A Comparism of Machine Learning Based Approaches in Predicting Agricultural Loan Defaulters Among Farmers in Lavun Local Government Area of Niger State

Zainab Olalere
Department of Computer Science
Federal University of Technology
Minna, Nigeria
olalerezainab5@gmail.com

Ishaq Oyebisi Oyefolahan
Department of Computer Science
Federal University of Technology
Minna, Nigeria
o.ishaq@futminna.edu.ng

Solomon Adelowo Adepoju
Department of Computer Science
Federal University of Technology
Minna, Nigeria
sa.adepoju@gmail.com

***Abstract*— Financial institutions in Nigeria have continuously extended generous loan offers to the manufacturing and industrial sector compared to the agricultural sector, due to the risk-benefit ratio difference attached to each. To aid the agricultural sector in Nigeria, the government established risk-sharing interventions in the agricultural sector with the aim of incentivizing the financial institutions towards issuance of credits to farmers. However, financial institutions still seek to reduce the leftover risk. This research was conducted in order to assist financial institutions reduce the risk of lending to farmers. A private agricultural loan dataset collected in Lavun Local Government Area in Niger state, Nigeria was used in this research to predict the likelihood of an agricultural loan default. Recursive feature elimination was used to reduce the features of the dataset from 60 to 44. Furthermore, machine learning algorithms of random forest, logistic regression, support vector machine, gradient boosting, and adaptive boosting were applied on the dataset. The results obtained shows that gradient boosting and random forest algorithms were the most effective in predicting agricultural loan defaults with precision and f1-score of 86.36% with 90.48% and 89.47% with 82.93% respectively. Improving the accuracy of the other machine learning models is proposed for further study.**

Keywords—agricultural loans, farm credit, risk-sharing, Lavun, Niger state

I. INTRODUCTION

Agriculture and food systems encompasses most of the activities people do for a living on Earth [1]. In Nigeria, the agricultural sector has employed over 70% of the total workforce, positioning the sector as an instrument for economic diversity and development [2], [3]. In the 1960s, Nigeria's export was predominantly from the agricultural sector with each region of the country playing a vital role: the north produced groundnuts, the south-east produced oil palm, and the south-west produced cocoa. However, the discovery of oil and the post-civil war rehabilitation and reconstruction contributed to the decline in agricultural exports from Nigeria [4]. Nevertheless, the rural communities within Nigeria embraced agriculture as a means of livelihood. Niger state, in the north-central region, is one of the states whose rural communities have continued with the mass production of agricultural produce, with yam and rice dominating other crop produce. For instance, rural households in Lapai, Kontagora, and Suleja regions of Niger state have recorded an average of 19 years of farming experience [5]. The need to harness the vast experience of these small-holder farmers towards improving the dwindling supply of agricultural produce due to the reduced focus given to agriculture and the burgeoning population necessitated the issuance of agricultural credits to the farmers. The agricultural credits are aimed at enabling the farmers purchase improved seeds, fertilizer, and hire mechanized farm tools.

Over the years, issuance of credit to various stakeholders in the sectors of the economy has been performed including the agricultural sector. Issuance of credit or loan involves giving an individual or a group a stipulated amount of money to enable the individual or group make purchase of goods or services with the aim of returning the borrowed money with the interest accrued to it. Due to the increased reliance on oil in Nigeria, most commercial banks give loans to industries in the oil and manufacturing sectors compared to the agricultural sector [4]. The varying credit allocation by banks is attributed to the risk of income and capital loss across the agricultural, manufacturing, and oil sectors. Since majority of the agrarian population are smallholder farmers who dwell in rural areas according to [2], obtaining credit from banks without sufficient collateral becomes an uphill task. Furthermore, the problems plaguing the agricultural sector in Nigeria – volatile commodity prices, disease outbreak, and climate change – makes the sector less attractive to formal financial institutions to offer credit facilities [4].

