**Earthline Journal of Mathematical Sciences**
ISSN (Online): 2581-8147
Volume 7, Number 2, 2021, Pages 315-332
https://doi.org/10.34198/ejms.7221.315332

# Detecting Electronic Banking Fraud on Highly Imbalanced Data using Hidden Markov Models

Abukari Abdul Aziz Danaa[1,*], Mohammed Ibrahim Daabo[2] and Alhassan Abdul-Barik[3]

[1] Department of Computer Science, Tamale Technical University, Tamale, Ghana
e-mail: azizdanaa@tatu.edu.gh

[2] Department of Computer Science, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana
e-mail: daabo2005@yahoo.com

[3] Department of Computer Science University for Development Studies, Tamale, Ghana
e-mail: barik75@yahoo.com

## Abstract

Recent researches have revealed the capability of Machine Learning (ML) techniques to effectively detect fraud in electronic banking transactions since they have the potential to detect new and unknown intrusions. A major challenge in the application of ML to fraud detection is the presence of highly imbalanced data sets. In many available datasets, majority of transactions are genuine with an extremely small percentage of fraudulent ones. Designing an accurate and efficient fraud detection system that is low on false positives but detects fraudulent activity effectively is a significant challenge for researchers. In this paper, a framework based on Hidden Markov Models (HMM), modified Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Synthetic Minority Oversampling Technique Techniques (SMOTE) is proposed to effectively detect fraud in a highly imbalanced electronic banking dataset. The various transaction types, transaction amounts and the frequency of transactions are taken into consideration by the proposed model to enable effective detection. With different number of hidden states for the proposed HMMs, simulations are performed for four (4) different approaches and their performances compared using precision, recall rate and F1-Score as the evaluation metrics. The study revealed that, our proposed approach is able to detect fraudulent transactions more effectively with reasonably low number of false positives.

Received: August 8, 2021; Accepted: September 1, 2021

2010 Mathematics Subject Classification: 62M05.

Keywords and phrases: evaluation, false positives, fraudulent, hidden Markov models, oversampling.

[*] Corresponding author

## 1. Introduction

E-banking is a form of banking where funds are transferred as exchange of electronic signals rather than cash, checks, or other types of paper documents [1]. Over the last few decades, E-Banking has redefined the way banking is conducted across the globe and the use of electronic payments platforms has continued to experience significant growth. It allows customers a 24-hour access to their accounts with the ability to transfer funds, perform on-line payments and apply for loans and other financial products virtually [2].

Fraud can be defined as any premeditated act of criminal deceit, trickery or falsification by a person or group of persons with the intention of altering facts, in order to obtain undue personal monetary advantage [3]. Unfortunately, fraud cases relating to cyber-crime perpetrated through E-banking resulted in an actual loss of GH¢14.31 million and therefore presents a unique challenge to individuals and financial institutions that offer those services [4]. To address this problem, financial institutions employ various fraud prevention tools such as real-time transaction authorization, transaction verification codes, transaction alerts, rule-based detection among others. Fraudsters however are adaptive, and given time, they devise several ways to circumvent such protection mechanisms [5]. There is therefore the need to implement enhanced technologies and systems that can detect fraud in real-time effectively in order to maintain the viability of these electronic payment systems where fraudsters constitute a very inventive and fast-moving fraternity. As preventive technology changes, so does the technology of criminals and the way they go about with their fraudulent activities [6]. While it is necessary to detect and possibly prevent fraudulent transactions, it is also very critical to ensure genuine transactions are executed successfully.

One of the most important techniques for intrusion/anomaly detection based on machine learning is using Hidden Markov Models (HMM) which are machine learning algorithms consisting of hidden states and observable outputs for modelling probability distributions over sequences of observations. The hidden state layer is a stable Markov chain and its state probability and state transition probability are decided from the initial state probability vector π and the state transition probabilities. Observable output layer is decided from the observed symbols probability matrix which is derived from the observed symbols of each hidden state [7].

