AN ENHANCED BANK CUSTOMERS CHURN PREDICTION MODEL USING A HYBRID GENETIC ALGORITHM AND K-MEANS FILTER AND ARTIFICIAL NEURAL NETWORK

1st **R. Yahaya**
Department of Computer Science
Federal University of Technology
Minna, Nigeria
rahmayahya5@gmail.com

2nd **O. A. Abisoye**
Department of Computer Science
Federal University of Technology
Minna, Nigeria
o.abisoye@futminna.edu.ng

3rd **S. A. Bashir**
Department of Computer Science
Federal University of Technology
Minna, Nigeria
bashirsulaimon@futminna.edu.ng

Abstract—Customer churn prediction is an important issue in banking industry and has gained attention over the years. Early identification of customers likely to leave a bank is vital in order to retain such customers. Predicting churning is a data mining tasks that require several data mining approaches. Churn prediction based on Artificial Neural Networks (ANNs) have been successful, however, they are affected by the noise or outliers present in such datasets. The effect of such noise, and number of training samples on churn prediction was investigated. Two filters were applied to the data, the Genetic Algorithm (GA) and Kmeans filter. The filtered data were used to train an ANN model and tested with a 30% unfiltered data. The performance show that the training performance improved when noise was filtered while the testing performance was affected by the unbalanced data caused by filtering.

Keywords—Customer Churn, Data Mining, Artificial Neural Network, K-means, Genetic Algorithm

I. INTRODUCTION

A customer is identified as a churner via his/her transaction history analysis. any Banking system customers are likely to churn due to poor customer services, unwarranted bank charges and other scenarios. Customer retention is often challenging for organizations as the cost of acquiring a new customer or subscriber is higher than retaining and old one. However, if an organization can easily predict customers that are likely to leave or unsubscribe form their service ahead, customer retention strategies will be directed entirely towards such customers. The bank plays a vital role in influencing a customer's satisfaction which leads to loyalty and continuous patronage, even referral. [1] suggested that monitoring customer behaviours can help organizations predict churners and lead to customer retention strategy creation. Churn prediction allows organizations to improve the efficiency of customer retention campaigns and mitigate the costs of churn. Several approaches have been used for Churn Prediction, including classification, clustering, association and rule-based approaches. Most researchers used two or more algorithms in churn detection, one improving on the other ([2]; [3]). [4] proposed a customer churn prediction model based on XGBoost and Multi-layer Perceptron which resulted in predicting churn better than other state-of-art prediction models. [5] proposed a hybrid model of classification and clustering techniques, experiments were done for each techniques and results were produced and compared, the hybrid model produced more accurate results in comparison to single model. Noise affects data analysis leading to wrong or incorrect results [6]. To deal with noise in data analysis, and achieve a good model, data has to be filtered. [7] proposed a noise filtering approach that combined Tomek-link with distance weighted KNN (TWK) while kmeans clustering have been applied for data filtering due to its successful clustering ability [8]. However,

kmeans clustering algorithms performance is affected by the initial cluster centers which are randomly selected.

In this paper, an enhanced bank customer churn prediction model using a Hybrid Genetic algorithm with K-means (GA-Kmeans) clustering and ANN is proposed. The GA searches for optimal cluster centers of the kmeans while the ANN is used for predicting churning. The effect of data filtering on the performance of ANN was also investigated in this paper. The remainder of the paper is structured as follows: Section 2 presents the review of related works, section 3 presents the materials and methods, while in section 4, the results and discussions were presented. Finally, in section 5, the conclusion and recommendations were presented.