The problem of credit financing of smallholder farmers necessitated the creation of credit risk guarantee frameworks that employs a risk sharing model which encourages financial institutions to lend to farmers easily. The Nigeria Incentive-Based Risk Sharing System for Agricultural Lending (NIRSAL) employs this model. Aside rendering technical assistance to farmers, modelling businesses to leverage market dynamics that benefits the agrarian population, and providing innovative insurance of agricultural products; NIRSAL offers Credit Risk Guarantee (CRG). The CRG is a framework designed to shield financiers, and investors who issue agribusiness loans against losses in a credit transaction through a risk sharing arrangement. The NIRSAL CRG covers the risk of default on the loan principal and accrued interest to the limit of a predetermined CRG rate. Another risk-sharing programme for smallholder farmer loans is the Anchor Borrowers' Programme (ABP). The ABP seeks to boost production of agricultural commodities and stabilizing the supply of inputs to agro-processors by providing farm

inputs (cash and labour). The programme, an initiative of the Central Bank of Nigeria (CBN), employs a risk-sharing model to encourage participation of financial institutions by absorbing 50% of the amount in default once it has been established that all means of loan recovery has been exhausted. Hence, the participating financial institutions bear the credit risk of the remaining 50% of the loan amount [3]. However, since financial institutions will want to bear a minimal amount of risk, it is imperative that the financial institutions obtain a way to predict loan defaulters while considering the peculiarities of farmers.

The aim of this study is to predict agricultural loan defaulters among smallholder farmers in Lavun, Niger State. Lavun was chosen due to the abundance of rice-based cropping enterprise in the area given the high consumption of rice in Nigeria. Most research focused on loans for other purposes except agriculture while others evaluated the indicators of agricultural loan default. Hence, this research bridges the gap of inadequate work into predicting agricultural loan defaults while considering the unique nature of farmers. The objectives of this paper are to:

i. Select the features in the private Nigerian agricultural loan dataset that are relevant to the prediction task.
ii. Perform machine learning classification tasks on the selected features.
iii. Present the result obtained from the classification task.
iv. Evaluate the result presented.

II. RELATED WORKS

The need to harness opportunities using resources that are not available at the moment usually informs the decision to seek for loans, either from individuals or financial institutions. Based on the need and required amount, a choice is made between obtaining loans from individuals within ones social circle or from established financial institutions. On the one hand, receiving loans from individuals confers benefits such as low to no interest rate on the borrowed cash. On the other hand, however, the amount borrowed may be small relative to the needed amount. Therefore, individuals approach financial institutions for credit. Compared to the peer-to-peer lending option which is based on a social trust model, financial institutions endeavor to limit the risk of the loan through various techniques including request for a collateral with a value greater or equal to the loan amount. Among other techniques is the use of machine learning algorithms to predict loan defaulters from previous loan datasets so as to minimize risk. Researchers have examined the prediction of loan defaulters using various statistical and machine learning methods.

A credit score model for airtime loans using machine learning was postulated by [6] using dataset obtained from ComzAfrica. In the research, machine learning algorithms – logistic regression, decision tree, and random forest – were applied to the ComzAfrica dataset of 1 January 2016 to 30 June 2017. The in-sample analysis of the algorithms yielded a uniform accuracy of 99.1% and specificity of 0.2%, 0.0%, and 0.8% for logistic regression, decision tree, and random forest algorithms respectively. Although the accuracy of the algorithms was high, the low specificity obtained indicates that the classifier incorrectly predicts default when considering those that actually defaulted.

Reference [7] utilized kaggle credit dataset with 12 attributes to predict loan default using logistic regression, naïve bayes, decision trees, and random forest algorithms. The results showed that logistic regression had the highest accuracy of 93.777%. The accuracy of random forest, naïve bayes, and decision trees were 93.44%, 89.86%, and 89.51%. Apart from the accuracy of the classifiers, other metrics were not measured.

Dataset with 64,000 tuples and 14 attributes was used in forecasting loan default in the research by [8]. The algorithms: logistic regression, gradient boosting, catboost classifier, and random forest, achieved accuracy and precision of 14.96% and 49%, 84.04% and 85%, 84.05% and 85%, and 83.51% and 86% respectively. All the algorithms except logistics regression achieved f1-score of 91%.

Logistics regression was used to determine the likelihood of a loan default in Bangladesh [9]. The authors were able to establish that logistics regression classifies 93.30% of the cases. Similarly, [10] used logistic regression model to predict loan defaults. The authors were able to achieve accuracy, precision, recall, and f1−score of 88.83%, 91.07%, 58.47%, and 71.22% respectively.

