Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)
IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

MACHINE LEARNING BASED CUSTOMER CHURN PREDICTION IN BANKING

Manas Rahman
Department of Computer Science
Central University of Kerala
Periye, Kasaragod
manasrahmanpadiyoor@gmail.com

V Kumar
Department of Computer Science
Central University of Kerala
Periye, Kasaragod
vkumar@cukerala.ac.in

***Abstract*—The number of service providers are being increased very rapidly in every business. In these days, there is no shortage of options for customers in the banking sector when choosing where to put their money. As a result, customer churn and engagement has become one of the top issues for most of the banks. In this paper, a method to predicts the customer churn in a Bank, using machine learning techniques, which is a branch of artificial intelligence is proposed. The research promotes the exploration of the likelihood of churn by analyzing customer behavior. The KNN, SVM, Decision Tree, and Random Forest classifiers are used in this study. Also, some feature selection methods have been done to find the more relevant features and to verify system performance. The experimentation was conducted on the churn modeling dataset from Kaggle. The results are compared to find an appropriate model with higher precision and predictability. As a result, the use of the Random Forest model after oversampling is better compared to other models in terms of accuracy.**

***Index Terms*—Customer churn in Bank, k-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest.**

I. Introduction

The market is very dynamic and highly competitive nowadays. It is because of the availability of a large number of service providers. The challenges of service providers are finding the changing customer behavior and their rising expectations. The raising aspirations of current generation consumers and their diverse demands for connectivity and innovative, personalized approaches are very distinct from previous generations of consumers. They are well educated and better informed of emerging approaches. Such advanced knowledge has changed their purchasing behavior, resulting in a trend of 'analysis-paralysis' over-analyzing the selling and purchase scenario, which ultimately helps them to improve their purchase decisions. Therefore, this is a big challenge for the new generation service providers to think of innovatively to fulfil and add values to the customers.

Corporations need to recognize their consumers. Liu and Shih [1] strengthen this argument by implying that increasing competitive pressures on organizations to develop innovative marketing approaches, to meet consumer expectations and enhance loyalty and retention. Canning [2] argues that offering more to all is no longer a viable sales strategy, and a market environment that continues to become more competitive needs an agenda that emphasizes on the most effective use of marketing capital. Technology has been used to help businesses to retain a competitive edge [3]. Data mining techniques [4] are a commonly used information technology for the extraction of marketing expertise and further guidance for business decisions.

It is very easy for customers to switch from one organization(Bank) to another for a better service quality or price rates. Organizations are convinced that recruiting new customers is far more expensive and hard than keeping existing clients [5]. But delivering reliable service on time and in budget to customers while maintaining a good working partnership with them is another significant challenge for them. They need to consider consumers and their needs to resolve these challenges. Among these, one of their primary emphasis will be on client churn. Customer churn takes place when clients or subscribers cease to engage incorporation with a company or service. For any organization, winning business from new clients means going via the sales pipeline, using their sales and marketing assets in the cycle. Customer retention, on the other hand, is usually more budget-effective, because they have already gained the confidence and loyalty of current customers. So, the need for a system that can efficiently predict customer churn in the early stages is really important for any organization. This paper aims to build a framework that can predict the client churn in the banking sector using some Machine learning techniques [6].

II. Literature Review

The analysis of the client churn in banking is a really broad area. In one of these studies, [7] pursue commercial bank client churn prediction based on the SVM model. For this work, a Chinese commercial bank consumer dataset that contains 50,000 customer information is selected. After preprocessing records, there are eventually 46,406 valid data records. Two types of SVM model is selected: linear SVM and SVM with radial basis kernel function. The predictive effect of the classification models was greatly improved by the undersampling approach. Due to the lopsided features of the actual commercial bank client churn dataset, the SVM model can not accurately predict churners and even the general assessment parameters can not calculate the predictive power of the model. The findings show that the integration of the random

sampling approach with the SVM model can substantially increase predictive capacity and help commercial banks to predict churners more precisely. But this study used a 1:10 proportion for churners to non-churners. In 1:1, the result is getting as 80.84% at maximum. This is the main drawback of this work.