II. LITERATURE REVEIW

There are several churn prediction models in banking industry and other financial institutions. They mostly applied data mining and machine learning approaches to solve the problems. [9] identified churn customers using some criteria before they unsubscribe from a service or leave the business. Customer churn prediction model was developed by analyzing historical behaviour data of defected customers, for early detection and retention purposes. They used random forests to build customer churn prediction models [10]. In a competitive world as ours, existing customer base and their data are priceless assets organizations boast of. According to [11], the cost of acquiring a new customer is usually high than retaining customer likely to churn, hence correctly identifying a churn customer through metrics such as recall, accuracy, precision and F1-score in customer churn detection will save the company from loss and enhance their customer retention strategies. Churn prediction challenges include; Capturing pattern of customer behaviour, especially in financial institutions. Past researches on churn prediction laid emphasis on predicting churn based on monthly, static or dynamic behaviour of customers. [12] claims they are unrealistic, as predicting churn based on monthly traits, might divert focus from churners who decide to unsubscribe at the beginning of the month, and using monthly traits to predict a customer's likelihood of churning might not take daily traits in cognizance, hence reducing the discriminative ability and performance of the model. K-means have been applied for churn prediction in combination with other models such as, C5.0 with and without misclassification cost in addition to logistic regression and ANN. Overall, C5.0 with misclassification cost surpassed all other models in terms of accuracy.

[13]) predicted the customer churn problem on a Nigerian bank datasets using WEKA tool for knowledge analysis. K-means clustering algorithm was used to cluster the data while, JRip algorithm was implemented in rule generation phase. Customer Relationship Management (CRM) helps employees or organization put their customer into consideration by offering them excellent services and satisfaction, which might in turn reduce churning. [14] developed a CRM model comprising of Genetic-based Data Mining (GDM) approach to counter some of the challenges in CRM. They used genetic algorithm and data mining in achieving their aim, by optimizing rules generated from C5.0 algorithm with a genetic algorithm to increase CRM classification time and accuracy, Genetic Algorithm reduced and improved upon C5.0 algorithm. The GDM model was able to find hidden data from a large chunk of data, equipping the researcher with information to serve customers better. [15] developed a robust classification model using Artificial neural network and further applied hyperparameter search space using Genetic Algorithm to detect suitable parameter settings. Results showed that applying hyperparameter optimization on the ANN classification models led to an improved rate of customer churn prediction. [16] designed a Multilayer Perceptron (MLP) model for churn prediction and results are further compared with Support Vector Machine, Naïve Bayes and Decision Tree. MLP-ANN outperformed other classifiers for both PCA and Normalize pre-processing techniques, finally used InfoGainAttribute to identify the highest factor attribute leading to customer retention. [17] applied feature selection aiding him remove irrelevant features which aided in improving the performance of the model and reduced training time and overfitting for model construction. [18] used neural network model within the software package (Alyuda) Neuro Intelligence for customer churn prediction in Banking industry he claimed that neural network was the best fit for pattern recognition, image processing, optimization

problems. From the review works, it is clear that ANN is one of the competitive models for churn prediction. Furthermore, the research works do not consider filtering the data before using them.

III. MATERIALS AND METHODS

The block diagram shown in Figure 1 shows the steps that were taking to achieve the aim of this research paper. It comprises of description of the data collection, description and preprocessing. It also involves the data filtering, model design, training, testing and performance evaluation.