The application of HMMs ranges from speech and image recognition, intrusion/fraud detection to motion/action analysis in videos among others and is generally characterized by the following [8];

1. The number of hidden states in the model denoted as *N*. The state at a specific time *t* is denoted by $q_t$.

2. The number of unique observation symbols denoted as *M*.

3. A transition probability between states denoted by a matrix $A = [a_{ij}]$, where:

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i). \tag{1}$$

Also, $$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N. \tag{2}$$

4. An emission probability matrix, $B = [b_j(k)]$, where

$$B_j(k) = P(V_k = q_t \mid S_j = q_t) \tag{3}$$

$$\sum_{k=1}^{M} b_j\,(k) = 1, \;\; 1 \le j \le N. \tag{4}$$

5. An initial probability for each state denoted by the vector $\pi \;=\; [\pi_i]$, where

$$\pi_i = P(q_1 = S_i), \quad \sum_{i=1}^{N} \pi_i \;= 1. \tag{5}$$

In recent decades, many research communities have been working toward HMM-based intrusion detection mainly because of its ability to detect new and unknown intrusions and usage in real-time applications by processing data streams on-the-fly. HMMs also allow for the usage of heterogeneous data sources as input, and visual representation of acquired knowledge relative to the other techniques of machine learning.

Over the past few years, the use of Electronic banking platforms has continued to experience significant growth and has redefined the way banking or E-commerce is conducted across the world [9]. On the other hand, fraudulent Electronic banking and E-commerce activities are becoming more and more sophisticated and challenging leading to massive financial losses. Effective and efficient detection of Electronic banking fraud is therefore regarded as one of the major challenges to all financial institutions, and is an increasing cause for concern [2].

According to the Bank of Ghana 2019 banking industry fraud report, fraud cases relating to cyber-crime perpetrated through electronic banking and mobile banking platforms accounts for the highest value of attempted fraud amounting to GH¢ 50.54

million with actual loss of GH¢14.31 million [4]. From available literature, majority of the works in the area of HMM-based fraud detection in Electronic banking focuses only on payments to merchants for  goods and services . Transaction amounts are mostly taken as observation symbols and the types of items purchased considered as the hidden states of the proposed Hidden Markov Models.  In related studies conducted by [10], [11], [12], [13], [14], and [15], techniques such as Neural Network , Bayesian Network , Dempster-Shafer theory, Support Vector Machine etc. are employed which incorporated other forms of electronic banking options such as remote funds transfers and deposits. However, all these proposed techniques perform classification based on a single transaction while relying on domain-expert features without considering a sequence of transactions to make a decision hence producing high levels of false positives.

A large number of false positives may translate into bad customer experience and may lead customers to take their business elsewhere. A major challenge in applying ML to fraud detection is presence of highly imbalanced data sets. In many available datasets, majority of transactions are genuine with an extremely small percentage of fraudulent ones. Designing an accurate and efficient fraud detection system that is low on false positives but detects fraudulent activity effectively is a significant challenge for researchers.

This proposed research seeks to develop and implement an improved fraud/intrusion detection system for both debit and credit transactions in electronic Banking using Hidden Markov Models by incorporating the various electronic banking platforms employed by customers, transaction amounts and the frequency at which these transactions occur.  To determine the transaction profile of customers, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) which is capable of discovering clusters of different shapes and sizes from a large amount of data containing noise and outliers was employed. Synthetic Minority Oversampling Technique (SMOTE) was also employed to handle the imbalanced class problem typical of Electronic banking datasets.

The rest of the paper is organized as follows: In Section 2, we present a review of related works. The methodology adopted for the study is outlined in Section 3. Detailed experimental results and discussion to establish the efficiency of the proposed approach is presented in Section 4. Finally, we conclude the paper with some discussions in Section 5.

## 2. Literature Review

Fraud Detection in Electronic Banking is understudied in literature perhaps due to security and data privacy concerns. We will begin by considering related works in electronic banking in general and then consider those specifically related to the use of credit cards which has been given considerable attention by researchers.

[10] presents a fraud detection system for online banking where differential analysis is used to obtain local evidence of fraud where a significant deviation from normal behavior indicates a potential fraud. The Dempster's rule of combination is applied to these evidences for final suspicion score of fraud. Their main contribution is a fraud detection method based on effective identification of devices used to access accounts and assessing the likelihood of being a fraud by tracking the number of different accounts accessed by each device. However, their system performs poorly for higher number of Hidden states and also when users' transaction patterns changes frequently.