In another study [8], a scientific study of the use of data mining in the extraction of information from repositories in the banking sector is presented. The findings show that customers who use more banking services (products) seem to be more loyal, so the bank can concentrate on those customers who use fewer than three products and sell them goods as per their needs. The database used consists of records on 1866 customers on the date of the study. The research is based on one method of churn prediction using a Neural network within the software package Alyuda NeuroInteligence. Which divides Data into three sets: training, validation, and testing set. Three forms of characteristics are described in the data analysis stage: the characteristics that to refuse, the characteristics that need, and the target characteristics to be measured. The model picks several hidden layers in the network design process. After training the network, the results are; the CCR % of validation is 93,959732. The study concluded that, because of the high proportion of retirees in the total number of customers (691/1886), the bank has very well-tailored programs for retirees, and the probability of competing is extremely small. The biggest downside of this work is that the neural network is relatively slow and tedious. Table I summarised the churn prediction in the banking system using 'Chinese Commercial bank data' and 'data from a small Croatian bank', and is also shown the drawbacks of the existing works, to overcome these drawbacks, this work proposed an ML-based customer churn prediction in banking system using dataset 'churn modeling data'.

The study [9] proposed a churn analysis model that helps telecommunication operators to predict customers that are most likely to get churned. The system uses machine learning strategies on a big data platform. The Area Under Curve (AUC) standard measure is used to assess the efficiency of the model. The dataset used for the study was provided by the Syriatel telecom company. The model has worked with 4 methodologies: Decision Tree, Random Forest, Gradient Boosted Machine Tree(GBM), and Extreme Gradient Boosting(XGBOOST). The Hortonworks Data Platform (HDP) was selected as the big data platform. Spark engines were used in almost all of the product's phases such as data analysis, function development, training, and software testing. The algorithm hyper-parameters were optimized with the aid of K-fold cross-validation. Since the target class is unbalanced, the sample for learning is rebalanced by taking a sample of data to balance the two classes. The study began with oversampling by multiplying the churn class to fit with the other class. A random under-sampling approach was also used, which decreases the sample size of the broad class to be compared with the second class. The training started on the Decision Tree algorithm and optimizing the hyper-parameter

TABLE I
PREVIOUS WORKS ON BANK CHURN PREDICTION

Authors	Title of the work	Year	Methodology	Remarks
B. He, Y. Shi, Q. Wan, and X. Zhao	Prediction of customer attrition of commercial banks based on svm model	2014	SVM Classification	Chinese commercial bank data, lesser accuracy in 1:1 ratio of churners to non-churners, 80.84% accuracy.
A. Bilal Zeri	Predicting customer churn in banking industry using neural networks	2016	Alyuda Neural network	small Croatian bank data, neural network is relatively slow and tedious, 93.30% accuracy.

depth and maximum number of nodes. In both Random Forest and GBM, the best results show that the best number of trees was 200 trees. And GBM got better results than DT and RF. The result showed that the best AUC value was 93.301% for XGBOOST on 180 trees. The models are tested by installing a new dataset for various times and without any constructive marketing intervention, XGBOOST also provided the best results with 89% AUC. The study hypothesized that the resulting decrease could be attributed to the non-stationary data model phenomenon, so the model needs to train per time.

III. METHODOLOGY

This work aims to predict customer churn in a commercial bank as early as possible using efficient data mining methods. A diagrammatic representation of the proposed model is given in Fig 1.

A. Dataset description

The dataset used in this analysis was obtained from Kaggle to model churns. The dataset includes information of 10000 bank clients, and the target parameter is a binary variable that represents whether the customer has left the bank or still a customer. Of this, 7963 were positive class(maintained) samples and 2037 were negative class(exited) samples. The target variable reflects the binary flag 1 when the client has a bank account closed, and 0 when the client is retained. The dataset contains 13 feature vectors(predictors) that were reported from customer data and transactions processed by the customer. The details of these features are given in Table II.

