Journal of Engineering Science and Technology
Vol. 17, No. 2 (2022) 0854 - 0873

SECURED LOAN PREDICTION SYSTEM USING ARTIFICIAL NEURAL NETWORK

MARION O. ADEBIYI[1], OLUWASEMILORE O. ADEOYE[2], ROSELINE O. OGUNDOKUN[3,]*, JULIUS O. OKESOLA[4], AYODELE A. ADEBIYI[5]

[1, 3, 5]Department of Computer Science, Landmark University Omu Aran, Nigeria
[4]Department of Computer Science, First Technical University, Nigeria
[1, 2, 5]Department of Computer and Information Sciences, Covenant University Ota, Nigeria
*Corresponding Author: ogundokun.roseline@lmu.edu.ng

Abstract

Loan approval is an essential factor that decides the loss or gains a financial institution would accrue at the end of a fiscal year. Banks are looking for ways to ensure that these loans are paid back within the specified period. Therefore, this study aims to develop a loan prediction system using Artificial Neural Network that will determine whether a loan is a good or bad one and whether a loan is a payable debt or bad debt. The system can also assist to predict whether a loan applicant would default in repayment or not. The study used an Artificial Neural Network algorithm to develop a loan prediction scheme. The system was designed and implemented using Python as the programming language, Hypertext Mark-Up Language (HTML), Cascading Style sheet (CSS) for the front end, and then PHP as the backend. The system also used the confusion matrix as the performance metrics to evaluate the system accuracy. The result shows that the system has 92% accuracy which showed that the developed system predicted well and can predict whether a loan applicant would default in repayment or not. The system can also predict whether a loan is a bad debtor payment one. The system was finally compared with other previous researches using the accuracy of the system and it was concluded that the proposed system performed better than the previous researches.

Keywords: Artificial neural network, Banking system, Confusion matrix, Loan approval, Loan prediction.

1. Introduction

Lending out cash is one of the critical responsibilities of a bank. Nowadays, there are many risks associated with bank lending currency, as a monetary institution, one of its duties is to minimize menaces from lending [1]. Of late, banks have progressively employed credit recording practices to assess the loan requests they got from clients. Credit counting is a way of mitigating menace by depicting the customer’s creditworthiness. The credit counting model is usually established by fiscal institutions and financial experts to help in the evaluation process of loan collection to categorize customers' loans into bad credit or good credit. Bad credit is one in which repayment of the loan at a due time is not always possible while good credit is that which is paid in due time [2]. As the banking processes continued to evolve and more people started banking, we saw so many drawbacks in the credit scoring model such as insufficient economic and manpower which is due to the ever-increasing population [2].

The accumulation of bad debt caused by bad credit has resulted in the bankruptcy of banks. This is because the borrower refuses to return money that was meant to serve the interests of depositors and meet their expenses. The depositors, in turn, start to withdraw their money from the bank leading to a shortage of cash in the bank and thereby resulting in a loss. If a similar situation happens in all banks, the entire economy would be affected. Key employees such as the manager are unfairly sacked due to the non-payment of loans [2]. The omen is usually on the manager whether to approve loans or not. Since the manager can’t properly predict whether the loan would be payable or not, he or she bears the brunt for their decision. Banks, therefore, tend to lay off staff to reduce the financial burden. Theorems can be formulated. With the knowledge or data accumulated we can be able to identify and analyse trends such as what age group hardly pays back loans. This information may be useful in formulating a hypothesis that can eventually lead to a theorem [3].

Banking Industry has accumulated a lot of data mostly from customer’s transactions, business financial statements, and payment records, etc. This data when effectively used and analysed can help to gain competitive advantage and to predict customer’s creditworthiness thereby avoiding risky transactions [3]. Data mining can be employed to perform a very key role to attain these efforts [3, 4]. Data mining is an uninterrupted procedure that syndicates business knowledge, machine learning approaches, and devices and great volumes of correct and important information to ensure the finding of non-intuitive perceptions concealed in the establishment’s company information. The information gotten from the trends can help an organization to formulate new strategies that would help the organization's relationship with its consumers and staff. Data mining techniques are classified into two main categories such as statistical methods and artificial intelligence [3, 4].

A statistical method, the most commonly used technique for loan prediction is the logistic regression and discriminant analysis. A logical regression model explains that the right function of the fitted likelihood of the event is a linear role of the perceived variables of the available explanatory values [3, 5]. Discriminant analysis is also known as fisher’s rule that deals with the linear blending of descriptive variables that distinguish superlative between a priori defined groups [3]. To accomplish this, one would have to make the best use of the between-group modification that is similar to the within collection variance. One of the advantages of the statistical methods is that they are not difficult to execute and are also capable of producing outcomes that can easily be interpreted [1]. Artificial intelligence methods include K-nearest neighbor [6, 7], Decision Trees, Neural Networks, Naïve Bayesian classifier, Genetic programming, and Support vector machine models. AI technique is mostly used whereas the dependent and independent variables reveal multifaceted non-linear relationship [4, 5]. The importance of the study cannot be overemphasized as the system developed will assist in rapid loan disbursement to customers. The prediction of loans makes the loan application process faster. Where organizations don't have to waste time analysing a loan applicant information and can save time.

A financial organization can gain a competitive advantage from the vast amount of knowledge and datasets gathered. These datasets can be evaluated and effectively utilized to help the bank create a unique service that cannot be easily copied. Risk can also be greatly minimized [8]. Risk happens where it is not known what the future result will be however where the many possible results may be expected with some level of certainty from the knowledge of past or existing events. Risk in bank loans includes credit risk, liquidity risk which is the risk that the customers may withdraw too much money at once and interest rate risk when the customers pay back the loan, the interest rate would not be sufficient to earn back the bank money [8].

Therefore, this study aims to develop a loan prediction system using Artificial Neural Network that will determine whether a loan is a good or bad one and whether a loan is a payable debt or bad debt. The system can also assist to predict whether a loan applicant would default in repayment or not.

The remaining part of this paper is structured as follows: Section 2 discusses the previous related research. Section 3 discussed the material and methods used in this study. The results and discussion about the findings discovered in this study were discussed in section 4 while section 5 concluded the paper with concluding remarks and possible future works.

2. Literature Review

This is the study of the computation that makes it possible to perceive, reason, and act. This means making computers comprehend human intelligence [9]. A machine is termed "intelligent" if it can do things that humans can do without being able to differentiate whether it's a human who did it or not. A good AI program always solves a real-life problem and also produces fresh opportunities. AI is used for many purposes from helping experts in various fields solve complex problems to helping providing answers to English questions. At the start of AI most of the applications were designed to show how practical AI is to the world but these days AI is being used by most business organizations to achieve strategic objectives. An intelligent software agent is one that has a reactive, proactive, and intelligent behavior. There are many areas of specialization of artificial intelligence which are game-playing, natural language processing (NLP), Expert systems, neural networks, and robotics [10, 11]. The neural network is usually designed for special programs such as data classification and pattern recognition [9].

