Hindawi
Advances in Fuzzy Systems
Volume 2020, Article ID 8581202, 12 pages
https://doi.org/10.1155/2020/8581202

*Research Article*

# Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks

**Peter Appiahene, Yaw Marfo Missah, and Ussiph Najim**

*Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

Correspondence should be addressed to Peter Appiahene; peter.appiahene@uenr.edu.gh

Received 3 February 2020; Revised 23 April 2020; Accepted 30 June 2020; Published 23 July 2020

Academic Editor: Zeki Ayag



The financial crisis that hit Ghana from 2015 to 2018 has raised various issues with respect to the efficiency of banks and the safety of depositors' in the banking industry. As part of measures to improve the banking sector and also restore customers' confidence, efficiency and performance analysis in the banking industry has become a hot issue. This is because stakeholders have to detect the underlying causes of inefficiencies within the banking industry. Nonparametric methods such as Data Envelopment Analysis (DEA) have been suggested in the literature as a good measure of banks' efficiency and performance. Machine learning algorithms have also been viewed as a good tool to estimate various nonparametric and nonlinear problems. This paper presents a combined DEA with three machine learning approaches in evaluating bank efficiency and performance using 444 Ghanaian bank branches, Decision Making Units (DMUs). The results were compared with the corresponding efficiency ratings obtained from the DEA. Finally, the prediction accuracies of the three machine learning algorithm models were compared. The results suggested that the decision tree (DT) and its C5.0 algorithm provided the best predictive model. It had 100% accuracy in predicting the 134 holdout sample dataset (30% banks) and a $P$ value of 0.00. The DT was followed closely by random forest algorithm with a predictive accuracy of 98.5% and a $P$ value of 0.00 and finally the neural network (86.6% accuracy) with a $P$ value 0.66. The study concluded that banks in Ghana can use the result of this study to predict their respective efficiencies. All experiments were performed within a simulation environment and conducted in R studio using R codes.

## 1. Introduction

The financial crisis that hit Ghana from 2015 to 2018 has raised various issues with respect to the efficiency of banks and the safety of depositors' in the banking industry. As measures to mitigate this financial crisis, the central bank (bank of Ghana) instituted some measures to reform the banking sector. The sole aim is to provide efficient banking services to the Ghanaian economy. It is also to make the local banks competitive globally. First and foremost, governments decided to avoid closure of distressed banks by recapitalizing them. For instance, the central bank directed a state owned bank (Ghana Commercial bank) to take over UT and Capital bank in 2017 [1]. Secondly, the bank of Ghana also avoided closure of banks in order to protect depositors' funds by merging five distressed banks, namely, Unibank, Biege bank, Construction bank, Royal bank, and Sovereign bank [2]. This consolidation process was envisaged to restore the financial viability of the distressed banks. Finally, the central bank also raised the minimum capital requirements of all commercial banks in Ghana to 400 million Ghana Cedis [3]. As part of measures to improve the banking sector and also restore customers' confidence, efficiency, and performance analysis in the banking industry has become a hot issue. This is because bank managers and other stakeholders want to detect and mitigate the underlying causes of inefficiencies within their banking operations. As two nonparametric models, DEA models share some similarities with machine learning algorithms. For example, DEA and machine learning algorithms both make assumptions about the

functional form that links its inputs to outputs. Bank's branch efficiency is also a comprehensive measure from various performance aspects using many financial variables [4]. This indicates that the relationship between the bank efficiency and multiple variables is highly complex and not straight forward. Machine learning algorithms have also been viewed as a good tool to approximate numerous nonparametric and nonlinear problems [5]. This means that the banking industry provides good opportunities for the applications of a combined DEA and machine learning models. There are also few literatures dealing with developing country bank branch efficiency using DEA and machine learning algorithms. This paper presents a combined DEA and three machine learning approaches in evaluating bank efficiency and performance using 444 Ghanaian bank branches. The results were also compared with the corresponding efficiency ratings obtained from CRS DEA. Finally, the prediction accuracies of the three machine learning algorithm models were compared. The motivation behind this study is the fact that the DEA property of unit invariant is similar to the property of scale preprocessing required by machine learning algorithms such as NNs. This validates the rationale to compare the results of pure DEA and DEA-machine learning algorithm model results.

The rest of the paper is organized as follows. Section 2 gives a brief review of related works on the topic. Section 3 presents the methodology and the framework used in the study. Section 4 gives both DEA and the three machine learning algorithm results analysis and further discussions. Finally, Section 5 presents our conclusions, recommendations, and future work suggested by the study.

## 2. Related Works

Nonparametric methods such as Data Envelopment Analysis (DEA) have been suggested in the literature [6–18] as a good measure of banks' efficiency and performance. For instance, [19] through the DEA model evaluated the marginal benefits of IT using 36 DMUs financial institution's data. The study suggested that for a collection of IT investment, IT impacted substantially on organizations revenues. In 2004, [20] also used DEA to assess the efficiency of 27 banks and suggested a positive impact of IT on the banks' efficiency. Chen et al. [21] in a study also used 27 DMUs of banks that suggested only three firms as efficient in the two efficient calculation phases. This work [22] assessed 40 Internet company's firms' performance using a DEA model which, according to them, can work well and can also be used to differentiate the causes of inefficiency. A study [23] attempted to evaluate the impact of ICT on the productivity of hotels in Portugal through Data Envelopment Analysis (DEA). The study did not only demonstrate how important ICT is in realizing advanced levels of productivity. It also discussed other explicit concerns which should be taken into consideration so that the positive returns of the investment in ICT can be achieved. Comparatively, DEA is a better way to arrange and evaluate data since it allows efficiency to change over time and requires no prior assumption on the specification of the best practice frontier [24]. It has also been reported in the literature [24] that DEA is a prominent method for performance analysis in the banking industry. However, the DEA frontier is very sensitive to the presence of outliers and statistical noise. It can hardly be used to predict the performance of other decision making units [24]. As a result, studies have started introducing machine learning recently as good substitutes to support in approximating efficiency frontiers for decision makers [24]. For example, [25] demonstrated how a machine learning algorithm, such as decision tree, was combined with DEA to predict the impact of IT on firms' performance. In another work [26], the authors used three decision tree algorithms, namely, C5.0, C4.5, and CART, to build the various decision tree predictive models. The study suggested that the C5.0 algorithm gave an accuracy of 100%, followed by the CART algorithm with an accuracy of 84.6% and, finally, the C4.5 algorithm with an accuracy of 83.34 on average. The study, therefore, recommended the usage of the C5.0 predictive model in predicting the financial performance of rural banks in Ghana. Chen et al. also [27] applied an innovative Data Envelopment Analysis method under a stochastic environment. The results of the study reveal that the overall efficiency level of the Chinese banks remains still low. This, according to the authors, was considerably determined by the contextual variables of the ownership structure and cost structure of the Chinese banks [27]. Another study [28] also used a novel approach, Synthetic Minority Oversampling Technique (SMOTE), to convert imbalanced data in a balanced form. The authors used Lasso regression to reduce the redundant features from the failure predictive model. The result of this study holds its application to various stakeholders like shareholders, lenders, and borrowers, etc. to measure the financial stress on banks [28]. A work [29] found a very similar performance for both models where random forest shows slight superiority to logistic regression. Both models yield an AUC of ~0.65, and from the results obtained, it indicates that they are able to correctly predict ~60% of both healthy and financially distressed companies ahead of time [29]. A study [30] also compared the accuracy of two approaches: traditional statistical techniques and machine learning techniques in an attempt to predict the failure of 3000 US banks. The empirical result of the study reveals that the artificial neural network and K-nearest neighbor methods were the most accurate. Finally, [4] utilized a Multinomial Logistic Regression to select the most significant predictor variables to build a neural network model and suggested that the models in each case yielded a favorable classification and prediction accuracy rate.