Reference [11] developed predictive models to assess loan risk using LightGBM, XGBoost, Logistic Regression and Random Forest. Random forest obtained the best results with an Area Under Receiver Operating curve of 89%. The use of logit model to predict micro-loan default in LendingClub dataset was examined by [12]. The feature selection process which was applied to the dataset selected 20 features with the greatest impact using correlation coefficient analysis. The logistics regression model achieved an accuracy of 92.9%. Other metrics were not evaluated.

Reference [13] applied deep neural network to predict loan default which was compared with logistic regression, decision tree, naïve bayes, and support vector machine algorithms. The authors used two distinct datasets with 79254 instances obtained from a medium-sized Turkish bank. For the loan performance data, the proposed deep neural network model achieved a weighted accuracy of 77.98%. Logistic regression, decision tree, naïve bayes, and support vector machine algorithms achieved weighted accuracy of 77.31%, 70.05%, 78.14%, and 57.04% respectively. The percentage of misclassified good loans and bad loans performance for deep neural network are 10.20% and 25.95% respectively. For the loan application data to discriminate between the creditworthy and non-creditworthy applicants, the proposed deep neural network model achieved a weighted accuracy of 85.69%. Logistic regression, decision tree, naïve bayes, and support vector machine algorithms achieved weighted accuracy of 78.01%, 82.34%, 77.93%, and 75.25% respectively. The percentage of misclassified good loans and bad loans applications for deep neural network are 15.45% and 13.92% respectively.

In a bid to improve the performance of loan default prediction methods, [14] explored comprehensive pre-processing, extraction, and selection of features in the dataset. The enhancement approach which utilized information gain, genetic algorithm, and particle swarm optimization for feature selection was tested using naïve bayes, decision tree, and random forest classifiers. It was established that the data

pre-processing methods improved classification accuracy and model performance.

OptiML was used by [15] to forecast credit non-payments using dataset from microcredit EKI. Three models were shortlisted for evaluation after executing OptiML on the dataset: decision forest, neural network, and a logistic regression model. The decision tree achieved accuracy, precision, and f-measure of 94.6%, 69.5%, and 0.0596 respectively. Similarly, the neural network model achieved accuracy, precision, and f-measure of 82.1%, 15.5%, and 0.2396 respectively. Furthermore, the logistic regression model achieved accuracy, precision, and f-measure of 94.7%, 66.0%, and 0.0463 respectively.

Reference [16] focused on forecasting loan defaults in online lending peer-to-peer systems using bidirectional long short term memory (BiLSTM). The dataset comprised of over 440000 online comments on about 6000 online P2P lending companies from Wangdaizhijia in China. The performance of the proposed model was compared with support vector machine, decision tree, deep neural network, and text convolutional neural network. The proposed method achieved precision, recall and f1 scores of 0.7964, 0.7740, and 0.8034 respectively.

Reference [17] applied logistic regression, random forest, decision tree, adaboost, XGboost, artificial neural network and support vector machine algorithms to predict loan defaults. Also, the Synthetic Minority Oversampling Technique (SMOTE) was employed to treat the imbalance between classes for the response variable. It was observed that XGBoost without implementation of SMOTE obtained the best result.

Reference [18] combined label propagation and transductive support vector machine (TSVM) with Dempster–Shafer theory for accurate default prediction of social lending using unlabeled data. The experiment was performed using the Lending Club dataset. The proposed method achieved an accuracy and f1-score of 76.79% and 86.47% respectively. In another perspective on loan repayment, ascertaining the likelihood of repayment of a credit card loan was examined by Ma (2020). The author applied XGBoost model to a dataset with 30,000 samples of credit-card billing information and repayment information. The proposed model achieved an Area Under Receiver Operating Curve (AUC) of 0.779.

Reference [20] applied a boosted decision tree model for forecasting loan default in peer-to-peer lending communities using the publicly available United States small business administration dataset and the Imperial College London Kaggle competition dataset. The dataset which consists of 899,164 data instances was used in the 80:20 ratio for training and testing. On applying decision tree and boosted decision tree model to the dataset, 99% and 98% accuracy were recorded.

Similarly, [21] proposed a binary particle swarm optimization with support vector machine to perform feature selection for the lending club dataset. For the classification task, extremely randomized tree and random forest were used as classifiers. Extremely randomized tree obtained a better accuracy of 64% compared to random forest. The extreme randomized tree classifier outperforms random forest in execution time up to 46%. In another work based on tree classifiers, [22] predicted loan default in peer-to-peer lending platform using a heterogeneous ensemble decision tree model based on gradient boosting decision trees, extreme gradient boosting, and light gradient boosting machine. The ensemble method obtained sensitivity, specificity, f1-score, and accuracy values of 0.9596, 0.1589, 0.8615, and 0.7185.