A decision support system (DSS) is a computerized information system intended to enable information specialists to choose one of the numerous alternative solutions to a problem. A decision support system (DSS) is very important for the success of any business. DSS can also be seen as a responsive and flexible graphical interface software system that helps decision-makers mostly people in a position of authority in an organization produce essential facts and information which is derived from raw data, journals, files, user inputs, knowledge stored in the knowledge base to spot a reoccurring trend and pattern to make decisions which would solve problems [12]. Decision support is for the most part utilized as a part of business management and planning procedure, medical service, military, and fundamentally any area in which management will experience complex decision situations [13].

Expert System: is a computer software program that comprises the knowledge and expertise of a human professional in a specific problem field [14]. ESs is a sub-domain of artificial intelligence research. Artificial Intelligence focuses on models that try to emulate how the human brain solves the problem and acquire knowledge. The driving force behind the expert system is to create a system that uses the undocumented knowledge of experts efficiently that even unskilled persons can make use of it. The expert system is being broken down into 2 main subsets that are, the inference engine and the knowledge base. For the Inference engine, the system combines the input data given by the user together with the data relationships. The inference engine thereafter establishes facts from these data and applies rules on it to infer new facts. The inference engine also performs explanation and debugging capabilities. A knowledgebase on the other hand involves the gathering of facts and establishing the connections which exist among them. A knowledgebase is built using a series of IF-THEN rules.

Data mining is a very important process in predictive analytics. Data mining means extracting knowledge or details from a very enormous number of data [15]. Data mining is mostly associated with knowledge discovery in the database (KDD) although data mining is contained in KDD. Data mining functionalities aim at identifying the different types of patterns that are seen in data mining tasks. Data mining tasks are in two categories [15].

i. Descriptive data mining defines and explains the whole properties or features of the data.

ii. Predictive data mining: it tries to make inferences on the data and also performs prediction.

Data mining is one of the sections of the method identified as Knowledge Discovery in Databases (KDD). Knowledge discovery in databases deals with the comprehension of important and fundamental patterns in datasets, thereby indicating to the entire process of intellectual exploration from datasets Stevens [16]. Data mining techniques are employed to obtain models from datasets and this data signifies a collection of details. Data mining techniques acquire from datasets, that is, they acquire to envisage the vital outcome from the certain input [16]. This kind of acquiring knowledge doesn't affect the compelling of the data to the workstations' memory, but it develops to modify its perfect to accomplish improved in the future [17].

Numerous studies had been examined in the field of loan prediction using ANN and some other algorithms. Several journals were reviewed and findings from the various journals assisted in the development of the loan prediction system, and they were discussed below.

Sudhakar and Reddy [8] conducted research using the decision tree technique to access the credit risk of a retail bank. This research was conducted to assist the banking sector by giving a lot of data to the credit decision making process by minimizing money and time wasted for evaluation of loan and minimizing the level of vulnerability suffered by loan officials by assisting them with knowledge gotten from historical data of loans using a decision tree.

Islam and Habib [18] postulated research to develop a model that would predict prospective business sectors in retail banking. Records of business clients of a retail bank were used for the research including records from the rural and urban territory of Bangladesh. These records were used to break down the main transactional determinants of clients and predicting of a prototype for likely subdivisions in a retail bank. A decision tree data mining algorithm was adopted for analyzing problems used to build the model and lastly executed it to test its performance using weka.

The key goal of the paper was to build a Credit Scoring Model that would be used by the Sudanese Banks. There was 2 data mining classification that was chosen which is Decision Tree (DT), Artificial Neural Network (ANN). Then Generic Algorithm (GA) and Principal Component Analysis (PCA) were used as feature selection techniques. Sudanese credit dataset and German credit dataset were applied for the evaluation of the techniques. The output of the classification revealed that ANN outperforms DT in most situations. In comparison, it was also seen that GA is better than the PCA technique as a feature selection. The German data set produced an accuracy of 80.67% while that of Sudanese was 69.74%. In conclusion. It was seen that ANN outperformed DT and its hybrid models which are PCA-DT and GA-DT.

Hamid and Ahmed [19] suggested a novel approach for grouping credit risk in the banking sector by using data mining. The data used for this model was gotten from banks to predict the status of loans. Three algorithms were employed to develop the projected models which are j48. bayesNet and Naïve Bayesian. The application that was used for implementation was weka and thereafter tested. The outcome of the research showed that j48 was the best based on accuracy. This research looked at the predicted behavior of five classifiers of credit risk prediction accuracy for different kinds of interference and how the precision could be improved by using classifier ensembles. Thereafter basing the outcome on four credit datasets and comparison with the performance of each classifier on predictive correctness at various attribute noise levels are presented. The experimental evaluation shows that the ensemble of classifiers techniques has the potential to improve prediction accuracy.

Srivastava et al. [20] suggested a prediction model that employed the Artificial Neural Networks (ANN) procedure of Machine Learning to accomplish loan nonpayment forecasting and liken it with the Logistic regression procedure. The authors prepared their archetypal on pre-documented data to estimate the of the debtor and they made effort to yield the greatest likely results.

Aslam et al. [21] projected a method that employed Vector Machines concerning the Loan Nonpayment Forecasting. When the study was Likened to other classifiers, it was deduced that the support vector machine achieved better than numerous previous methods when it came to throughput and arithmetical implementation where huge data were linked with manifold descriptive variables. The precision of the projected system was 81%.

Obare et al. [22] employed a logistic regression model to assess the individual loan nonpayment in Kenya. This research utilized an arithmetical examination procedure that is it deals with nonpayment of solitary loans as a debtor's characteristics. The prototype had a train data precision of 0.7727 and test data of 0.7333. The logistic regression model with the train and test statistics showed a precision of 0.8440 and 0.8244 respectively. The biggest downside of this model was that it has a high level of false positives.

Tariq et al. [23] presented an all-inclusive investigation and established a procedure to envisage the loan nonpayment. This investigation employed procedures termed KDD, CRISP-DM, and SEMMA. Built on specific constraints, the superlative scheme was carefully chosen, elucidated, and recommended because of its momentous physiognomies concerning the estimation of the loan nonpayment in the fiscal subdivision. This scheme has a precision of 78% and it flops since its ROC score and zone weren't upright.

Kumar et al. [24] suggested the Neural Network method for the assessment of Loan Nonpayment. The author recommended a basis to merge a neural network method that was employed to conjecture the nonpayment loans. The estimation was made in terms of the fiscal and public particulars offered by the probable borrower.