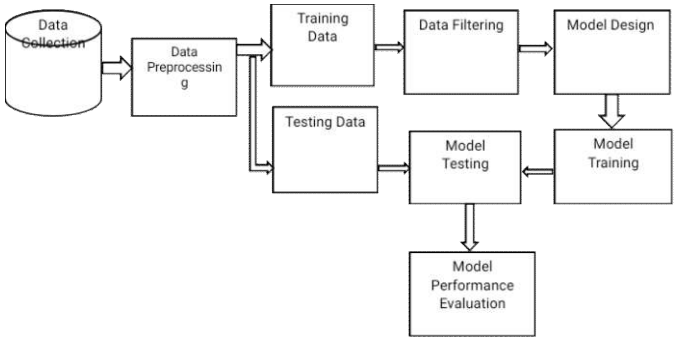


Figure 1: Proposed Methodology

A. DATA COLLECTION AND DESCRIPTION

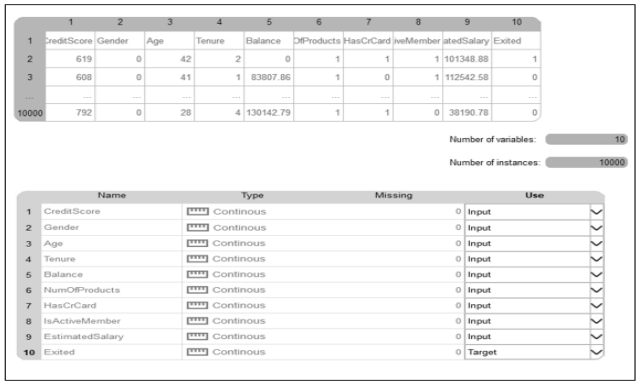
The Dataset for this research was obtained from Kaggle database. The dataset is a bank dataset used for churn prediction challenge. It comprises of 10,000 bank records with ten (10) attributes of each customer. Table.1 shows a sample of the dataset; 80% non-churn and 20% churn samples. Figure 2 shows the sample dataset with its 10 attributes.

B. DATA PRE-PROCESSING

The dataset will be pre-processed to put the data in appropriate state using the following steps:

i. Filling of any missing value(s) using average filling techniques.
ii. Manual attribute selection.
iii. Data Partitioning: Partitioning the data into training and testing ratio (70:30) respectively.

In this paper, no missing values was found and four attributes were manually removed. The removed attributes are the Row number, CustomerId, Surname and Geography. The dataset was partitioned into training and testing in the ratio 70:30 respectively. Figure 2 shows a sample of the dataset with their attributes.



	1	2	3	4	5	6	7	8	9	10
1	CreditScore	Gender	Age	Tenure	Balance	OfProducts	HasCrCard	iveMember	atedSalary	Exited
2	619	0	42	2	0	1	1	1	101348.88	1
3	608	0	41	1	83807.86	1	0	1	112542.58	0
...	...	...	...	...	...	...	...	...	...	...
10000	792	0	28	4	130142.79	1	1	0	38190.78	0

Number of variables: 10
Number of instances: 10000

	Name	Type	Missing	Use
1	CreditScore	Continous	0	Input
2	Gender	Continous	0	Input
3	Age	Continous	0	Input
4	Tenure	Continous	0	Input
5	Balance	Continous	0	Input
6	NumOfProducts	Continous	0	Input
7	HasCrCard	Continous	0	Input
8	IsActiveMember	Continous	0	Input
9	EstimatedSalary	Continous	0	Input
10	Exited	Continous	0	Target

Figure 2: Dataset sample

C. DATA FILTERING

The training dataset was filtered to reduce noise or outliers that may affect the model's performance. An optimized K-means clustering filtering algorithm using Genetic Algorithm (GA) as an optimizer was proposed. It has been established that K-means clustering is sensitive to initial cluster centers which are generated randomly. In this paper, GA was used to search for optimal initial cluster centers. The process of data filtering using GA-kmeans is shown in the flowchart in Figure.3.