## 3. Basics of Data Envelopment Analysis (DEA)

Data Envelopment Analysis is a nonparametric method that produces a comparative ratio of weighted outputs to inputs for each Decision Making Unit (DMU) under consideration [31–33]. This study presumes that there are $n$ DMUs to be evaluated and in this case ($n = 444$). Each DMU consumes $m$ different inputs.

Specifically, DMU$j$ consumes the amount$x_{ij}$ of input $i$ and generates a quantity $y_{rj}$ of output $r$. The further adopts

that $x_{ij} > 0$ and $y_{rj} > 0$. The input-oriented efficiency of a specific $DMU_0$ under the postulation of Variable Returns to Scale (VRS) can be deduced from the following primal-dual linear programs, the BCC model proposed by [31].

The BCC Envelopment model is as follows:

$$\begin{aligned} &\min \\ &\theta, \lambda, s^+, s^- z_0 = \theta - \varepsilon.\overrightarrow{1}s^+ - \varepsilon.\overrightarrow{1}s^- \\ &\text{subject to } Y\lambda - s^+ = Y_0, \\ &\theta X_0 - Y\lambda - s^- = 0, \\ &\overrightarrow{1}\lambda = 1, \\ &\lambda, s^+, s^- \geq 0, \end{aligned} \tag{1}$$

where $s^+$ and $s^+$are the slacks in the system.

BCC Multiplier form is as follows:

$$\begin{aligned} &\max \\ &\mu, \nu \quad w_0 = \mu^T Y_0 + u_0. \end{aligned} \tag{2}$$

This is also subject to $\nu^T X_0 = 1$ and

$$\begin{aligned} \mu^T Y - \nu^T X + u_0 \overrightarrow{1} &\leq 0 \\ \mu^T &\leq -\varepsilon.\overrightarrow{1} \\ \nu^T &\leq -\varepsilon.\overrightarrow{1} \\ u_0 &\text{ free.} \end{aligned} \tag{3}$$

Determining a DEA involves solving $n$ linear programming tasks of the above model, one for each DMU. The ideal estimation of the variance $\theta$ determines the corresponding diminution of all inputs for DMU0 that will transfer it onto the frontier, which is the envelopment surface defined by the efficient DMUs in the sample. DMU0 is DEA efficient as far as there exists an ideal solution $\mu^*, \nu^*$ of (3) with $\mu^* > 0$ and $\nu^* > 0$, and an ideal solution $(\theta^*, \lambda^*)$ of equation (1) such that

$$z_0^* = w_0^* = 1, \tag{4}$$

where $z_0^*$is the response for the BCC Envelopment form and $w_0^*$is the response for the BCC Multiplier form.

From this end and beyond, the optimal value is denoted by $^*$. The condition on $\mu^* > 0$ and $\nu * > 0$ assures that DMU0 is an efficient frontier and that slack values of all constraints in the (1) are 0 because of the complementary slackness proposition for dual programs. This model allows Variable Return to Scale (VRS) proposed by [31].

If the convexity constraint ($\overrightarrow{1}\lambda = 1$) in (1) and the variable $u_0$in (3) are taken out, the feasible region is increased, which results in the decrease in the number of efficient DMUs, and all DMUs are operating at Constant Return to Scale (CRS), and resultant DEA model is CCR also proposed by [32].

*3.1. Decision Tree Algorithm.* The decision tree is one of the topmost machine learning algorithms which suggest a graphical or a diagrammatic illustration of a technique for classifying, predicting, and evaluating an item of importance or concern. It is an easy and commonly used classification method. It deals with decision analysis by employing a tree-like structure of decisions and its relative potential outcomes [34]. It has nodes where at each node in a decision tree, an attribute must be selected to divide the node's instances into subgroups.

Decision Tree accepts input set of well-ordered data, and output shaft, which is delivered in which each end node (leaf) is a decision (a class) and each nonend node (middle) shows a test [35]. The most common algorithms used in decision tree are ID3, CART, CHAID, and C4.5 with its extension C5.0. For this study, the C5.0 which is an extension of the ID3 and successor of the C4.5 algorithm proposed by Ross Quinlan in 1994 [36] was adopted for the study and implemented in R studio using R codes with package C5.0 [37].

*3.2. Random Forest Algorithm.* Random forest (RF) algorithm is for classification and prediction developed by Breiman in 2001 and cited by [38] that utilizes an ensemble of classification trees [39–41]. RF is an ensemble machine learning algorithm [39]. The fundamental principle of the RF algorithm is that constructing a smaller DT with limited characteristics is an inexpensive process in terms of computation [39]. Thus, it is possible to construct numerous small, weak decision trees in parallel and merge these smaller trees to form one strong learner by using their mean performance or even or selecting the popular one. In terms of application and practicability, RF algorithms are considered to be more precise learning algorithms to date [39].

The RF algorithm adopted for this study was Leo Breiman and Adele Cutler random forest algorithm [39, 42]. This was implemented in R studio using R codes with "randomForest" package [43, 44]. A random forest model has a better ability in modeling and predicting. An important feature of Breiman's algorithm, according to [45], is the variable importance calculation.

*3.3. Artificial Neural Network.* Artificial neural network (ANN) is a type of Artificial Intelligence (AI) technique that mimics the behavior of the human brain [46, 47]. A neural network is a massively parallel distributed processor made up of simple processing units that have a natural tendency for storing experiential knowledge and making it available for use. ANNs can be grouped into two major categories: feed-forward and feedback (recurrent) networks. In the former network, no loops are formed by the network connections, while one or more loop may exist in the latter. The most commonly used family of feed-forward networks is a layered network in which neurons are organized into layers with connections strictly in one direction from one layer to another [48].

The basic system of NN without the hidden layer consists of only two layers: the input and output layer. This is normally called the skip layer because it is made up of a straight forward linear regression modeling in a NN design. The input layer communicates directly with the output layer without involving the hidden layer. This study adopted the backpropagation algorithm for building the neural network model for predictions. The linear combination functions and

sigmoid transfer functions were used. The S-shaped or binary sigmoidal function is, by far, the most common transfer function [49]. The formula for the sigmoid is given: $\text{Sigmoid}(x) = (1/1 + e^{-x})$.