A Taiwan credit dataset was employed in forecasting loan defaults by [23]. The classification task was performed using the bagging ensemble method with REP tree algorithm, linear regression, and decision stump. The proposed work obtained an accuracy of 81% when REP Tree was used compared to the base learners. In another work, [24] examined calculation of a bank's customer credit worthiness using Microsoft Azure machine learning studio. The proposed method was compared against three algorithms: bayes point, logistic regression, and decision tree. The proposed method achieved accuracy, true positive, recall, and prediction rate of 82.20%, 1360, 0.411, and 0.110 respectively.

Reference [25] presented a credit scoring model used by two microfinance institutions: one in Bosnia, the other Herzegovina. Data preprocessing was performed using Oracle data miner on the dataset which has 87531 records with over 60 attributes. The GLM algorithm in the Oracle data miner software was used to perform the classification task. Results obtained showed that GLM achieved an overall accuracy of 98.2046% and average accuracy of 98.7185%.

In this work, comparison of various machine learning algorithms to predict agricultural loan default is performed. The domain of agriculture was selected because agriculture is one of the major sectors in Nigeria that provides immense economic and job opportunities for the Nigerian masses. Also, given that the agricultural sector is one of the most underserved sectors, credit-wise, in Nigeria, it is imperative that research into identifying possible loan defaults in the sector given the high-risk rating banks accord the sector.

III. METHODOLOGY

The prediction of agricultural defaults was identified as a classification problem, hence, five machine learning classifiers – random forest, support vector machine, gradient boosting, Ada boosting, and logistic regression - were used on the dataset. Nominal features were converted to binary and ordinal numbers based on the characteristic of the feature. Feature selection was performed on the dataset to reduce the number of attributes and obtain good result. The classifiers were evaluated using accuracy, precision, recall, f1-score, and the Area Under Receiver Operating Curve (AUC) metrics.

A. Dataset

The dataset used in this work was obtained from a financial institution that liaised with farmers in Lavun local government area of Niger state in Nigeria. The financial institution facilitated the issuance of agricultural loans to farmers in Lavun who predominantly cultivate rice. The dataset contains 174 unique loan instances with labels indicating a loan default. Due to the number of attributes, 60 features, captured in the dataset relative to the number of unique instances, the dataset was subjected to feature selection process to enhance the prediction metrics. The dataset adhered to a uniform distribution pattern. That is, there were equal number of loan default and non-defaulting instances in the dataset.

B. Tool

The python programming language of version 3.8.0 was used to perform the classification task. The machine learning and feature selection algorithms used in this research were obtained from the sklearn python machine learning library of version 0.22.2.post1.

C. Feature Selection

The dataset contained 64 features and 174 instances. The performance of the machine learning algorithms to be explored is hinged on the use of relevant features for model training. Therefore, recursive feature elimination and cross-validation (RFECV) was used to select the best features. The feature selection process picked 44 features as the useful attributes in forecasting loan defaults from 60 features as shown in Table I.

1) Recursive Feature Elimination with Cross-Validation (RFECV)

Recursive Feature Elimination with Cross-Validation (RFECV) is a feature selection algorithm which trains a classifier on all the features in a dataset before selecting subsets of the dataset's features at each iteration. RFECV then cross-validates the features chosen against the classifier so as to mitigate the stochastic nature of machine learning classifiers. RFECV was used to select features by recursively exploring smaller sets of features continuously in a cross-validation loop to obtain the optimal feature count. Random forest was used as the estimator for the RFECV. The choice of random forest, as an estimator, was based on its ability to perform classification tasks on datasets without any need for data normalization. 5-fold cross-validation was also performed. Algorithm 1 describes the operations of RFECV.

Algorithm 1: Recursive Feature Elimination with Cross-Validation

Train random forest classifier on training set of data
Perform 5-fold cross-validation
Calculate variable rankings
For each subset size S_i, I = 1,2,…,S do
    Keep S_i most important variables
    Calculate random forest classifier performance
    Perform 5-fold cross-validation
end
Calculate the profile performance over S_i
Determine the appropriate numbers of features

D. Classification

Classification of the instances in the dataset was performed using random forest, support vector machine, gradient boosting, Ada boosting, and logistic regression algorithms. The overview and configuration details of the classifiers is given as thus.