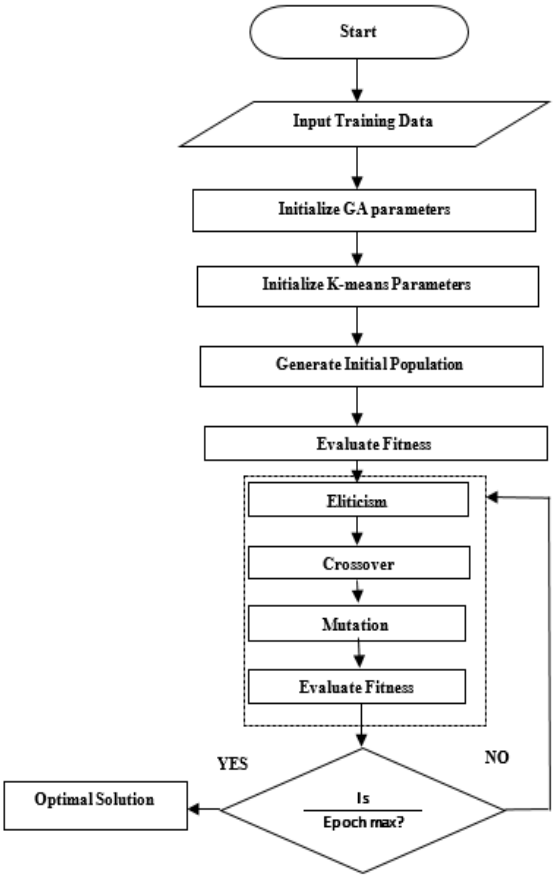


Figure 3: GA-Kmeans Flowchart

D. MODEL DESIGN:

An Artificial Neural Network (ANN) model was designed for bank customers churn prediction using MatLab software. The model is a feedforward backpropagation model with 9 input nodes corresponding to the number of inputs, a single hidden layer with 10 hidden nodes and an output layer with one node corresponding to the desired output. Figure 4 shows a designed ANN model. The first layer uses the tansig activation function while the second layer uses the logsig activation function. The scaled conjugate gradient (SCG) training algorithm was used.

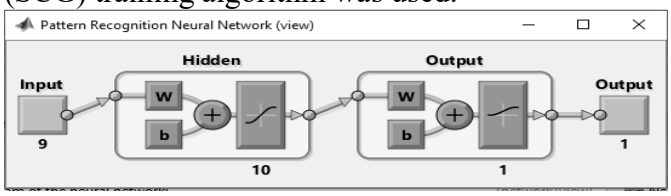


Figure 4: A designed ANN model

Mathematically, the designed ANN model is represented as follows:

Let $Y\{1,t\}$ be the output of the model and $X\{1,t\}$ be the input of the model where, t is the time step. The output of layer 1, layer 2 and the output of the model can be defined as follows:

$$L_1 = \Phi_1(\alpha(b_1, 1, Q) + W_1 * \beta) \quad (1)$$

$$L_2 = \Phi_2(\alpha(b_2, 1, Q) + W_2 * L_1 \quad (2)$$

$$Y\{1,t\} = L_2 \quad (3)$$

Where, Φ_1 is the tansig activation function and Φ_2 is the logsig activation function, α is the repmat function in Matlab and β is the minmax mapping function between $X\{1,t\}$ and -1.

E. MODEL TRAINING AND TESTING

Three models were trained using the designed ANN model. The first model was trained with unfiltered data which has 7000 samples, the second model was trained using 3703 samples from kmeans filter while the third model was trained using 5281 samples from GA-kmeans filter. The 30% reserved data (3000 samples) were used to evaluate the performance of the model and determine the effect of the filtering techniques on the churn prediction.

F. PERFORMANCE EVALUATION

The performance of the developed models were evaluated using the following evaluation metrics:

1) *Accuracy*

The accuracy of a model is the sum of the correctly classified positive instances and the correctly classified negative instances relative to the total number of correctly and incorrectly classified instances and is given as;

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (4)$$

2) *Sensitivity*

The sensitivity of a model measures the percentage of correctly classified positive instances and is given as;

$$TPR = \frac{TP}{TP+FN} \quad (5)$$

3) ***Specificity***

The specificity of a model measures the percentage of correctly classified negative instances and is given as;

$$SPC = \frac{TN}{FP+TN} \times 100\% \quad (6)$$

4) ***Precision***

Precision is the fraction of instances that were correctly classified and is given as;

$$PPV = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

5) ***Mean Squared Error (MSE)***

The MSE measures the mean of the squared difference between the model output and target output. It is given as;

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (T_i - P_i)^2 \quad (8)$$

Where, TP (True positive) is correctly classified positive instances, TN (True negative) is correctly classified negative instances, FP (False positive) is incorrectly classified negative instances and FN (False negative) is incorrectly classified positive instances. N is the number of instances, T_i and P_i are the target and predicted values of the i[th] sample respectively. V

IV. RESULTS AND DISCUSSION

Three categories of results were presented. The Filtering results, model training results and model testing results. The filtering results presented comprises of the kmeans clustering result and optimized kmeans clustering using GA.