[16] considered transaction amounts and purchases types as the emission symbols and hidden states respectively of the proposed HMM for online banking FDS. The model is trained with the normal behavior of an account holder using Baum-Welch algorithm and a One-time-Password is sent to the Customers contact number for authorization if an incoming transaction violates the behavior sequence. Although, the accuracy of their system was close to 72 percent over a wide variation in the input data, False Positives was still high especially when the transaction data is highly skewed. A fraudulent transaction could still go through if a fraudster has access to a customer's phone.

[11] incorporates several advanced data mining techniques for online banking fraud detection by building a contrast vector for each transaction based on its customer's historical behavior sequence. A novel algorithm, Contrast Miner, was introduced to efficiently mine contrast patterns and distinguish fraudulent from genuine behavior, followed by an effective pattern selection and risk scoring that combines predictions from different models. Results from experiments on large-scale real online banking data demonstrated that the proposed system achieves substantially higher accuracy and with lower false positives by incorporating domain knowledge and traditional fraud detection methods.

[12] rather modeled the sequence of operations in online banking transaction processing using HMMs and described how it could be used for the detection of frauds.

The observation sequence length is fixed to two (2) whilst changing sequence length for training i.e., changing dataset length from 10 to 80 with difference of ten. Simulation results revealed that, although the complexity of the system also increases for increased observation sequence length, the accuracy of the proposed system is close to 60% with reduced false Positive rate.

The work done by [14] employed HMMs and k-means algorithm for detecting fraud in online banking transactions. In their proposed model, a variable is used to keep the number of transactions within a period of time before and after each transaction as well as the quantified amounts as the observation symbols. If an incoming transaction is not accepted by the trained HMM with sufficiently high probability, it is considered fraudulent. The feasibility of their proposed model is demonstrated through simulation experiments using real-world bank transaction data. In the case of enough historical transactions, their model performs well for low, medium frequency and amount of user groups. An efficient Prior determination of the number of clusters is considered a major challenge in their proposed approach.

Specifically on fraud detection relating to the use of Credit Cards, [17] considered purchase types and transaction amounts as hidden states and observation symbols respectively in their proposed HMM. In order to estimate the model parameters, the K-means clustering algorithm is employed to determine the spending profile of cardholders. An incoming transaction is considered fraud if it is not accepted by the HMM with a significantly high probability. Experimental results revealed that, their proposed model recorded an accuracy close to eighty (80) percent over a wide variation of the data. An efficient prior determination of the number of clusters and significant number of false positives were considered the major challenges in their proposed approach.

[18] performed a comparative analysis of intrusion detection models on highly skewed credit card data based on Decision Trees, Random Forest, Support Vector Machines (SVM) and logistic regression. The original sample was randomly partitioned into k-equal sized subsamples where a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples used as training data. With the four basic metrics employed, namely True positive (TPR), True Negative (TNR), False Positive (FPR) and False Negative (FNR) rates, Simulation results using dataset provided by ULB machine learning revealed that, Logistic regression and Random forest shows the most precise and high accuracy in the area of credit card fraud detection but requires very large dataset for training and also suffers from the imbalanced dataset problem even after preprocessing.

In order to reduce the number of false positives, [19] proposed a model based on automated feature engineering to automatically derive behavioral features based on the historical data of a credit card associated with a transaction. A total of 237 features for each transaction was generated, and a random forest was then employed to learn a classifier. One important feature of their proposed model is that, it also utilizes the distance between two locations transactions on an account has occurred and whether they occurred in person or remotely is established. The proposed model was tested on data from a large multinational bank and compared to existing solutions and revealed that, on an unseen data of 1.852 million transactions, false positives was reduced by about 54%. However, since their models Perform classification based on a single transaction there was a performance degradation when transaction pattern of users changes frequently.

## 3. Methodology

There is generally a very limited number of public datasets on electronic banking for research purposes mainly due to personal and security concerns. In this research, a Kaggle provided dataset of simulated mobile based transactions is adopted. As detailed in Table 1, the dataset is highly imbalanced due to the fact that only 8,312 transactions out of the almost 6 million transactions are labeled as fraud.