B. Data Preprocessing

Preprocessing the data is a significant phase in the process of data mining. Since they have a direct effect on task success rate. It must deal with irrelevance, noisiness, and unreliability of data. And if necessary, the data conversion too. Predictors descriptions after preprocessing are listed in Table III. These are the attributes taken for deciding on churn prediction in this study.

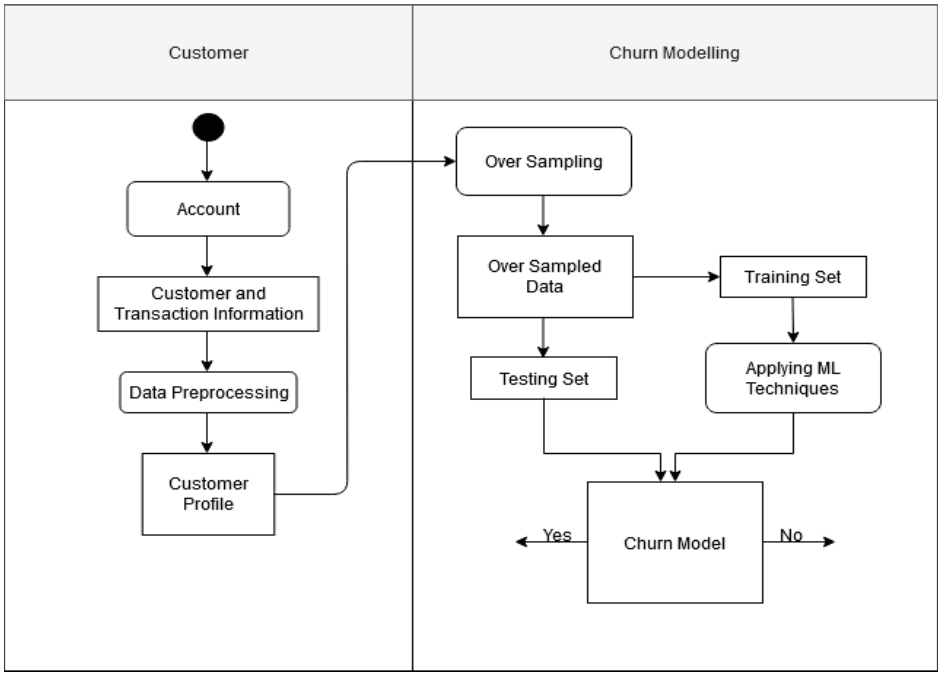

Fig. 1. Activity diagram of the proposed system

TABLE II
DATASET DESCRIPTION

Feature Name	Feature Description
Row number	Row numbers from 1 to 10000.
Customer Id	Unique Ids for bank customer identification.
Surname	Customer's last name.
Credit Score	Credit score of the customer.
Geography	The country from which the customer belongs.
Gender	Male or Female.
Age	Age of the customer.
Tenure	Number of years for which the customer has been with the bank.
Balance	Bank balance of the customer.
Num of Products	Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.).
Has Cr Card	Binary flag for whether the customer holds a credit card with the bank or not.
Is Active Member	Binary flag for whether the customer is an active member with the bank or not.
Estimated Salary	Estimated salary of the customer in Dollars.
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained.

1) Irrelevancy: Data or features which have no impact on the topic of discussion shall be considered as irrelevant. Keeping such attributes may sometimes affect the performance of classifiers. While considering the churn dataset, the features named Row number, Customer Id, Surname, and Geography has nothing to do with the prediction. So these features have been neglected manually in this study.

2) Transformation: Data transformation is the practice of turning data from one form into another. Properly structured and validated data enhance the quality of data and protect applications against possible minefields such as null values, unwanted duplicates, incorrect indexing, and incompatible formats. In this work the following data transformation is carried out;

- Gender: Female -> 0 and Male -> 1

C. Feature Selection

In machine learning, the process of identifying a subset of appropriate predictors to be used in model building is called feature selection. The selection phase of the feature is very critical as it will help to shorten the training time, escape the high-dimensionality curse, and above all simplify the model.