The codes for the NN model building was written in R codes using the RMiner studio with the "neuralnet" package [50]. This study adopted the use of only one hidden layer with five (5) Neurons in a three (3) layer network. This number of hidden neurons was chosen based on the equation Nh = Ni-1 proposed by [51] and cited by [52], where Nh is the number of hidden neurons to be used and Ni is also the number of input neurons. For this study, the number of inputs was 6 which implies that Ni = 6.

*3.4. List of Performance Measures.* There are so many metrics for evaluating machine learning algorithms, but for the purpose of this study, we would focus on the following:

*Classification Accuracy*. Classification accuracy is actually the meaning of the term accuracy in machine learning performance measure [53]. Mathematically, it is defined as the ratio of the number of predictions done correctly by the machine learning algorithm to the total data set:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{the total number of predictions made}}. \quad (5)$$

*Confusion Matrix*. Confusion matrix also presents a matrix as output and defines the comprehensive performance of the model. The confusion Matrix forms the basis for the other types of metrics.

*True Positive Rate (Sensitivity)*. True positive rate is defined as TP/(FN + TP). This represents the proportion of positive data points that are correctly considered as positive, with respect to all positive data points:

$$\text{true positive rate} = \frac{\text{true positive}}{\text{false negative} + \text{true positive}}. \quad (6)$$

*False Positive Rate (Specificity)*. False Positive Rate is defined as FP/(FP + TN). This represents the proportion of negative data points that are incorrectly considered as positive, with respect to all negative data points:

$$\text{false positive rate} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}}. \quad (7)$$

*Kappa*. A measure of how closely the occurrences predicted by a predictive model or classifier correspond to the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy.

## 4. Methodology

*4.1. Proposed Framework of the Study.* This framework suggested by this study was used to build the predictive models. It consists of three different stages: Data collection stage (Stage I), Data preprocessing stage (Stage II), and the predictive model development stage (Stage III). This means that the dataset for the model development goes through three different stages, Stage I, Stage II, and Stage III (Figure 1).

At the data collection stage (Stage I), the raw data were collected from the banks. After the collection of the data, it is preprocessed in Stage II before it finally enters the stage where the predictive model development takes place, thus Stage III. It is during the preprocessing stage where the entire dataset is organized, transformed, or Encoded into a form that can easily be used by the model. In this case, the financial data, such as the Cedi value of IT expenditure (I), fixed asset (A), total deposit (D), profit (R), rate of performing loans (%PL), and finally the number of employees (E) from banks was used to calculate the efficiencies of the various banks (DMUs) using the CRS technology. The efficiencies of the banks at both deposit and investment stages were classified into classes (Class A: efficient and Class B: inefficient) based on their efficiency scores (efficiency score of 1 unit or 100%). In real life situations, it is very difficult to have units or departments attaining 100% efficiency and bank branches in Ghana are not an exception [4].

This is also evident in the Ghanaian banking sector, where the central bank (bank of Ghana) always has a minimum capital requirement for banks and other financial institutions operating in Ghana [3, 54, 55]. This means that there is always a "cutoff point" for banks to meet in order to be efficient and remain competitive in the banking sector. Based on this, the authors also inferred and considered banks with an efficient value of 80% or more as efficient. The study, therefore, adopted and used the efficiency "cutoff point" (DMU efficiency $\geq 0.8$) suggested by [4, 25] and considered efficient bank as one with an efficiency value $\geq 0.8$. This efficiency classification (Class A and Class B) was used as the response variable.

Now the banks, financial data, such as the IT expenditure, fixed assets, etc., were used as predictor variables to predict the efficiency scores (classes) of each bank branch for building the models.

The efficiency classes and the predictor variables formed the dataset of our models. This final dataset for building the model is then randomly divided into two, 70% to be used to train or build and also validate the model using K-folds cross-validation. The remaining 30% (test dataset) was used to test the models. Thus, in the case of the banks, 70% of the DMUs which were selected randomly from the total dataset, were used to build and validate the model. This model was used to predict the efficiency of the other 30% of banks (DMUs). During the model building, the dataset also goes through rule extraction and finally building the classifier.

*4.2. Two-Stage DEA Model for Efficient Analysis.* In this DEA model, the various units under consideration were bank branches in Ghana whose performance or productivity measures were grouped into inputs and outputs. Using the bank as an example to derive our model shown in Figure 2, the banks business process and activities are viewed as a dual role process. The first stage (Deposits Stage) of the model

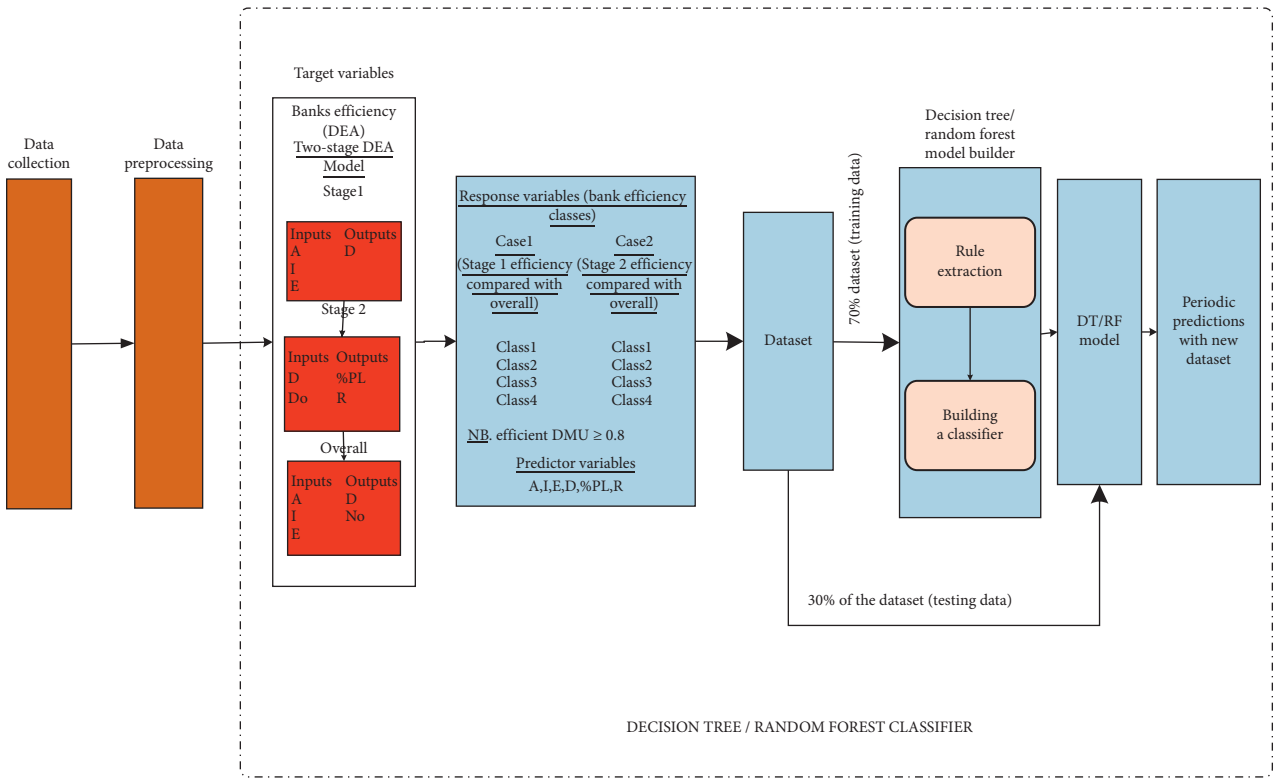


Figure 1: The research framework (the authors' construct).