In estimating the probability of nonpayment, Kwofie et al. [25] utilized data from a microfinance company to estimate the effectiveness of logistic regression. The author employed predictors, for instance, age, family status, gender, years of education, years of industry, and base capital. The prognosticators which were pertinent in the method were marital position, years of business experience including base capital. This method was operative at around 91 percent nevertheless the crucial weakness was that the variation stated in the outcome was insignificant.

Jayadev et al. [26] made effort to employ the utilization of machine learning to train the archetypal and set the classifier as LSVM and the competence measured as RMSE. The weakness was that its lessened productivity if the descriptive variables were a smaller amount than 10. He also made effort to train the model utilizing ANN. In this, the Neural Network classifier was 93 percent precise.

Khandani et al. [27] projected logistic regression as 86 percent precise. Its limitation was that owing to the linearity communication between them, its misplaced productivity in data with a huge number of variables.

Hassan and Abraham [28] make effort to train their proposed method by employing ANN. They employed an optimizer named "Adam," which had 2 impervious coatings of 20 nodes respectively and set the number of periods to 1000. The scheme was 93 percent precise, nevertheless, the method was over proficient and didn't accomplish well in the test dataset.

3. Material and Method

This section discussed how data used in this study were gathered and the method used for the designer of the proposed system.

3.1. Data collection

Requirement's elicitation was done by collecting user information from the Igboora Micro Finance Bank. The data collected was mainly derived from the past credit history of customers who had collected a loan from the bank. The data used was the Igboora microfinance bank data set containing 14 columns and 1057 rows. The variables contained in the dataset are the value of the loan, date granted, interest rate, maturity date, text66, security, security value, balance, purpose, categorization, marital, sex, age, IsPaid. For the smooth running of the application, the following libraries have to be installed and imported. They are TensorFlow, theano, and Keras.

Many factors determine whether a loan would be paid back, in the traditional approach the 5 Cs of credit scoring was used which includes:

i. Collateral: This refers to an asset o of great importance and value which when sold could cover the value of the money loaned.

ii. Credit History: This is also known as Character. The records of any individual or organization wanting to borrow a loan are checked. If the person/organization were able to pay back within the specified period such a person has a good credit history, otherwise a bad credit history.

iii. Capital: The amount that is to be collected s also a great determinant if the loan would be paid back. If an individual borrows $500,000 it is easier to pay back than someone who collects a loan of $1,000,000.

iv. Condition: The purpose for which loans are being collected differs from person to person. A person that collects a loan for personal upkeep or hospital bills is likely not to pay back the loan compared to someone who collects it for a business that would yield much greater rewards because it involves investment. The condition of the loan could also include interest rate, amount of principal, and so on.

v. Capacity: The income of the loaner is also a great determinant. Someone who has a stable job, and a good wage/salary are more likely to pay back the loan than someone who doesn’t. The length of time a person has spent on a job should also be checked.

3.2. System design

In this sub-section, the way the equipment would interact and assist in selecting the right technologies for the system was discussed. This also involves using the specific requirement of the bank (that is used as a case study in this research) in designing the elements of the system such as its modules, UMLs, components, and interfaces. Various UMLs were used in this study for instance Use cases, Activity diagram, Class diagram, and Sequence diagram.

Figure 1 shows the use case diagram for the administrator of the loan prediction system that shows the various actions performed by the administrator of the loan prediction system. Figure 2 shows the use case diagram for a manager of a loan prediction system, and this shows all the tasks performed by the manager in the system. Lastly, Fig. 3 shows the use case diagram for a system admin of a loan prediction system.

Figure 4 shows the class diagram of the loan prediction system, and this is the visual representation that shows how object classes of the proposed system interact with other classes and also the associations among the classes of the loan prediction system. This also defines all the packages, interfaces, and classes that make up the proposed system and how these components interrelate with each other.

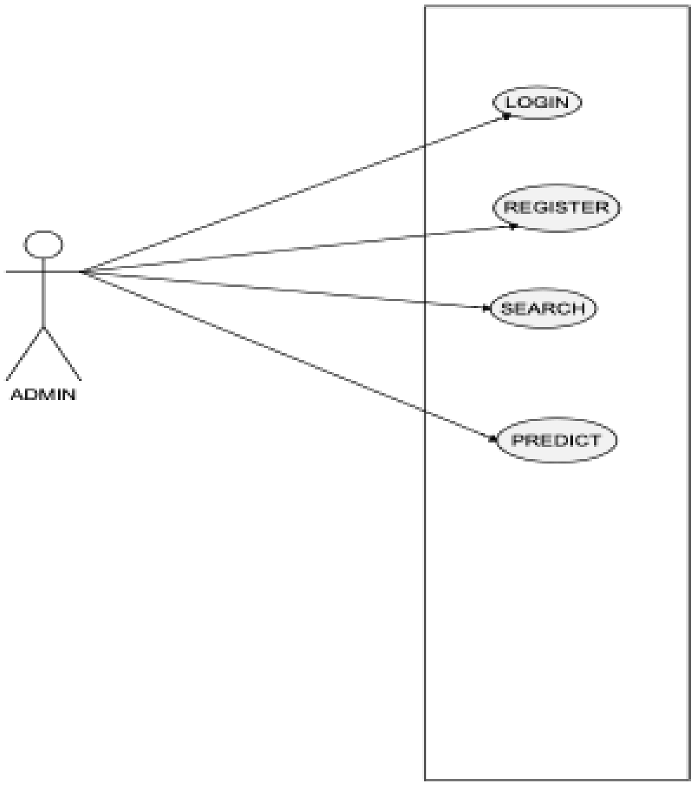

Fig. 1. A use case diagram for the admin of a loan prediction system.

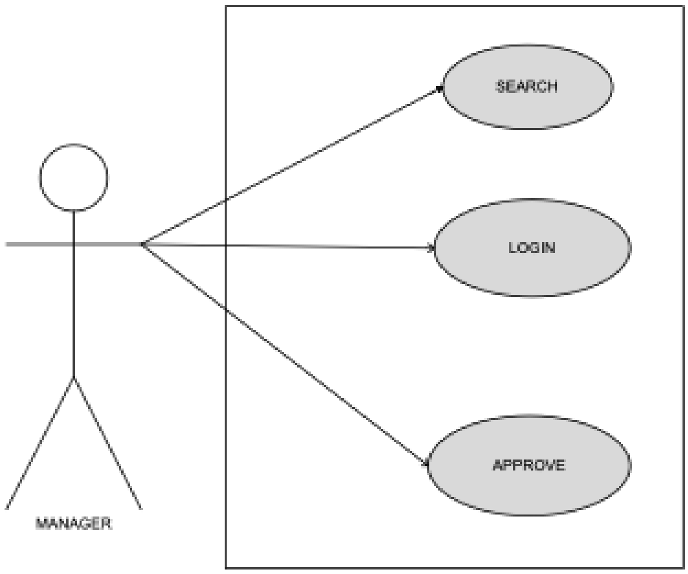

Fig. 2. A use case diagram for a manager of a loan prediction system.