E. Random Forest

Random forest is an ensemble learning method which performs classification operations by developing multiple decision trees. In this work, a random forest classifier from the sklearn library of python was used. The maximum depth of the tree was specified to be 5. The maximum depth specification was established due to poor performance of the classifier when the nodes are expanded until all leaves contain less than the minimum sample split.

F. Support Vector Machine

A support vector machine (SVM) classifier constructs a hyperplane or collection of hyperplanes in a high or infinite dimensional space for classification operations. In addition to linear classification, the SVM uses the kernel trick to classify non-linear data. The radial basis function kernel was used to classify loan defaulters using gamma and C parameters of 3 and 0.01 respectively. The kernel trick was applied due to the non-linear nature of the data. That is, there is no concrete feature to use in distinguishing a loan default instance from a non-default instance.

G. Gradient Boosting

Gradient boosting is an ensemble classification algorithm established on decision trees. Generally, gradient boosting performs better than random forest on the same dataset. In this research, the maximum depth of the decision tree is 6.

H. Adaptive Boosting

The adaptive boosting (Ada boost) classifier fits a classifier on the main dataset, then, additional copies of the classifier are applied on the same dataset. Then, the weights of incorrectly classified instances are adjusted to ensure that subsequent classifiers may focus on more difficult case.

I. Logistic Regression

Logistics regression is a statistical model which utilizes logistic function to model a binary dependent variable. The binary logistic regression model was used in this work. The liblinear solver for the optimization problem was employed because liblinear works well for small datasets and it handles the one-versus-rest schemes.

J. Performance Metrics

Measuring the performance of the classifiers used in this work is essential towards guiding further research in this domain and also, choosing the appropriate classifier for industrial application. Therefore, five performance metrics – accuracy, area under receiver operating characteristic curve (AUC), precision, recall, and f1-score - were employed in measuring the effectiveness of each classifier. Since loan default is being predicted, a loan default is tagged the positive class while a non-defaulting loan is of the negative class.

K. Accuracy

The accuracy of a classifier refers to the number of correctly classified instance in the dataset. That is, the sum of true positive instances with that of the true negative instances divided by the total number positive and negative instances. Here, true positive (TP) refers to the number of loan default that were correctly classified as loan defaults while true negative (TN), refers to the number of non-defaulting loans that were correctly classified as non-default loans. Furthermore, the total number of loan defaults is defined as positive (P) while the total number of non-defaulting loans is defined as negative (N) The formula for calculating accuracy is given in (1).

$$\frac{TP+TN}{P+N} \qquad (1)$$

L. Area Under the Receiver Operating Characteristics

Area Under the Receiver Operating Characteristics (AUC) refers to the degree of the classifier's separability based on the Receiver Operating Characteristic (ROC) curve. ROC curves show the trade-off between true positive rate and false positive rate. Here, the true positive rate refers to the rate at which a defaulting loan is classified as such while the false positive rate defines the rate at which non-defaulting loans are classified as loan defaults. Higher AUC means the classifier is better at predicting loan defaults and non-defaulting loans appropriately.

M. Precision

Precision, a measure of exactness, refers to the percentage of correct predictions among the test data. It measures the exactness of the classifier. The formula used in calculating precision is given in (2). Here, false positive (FP) refers to the non-defaulting loans mistakenly classified as defaulting.

$$\frac{TP}{TP+FP} \tag{2}$$

N. Recall

Recall, also known as sensitivity, is defined as the number of positive cases that were correctly identified. It measures the completeness of the classifier. The formula for calculating recall is given in (3). Here, false negative (FN) refers to the defaulting loans which were mistakenly classified as non-defaulting.

$$\frac{TP}{TP+FN} \tag{3}$$

O. F1-Score

F1-score, also known as f-score, is the harmonic mean of the precision and recall score. In other words, it conveys the balance between the precision and the recall of a classifier. A model with the best performance shows maximum f1-score. The formula for calculating the f1-score of a model is given in (4).