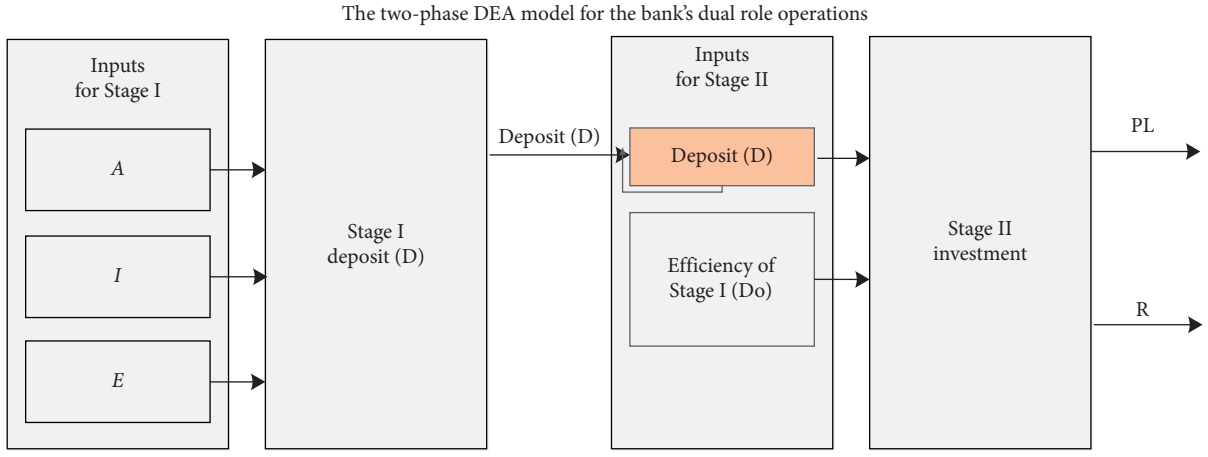


Figure 2: The proposed dual role DEA model adapted from [33].

consists of a collection of funds (Deposits) in Ghana Cedi as an intermediate measure of customers using their fixed asset, a number of workers (Employees at each unit), and IT infrastructure. In the next stage (Investment Stage), these banks use the deposits accumulated in stage I and their stage I efficiency scores invest the deposits into securities and also give loans to its customers. Returns (Profits) generated from the investment in securities and percentage of performing loans [56–58], which is a good indicator of risk status, were used as two outputs in stage II.

The DEA model used at the deposit stage I produce stage I efficiency for each DUM and the DEA applied at the Investment stage II also gives stage II efficiency for each DMU. DEA models employed at the overall stage finally give the overall efficiency for each DMU. The proposed DEA framework for the banks dual role operation is depicted in Figure 2.

The various variables used at the deposit stage, investment stage, and finally, the overall stages are stated as follows:

Deposit Stage, I

Input:

Fixed Assets (billions of GH) denoted as A
Total IT expenditure (billions of GH) denoted as I
Total number of Employees denoted as E

Output:

Deposit

Investment Stage, II

Input:

The efficiency of the stage I denoted as Do
The deposit also denoted as D.

Output:

Percentage of Performing loans (PL)
Profit accrued from investing in securities (R)

Overall Stage

Input

Efficiency of stage 2 denoted as No
Fixed Assets (billions of GH) denoted as A
Total IT expenditure (billions of GH) denoted as I
Total number of Employees denoted as E

Output:

Percentage of Performing Loans (PL) which is equal to the percentage of nonperforming loans and 100%
Profit accrued from investing in securities (R)

After using the classical CCR model of DEA, the efficiency score for each DUM in each stage $w$ resulted in the following three types of efficiency score:

Efficiency of stage I, Do
Efficiency of stage II, No
Overall efficiency, G

*4.3. Dataset and Sample.* At this point in the study, the authors demonstrated how DEA was used to assess the efficiency of firms using bank branches in Ghana (DMUs) as a case study. The dataset collected for the study contains 444 bank branches (DMUs). The Cedi value of the banks' IT expenditure, the Cedi value of fixed assets, the number of staff at each branch, profits generated from investing deposits, percentage of performing loans on the various bank branches, and the Cedi value of the total deposit were obtained from the various DMUs. Specifically, the audited 2016 financial statements from each bank were used.

For each bank branch, the technical efficiencies in both stages and their corresponding overall efficiencies were analyzed using CCR DEA technology. The efficiency of the stage I was calculated using IT expenditure (GH ¢), Fixed Asset (GH ¢), and the number of employees at each DMU as inputs with deposit as the main output. With respect to stage II, the stage I efficiency (Do) and the deposit (GH¢) realized from stage I were used as input with banks profit after all the necessary deductions and the percentage of performing loans as outputs.

The efficiency of the overall stage was also calculated using fixed assets, the number of employees IT expenditure, and efficiency of stage II (No) whiles their outputs were the banks profit after all the necessary deductions and the percentage of performing loans. The efficiency of each DMU at each stage was calculated using the DEA Two-Phase BuildHull Algorithm which was implemented in Rminer studio Version 1.2–5 [59] with its package Robust Data Envelopment Analysis (rDEA). The overall efficiency score of each DMU was categorized as either efficient-Class A or inefficient-Class B.

*4.3.1. Predictor Variables.* These are variables also called independent variables or experimental variables employed in statistical analysis to forecast or predict another variable called target or dependent variable [38, 60, 61]. For this study, the predictor variables were fixed assets, IT expenditure, number of employees, total deposits, percentage of performing loans, and profits accrued from investing the deposit.

*4.3.2. Response Variables.* The response variables (overall efficiency scores of the bank branches) that were categorized into efficient (Class A) or inefficient (Class B) were used as the response variable for the predictive models.

## 5. Results and Discussion

*5.1. Bank Efficiency Scores and Classes Using the Adapted DEA Two-Phase BuildHull Algorithm.* For each DMU (Bank), the technical efficiencies in both stages (Deposit Stage, Investment Stage) and their corresponding overall efficiencies were analyzed using CCR DEA Two-Phase BuildHull algorithm proposed by [62]. Using the 444 bank branches (DMUs), the efficiency of each bank branch at each stage was analyzed using the following scenarios:

Scenario 1: When an efficient unit is defined as a unit with an efficiency score of 1 unit or 100%.