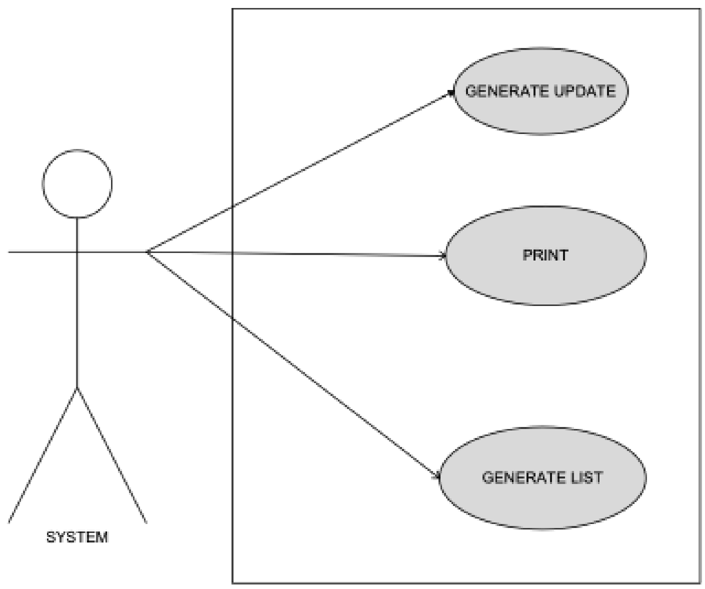

Fig. 3. A use case diagram for a system admin of a loan prediction system.

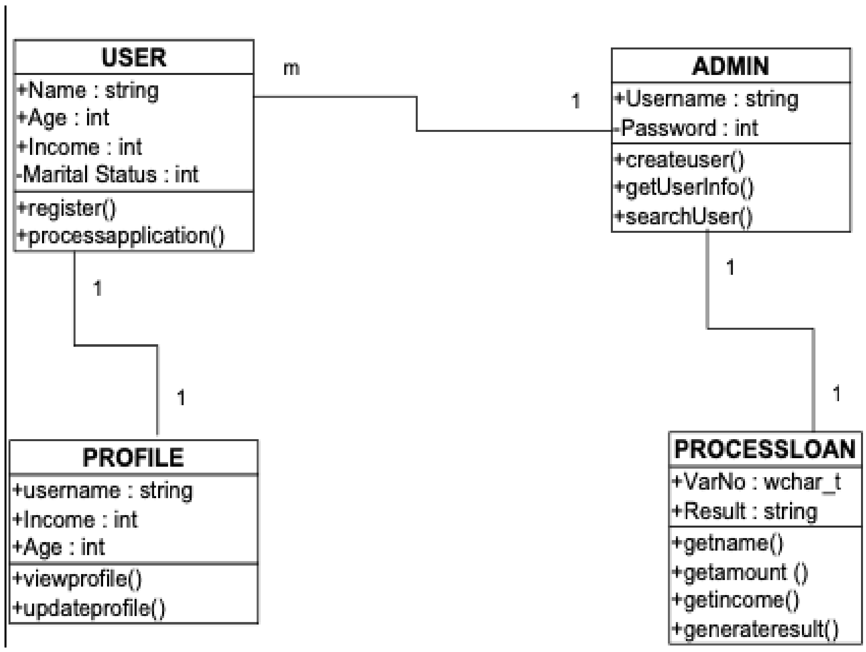

Fig. 4. Class diagram of a loan prediction system.

Figure 5 shows the sequence diagram of the loan prediction system which shows and model the various communication exchange among the actors and the objects within the loan prediction system and also the interactions between system components.

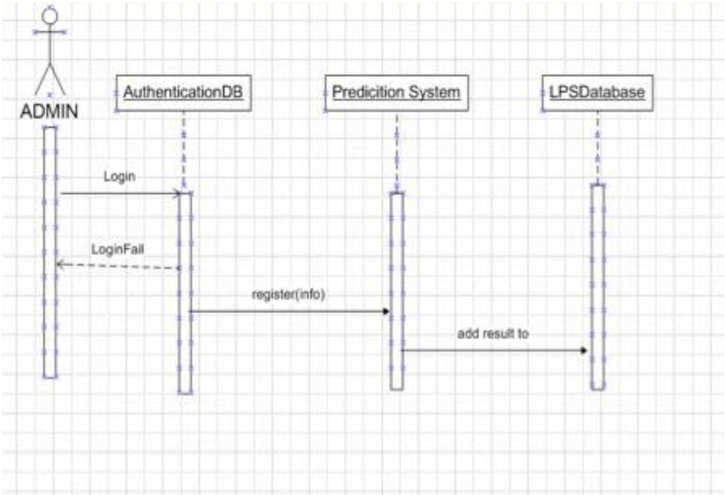

Fig. 5. Sequence diagram of the loan prediction system.

Figure 6 shows the activity diagram of the proposed system, and this is used to model the processing of data, and this shows the activities involved in the system process. This also helps to show the flow of activities in the loan prediction system and the sequence from one activity to the other.

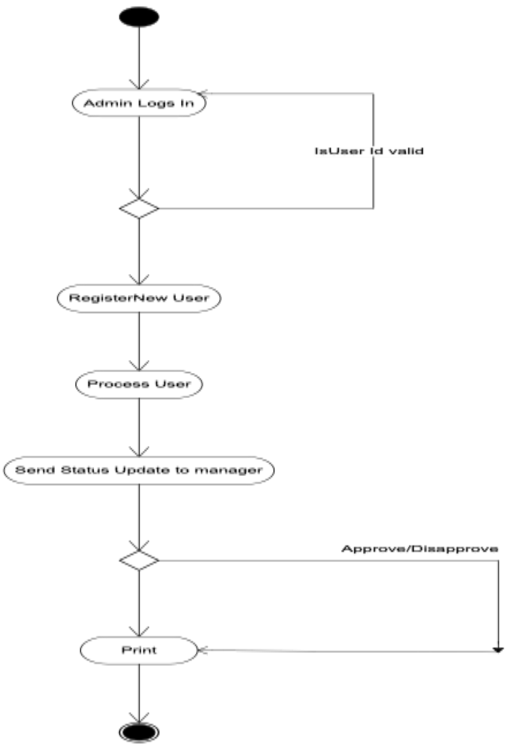

Fig. 6. Activity diagram of the loan prediction system.