$$\frac{2 \text{ x precision x recall}}{\text{precision+recall}} \tag{4}$$

TABLE I: FEATURES OF THE DATASET AND SELECTED FEATURES INDICATED BY ASTERISK *

Sex	Yield sale to offtakers*	Yield sale to farmgate*	Yield sale to self*
Marital status*	Yield wasted*	Production method*	Farming cost*
Age range*	Awareness of credit*	Challenge accessing credit*	Count of credit access to farm
Farming regularity	Challenge repaying credit*	Bank account	Type of account*
Farming duration*	Account opening facilitator*	Account opening date*	Reason for opening account*
Farming system	Frequency of saving*	Withdrawal frequency*	Number of dependants*
Association link*	Household size*	Male children between 6 to 18 years*	Female children between 6 to 18 years*
Education level	Number of male children attending school*	Number of female children attending school*	Hunger due to inadequate food*
Farm size*	Main source of income	Total income per month*	Number of full time female workers*
Cultivation interval	Number of part time male workers*	Number of part time female workers*	Number of full time male workers
Crop type	Type of farm animals*	Food shortage*	Drinking water treatment*
Locality seasons*	House roofing material*	House building material*	Access to electricity*
Reason for not cultivating multiple seasons*	Access to agricultural insurance	Awareness of agricultural insurance	Account opening balance at start of season*
Season 1 yield gain	Access to health facilities	Toilet facility*	Mechanized farm tools*
Season 2 yield gain	Personal possession	Source of farm funding*	Media accessed by farmers*

IV. Results and Discussions

The machine learning algorithms selected for the classification process were applied to the dataset which had the selected features. The performance of the models was measured using accuracy, precision, recall, AUC, and f1-score metrics. Also, the receiver operating characteristic curve was plotted to show the performance of each model at all classification thresholds. Table II shows the accuracy, AUC score, precision score, recall score, and f1-score of the models applied to the agricultural loan dataset. Fig. 1 illustrates the performance of each model against each metric.

TABLE II: CLASSIFIER PERFORMANCE METRIC EVALUATION

Classifiers	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-score**
Gradient Boosting	88.57%	87.5%	86.36%	95%	90.48%
Ada Boost	80%	80.83%	88.24%	75%	81.08%
Random Forest	80%	80.94%	89.47%	77.27%	82.93%
SVM	80%	80.59%	73.68%	87.5%	79.99%
Logistic Regression	82%	84.52%	72.22%	92.86%	81.25%

For the model accuracy, gradient boosting and logistic regression model obtained the highest accuracy of 88.57% and 82% respectively. The other models – Ada boost, random forest, and SVM – obtained an accuracy of 80%. Based on the accuracy of the models examined, gradient boosting was able to classify most of the test instances correctly compared to the other four models. Fig. 1 shows the classification accuracy of the models.

Similarly, gradient boosting and logistic regression model achieved the highest AUC scores of 87.5% and 84.52% respectively. This means that the gradient boosting and logistic regression model had better separability than ada boost, random forest, and SVM. In other words, the models were able to tag more defaulting loans as defaults and non-defaulting loans as non-defaults compared to the other models.

Interestingly, random forest and ada boost models attained the best precision scores of 89.47% and 88.24% respectively. That is, the random forest and ada boost models classified a lesser number of non-defaulting loans as defaults (false positive) while predicting a greater number of defaulting loans as defaults (true positive). Although logistic regression

and gradient boosting models had the best accuracy and AUC scores, they performed poorly in labelling non-defaulting loans. In other words, logistic regression and gradient boosting models labelled more non-defaulting loans as defaults compared to random forest and ada boost models.

Gradient boosting and logistic regression models proved efficient in predicting agricultural loan defaults by attaining recall scores of 95% and 92.86% respectively. In other words, gradient boosting and logistic regression models were able to predict agricultural loan defaulters better than other classifiers. SVM followed with a recall score of 87.5%. Random forest and Ada boost trailed behind with recall scores of 77.27% and 75% respectively. When the recall scores and precision scores of random forest and Ada boost models are juxtaposed, it can be deduced that although the models classified lesser number of non-defaulting loans as defaults, they could not identify agricultural loan defaulters effectively.

Consequently, the f1-score of 90.48% was obtained by gradient boosting model, thus, making the model the most effective in predicting agricultural loan defaulters. Random forest which performed averagely in most of the metrics, apart from the precision metric, proved to be the second most effective model with an f1-score of 82.93%. Logistic regression model positioned itself as a competitive model by obtaining an f1-score of 81.25%. Ada boost and SVM models attained f1-scores of 81.08% and 79.99% respectively.