A. Filtering Results

Figure 6 shows the confusion matrices of kmeans clustering and GA-kmeans clustering respectively. Kmeans clustering performance shown in Figure 6(a) shows that the TP, TN, FP, and FN are 3299, 404, 2258 and 1039 respectively. Similarly, for GA-kmeans clustering (Figure 6(b)), the TP, TN, FP, and FN are 5131, 150, 426 and 1293 respectively.

a) Kmeans clustering b) GA-Kmeans clustering

Figure 6: Confusion matrices of kmeans and GA-kmeans

The result show that 47.1% of the training data were filtered as noise or outiers when kmeans clustering was used while, when GA-kmeans was used, the only 24.6% of the training data was filtered out. This indicated that the optimized kmeans clustering improved the detection of

outliers due to the optimal initialization of initial cluster centers. Figure 7 shows the convergence curve of the GA used fpr optimizing the kmeans. The curve shows the minimum fitness value obtained and the iteration where convergence too place. The optimum fitness value obtained is 0.2456 at the 11th iteration.

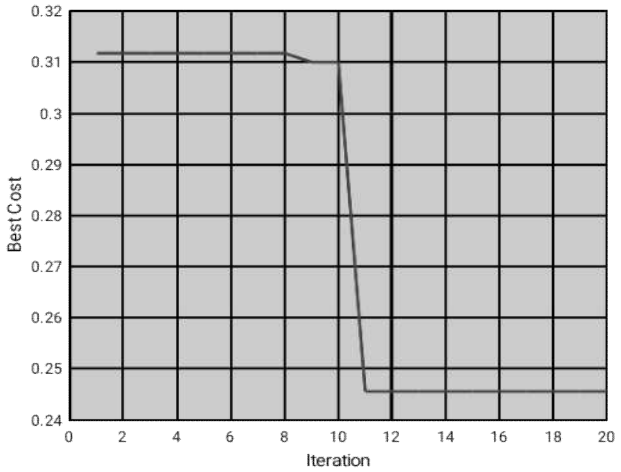


Figure 7: Convergence Curve

Table 1 is the summary of results obtained by the two clustering techniques evaluated in this work. The result shows the MSE, accuracy, sensitivity, specificity and number of samples. The optimized GA-kmeans filtering achieved an accuracy of 75.4% with an MSE of 0.2456 while kmeans filtering achieved an accuracy of 52.9% with an MSE of 0.4710. The training data after filtering using GA-kmeans is 5281 representing 5131 normal customers and 151 churn customers. For kmeans, 3703 samples were obtained after training representing 3299 normal and 404 churn customers.

Table 1: Filtering Results Summary

Clustering Algorithm	MSE	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Samples
Kmeans	0.2456	75.4	92.30	10.4	5281
Kmeans	0.4710	52.9	59.40	28.00	3703

B. Training Results

Three models were trained and compared with a view of identifying the best model. The models are ANN model, kmeans-ANN model and GA-kmeans-ANN model. The training results obtained are shown in the confusion matrices in Figure 8 and a summary of the performances of the models are calculated and shown in Table 2. The results show that training the churn prediction models without data filtering (ANN model) obtains a precision and accuracy of 86.7% and 85.8% respectively. Also, kmeans-model obtained a precision and accuracy of 99.99% and 99.90% while the GA-kmeans-ANN model obtained a 100% precision and accuracy. The result indicated that there is a significant increase (14.2%) in training accuracy as the noise of the training data reduced. Furthermore, the improvement in training accuracy for GA-kmeans-ANN against kmeans-ANN indicated that more noiseless training data improves the performance of the models.