4. Results and Discussion

In this session, the implementation details for the Loan Prediction System were explained and displayed. Implementation is the fourth phase of system design and implementation. The loan prediction system has been developed to ensure a flexible, easy, and interactive interface for the users. This section also shows the entire process involved in the development of our artificial neural network from scratch from data preparation up to model evaluation.

4.1. Data preparation

i. **Encoding the categorical data (converting the string to integer):** Due to the nature of the algorithm, the authors are going to transform the categorical variables in the dataset to numerical variables to be able to be processed. Encoding is done by creating a simple dummy variable of each categorical variable. We do this by assigning a unit to each category. In our dataset we have a row which contains categorical variables such as marital status (married, single, divorced, separated, widowed), units are assigned to each of the categories for married 0 can be assigned, for single 1 can be assigned, for divorced 2 can be assigned, etc. The function that does this is the LabelEncoder() function as shown in Fig. 7.

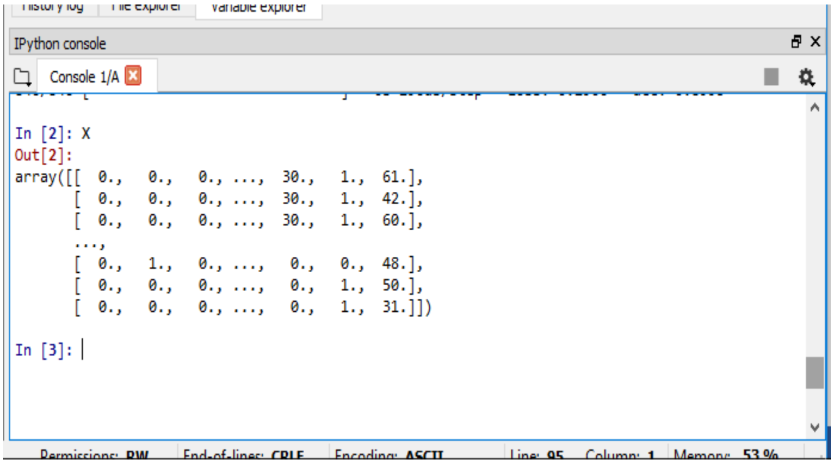

Fig. 7. Conversion of categorical variables to numerical values (source: personal code)

ii. **OneHotEncoding:** This is the next process after assigning units to the various categories. The problem with label encoding is it assumes that the higher the categorical data he more important it is. Figure 8 shows the impact of OneHotEncoder on the dataset. For example, it would assume divorced is of more importance than married because it has a higher dummy variable. This is why we need the OneHotEncoder() function which creates binary variables for each of the categories. For example, in our data set, we have a categorical set that contains 5 categories, which means 5 binary variables would be created one column representing the variables individually.

X - NumPy array

	0	1	2	3	4	5	6	7
0	0	0	0	0	1	30	1	61
1	0	0	0	0	0	30	1	42
2	0	0	0	0	1	30	1	60
3	0	0	0	0	1	30	1	37
4	0	0	0	0	1	30	1	59
5	0	1	0	0	1	30	1	41
6	0	1	0	0	1	30	1	26
7	0	0	0	0	1	0	1	40
8	0	0	1	0	0	30	0	36
9	0	0	0	0	1	30	0	63
10	0	1	0	0	1	30	0	40
11	0	0	0	0	1	30	1	52
12	0	0	0	0	0	0	0	38
13	0	1	0	0	1	30	0	45
14	0	0	0	0	0	27	1	29
15	0	0	0	0	1	30	0	54
16	0	0	0	0	1	30	1	49
17	0	0	0	0	1	30	0	30
18	1	0	0	0	1	30	0	24

Format Resize ☑ Background color

Fig. 8. Impact of OneHotEncoder on the dataset (source: personal code).

iii. **Splitting the data set into training and test data:** The data was split into two. First, the training dataset and secondly, the test data set in the ratio 1:4. The training data set was used to teach the model while the test was used to validate the result of the trained model.

iv. **Feature scaling/normalization:** Normalization helps to rescale the various numerical attribute to values between the range of 0 and 1 or **-**1 and +1 that is the maximum value being 1 and the minimum value is 0. Data normalization is needed when the data we are dealing with has a lot of varying scales this is done to eliminate the influence of one variable over the other thereby making the data set consistent with the other data used in the system (in the Igboora bank data set we have the value of the loan (2,000,000 - 20,000), age (21- 65) and interest rate (0-30). Figure 9 shows the impact of feature scaling on the dataset.

X_test - NumPy array

	0	1	2	3	4	5	6	7
0	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	-1.20212	1.07636
1	-0.174501	1.78553	-0.19503	-0.155606	0.600107	0.485506	0.831865	-0.620508
2	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	0.831865	-0.389117
3	-0.174501	1.78553	-0.19503	-0.155606	-1.66637	0.485506	0.831865	1.61627
4	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	-1.20212	-0.234856
5	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	0.831865	-0.697638
6	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	0.831865	-1.31468
7	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	0.831865	-0.466247
8	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	0.831865	-0.697638
9	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	-2.18915	-1.20212	-0.00346449
10	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	-1.20212	0.459318
11	-0.174501	1.78553	-0.19503	-0.155606	-1.66637	-2.18915	-1.20212	1.53914
12	-0.174501	-0.560058	-0.19503	-0.155606	-1.66637	0.485506	0.831865	-1.6232
13	-0.174501	-0.560058	5.12741	-0.155606	-1.66637	0.485506	-1.20212	-0.234856
14	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	-2.18915	-1.20212	-1.16042
15	-0.174501	-0.560058	-0.19503	-0.155606	-1.66637	0.485506	0.831865	-1.31468
16	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	-1.20212	0.536449
17	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	0.485506	-1.20212	1.15349
18	-0.174501	-0.560058	-0.19503	-0.155606	0.600107	-2.18915	-1.20212	1.69341

Format Resize ☑ Background color

Fig. 9. Impact of feature scaling on the dataset (source: personal code).