For banks' efficiency in terms of utilizing their resources to collect deposits from customers, only 14 (3.15%) bank branches were efficient (had 100% efficiency). Just 33(7.43%) bank branches had an efficiency score of between 80% to 99%, 1 (0.23%) bank branches also had efficiency score of between 70 and 79, 21 (4.73%) had efficiency score of between 60 and 69, 19 (4.28%) had between 50 to 59, and finally 356 (80.18%) had an efficiency score below 50%. This 356 (80.18%) number of bank branches confirms the fact that a lot of Ghanaian banks are not efficient in using their resources to collect deposits from customers as most banks were struggling to meet the minimum capital requirements set by the central bank (bank of Ghana) in 2017 [3, 54, 55]. The result of the deposit stage efficiency is also shown in Figure 3.

For banks' efficiency with respect to investing customers deposits shown in Figure 4, only 1 bank (DMU200) was efficient in investing the deposit to generate profit for the

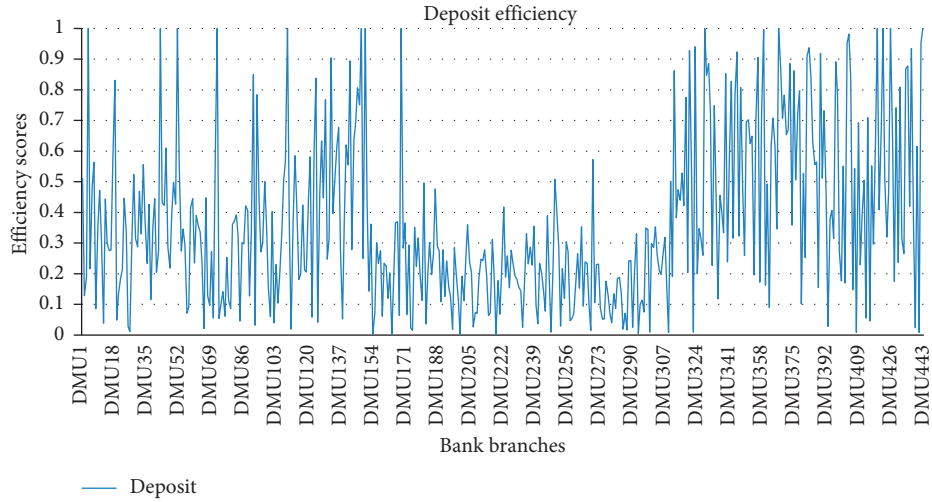


Figure 3: A graph showing the deposit efficiency scores of the 444 DMUs (the authors' construct).

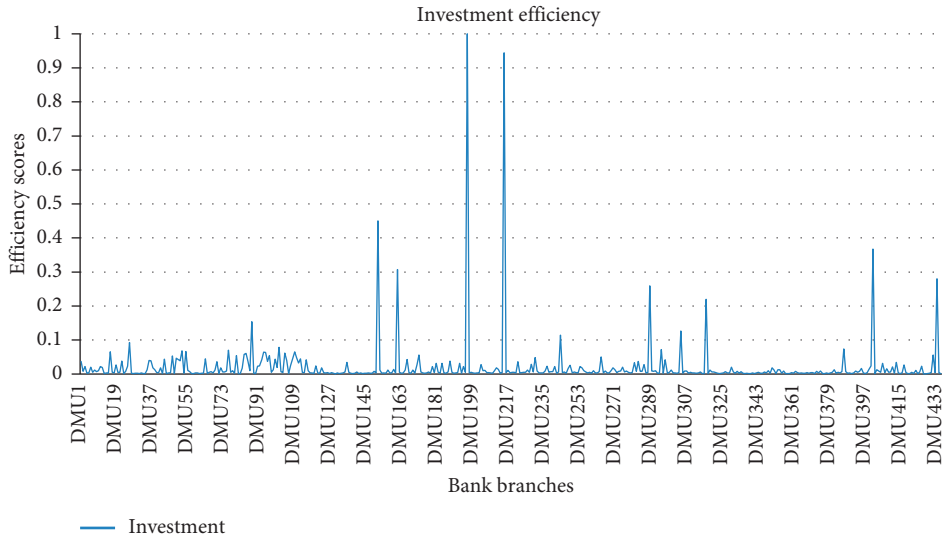


Figure 4: A graph showing the investment efficiency scores of the 444 DMUs (the authors' construct).

banks while only 1 bank (DMU219) had an efficiency score of between 80% to 100%. This result suggests that close to 99.5% of Ghanaian bank branches that were considered for the study were not efficient in investing. This also confirms reports of crises that have hit the Ghanaian banking industry with issues such as an alleged manager and board of directors squandering depositors' money without investing them [63]. This has also led to about seven (7) universal banks collapsing in 2017 and 2018 [1, 2].

For overall efficiency in the entire banking operations also shown in Figure 5, 79 (17.79%) bank branches were efficient (had a 100% efficiency score) with the majority (290 representing, 65.32%) of them having an efficiency score of between 80% and 99%. 4 (0.9%) bank branches had an efficiency score of between 70 and 79%, 32 (7.21%) had an efficiency score of between 60 and 69%, and finally 39 (8.78%) branches had between 50 to 59%. In terms of overall efficiency, there was no bank branch that had less than 50% efficiency score evident in Figure 5.

This analysis means that even though most bank branches in Ghana do not experience a higher percentage of efficiency in collecting deposits and investing the deposit, they still enjoy the highest overall efficiency. The results suggest that banks in Ghana should identify ways of improving their efficiencies in both the deposit stage and investment stage and should not only rely on their overall efficiency scores as a means of measuring their performance and success.

*5.2. Machine Learning Algorithm Results and Discussion.* In this study, machine learning algorithms were used in order to identify the best performing classification models. Three types of machine learning algorithms were employed: decision tree, random forest, and artificial neural network. To determine how accurate models were with real world data, we held back a subset of the dataset for testing purposes. Thus, the data set was split into training and

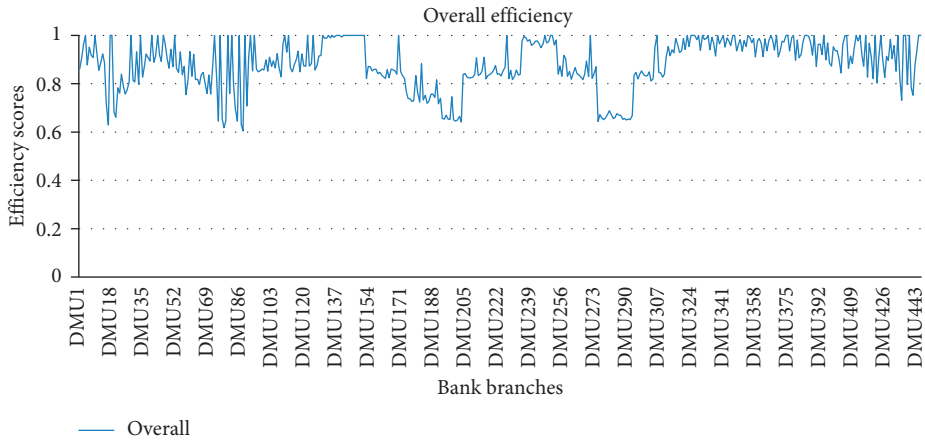


Figure 5: A graph showing the overall efficiency scores of the 444 DMUs (the authors' construct).

validation (70%) and 30% for testing. For the performance analysis, the test dataset was used for assessment.