**Table 1:** Details of the Paysim Dataset adopted for the study.

| Transaction Type | # of Genuine Transactions | # of Fraudulent Transactions | Total |
|---|---|---|---|
| **TRANSFER** | 528812 | 4097 | 532909 |
| **CASH-OUT** | 2233384 | 4116 | 2237500 |
| **CASH-IN** | 1399284 | 0 | 1399284 |
| **DEBIT** | 41432 | 0 | 41432 |
| **PAYMENTS** | 2151494 | 0 | 2151494 |
| **TOTAL** | 6354407 | 8213 | 6362620 |

‘CASH IN’ and ‘CASH OUT’ represents an increase in account balance of a customer as a result of cash inflow and a decrease in account balance as a result of cash outflow respectively. ‘TRANSFER’ refers to movement of money between users whilst ‘PAYMENT’ represents the settlements made for goods and services to merchants. ‘DEBIT’ as used in this context signifies the sending of money from a mobile service (electronic wallet) to a bank account.

### 3.1. Data pre-processing

To effectively evaluate the performance of our proposed models on the highly class imbalanced dataset, Synthetic Minority Oversampling Technique (SMOTE) is employed to generate virtual training records by linear interpolation for the fraudulent transactions by randomly selecting one or more of the k-nearest neighbors for each specific fraudulent transaction. After the oversampling process, the data is reconstructed and then the proposed Hidden Markov Models is applied on the processed data. Specifically, the sampling rate is set to 73000 %.

The proposed SMOTE technique as adopted in this study is presented in Algorithm 1.

**Algorithm 1:** The *SMOTE* algorithm

*Procedure SMOTE* $(f, R, k)$

**Input:** Number of Fraudulent Transactions $(f)$; Amount of SMOTE $R$ %; Number of nearest neighbors $k$

**Output:** $(R/100) * f$ synthetic Fraudulent Transactions

1: The number of Fraudulent Transactions is set to $f$

2: For each $y \in f$, the k-nearest neighbors of $y$ are obtained by calculating the Euclidean distance between $y$ and every other sample in set $f$.

3: The sampling rate $R$ is set according to the imbalanced proportion.

4: For each $y \in f$, $R$ examples are randomly selected from its k-nearest neighbors, and they construct the set $f_1$.

5: For each example $y_t \in f_1$ $(t = 1, 2, 3, \dots, R)$, generate a new sample as in equation 6 below:

$$y' = y + rand(0,1) * |y - y_t|. \tag{6}$$

### 3.2. Identifying transaction profile of customers

For optimal training of the our proposed Hidden Markov Models, a modified Density Based Spatial Clustering of Applications with Noise (DBSCAN) and the K-means clustering algorithms are executed on each customer's previous transactions by considering the amount and frequency of transactions. K-means is an unsupervised learning algorithm for grouping a given set of data based on their similarity where the numbers of clusters are fixed a priori. The grouping is performed by minimizing the sum of squares of distances between each data point and the centroid of the cluster to which it belongs to [20]. The DBSCAN clustering technique however filters out outliers and discovers clusters of arbitrary shapes [21]. We modified the DBSCAN algorithm by adding a step that computes the centroid of each cluster later to be used to dynamically convert an incoming transaction into an observation symbol in the fraud detection process.

The proposed DBSCAN technique as adopted in this study is presented as in Algorithm 2.

**Algorithm 2**: The DBSCAN algorithm.

DBSCAN (dataset, *d*, minpts)

**Input**: A set of points, *dataset*, distance threshold *d*, and the minimum number of points required to form a cluster, min*pts*.

**Output**: A set of clusters representing the various observation symbols

```
$n = 1$, #initialise the cluster index to 1
For each unvisited point $pt$ in dataset, mark $pt$ as visited
Find the neighboring points, $N$ of $p$
If $|N|$ >=minpts then $N = N U N'$
if $p'$ is not a member of any cluster, Mark it as noise.
Compute the centroid $\hbar$ of each cluster using (7) below
```

$$\hbar = \frac{1}{n_i} \sum_{x_j \in c_i} x_j, \tag{7}$$

where $n_i$ is the number of points in cluster $c_i$.