a) ANN b) Kmeans-ANN c) GA-Kmeans-ANN

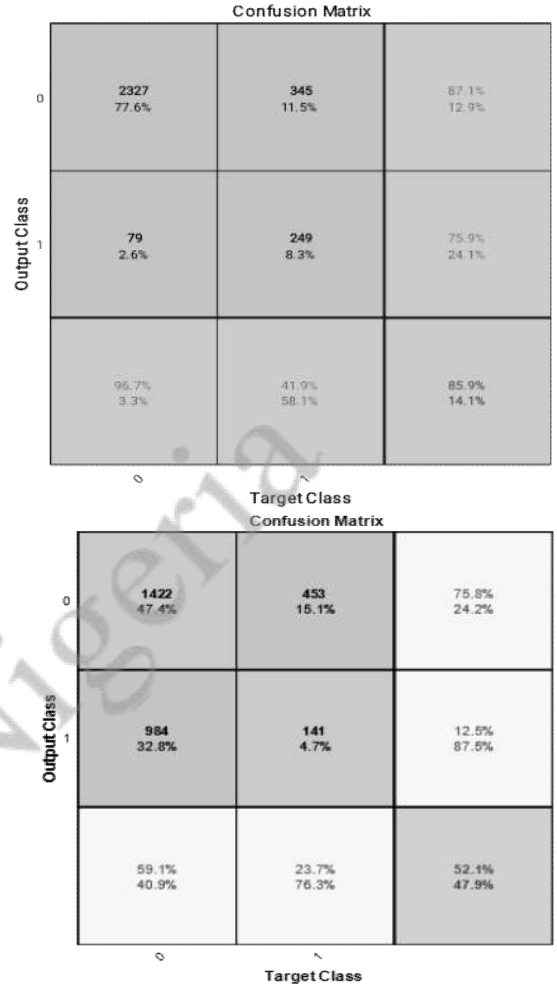
Figure 8: Confusion matrices for ANN, kmeans-ANN and GA-kmeans-ANN

Table 2: Performances of Trained Models

C. Testing Results

Figure 9 show the confusion matrices of the testing carried out on the training models for ANN and Kmeans-ANN respectively and Figure 10 shows the confusion matrix for GA-kmeans-ANN model. Table 3 shows the results calculated from the confusion matrices for the tests. The test

results show that the model trained with ANN performs better in terms of precision, accuracy, sensitivity and specificity followed by the GA-kmeans-ANN model and lastly the kmeans-ANN model. The accuracy of the ANN model is 85.9%, 76.6% for GA-kmeans-ANN and 52.10% for kmeans-ANN model.



a) ANN b) Kmeans-ANN

Figure 9: Confusion matrices for ANN and kmeans-ANN

Models	Precision (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
ANN	86.70	85.80	96.90	42.70
Kmeans-ANN	99.99	99.90	100.00	99.50
GA-Kmeans-ANN	100.00	100.00	100.00	99.30