1) mRMR: Minimum Redundancy Maximum Relevance (mRMR) is one of the feature selection methods of filter type. For classification problems, it will rate features sequentially using the mRMR algorithm [10]. The filter type feature selection algorithm evaluates feature significance based on feature characteristics, such as the variance of feature and

TABLE III
PREPROCESSED DATASET DESCRIPTION

Feature Name	Min	Max	Mean	SD
Credit Score	350	850	650.5288	96.6533
Gender	0	1	0.5457	0.4979
Age	18	92	38.9218	10.4878
Tenure	0	10	5.0128	2.8922
Balance	0	250898.09	76485.8893	62397.4052
Num Of Products	1	4	1.5302	0.5817
Has Cr Card	0	1	0.7055	0.4558
Is Active Member	0	1	0.5151	0.4998
Estimated Salary	11.58	199992.48	100090.2399	57510.4928

reaction relevance. The selection of features will be part of the preprocessing phase of the data. Hence, the filter type feature selection is uncorrelated with the training algorithm.

2) Relief: It is also one of the filter type feature selection that will rank features using the Relief algorithm [11]. This algorithm works best to estimate the significance of features for distance-based supervised models that use pair distances between observations to predict the response. What it is doing is rank the predictors based on importance using the specified number of nearest neighbors. And the result will be the predictor numbers listed according to their ranking.

D. Over Sampling

In data processing, oversampling and undersampling are strategies used to configure the class distribution of given data. Since the data is highly imbalanced(7963 positive class samples and 2037 negative class) and the size of the available data sample is small, this study will make use of the oversampling technique. Because if undersampling is preferred, the size of data will decrease in a way that enough data will not be there to build the model. Hence, this study is using the random oversampling by resampling the minority class(negative class).

E. Classification

The classification methods were applied over the preprocessed data. KNN, SVM, Decision Tree(DT) and RF classifiers are used for comparison of results. And also the comparison of results of different classifiers has been carried out over the selected features by different feature selection methods.

1) k-Nearest Neighbor (KNN): The KNN method is one of the easiest and most efficient non-parametric ways of classification, based on supervised learning [12]. KNN works by identifying the k nearest samples from an existing dataset and when a new unknown sample appears, classify the new sample in the most similar class. That is, the classification algorithm determines the test sample group by the k training samples that are the nearest neighbors to the test sample and assign it to the class with the highest likelihood.

2) Support Vector Machine(SVM): Support Vector Machine is an efficient, supervised machine learning algorithm derived from Vapnik's theory of statistical learning [13], [14], [15]. This has proved its success in the fields of classification [16], regression [17], time series prediction, and estimation in geotechnical practice and mining science [18]. SVM's main objective is to find an efficient distinguishing hyperplane that precisely categorizes data points and as far as possible, and distinguishes the points of two classes by reducing the possibility of misclassifying the training samples and unknown test samples. This implies that there is the maximum distance between two classes and the separating hyperplane. The Linear support vector machine(LSVM) model is used in this work. LSVM was originally developed to deal with binary class problems [19].

3) Decision Tree(DT): A decision tree is a procedure that slices a collection of data into various branch-like segments [20]. A tree of decisions is easy to read. This advantage makes explanations for the model simple. While another algorithm (like a neural network) can generate a much more accurate model in a given scenario, a decision tree could be trained to predict the neural network's predictions, thus opening up the neural network's "black box". Another benefit is that, in the correlation between the target variables and the predictor variables it can model a high degree of nonlinearity. A decision tree is composed of two major strategies [21]; Tree creation and Classification.

4) Random Forest(RF): Breiman [22] presented RF as an ensemble classifier for tree learners. The method employs several decision trees so that each tree relies on the values of an individually selected random vector with the same distribution for all trees. Right choice for the tendency of decision trees to overfit their training collection. In short, Random forests are actually a way to combine many deep decision trees which are learned on various sections of the same dataset with the target of decreasing the variance. The real advantage of using RF is it comes with quite high dimensional data, with no need to perform dimensionality reduction and feature selection. The training rate is also higher and ease to use in parallel models.