*5.2.1. Decision Tree Prediction Analysis.* With the 134 banks (30%) that were used as test dataset, the decision tree model predicted all of them correct (100% accuracy) with a kappa value of 1 and $P$ value of 1.1e-11 which shows how significant the model was. The confusion matrix and detailed statistics of the prediction are shown as follows:

```
Confusion Matrix and Statistics
Reference
Prediction Class A Class B
Class A 111 0
Class B 0 23
Accuracy: 1
95% CI: (0.9728, 1)
No Information Rate: 0.8284
Pvalue [Acc > NIR]: 1.1e − 11
Kappa: 1
Mcnemar s TestPvalue: NA
Sensitivity: 1.0000
Specificity: 1.0000
Pos Pred Value: 1.0000
Neg Pred Value: 1.0000
Prevalence: 0.8284
Detection Rate: 0.8284
Detection Prevalence: 0.8284
Balanced Accuracy: 1.0000
 Positive  Class: Class A
```

*5.2.2. Random Forest Prediction Analysis.* With respect to the random forest predictions using the randomly selected 134 (30%) dataset, the random forest predicted 132 (98.5% accuracy) out of 134 correct with a kappa value of 0.95 and $P$ value of 0.00. The confusion matrix and detailed statistics of the prediction are shown as follows:

```
Type of random forest: classification
Number of trees: 750
No. of variables tried at each split: 6
OOB estimate of error rate: 2.26%
Confusion Matrix and Statistics
Reference
Prediction Class A Class B
Class A 108 2
Class B 0 24
Accuracy: 0.9850746
95% CI: (0.9471253, 0.9981873)
No Information Rate: 0.8059701
Pvalue [Acc > NIR]: 0.0000000001537697
Kappa: 0.9508437
Mcnemar s TestPvalue: 0.4795001
Sensitivity: 1.0000000
Specificity: 0.9230769
Pos Pred Value: 0.9818182
Neg Pred Value: 1.0000000
Prevalence: 0.8059701
Detection Rate: 0.8059701
Detection Prevalence: 0.8208955
Balanced Accuracy: 0.9615385
 Positive  Class: Class A
```

*5.2.3. Neural Network Prediction Analysis.* The neural network model using the 134 test dataset also predicted 116 (86.6% accuracy) banks efficiency classes correct, but with a very low kappa value of −0.014 and poor $P$ value 0.66 as compared to the other two models:

```
Confusion Matrix and Statistics
```

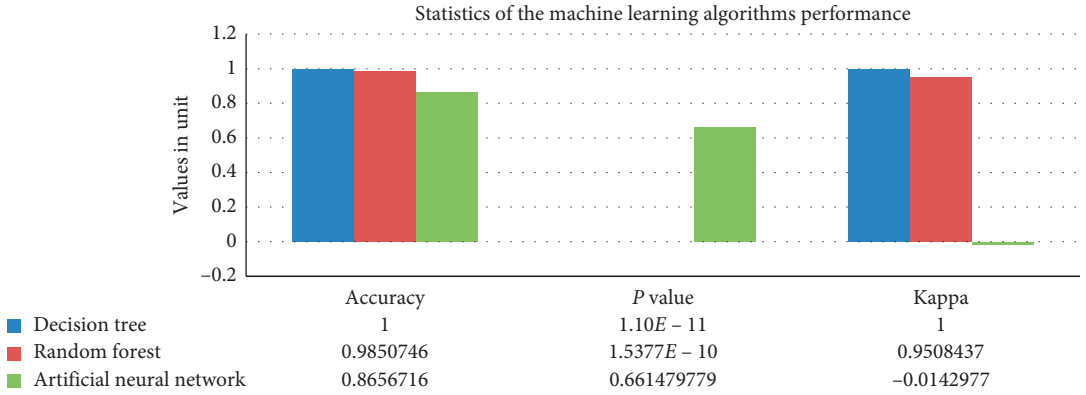


Figure 6: The graph shows the performance of the three models (the authors' construct).

```
Reference
Prediction 1 2
1 116 17
2 1 0
Accuracy: 0.8656716
95% CI: (0.796034, 0.9184046)
No Information Rate: 0.8731343
Pvalue [Acc > NIR]: 0.661479779
Kappa: −0.0142977
Mcnemar s TestPvalue: 0.000406952
Sensitivity: 0.9914530
Specificity: 0.0000000
Pos Pred Value: 0.8721805
Neg Pred Value: 0.0000000
Prevalence: 0.8731343
Detection Rate: 0.8656716
Detection Prevalence: 0.9925373
Balanced Accuracy: 0.4957265
 Positive  Class: 1
```

*5.3. Comparative Analysis of the Machine Learning Algorithms.* To estimate the three machine learning algorithm models used in the study, performance, the overall accuracy, Kappa, sensitivity, specificity, and the level of significance using their $P$ values were considered as the evaluating measures. 10-fold cross-validation (CV) was applied to check overfitting and performance of all predicting models. The mean values of the10-fold CV for each measure are given in Figure 6 and the range of these values from all predicting models is also given.

For the test dataset, the DT model performed better than the other two models, but the difference in measures between DT and RF was very small. However, NN had the lowest accuracy. After the analysis of the results of the three algorithms, the following were suggested by the study:

For predicting the overall efficiency and performance (where DEA score of 0.8–1 is classified as an efficient bank) of banks, the DT was the best. It gave the highest accuracy (100%) in predicting the overall efficiency of each bank with a kappa value of 1 and $P$ value of $1.1e-11$.

This was followed by RF with 98.5% accuracy with a kappa value of 0.95 and $P$ value of 0.00.

The last algorithm in terms of prediction accuracy was the NN which had an accuracy of 86.6% with a very low kappa value of −0.014 and poor $P$ value 0.66 as compared to the other two models.

## 6. Conclusion

In this study, the authors combined DEA with three machine learning algorithms for analysis and predicted the efficiency of bank branches in Ghana. The DEA and its Two-Phase BuildHull algorithm were implemented in R studio using R codes to assess the efficiency of the 444 bank branches at the deposit stage and investment stage. The overall stage efficiency of each bank was also calculated and categorized as efficient (Class A) or inefficient (Class B) using the adopted "cutoff point" of 0.8 units or 80%. This efficiency class designated by the CCR DEA was used as the response variable.

For the predictive models, we utilized three popular machine learning algorithms and compared them to each other using several performance metrics. Four hundred and forty-four (444) commercial bank branches in Ghana were involved in this study were 70% banks branches dataset were randomly selected to train and validate each of the three models. The proposed models were used to predict the efficiency of the remaining 30% bank branches. The best performed machine learning algorithm models (in terms of several performance measures) were determined using a holdout sample data set.