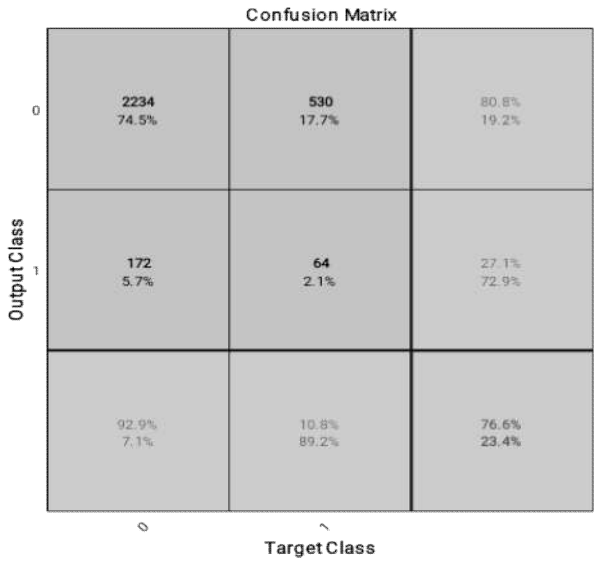


Figure 10: Confusion matrices for GA-kmeans-ANN

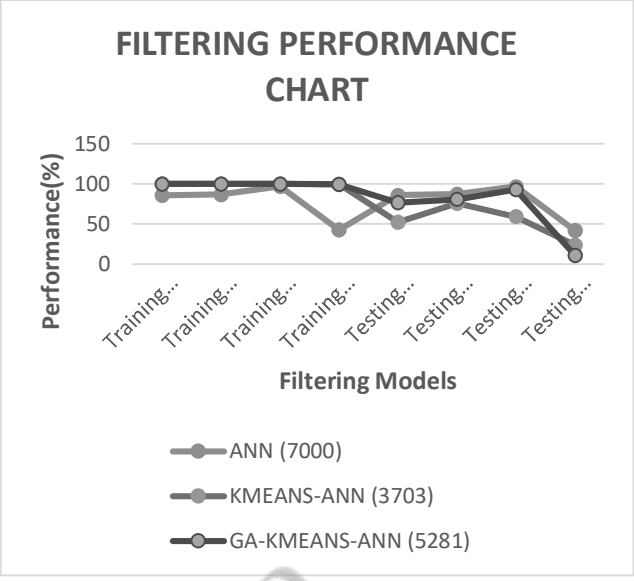


Table 3: Model Testing Results

Models	Precision (%)	Accuracy (%)	Sensitivity	Specificity
ANN	87.10	85.90	96.70	41.90
Kmeans-ANN	75.80	52.10	59.10	23.70
GA-Kmeans-ANN	80.8	76.60	92.90	10.80

D. Discussion of Results

Table 4 shows the combined training and testing results obtained during the experiments. It includes the Number of Training Samples (NTS), accuracy, precision, sensitivity and specificity for ANN, kmeans-ANN and GA-kmeans-ANN respectively. From the results as shown in the model comparison chart in Figure 11, the highest training sample results in the lowest training performance as shown in the blue curve. This is attributed to the noise or outliers present in the training data. In contrast, as the training sample increases, the test performance reduces as shown in the green, red and blue curves respectively. This is attributed to the balance in training and testing samples. The lesser the difference, the better the testing result.

Table 4: Performance Evaluation Results

MODELS	NTS	Accuracy		Precision		Sensitivity		Specificity	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
ANN	7000	85.8	85.9	86.7	87.1	96.9	96.7	42.7	41.9
KMEANS-ANN	3703	99.9	52.1	99.99	75.8	100	59.1	99.5	23.7
GA-KMEANS-ANN	5281	100	76.6	100	80.8	100	92.9	99.3	10.8

V. CONCLUSION

In this paper, an enhanced bank customer churn prediction model is proposed using optimized (GA-Kmeans) filtering and Artificial Neural Network (ANN). The effects of data filtering on the performance of ANN models for bank customers churn prediction. The dataset was first preprocessed by manually removing attributes that are not useful and partitioning of the data into training and testing in 70:30 respectively. The training data were filtered using kmeans clustering technique and

optimized GA-kmeans clustering technique. The effect of the clustering techniques were evaluated and compared with un-filtered data. The results show that the training performance improved as the noise in the data reduces while the testing results were not improved with filtering. This is because the imbalance between the positive and negative classes affected the testing results. To improve the result of the developed model, the classes of the filtered data will be balanced using an appropriate data balancing technique and the performance will be compared with unbalance dataset.

REFERENCES

[1] Briker, Vitaly; Farrow, Richard; Trevino, William; and Allen, Brent (2019) "Identifying Customer Churn in After-market Operations using Machine Learning Algorithms," SMU Data Science Review: Vol. 2: No. 3, Article 6.

[2] Gde, K., Karvana, M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. *2019 International Workshop on Big Data and Information Security (IWBIS)*, 33–38

[3] Jeyakarthic, M., & Venkatesh, S. (2020). *An Effective Customer Churn Prediction Model Using Adaptive Gain with Back Propagation Neural Network in Cloud Computing Environment an Effective Customer Churn Prediction Model Using Adaptive Gain with Back Propagation Neural Network In Cloud Computing E*.

[4] Tang, Q., Xia, G., & Zhang, X. (2020). A Customer Churn Prediction Model Based on Xgboost and MLP. *Ieeexplore.Ieee.Org*. Https://Ieeexplore.Ieee.Org/Abstract/Document/9103818/

[5] Vijaya, E. S. J. (2019). Hybrid PPFCM-ANN Model: An Efficient System for Customer Churn Prediction Through Probabilistic Possibilistic Fuzzy Clustering and Artificial Neural Network. *Neural Computing and Applications*, *31*(11), 7181–7200.

