 International Conference on Advances in Technology and Engineering (ICATE) (pp. 1-9). IEEE.

[209] Xia, J., Wang, Z., Yang, D., Li, R., Liang, G., Chen, H., ... & Pan, Z. (2022). Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. Computers in Biology and Medicine, 143, 105206.

[210] Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). IEEE.

[211] Zhao, W., Wang, L., & Zhang, Z. (2020). Artificial ecosystem-based optimization: a novel nature-inspired meta-heuristic algorithm. Neural Computing and Applications, 32(13), 9383-9425.

[212] Haklı, H., & Uğuz, H. (2014). A novel particle swarm optimization algorithm with Levy flight. Applied Soft Computing, 23, 333-345.

[213] Kołodziejczyk, J., & Tarasenko, Y. (2021). Particle Swarm Optimization and Levy Flight integration. Procedia Computer Science, 192, 4658-4671.

[214] Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. The American Statistician, 52(2), 181-184.

[215] Painsky, A., & Wornell, G. (2018, June). On the universality of the logistic loss function. In 2018 IEEE International Symposium on Information Theory (ISIT) (pp. 936-940). IEEE.

[216] Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. Advances in engineering software, 95, 51-67.