IV. RESULTS AND DISCUSSIONS

When the preprocessing of the data has been completed, the data will be in the operational form. And the 10 features which are obtained after preprocessing is taken for the remaining study. Among that, 70% of data will be used for training and the remaining 30% will use for testing as random. The classifiers will be used alone and along with the specified feature selection methods. And each model is evaluated by the accuracy which is obtained after a 10 fold cross-validation. And the random confusion matrix was also produced for each model. The performance of classifiers varies when using different feature selection methods. The features selected in each feature selection method and the classifiers parameter details will describe in the following paragraphs.

For KNN, the k-value is set to 5. That is the nearest five neighbors are considered for classifying the new data. By reducing the neighbors than 5, sometimes the accuracy is increasing and wise versa. But, since the data is taking as randomly for the classification it is not a good practice to select fewer neighbors. But when the number of neighbors is greater than 5, the result is highly decreasing. Hence, the value of k is selected as 5(in which the accuracy and the change

is optimized). And the distance measure used is Euclidean distance. For SVM, the linear kernel function is used(LSVM). In the case of RF, the number of trees in the forest is set as 100. All these parameters are selected based on the optimization of classification accuracy.

The results of various classification techniques with and without oversampling(without feature selection) are given in table IV. It shows that the DT and RF classifiers accuracy increased after oversampling, but there is no change in KNN accuracy with regard to oversampling and SVM accuracy reduced with oversampling, this indicates that SVM is not suitable for huge amounts of data.

The best 6 features selected by MRMR method are Number of Products, Is Active Member, Gender, Age, Balance, Tenure. The accuracies of the various classifiers using MRMR selection method are shown in the table V, the accuracy of KNN is increased compared to KNN without MRMR. The SVM accuracy is almost similar to SVM without MRMR, the DT and RF accuracies are decreased a little bit compared to previous models.

The best features selected by Relief method are Number of Products, Age, Balance, Tenure, Gender, Has CrCard. The accuracies of the various classifiers are shown in table VI, the KNN accuracy is increased compared to KNN without feature selection and SVM remains the same, but DT and RF accuracies decreased little bit compared to previous models.

While resampling the negative class samples using oversampling(making negative class samples count the same as positive class), the data imbalance problem will be solved. Another finding is that resampling is decreasing the SVM score. By resampling, the actual data size is increasing. Hence, the SVM is unable to perform the required classification. The KNN seems to maintain almost the same accuracy after resampling. But tree classifiers DT and RF, are increasing the accuracy, it is because the tree classifiers will improve the accuracy when the amount of data is higher and balanced.

While applying feature selection methods, the score of KNN in increasing by a little. SVM accuracy remains almost the same after feature selection also. But in DT and RF, the scores are decreasing a little. It is because the tree classifiers are handling each feature more reliably. Hence, the reduction in features will effect this reliability and it will reduce the accuracy. In short, the RF after oversampling is giving higher accuracy than KNN, SVM, and DT in this study. And the feature selection does not affect tree classifiers. Even if the results are not improved after feature selection, feature ranking is done. Among the features under consideration, the "NumOf Products" is the feature that has higher significance in this study. And as a conclusion, the people with more number of bank products like mobile banking, internet banking, savings account, fixed deposits, etc., are less likely to be churned. Hence the bank needs to focus on the people who are using fewer products.

TABLE IV
RESULTS BY APPLYING CLASSIFIERS DIRECTLY

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	81.65	81.37
SVM	79.63	70.36
DT	78.99	91.98
RF	85.18	95.74

TABLE V
RESULTS AFTER MRMR FEATURE SELECTION

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	83.97	82.57
SVM	79.63	69.96
DT	78.32	91.73
RF	83.66	92.95

TABLE VI
RESULTS AFTER RELIEFF FEATURE SELECTION

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	82.15	80.99
SVM	79.63	69.53
DT	77.61	90.74
RF	81.75	92.19

V. CONCLUSION

While the banking sector is considered, like any other organization, customer engagement has become one of the primary concerns. To resolve this crisis, banks need to identify customer churn possibilities as quickly as possible. There are various studies ongoing in banking churn prediction. Different entities measure the churn rate of customers in various ways using different bits of data or information. The need for a system that can forecast the client churning in banking in a generalized way in the early stages is really important. The system needs to works with fixed and potential data sources that are independent of any service provider. And also the model must be in a form in which; can use minimal information and can give maximum throughput for the prediction. This study focus to fulfil these needs.