Spending profiles of accountholders are determined at the end of the clustering step. Let $\psi_i$ be the percentage of total number of transactions of an accountholder, then, the spending profile $\rho$ of an account holder, $\vartheta$ is determined as in (8):

$$\rho(\vartheta) = \arg \underset{i}{Max}(\psi_i). \tag{8}$$

The cluster number to which most of the transactions of the account holder belongs to represents the spending profile of the account holder. The computed centroids are used to generate the observation symbol for a new transaction $\emptyset$ (denoted by $\emptyset_m$) is defined as in (9).

$$\phi_m = v_{arg_i} min|m - n_i|. \tag{9}$$

The i^th^ transaction on account $A_k$ denoted as $P_{i,y}^{A_k}$ is suspected to be an outlier if it does not belong to any cluster in the set $C'$ where *y* refers to the frequency of transaction. If the average distance of the amount $p$ of an outlier transaction $P_{i,y}^{A_k}$ from the set of existing clusters in $C'$ is $W_a$, then its level of deviation $o_l$ is given as in (10):

$$o_l = \begin{cases} \frac{W_a - \epsilon}{W_a} & if\ |N_\epsilon(p)| < MinPts \\ 0 & otherwise. \end{cases} \tag{10}$$

The key idea of the modified DBSCAN algorithm is that for each point *p* in a cluster $C_i$, there are at least a minimum number of points (*MinPts*) in the e-neighborhood of that point $p$ denoted as $N_\varepsilon(p)$ i.e., the density in the e-neighborhood has to exceed some threshold. The proposed K-means algorithm as adopted in this study is presented in Algorithm 3.

**Algorithm 3:** K-Means Clustering Algorithm

```
1: Specify the number of clusters to assign
2: Randomly initialize k centroids
3: Repeat
4:  Expectation: Assign each point to its closest centroid
5:  Maximization: Compute the new Centroid (Mean) of each cluster
6: Until the Centroid Positions do not change
```

Specifically, for this research, the set $C$ = {low-frequency low-amount, low-frequency medium-amount, low-frequency high-amount, medium frequency low-amount, medium-frequency medium-amount, medium-frequency high-amount, high-frequency low- amount, high-frequency medium-amount, high-frequency high-amount} denotes the clusters.

The set $R$ = {transaction_amount, frequency_of_transaction} represents the set of attributes used to generate these clusters.

To compute the probability of an observed sequence, $\sigma = (\sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_{T-1})$ with respect to our Hidden Markov Model $\lambda$, where $\upsilon = (\upsilon_0, \upsilon_1, \upsilon_2, \ldots, \upsilon_{T-1})$ represents the various hidden states, the definition of the emission transition Matrix is defined as in (11);

$$P(\sigma|\upsilon,\lambda) = b_{\upsilon 0}(\sigma_o) b_{\upsilon 1}(\sigma_1) \ldots b_{\upsilon T-1}(\sigma_{T-1}). \tag{11}$$

The Initial transition vector, π and State Transition Matrix, A are also defined as in (12) and (13);

$$P(\upsilon|\lambda) = \pi_\upsilon a_{\upsilon,\upsilon 1} \ldots a_{\upsilon T-2,\upsilon T-1} \tag{12}$$

and

$$P(\sigma,\upsilon|\lambda) = P(\sigma|\upsilon,\lambda)P(\upsilon|\lambda). \tag{13}$$

By summing over all possible state sequences, (14) through (16) is obtained

$$P(\sigma|\upsilon) = \sum P(\sigma,\upsilon|\lambda)_X \tag{14}$$

$$= \sum P(\sigma|\upsilon,\lambda)P(\upsilon|\lambda)_\upsilon \tag{15}$$

$$= \sum \pi_{\upsilon 0} b_{\upsilon 0}(\sigma_o) a_{\upsilon,\upsilon 1} \ldots a_{\upsilon T-2,\upsilon T-1} b_{\upsilon T-1}(\sigma_T - 1)_\upsilon. \tag{16}$$

The probability of the observation sequences denoted as $e_t(i)$, where the system is in state $q_i$ at time *t* is defined in (17).

$$e_t(i) = P(\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_t, \upsilon_t = q_i|\lambda). \tag{17}$$

$e_t(i)$ is then calculated recursively as in (18) to (20):