4.2. The system algorithm

i. **Initializing the ANN:** This process involves importing the Keras library and using it to start the algorithm as shown below.

```
import Keras
from Keras.models imports Sequential
from Keras.layers import Dense
classifier = Sequential ()
```

ii. **Adding the input layer:** The input layer is where each customer details would be entered and to do this, we would be using the following module under the dense function; units, kernel_intializer, activation, input dim. The unit's parameter specifies the no of nodes that would be going into the first hidden layer. The kernel_initializer is in charge of assigning weights to the different input nodes although weights initialization is very random at this stage and is usually very near to zero (0). The activation function is used on the summation of all the inputs nodes with the various associated weights to determine the frequencies by which the signals would be passed by the neuron. The best activation function to select is the one closest to 0 that's why the relu function was selected for the first input and hidden layer.

```
# Adding the input layer and the first hidden layer
classifier.add(Dense(units= 4, kernel_initializer ='uniform', activation = 'relu', input_dim = 8))
# Adding the second hidden layer
classifier.add(Dense(units= 4, kernel_initializer ='uniform', activation = 'relu'))
# Adding the output layer
```

iii. **Compiling the ANN:** This is where the bulk of the work is done that is the algorithm is compiled here. The optimizer is the one that helps us select the algorithm that can help us find the most optimal weights for each of the inputs nodes and we know that weights are assigned according to the level of importance of each node making the optimizer a very important parameter. The Stochastic Gradient Descent (SGD) algorithm is mostly used and has several types. The adaptive moments SGD also known as "Adam" is the algorithm that we are going to use. Thereafter we have the loss parameter which helps to minimize the loss/error made by the model. The binary_crossentropy was used because our dependent variable contained just two attributes which are 0 and 1if, not we would have used the categorical_crossentropy.

```
classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
```

iv. **Fitting the ANN:** Firstly we have to fit the training set of the independent and the dependent variable to the model. Epochs represent the number of times the steps would be repeated and choosing the epoch is purely random there is no particular rule to determine the epochs.

```
classifier.fit(X_train, y_train, batch_size = 10, epochs = 50)
```

v. **Predicting the test set results:** This is where we would predict whether the customer would pay back the loan or not and thereafter converts it to probability.

```
y_pred = classifier.predict(X_test)

y_pred = (y_pred > 0.5) y_pred = (y_pred > 0.5)
```

vi. Model Evaluation: This is where the model is evaluated, and the confusion matrix is used in this study.