The results suggested that the decision tree and its C5.0 algorithm predicted all the 134 holdout sample dataset (30% banks). Thus the DT had an accuracy of 100% with a Kappa value of 1 and $P$ value of 0.00 which shows how significant the DT model was. The next best performing predictive model was the random forest algorithm with a predictive accuracy of 98.5% with a kappa value of 0.95 and $P$ value of

0.00. Finally, the random forest algorithm predictive model was followed by the neural network model, which also predicted 116 (86.6% accuracy) out of 134 banks efficiency classes correct, but with a very low kappa value of −0.014 and poor $P$ value 0.66 as compared to the other two models.

Overall, these results of the study may have important implications for Ghanaian banks. In this analysis, we determined the efficiency of each bank at stage I (deposit efficiency), stage II (investment efficiency), and finally, the overall efficiency of each bank. According to our analysis and findings, most banks (369 representing 83.1%) in Ghana were efficient in their overall banking operations using the "cutoff point." Even though a lot of these banks were efficient in their overall banking operations, their efficiency in collecting deposits (47 banks representing 10.59%) and especially investing the deposit (only 2 banks representing 0.45%) was poor. The study, therefore, suggests to bank managers and other stakeholders in Ghana to take a second look at their efficiency and performance in collecting deposits and investing the deposit. This means that managers and other stakeholders should not only depend or over-rely on their individual overall efficiency. The study concluded that banks in Ghana can use the result of this study to predict their respective efficiencies. Thus, the use of the decision tree predictive model as it was the best performing predictive model. Future studies can look at combining DEA with other topmost machine learning algorithms to predict the efficiency of the banks and the results compared with this study. Other factors that can also impact on banks' efficiency and performance, such as liquidity ratio, can also be taken into consideration as predictor variables in future studies.

## Data Availability

The data for the study were obtained from various banks in Ghana using their annual financial statements.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] Y. E. Addison, "Banking crisis: BoG's roadmap for clearing UT, capital bank mess," 2018, https://citinewsroom.com/2018/08/14/banking-crisis-bogs-roadmap-for-clearing-utcapital-bank-mess/.

[2] Y. E. Addison, "BoG collapses 5 banks into consolidated bank Ghana Ltd.," 2018, https://www.ghanaweb.com/GhanaHomePage/NewsArchive/BoG-collapses-5-banks-into-Consolidated-Bank-Ghana-Ltd-673691.

[3] Y. E. Addison, "19 banks meet new capital requirement," 2018, https://www.graphic.com.gh/business/business-news/19-banks-meetnew-capital-requirement.html.

[4] P. Appiahene and Y. A. W. M. Missah, "Predicting the operational efficiency of banks in the presence of information technology investment using artificial neural network," in *Proceedings of the Academics World 132nd International Conference*, pp. 6–11, Florence Italy, May 2019.

[5] L. Martin, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: a literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[6] N. Hamid, N. A. Ramli, and S. A. S. Hussin, "Efficiency measurement of the banking sector in the presence of non-performing loan," *AIP Conference Proceedings*, vol. 1795, 2017.

[7] F. Sufian, F. Kamarudin, and A. m. Nassir, "Determinants of efficiency in the Malaysian banking sector: does bank origins matter?" *Intellectual Economics*, vol. 10, no. 1, pp. 38–54, 2016.

[8] Y. Ascarya Diana, "Comparing the efficiency of islamic banks in Malaysia and Indonesia," in *Proceedings of the Pada International Conference on Islamic Banking & Fianace*, (IICiBF): Research and Development: The Bridges between Ideals and Realities, Kuala Lumpur, Malaysia, September 2007.

[9] S. A. H. Havidz and C. Setiawan, "bank efficiency and non-performing financing (NPF) in the Indonesian islamic banks," *Asian Journal of Economic Modelling*, vol. 3, no. 3, pp. 61–79, 2015.

[10] D. Ascarya Yumanita, N. A. Achsani, and G. S. Rokhimah, "Measuring the efficiency of islamic bank in Indonesia and Malaysia using parametric and nonparametric approach," in *Proceedings of the 3rd International Conference on Islamic Banking and Finance*, Jakarta, Indonesia, February 2010.

[11] E. Grmanová and E. Ivanová, "Efficiency of banks in Slovakia: measuring by DEA models," *Journal of International Studies*, vol. 11, no. 1, pp. 257–272, 2018.

[12] I. Jemric and B. Vujcic, "Efficiency of banks in Croatia: a DEA approach," *Comparative Economic Studies*, vol. 44, no. 2-3, pp. 169–193, 2002.

[13] G. E. Halkos and D. S. Salamouris, "Efficiency measurement of the Greek commercial banks with the use of financial ratios: a data envelopment analysis approach," *Management Accounting Research*, vol. 15, no. 2, pp. 201–224, 2004.

[14] J. Titko, J. Stankevičienė, and N. Lāce, "Measuring bank efficiency: DEA application," *Technological and Economic Development of Economy*, vol. 20, no. 4, pp. 739–757, 2014.

[15] D. Shibu and I. C. Ayekpam, "A study on efficiency on Assam gramin vikash bank branches," *Indian Journal of Applied Research*, vol. 7, no. 5, pp. 115–118, 2018.

[16] E. Aggelopoulos and A. Georgopoulos, "Bank branch efficiency under environmental change: a bootstrap DEA on monthly profit and loss accounting statements of greek retail branches," *European Journal of Operational Research*, vol. 261, no. 3, 2017.

[17] A. E. LaPlante and J. C. Paradi, "Evaluation of bank branch growth potential using data envelopment analysis," *Omega*, vol. 52, pp. 23–41, 2014.

[18] T. Emmanuel, "Data envelopment analysis and its use in banking," *Interfaces*, vol. 29, no. 3, pp. 1–13, 1999.

[19] C. H. Wang, R. D. Gopal, and S. Zionts, "Use of data envelopment analysis in assessing Information technology impact on firm performance," *Annals of Operations Research*, vol. 73, pp. 191–213, 1997.

[20] Y. Chen and J. Zhu, "Measuring information technology's indirect impact on firm performance," *Information Technology and Management*, vol. 5, no. 1/2, pp. 9–22, 2004.

[21] Y. Chen, L. Liang, F. Yang, and J. Zhu, "Evaluation of information technology investment: a data envelopment analysis approach," *Computers & Operations Research*, vol. 33, no. 5, pp. 1368–1379, 2006.

[22] X. Cao and F. Yang, "Measuring the performance of Internet companies using a two-stage data envelopment analysis model," *Enterprise Information Systems*, vol. 5, no. 2, pp. 37–41, 2011.

[23] C. M. L. Paço and J. M. C. Pèrez, "Assessing the impact of information and communication technologies on the Portuguese hotel sector: an exploratory analysis with data envelopment analysis," *Tourism & Management Studies*, vol. 11, no. 1, pp. 35–43, 2015.

[24] D. Dash, Z. Yang, and L. Liang, "Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank," *Expert Systems with Applications*, vol. 31, no. 1, pp. 108–115, 2006.