1. Let $e_0(i) = \pi_i b_i(\sigma_0)$ (18)

2. For $i = 0, 1, 2, \ldots, N-1$ and $t = 1, 2, \ldots, T-1$, we compute

$$e_t(i) = \sum_{j=0}^{N-1} [e_{t-1}(j) a_{ji}]\, b_i(\sigma_t) \tag{19}$$

3. Equation (20) is obtained from (19)

$$P(\sigma|\lambda) = \sum_{i=0}^{N-1} e_{T-1}(i). \tag{20}$$

### 3.3. Training the proposed HMMs

The transaction amounts are categorized into a Low $(l) = (0;\ 100]$, Medium $(m) = (100;\ 500]$, and High $(h) = (500;\ \text{Transaction Limit}]$ values. The frequency at which these transactions occur on a particular is also categorized into a Low (Less than 5 times a month), Intermediate (Between 5 and 10 times a month), and High (at least 10 times a month) are also considered by our proposed model. For example, if an accountholder performs about seven (7) transactions with the month with an average value of say 300, then the corresponding observation symbol is medium-frequency medium-amount (mm).

The various transaction types are considered the internal states whilst the transaction amounts combined with the frequency at which they occur denoted as $\{ll, lm, lh, ml, mm, mh, hl, hm, hh\}$ represents the observation symbols of our proposed Hidden Markov Model.

After formulating the hidden states and observation symbol, a hybrid optimization algorithm as presented in Algorithm 4 comprising the Baum-Welch, Particle Swam and Genetic Algorithms is used to effectively train the proposed models.

**Algorithm 4:** A hybrid algorithm for optimizing the parameters of the proposed HMMs

```
1. Initialize the parameters $(A, B, \pi)$ using the spending profile of the customer
2. $\alpha_t(i) = \sum_{j=0}^{N-1}[\alpha_{t-1}(j)a_{ji}]\, b_i(O_t)$        (21)
3. $\beta_t(i) = \sum_{j=0}^{N-1}[\beta_{t+1}(j)a_{ij}]\, b_j(O_{t+1})$        (22)
4. $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$        (23)
5. $\gamma_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$        (24)
6. For $0 \leq i,j \leq N-1$
7. $\pi_i = \gamma_0(i)$        (25)
8. $a_{ij} = \sum_{t=0}^{T-2}\gamma_t\,(i,j)/\sum_{t=0}^{T-2}\gamma_t\,(i)$        (26)
9. $b_j(k) = \sum_{\substack{t\epsilon\{0,1,\ldots T-1\}\\ O_t=k}}\gamma_t\,(j)/\sum_{t=0}^{T-1}\gamma_t\,(i)$        (27)
10. Go to 6
11. After 100 iterations, each solution becomes a chromosome for the genetic procedure and the fitness function $P(O|\lambda)$ is applied.
12. A multiple point crossover and mutation is performed to select the best 50 solutions for the next generation which are then positioned as particles in a search space using the PSO technique.
13. The position and velocity of each particle during each iteration is updated using;
14. $X_i(t+1) = X_i(t) + V_i(t)$        (28)
15. $V_i(t+1) = wV_i(t) + r_1\,\mathrm{U}([0,1])\,(X + i\,(t) - X_i(t))$
        $+ r_2\,\mathrm{U}([0,1])\left(\hat{x}_i(t)\text{- } X_i(t)\right).$        (29)
```

16. Compare the best solution each particle and the best position of the entire group and make appropriate adjustments

17. Termination Criteria Reached? If yes go to 18, otherwise go to 13

18. Output A, B and π

### 3.4. Fraud detection

To effectively classify an incoming transaction as fraudulent or otherwise, sequence of observation symbols, say $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_r$ are extracted from the training data of an account holder and its probability of acceptance, $\partial_1$ is computed by the model as in (30) by employing (20)

$$\partial_1 = p(\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_r | \lambda). \tag{30}$$

An incoming transaction occurring at time is converted to an observation symbol denoted as $\sigma_{t+1}$ using (9) is used to replace the first observation symbol, $\sigma_1$ and its probability of acceptance by the model denoted as $\partial_2$ is also computed as in (31).

$$\partial_2 = p(\sigma_2, \sigma_3, \sigma_4, \ldots \sigma_{r+1} | \lambda). \tag{31}$$

The newly generated transaction is classified as fraud and if the difference between $\partial_1$ and $\partial_1$ denoted as $\varDelta\partial$ is above a predefined threshold ($\varepsilon$) as in (32).

$$\varDelta\partial/\partial_1 \geq \varepsilon. \tag{32}$$

A genuine transaction is added to the sequence permanently to contribute to determining the validity or otherwise of the next transaction since transaction behavior of an accountholder could be dynamic. Otherwise, the transaction is declined, and the symbol is discarded.