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
```

Figure 10 shows the final result of the training and test data, and the accuracy of our model can be calculated with this result. This is discussed below:

The confusion matrix takes the predictable classes as arguments to get the prediction with; (TP = 170; TN = 24; FP = 11; FN = 7).

True Positive (TP): positive observation and predictions.

False Negative (FN): positive observation and negative prediction.

True Negative (TN): negative observation and negative prediction.

False Positive (FP): negative observation and positive prediction.

Fig 10 shows the result of confusion matric for the system

The accuracy is calculated as shown below;

Accuracy = TP+TN/TP+TN+FP+FN

= 170+24/212 = 0.9151

This means our model is 91.51% accurate which is approximate to 92%.

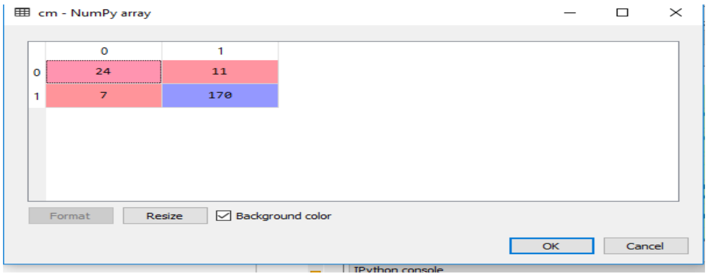

Fig. 10. Result of the Confusion matrix (source: personal code).

4.3. Implementation tools used for the system development

i. PYTHON Programming: This is a free open source, friendly and easy high-level general-purpose programming language which was used to program the neural network.

ii. SPYDER.: Spyder is embedded inside an anaconda and was chosen because it is one of the best ideas for coding with python. Spyder full name scientific python development is very good in data mining because of the scientific components/libraries which are embedded inside of it. Numpy is a library in spyder which can is useful when dealing with multi-dimensional arrays or very high computation functions. It also has Matplotlib which is useful when plots are to be embedded inside an application.

iii. SUBLIME: This was the IDE that was used to program the front end which is HTML, CSS, and BootStrap. It is very good because it allows several plugins to be embedded inside of it easily. It is a free code editor that was developed using python APIs.

iv. HTML: HyperText Markup Language is the programming language that is used for building web pages and applications. When combined with Javascript and CSS it forms a very strong technology that the world wide web can use. HTML elements and tags are what makeup HTML pages. All HTML elements are embedded inside tags and all HTML elements must have a header and a body. Lastly, all HTML documents must have this declaration (<!DOCTYPE html>) or else various browsers would revert it to quirk modes for rendering.

4.4. Program modules and interfaces

This subsection shows the implementation result interfaces of the proposed system. Figure 11 shows the home page of the system and a summary of the entire system. It also helps in the navigation of the loan prediction system developed. Here you have buttons such as About us, Contact, Sign in, and Register. When you click on each of the buttons, it navigates you to another page.

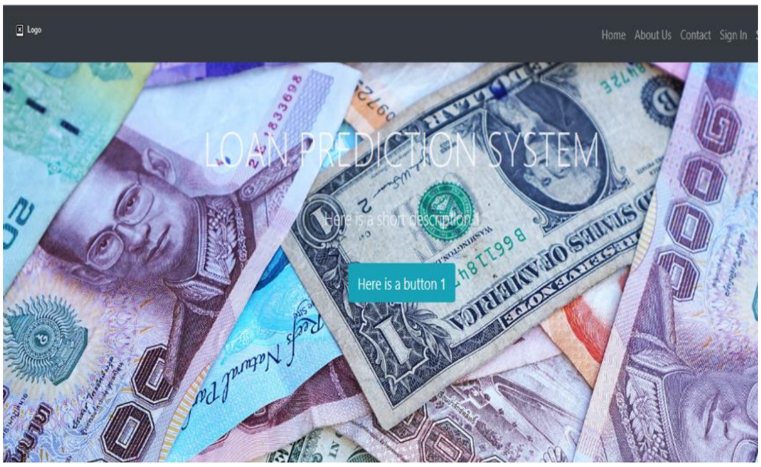

Fig. 11. Developed system home page interface.

Figure 12 shows the about us interface that is the Loan Prediction System which gives a summary of what the Loan Prediction System is all about. This interface is where the customers get to know what the bank is into that is their activities.

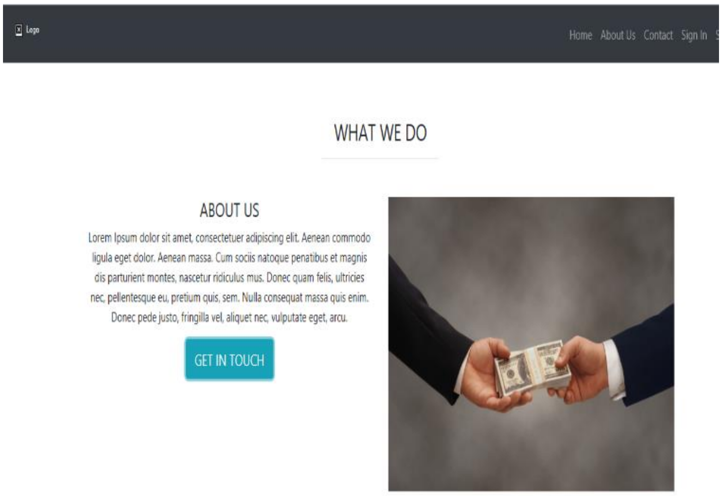

Fig. 12. The system about the Loan Prediction System (LPS) interface.

Figure 13 shows the contact section of the system which provides details on how to reach the developers of the system in case of any malfunction or update with the system. If a customer has one or two complaints or inquiries to make, this is where the contact information is acquired or gotten.

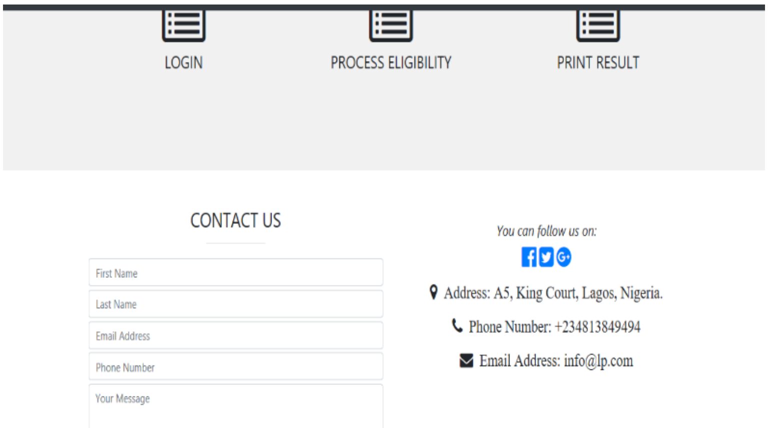

Fig. 13. The system contact details of the LPS.

Figure 14 shows the registration form for the system which is a form where customers' information would be inputted. This is the page where new users or customers come and register their details for the application of a loan from the bank. Registered users can also come back each time to log in and use the developed system over and over again.

Figure 15 shows the result page of the system. This is the page that displays the final result of the implementation, and this is where the result of the prediction is displayed as well. The prediction made after implementation here is that "there is a probability the user will pay back the loan".

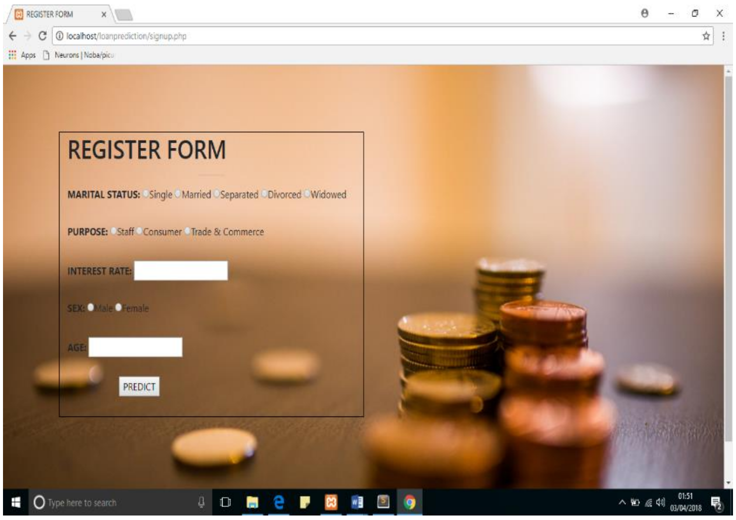

Fig. 14. The system register page.

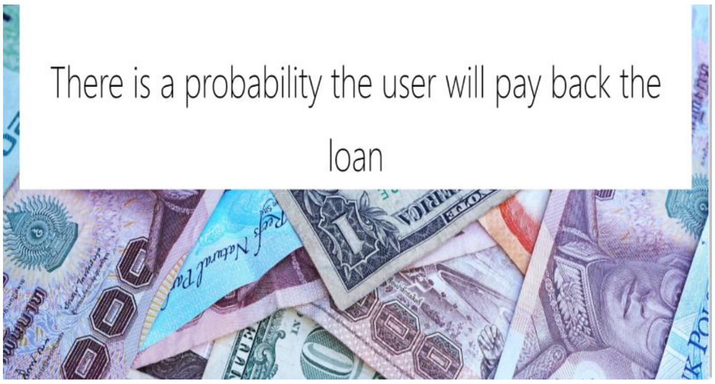

Fig. 15. Image showing the result.

4.5. Performance evaluation

Many researchers have worked in the field of loan prediction and estimation. The comparison with previous researches was shown in Table 1. Table 1 shows that the proposed system has an accuracy of 92% which makes the system an effective and accurate one for loan prediction.

Table 1. Comparative analysis with previous researches.

Previous Researches	Accuracy
Proposed System, 2020	92%
Tariq et al. [23]	78%
Obare et al. [22]	84%
Khandani et al. [27]	86%
Kwofie et al. [25]	91%
Jayadev et al. [26]	93%
Hassan and Abraham [28]	93%

5. Conclusion

The emergence of big data in the world has created avenues to utilize these data. The loan prediction system has been developed to aid decision-makers such as managers in the financial industry to make an important decision which is whether a customer requesting the loan would pay back or not. The Market sensei software was reviewed which helps to predict stock prices and alerts a customer when a stock price goes up. It was developed using an Artificial Neural Network algorithm which takes the stock as the input and also graphically shows the prediction. This system is very effective but has its limitations but only covers closed-end credit. This study covers the aspect where the human being cannot easily make decisions in the loan application process which is why this study is very important because it can determine whether a bank would go bankrupt or not.

The use of a decision support system and the expert system has helped to solve major issues in the banking industry and the world at large. This study emphasizes the use of a decision support system in helping managers make decisions about loan approval and also increasing the accuracy of the decision. This is therefore a very valuable asset to the financial sector. The result shows that the system has 92% accuracy which showed that the developed system predicted well and can predict whether a loan applicant would default in repayment or not. The system can also predict whether a loan is a bad debtor payment one. The system was finally compared with other previous researches using the accuracy of the system and it was concluded that the proposed system performed better than the previous researches.