Overall, the gradient boosting model is the most effective model in predicting agricultural loan defaulters among smallholder farmers in Lavun local government Niger state, Nigeria.

V. CONCLUSION

In this work, the prediction of agricultural loan defaulters using five machine learning algorithms – random forest, SVM, gradient boosting, Ada boosting, and logistic regression – was performed on the dataset from Lavun local government area of Niger state in Nigeria. The results achieved shows that gradient boosting and random forest algorithms were the most effective in predicting agricultural loan defaults with precision score of 86.36% and 89.47% respectively. Furthermore, the f1-score of 90.48% and 82.93% was obtained by gradient boosting and random forest respectively.

Further studies on the improvement of the accuracy of the other machine learning models is proposed.

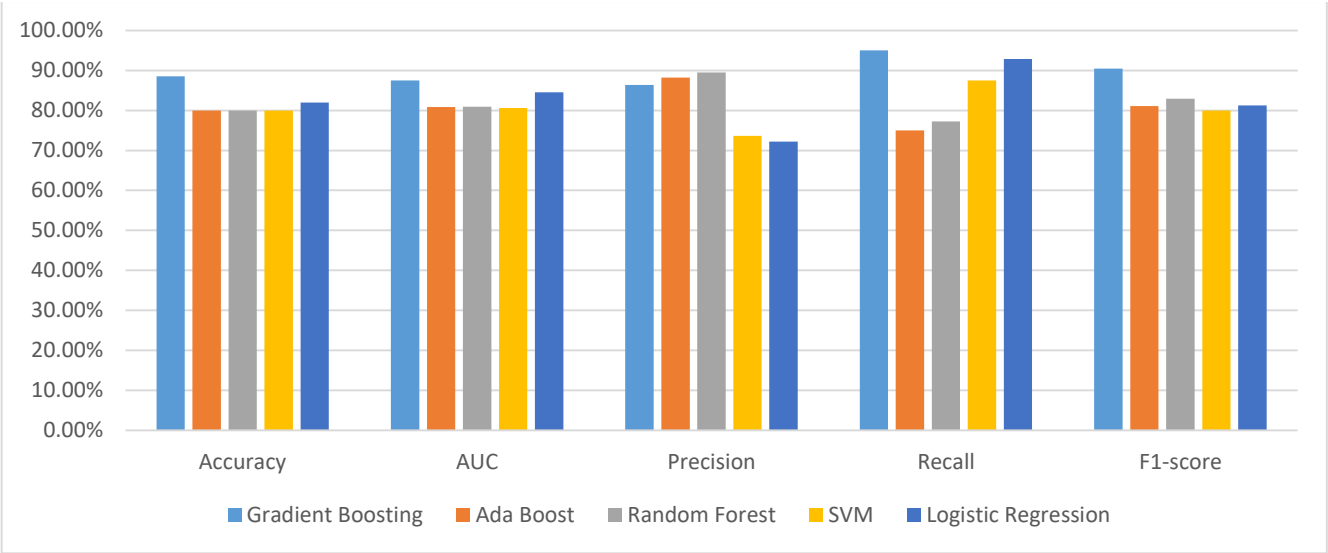

FIGURE 1: PERFORMANCE METRICS OF THE MACHINE LEARNING ALGORITHMS

REFERENCES

[1] J. R. Porter, A. J. Challinor, C. B. Henriksen, S. M. Howden, P. Martre, and P. Smith, "Invited review: Intergovernmental Panel on Climate Change, agriculture, and food—A case of shifting cultivation and history," *Global Change Biology*, vol. 25, no. 8. pp. 2518–2529, 2019.

[2] M. O. Adenekan and E. O. Augustus, "AGRICULTURAL TRANSFORMATION IN NIGERIA FOR SUSTAINABLE FOOD SECURITY," *J. Glob. Biosci.*, vol. 10, no. 1, pp. 8230–8242, 2021.

[3] G. O. Evbuomwan and L. U. Okoye, "Evaluating the Prospects of the Anchor Borrowers ' Programme for Small Scale Farmers in Nigeria," vol. 2, no. July, pp. 1–10, 2017.

[4] M. Sulaimon, "Agricultural credit guarantee scheme fund (ACGSF) and agricultural performance in Nigeria: A threshold regression analysis," no. 105564, 2021.