Due to the highly imbalanced nature of the dataset, precision(p), recall(R) and F1-scores (F) as presented in equations (33) to (35) respectively are used as evaluation metrics (Wedge et al. [19]). $Tp, Fp, Fn$ represent True Positives, False Positives and False Negatives respectively.

$$p = Tp/(Tp + Fp) \tag{33}$$

$$R = Tp/(Tp + Fn) \tag{34}$$

$$F = 2 * (P * R)/(P + R). \tag{35}$$

Precision quantifies the number of correct positive predictions made whilst Recall refers to the number of correct positive predictions made out of all possible positive predictions. F-Measure however provides a way to combine both precision and recall into a single measure.

## 4. Simulation Results and Discussion

For different number of hidden states, four (4) sets of simulations were performed in two (2) stages using Python programming and their performance compared. For all the four sets of experiments, the proposed hidden Markov models were executed in the second stage. In the first stages of the first and second set of experiments, K-means and the modified DBSCAN algorithms were executed respectively. In first stage of the third set of experiment, both SMOTE and K-means techniques were employed whereas SMOTE and the modified DBSCAN clustering techniques were executed during the last set of experiments. The dataset was loaded and divided into two, 80% of it is used for training and evaluation whilst the rest is held back for validation.

### 4.1. Precision comparison

The precision of the four approaches is presented for different number of hidden states and presented in Figure 1. It is very clear that our proposed approach (SMOTE+DBSCAN+HMM) performed better for the various hidden states. Applying only the modified DBSCAN clustering technique with Hidden Markov Models performed relatively better than that of employing K-Means. It is also worth noting that, relatively higher values of precision scores were recorded when the SMOTE technique is adopted.

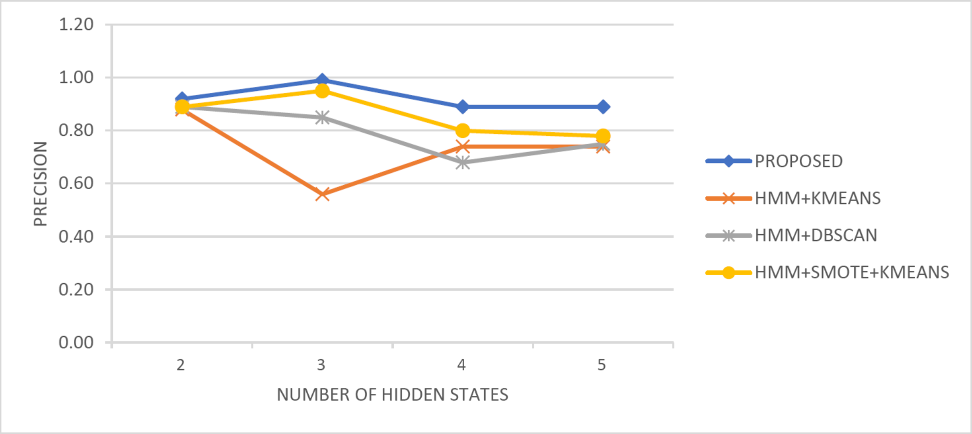

**Figure 1:** Comparison of the precision of the four (4) different approaches for different number of hidden states.

### 4.2. Recall rates comparison

A comparison of the Recall rates of the four (4) different approaches for different numbers of hidden states are presented in Figure 2. It is evident that approaches that employed the SMOTE technique appear to perform relatively better. Similarly the modified DBSCAN clustering technique performed better as compared to the K-means.

It can also be observed that, for higher values of N, all the approaches performed well except for those that employed Only K-means and Hidden Markov models without handling the class imbalance classification.