The system still has a lot that can be added to it for improvement. The G.P.U could be changed to the NAVIDIA G.P.U for efficiency and ease of use. The system could also include a progress bar that shows the progress of the loan approved and sends feedback to the manager when payment has been completed. The system should be able to explain why the user would not be able to pay back the loan and also sends feedback to the user. It should also be available to the users on the web so they can see for themselves why their loan was not approved. Other technical issues relating to security and integrity of data should be considered.

Abbreviations

ANN	Artificial Neural Network
Cart	Classification and Regression Trees
CSS	Cascading Style Sheet
DT	Decision Tree
GA	Generic Algorithm
GPU	Graphic Processing Unit
HTML	Hypertext Mark-Up Language
KDD	Knowledge Discovery Database
KNN	K-Nearest Neighbour
LPS	Loan Prediction System
NLP	Natural Language Processing
PCA	Principal Component Analysis
UML	Universal Modelling Language

References

1. Ince, H.; and Aktan, B.A. (2009). Comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, 10(3), 233-240.
2. Lee, T.-S.; Chiu, C.-C.; Chou, Y.-C.; and Lu, C-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, 50(4), 1113-1130.
3. Yeh, I.-C.; and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of the probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480.
4. Oladele, T.O.; Ogundokun, R.O.; Kayode, A.A.; Adegun, A.A.; Adebiyi, M.O. (2019). Application of data mining algorithms for feature selection and prediction of diabetic retinopathy. In: Misra S. et al. (eds.) *Computational science and its applications* – ICCSA 2019. ICCSA 2019. Lecture Notes in Computer Science, vol 11623. Springer, Cham.
5. Omolewa, O.T.; Oladele, A.T.; Adeyinka, A.A.; and Oluwaseun, O.R. (2019). Prediction of student's academic performance using k-Means clustering and multiple linear regressions. *Journal of Engineering and Applied Sciences*, 14(22), 8254-8260.
6. Noh, Y.-K.; Park, F.C.; and Lee, D.D. (2013). K-nearest neighbor classification algorithm for multiple-choice sequential sampling. *Proceedings of the Annual Meeting of the Cognitive Science and Society* (*COGSCI* 2013), 35, 3151-3156.
7. Adebiyi, M.O.; Adigun, E.B.; Ogundokun, R.O.; Adeniyi, A.E.; Ayegba, P.; and Oladipupo, O.O. (2020). Semantics-based clustering approach for similar research area detection. *TELKOMNIKA* (*Telecommunication Computing Electronics and Control*), 18(4), 1874-1883.
8. Sudhakar, M.; and Reddy, C.V.K. (2016). Two-step credit risk assessment model for retail bank loan applications using decision tree data mining technique. *International Journal of Advanced Research in Computer Engineering and Technology*, 5(3), 705-718.
9. Winston, P.H. (1992). *Artificial intelligence*. (3rd ed.). Boston, MA: Addison-Wesley Longman Publishing Co Inc.
10. Ogundokun, R.O.; Abikoye, O.C.; Adegun, A.A.; and Awotunde, J.B. (2020). Speech recognition system: Overview of the state-of-the-arts. *International Journal of Engineering Research and Technology*, 13(3), 384-392
11. Adegun, A.A.; Ogundokun, R.O.; Ogbonyomi, S.; and Sadiku, P.O. (2020). Design and implementation of an intelligent gaming agent using A* algorithm and finite state machines. *International Journal of Engineering Research and Technology*, 13(2), 191-206.
12. Turban, E.; Aronson, J.E.; and Liang, T.P (2005). *Decision support system and intelligent systems*. Berkley, UC: Pearson/Prentice Hall.
13. Druzdzel, M.J.; and Flynn, R.R (1999). *Decision support systems. In Bates, M.J.; and Maack, M.N.* (*eds.*) *Encyclopedia of Library and Information Sciences*. (3rd ed.). Boca Raton, CRC Press.

14. Kulkarni, A.; and Purwar, S. (2009). Wavelet-based adaptive backstepping controller for a class of nonregular systems with input constraints. *Expert Systems With Applications*, 36(3), 6686-6696.
15. Han, J.; Kamber, M.; and Pei, J.(2012). *Data mining: concept and technique*. (3rd ed.). Waltham, MA: Morgan Kaufmann Publishers, 205-586.
16. Stevens, D. (2014). *Predicting real estate prices using text mining: Automated real estate description analysis*. HAIT Master Thesis. Tilburg University School of Humanities, Department of Communication and Information Sciences, Tilburg center for Cognition and Communication (TiCC), Tilburg, The Netherlands.
17. Witten, L.; Frank, E.; and Hall, M. (2017). *Data mining: Practical machine Learning Tools and Techniques. San Francisco*, CA: Morgan Kaufmann.
18. Islam, M.R.; and Habib, M.A. (2015). A data mining approach to predict the prospective business sector for lending in retail banking using a decision tree. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 13-22.
19. Hamid, A.J. and Ahmed, T.M. (2016). Developing a prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal*, 3(1), 1-8.
20. Srivastava, S.; Saranya, G.; Pratap, A.; Agrawal, R.; and Jain, A. (2020). Loan default prediction using artificial neural networks. *International Journal of Advanced Science and Technology*, 29(6), 2761-2769.
21. Aslam, U.; Aziz, H.I.T.; Sohail, A; and Batcha, N.K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience*, 16(8), 3483-3488.
22. Obare, D.M.; Njoroge, G.G.; and Muraya, M.M. (2019). Analysis of individual loan defaults using logit under supervised machine learning approach. *Asian Journal of Probability and Statistics*, 3(4), 1-12.
23. Tariq, H.I.; Sohail, A; Aslam, U.; and Batcha, N.K. (2019). Loan default prediction model using sample, explore, modify, model, and assess. *Journal of Computational and Theoretical Nanoscience*, 16(8), 3489-3503.
24. Kumar, M.; Goel, V.; Jain, T.; Singhal, S.; and Goel L.M. (2018). Neural network approach to loan default prediction. *International Research Journal of Engineering and Technology* (*IRJET*), 5(4), 4231-4234.
25. Kwofie, C.; Owusu-Ansah, C.; and Boadi C. (2018). Predicting the probability of loan-default: an application of binary logistic regression. *Research Journal of Mathematics and Statistics*, 7(4), 46-52.
26. Jayadev, M.; Shah, N.M.; and Vadlamani, R. (2016). Predicting educational loan defaults: Application of machine learning and deep learning models. *Proceedings of the* 4*th International Conference on Business Analytics and Intelligence* 2016, Indian Institute of Management Bangalore.
27. Khandani, A.E.; Kim, A.J.; and Lo, A.W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767-2787.
28. Hassan, A.K.I.; and Abraham, A. (2013). Modeling consumer loan default prediction using ensemble neural networks. *Proceedings of* 2013 *International Conference on Computing, Electrical and Electronic Engineering* (*ICCEEE*), Khartoum, Sudan, 719-724.