[6] Hemalatha, M., & Mahalakshmi, S. (2020). *Predicting Churn Customer In Telecom Using Peergrading Regression Learning Technique*. *6*, 1025–1037.

[7] Li, Y., Wei, J., Kang, K., & Wu, Z. (2019). *An Efficient Noise-Filtered Ensemble Model for Customer Churn Analysis In Aviation Industry*. *37*, 2575–2585.

[8] Mamman J., Aibinu A. M, Abdullahi B. U, Abdullahi I. M, (2015) "Diabetic classification using cascaded data mining technique", International Journal of Computer Trends and Technology, vol. 22, number 2, April 2015.

[9] Arivazhagan, B., & Sankara, S. D. R. S. (2020). Customer Churn Prediction Model Using Regression with Bayesian Boosting Technique in Data Mining. *Ijaema.Com*, *XII*(V), 1096–1103.

[10] Li, W., & Zhou, C. (2020). Customer Churn Prediction in Telecom Using Big Data Analytics. *IOP Conference Series: Materials Science and Engineering*, *768*(5).

[11] Amornvetchayakul, P., & Phumchusri, N. (2020). Customer Churn Prediction for A Software-As-A-Service Inventory Management Software Company: A Case Study in Thailand. *2020 IEEE 7th International Conference on Industrial Engineering and Applications, ICIEA 2020*, 514–518.

[12] Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic Behavior Based Churn Prediction in Mobile Telecom. *Expert Systems with Applications*, 113779.

[13] Kolajo, T., & Adeyemo, A. B. (2015). *Computing, Information Systems & Development Informatics Journal*. *April 2012*.

[14] Makinde, A. S., Oguntuase, A., Vincent, O. R., Acheme, I. D., & Akinwale, A. T. (2020). An Improved Customer Relationship Management Model for Business-To-Business E-Commerce Using Genetic-Based Data Mining Process. *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*.

[15] Fridrich, M. (2017). *Hyperparameter Optimization of Artificial Neural Network in Customer Churn Prediction Using Genetic Algorithm*. *8527*(1), 9–21.

[16] NNA Sjarif, NF Azmi, HM Sarkan, SM Sam, and M. O. (2020). Predicting Churn: How Multilayer Perceptron Method Can Help with Customer Retention in Telecom Industry

Predicting Churn: How Multilayer Perceptron Method Can Help with Customer Retention in Telecom Industry. *IOP Conference Series: Materials Science and Engineering*, 0–5.

[17] Wadikar, D. (2020). Customer Churn Prediction. *Masters Dissertation. Technological University Dublin.*

[18] Zoric, A. B. (2016). *Predicting Customer Churn in Banking Industry Using Neural Networks*. *14*(2), 116–124.

[19] Abbasimehr, H., & Alizadeh, S. (2013). *A Novel Genetic Algorithm Based Method for Building Accurate and Comprehensible Churn Prediction Models*. *2*(4), 1–14.

[20] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. *Journal of Big Data*, *6*(1).

[21] Ali, I. (2019). Churn Prediction in Banking System Using K-. *2019 International Conference on Electrical, Communication, And Computer Engineering (ICECCE)*, *July*, 1–6.

[22] Amuda, K. A., & Adeyemo, A. B. (2019). *Customers Churn Prediction in Financial Institution Using Artificial Neural Network*. Http://Arxiv.Org/Abs/1912.11346

[23] Jha N., Parekh D., Mouhoub M., Makkar V. (2020) Customer Segmentation And Churn Prediction In Online Retail. In: Goutte C., Zhu X. (Eds) Advances In Artificial Intelligence. Canadian AI 2020. Lecture Notes in Computer Science, Vol 12109. Springer, Cham. Https://Doi.Org/10.1007/978-3-030-47358-7_33

[24] Liu, Y., & Zhuang, Y. (2015). *Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in The Context of Big Data*. *June*, 87–93.