[25] D. Wu, "Detecting information technology impact on firm performance using DEA and decision tree," *International Journal of Information Technology and Management*, vol. 5, no. 2-3, pp. 162–174, 2006.

[26] E. Awoin, P. Appiahene, F. Gyasi, and A. Sabtiwu, "Predicting the performance of rural banks in Ghana using machine learning approach," *Advances in Fuzzy Systems*, vol. 2020, Article ID 8028019, 7 pages, 2020.

[27] Z. Chen, R. Matousek, and P. Wanke, "Chinese bank efficiency during the global financial crisis: a combined approach using satisficing DEA and Support Vector Machines," *The North American Journal of Economics and Finance*, vol. 43, 2017.

[28] S. Shrivastava, P. M. Jeyanthi, and S. Singh, "Failure prediction of Indian Banks using SMOTE , Lasso regression , bagging and boosting Failure prediction of Indian Banks using SMOTE , Lasso regression , bagging and boosting," *Cogent Economics and Finance*, vol. 8, no. 1, 2020.

[29] G. N. Zhang and F. Ye, *Predicting Financial Distress in Norway Using Logistic Regression and Random Forest Models*, Norwegian School of Economics, Bergen, Norway, 2019.

[30] H. H. Le and J. Viviani, "Predicting bank failure: an improvement by implementing machine learning approach on classical financial ratios," *Research in International Business and Finance*, vol. 44, no. 16, 2017.

[31] R. D. Banker, A. Charnes, and W. Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Manage. Sci.*vol. 30, no. 9, 1984.

[32] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.

[33] P. Appiahene, Y. M. Missah, and U. Najim, "Evaluation of information technology impact on bank' s performance: the Ghanaian experience," *International Journal of Engineering Business Management*, vol. 11, pp. 1–10, 2019.

[34] J. Hu, Z. Yang, Q. Wang, and S. Yang, "A hybrid modified DEA efficient evaluation method in electric power enterprises," in *Proceedings of the 2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS)*, pp. 283–287, Jinzhou, China, August 2016.

[35] B. Hssina, A. Merbouha, and B. Bouikhalene, "Predicting learners' performa nce in an E-learning platform based on decision tree analysis," in *Proceedings of the International Arab Conference on Information Technology (ACIT'2016)*, pp. 1–5, Beni-Mellal, Morocco, 2016.

[36] R. Pandya and J. Pandya, "C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning," *International Journal of Computer Applications*, vol. 117, no. 16, pp. 18–21, 2015.

[37] M. Kuhn, S. Weston, N. Coulter, and M. Culp, *C5. 0: C5. 0 Decision Trees and Rule-Based Models*, CRAN, Germany, 2018, https://cloud.r-project.org/package=C50.

[38] D.-J. CHi, C.-C. Yeh, and M.-C. Lai, "A hybrid approach of dea, rough set theory and random forests for credit rating," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 8, pp. 4885–4897, 2011.

[39] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Machine Learning*, Springer, Berlin, Germany, 2011.

[40] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[41] A. Cutler, *Random Forests for Regression and Classification*, Ovronnaz, Utah State University, Salt Lake City, UT, USA, 2010.

[42] L. Breiman, "Random forests," in *Machine Learning*, pp. 5–32, Springer, Berlin, Germany, 2001.

[43] S. R. ColorBrewer and A. Liaw, *Package Random Forest*, University of California, Berkeley, CA, USA, 2018.

[44] D. Tang, "Random Forest," pp. 1–9, 2016.

[45] F. Livingston, *Implementation of Breiman' S Random Forest Machine Learning Algorithm*, Springer, Berlin, Germany, 2005.

[46] M. Lam, "neural network techniques for financial performance prediction: integrating fundamental and technical analysis," *Decision Support Systems*, vol. 37, no. 4, pp. 567–581, 2004.

[47] P. Appiahene and Y. M. Missah, "Predicting the operational efficiency of banks in the presence of information technology investment," in *Proceedings of the International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pp. 6–11, Zakopane, Poland, June 2019.

[48] M. H. Al Shamisi, A. H. Assi, and H. A. N. Hejase, "Using MATLAB to develop artificial neural network models for predicting global solar radiation in Al ain city – uae," *Engineering Education and Research Using MATLAB*, IntechOpen, London, UK, 2011.

[49] G. K. F. T. Ã and K. K. W. Yau, "Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, pp. 1761–1768, 2007.

[50] S. Fritsch, F. Guenther, and M. F. Guenther, *Package "Neuralnet"*, The Comprehensive R Archive Network, CRAN, Germany, 2016, https://github.com/bips-hb/neuralnet.

[51] S. Tamura and M. Tateishi, "Capabilities of a four-layered feedforward neural network: four layers versus three," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 251–255, 1997.

[52] T. Vujičić, T. Matijevi, and Š. Zoran, "Comparative analysis of methods for determining number of hidden neurons in artificial neural network," in *Proceedings of the Conference on Information and Intelligent Systems*, pp. 219–223, London, UK, 2016.

[53] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *Int. J. Comput. Trends Technol.*vol. 48, no. 3, 2017.

[54] Y. E. Addison, "BoG enforces banks' capital requirement directive," 2018, https://citinewsroom.com/2018/07/02/bog-enforces-bankscapital-requirement-directive/.

[55] Y. E. Addison, "Minimum capital requirement for banks pegged at ¢400 million," 2017, https://www.myjoyonline.com/business/2017/September-8th/minimum-capitalrequirement-for-banks-to-reach-400-million.php.

[56] K. Greenidge and T. Grosvenor, "Forcasting non-performing loans in Barbados," *Journal of Business, Finance and Economics in Emerging*, vol. 5, 2010.

[57] R. Attah, *Ghanaian Bank Performance and Ownership, Size, Risk, and Efficiency*, Walden University, Minneapolis, MN, USA, 2017.

[58] N. Zelenyuk and V. Zelenyuk, *Drivers of Efciency in Banking: Importance of Model Specifications*, School of Economics, University of Queensland, Brisbane, Australia, 2015.

[59] J. Simm, G. Besstremyannaya, and M. J. Simm, *Package "rDEA"*, CRAN, Russia, 2016, https://github.com/jaak-s/rDEA.

[60] M.-C. Tsai, S.-P. Lin, C.-C. Cheng, and Y.-P. Lin, "The consumer loan default predicting model - an application of DEA-DA and neural network," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11682–11690, 2009.

[61] A. M. El-habil, "An application on multinomial logistic regression model," *Pakistan Journal of Statistics and Operation Research*, vol. 8, no. 2, pp. 271–291, 2012.

[62] S. Mehrabiana, "Using non-archimedean DEA models for classification of DMUs: a new algorithm," *International Journal of Data Envelopment Analysis*, vol. 1, no. 4, pp. 247–257, 2013.

[63] I. Abubakar, *Shareholders, Directors of Defunct UT, Capital Bank Engaged in "willful Deceit"-BoG*, 2018, https://www.myjoyonline.com/business/2018/August-7th/shareholders-directors-ofdefunct-ut-capital-bank-engaged-in-willful-deceit-bog.php.