[5] M. Mustapha, "Food Insecurity and Coping Strategies among Rural Households in Niger State, Nigeria," *Lapai J. Econ.*, vol. 3, no. 1, pp. 92–107, 2019.

[6] B. Dushimimana, Y. Wambui, T. Lubega, and P. E. McSharry, "Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans," *J. Risk Financ. Manag.*, vol. 13, no. 8, p. 180, 2020.

[7] A. Datkhile, K. Chandak, S. Bhandari, H. Gajare, and M. Karyakarte, "Statistical Modelling on Loan Default Prediction Using Different Models," *IJRESM*, vol. 3, no. 3, pp. 3–5, 2020.

[8] B. Patel, H. Patil, J. Hembram, and S. Jaswal, "Loan default forecasting using data mining," in *2020 International Conference for Emerging Technology, INCET 2020*, 2020, pp. 7–10.

[9] M. Aslam, S. Kumar, and S. Sorooshian, "Predicting likelihood for loan default among bank borrowers," *Int. J. Financ. Res.*, vol. 11, no. 1, pp. 318–328, 2020.

[10] E. Elakkiya, K. Radhaiah, and G. M. Rayalu, "Logistic regression models for prediction loan defaults-qualtitative data analysis," vol. 9, no. 8, pp. 6027–6034, 2020.

[11] A. Coşer, M. M. Maer-Matei, and C. Albu, "Predictive models for loan default risk assessment," *Econ. Comput. Econ. Cybern. Stud. Res.*, vol. 53, no. 2, pp. 149–165, 2019.

[12] T. Deng, "Study of the prediction of micro-loan default based on logit model," in *Proceedings - 2019 International Conference on Economic Management and Model Engineering, ICEMME 2019*, 2019, pp. 260–264.

[13] S. Bayraci and O. Susuz, "A Deep Neural Network (DNN) based classification model in application to loan default prediction," *Theor. Appl. Econ.*, vol. XXVI, no. 4, pp. 75–84, 2019.
[14] A. Al-Qerem, G. Al-Naymat, and M. Alhasan, "Loan default prediction model improvement through comprehensive preprocessing and features selection," in *Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019*, 2019, pp. 235–240.
[15] E. Zoran, "Predicting Default Loans Using Machine Learning (OptiML)," in *27th Telecommunications Forum TELFOR*, 2019, vol. 7, pp. 1–27.
[16] X. Fu, T. Ouyang, J. Chen, and X. Luo, "Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks," *Inf. Process. Manag.*, vol. 57, no. 4, p. 102236, 2020.
[17] N. Madane and N. Siddharth, "Loan Prediction Analysis Using Decision Tree," *J. Gujarat Res. Soc.*, vol. 21, no. 14, pp. 214–221, 2019.
[18] A. Kim and S. B. Cho, "An ensemble semi-supervised learning method for predicting defaults in social lending," *Eng. Appl. Artif. Intell.*, vol. 81, no. December 2017, pp. 193–199, 2019.
[19] Y. Ma, "Prediction of Default Probability of Credit-Card Bills," *Open J. Bus. Manag.*, vol. 08, no. 01, pp. 231–244, 2020.
[20] A. Semiu and A. A. R. Gilal, "A boosted decision tree model for predicting loan default in P2P lending communities," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 1257–1261, 2019.
[21] N. Setiawan, Suharjito, and Diana, "A comparison of prediction methods for credit default on peer to peer lending using machine learning," in *Procedia Computer Science*, 2019, vol. 157, pp. 38–45.
[22] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Phys. A Stat. Mech. its Appl.*, vol. 534, p. 122370, 2019.
[23] A. Motwani, G. Bajaj, and S. Mohane, "Predictive Modelling for Credit Risk Detection using Ensemble Method," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 6, pp. 863–867, 2018.
[24] A. Motwani, P. Chaurasiya, and G. Bajaj, "Predicting Credit Worthiness of Bank Customer with Machine Learning over Cloud," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 7, pp. 1471–1477, 2018.
[25] J. Nalić and A. Švraka, "Using data mining approaches to build credit scoring model: Case study - Implementation of credit scoring model in microfinance institution," in *2018 17th International Symposium on INFOTEH-JAHORINA, INFOTEH 2018 - Proceedings*, 2018, vol. 2018-Janua, no. March, pp. 1–5.