The purpose of this study is to build the most appropriate model to predict client churn in a Bank in the early stages. The study only used a small amount of data (10000 samples), and also highly imbalanced. But real commercial bank data would be much larger. By oversampling, both of these headaches up to a certain degree can be resolved. The model examined KNN, SVM, Decision Tree, RF classifiers under different conditions for this study. A better result is achieved when using the RF classifier together with oversampling(95.74%). Feature selection methods have nothing to do with tree classifiers(Decision Tree and RF). As the result indicates, feature reduction(feature selection) is decreasing the prediction score of tree classifiers. Another observation is that unlike other classifiers, in SVM, oversampling is decreasing the score. It's because the Bank dataset is lopsided. Hence, SVM unable to handle the data well enough.

REFERENCES

[1] D.-R. Liu and Y.-Y. Shih, "Integrating ahp and data mining for product recommendation based on customer lifetime value," *Information & Management*, vol. 42, no. 3, pp. 387–400, 2005.
[2] G. Canning Jr, "Do a value analysis of your customer base," *Industrial Marketing Management*, vol. 11, no. 2, pp. 89–93, 1982.
[3] R. W. Stone and D. J. Good, "The assimilation of computer-aided marketing activities," *Information & management*, vol. 38, no. 7, pp. 437–447, 2001.
[4] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.
[5] M.-K. Kim, M.-C. Park, and D.-H. Jeong, "The effects of customer satisfaction and switching barrier on customer loyalty in korean mobile telecommunication services," *Telecommunications policy*, vol. 28, no. 2, pp. 145–159, 2004.
[6] I. T. J. Swamidason, "Survey of data mining algorithm's for intelligent computing system," *Journal of Trends in Computer Science and Smart Technology*, vol. 01, pp. 14–23, 09 2019.
[7] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on svm model," *Procedia Computer Science*, vol. 31, pp. 423–430, 2014.
[8] A. Bilal Zorić, "Predicting customer churn in banking industry using neural networks," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 14, no. 2, pp. 116–124, 2016.
[9] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, p. 28, 2019.
[10] Y. Jiang and C. Li, "mrmr-based feature selection for classification of cotton foreign matter using hyperspectral imaging," *Computers and Electronics in Agriculture*, vol. 119, pp. 191–200, 2015.
[11] L. Beretta and A. Santaniello, "Implementing relieff filters to extract meaningful features from genetic lifetime datasets," *Journal of biomedical informatics*, vol. 44, no. 2, pp. 361–369, 2011.
[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[14] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
[15] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.
[16] J. Raj and V. Ananthi, "Recurrent neural networks and nonlinear prediction in support vector machines," *Journal of Soft Computing Paradigm*, vol. 2019, pp. 33–40, 09 2019.
[17] P. G. Nieto, E. F. Combarro, J. del Coz Díaz, and E. Montañés, "A svm-based regression model to study the air quality at local scale in oviedo urban area (northern spain): A case study," *Applied Mathematics and Computation*, vol. 219, no. 17, pp. 8923–8937, 2013.
[18] S.-G. Cao, Y.-B. Liu, and Y.-P. Wang, "A forecasting and forewarning model for methane hazard in working face of coal mine based on ls-svm," *Journal of China University of Mining and Technology*, vol. 18, no. 2, pp. 172–176, 2008.
[19] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.
[20] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
[21] N. B. Amor, S. Benferhat, and Z. Elouedi, "Qualitative classification with possibilistic decision trees," in *Modern Information Processing*. Elsevier, 2006, pp. 159–169.
[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.