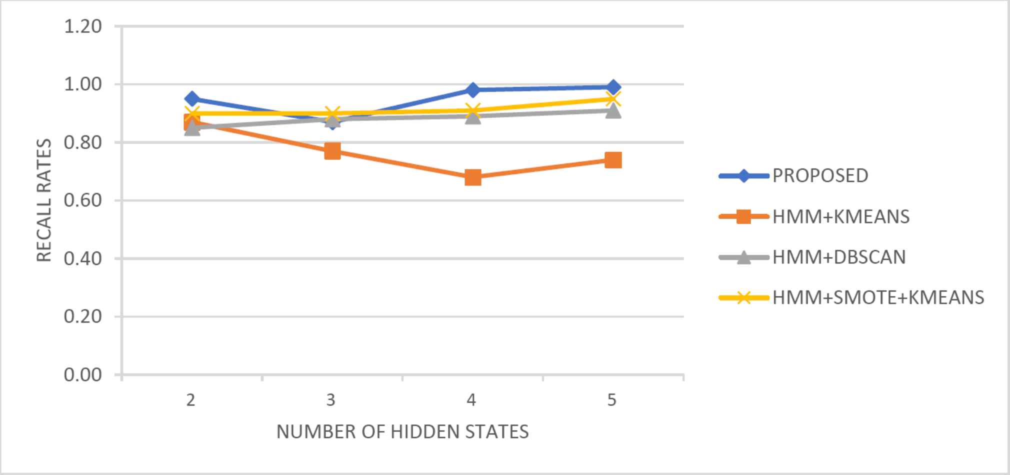

**Figure 2:** Recall rates of the four (4) approaches for different number of Hidden states.

### 4.3. F-measure comparison

In Figure 3, the F1-score for the four approaches are presented various Hidden states. It is observed that, higher F1-scores are obtained when the modified DBSCAN clustering technique is used as compared to using the K-means. Also, approaches that incorporated the SMOTE technique performed better. Employing both SMOTE and the modified DBSCAN clustering algorithms appears to perform relatively better than the other. All approaches performed relatively better when the number of hidden states is 3.

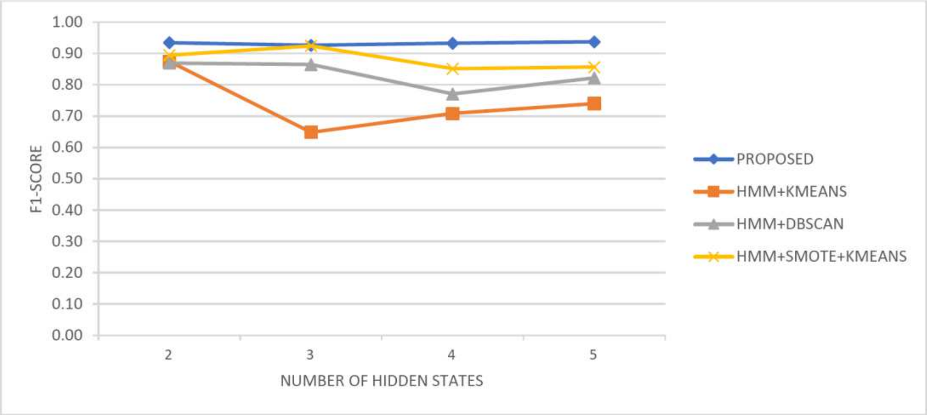

**Figure 3:** F1-scores of the four (4) different approaches for various number of Hidden states.

## 5. Conclusion

In this research, an improved electronic banking fraud detection framework based on Hidden Markov Models (HMM) and modified Density Based Spatial Clustering of Applications with Noise (DBSCAN) is proposed and implemented. The Synthetic Minority Oversampling Technique (SMOTE) is also employed due to the highly class imbalance nature of the dataset adopted. With different numbers of hidden states, simulations were performed two stages for four (4) different approaches in Python and their performance compared. For all the four sets of experiments, the proposed hidden Markov models were executed in the second stage. In the first stages of the first and second set of experiments, K-means and the modified DBSCAN algorithms were executed respectively. In first stage of the third set of experiment, both SMOTE and K-means techniques were employed whereas SMOTE and the modified DBSCAN clustering techniques were executed during the last set of experiments.

Generally, our proposed approach (SMOTE+DBSCAN+HMM) performed relatively better for all the various hidden states in terms of precision, recall and F1-Scores. Employing the modified DBSCAN clustering technique to determine the spending profile of customers and subsequently performed relatively better than using the K-Means algorithm since it filters out most of the easily recognizable fraudulent transactions before the proposed HMMs are applied. It is also evident from the simulation analysis that, the SMOTE technique effectively handles the class imbalance classification necessary to achieve improved performance.

## References

[1] M. A. Ali, N. Hussin and I. A. Abed, E-banking fraud detection: a short review, *Int. J. Innov. Creat. Chang.* 6(8) (2019), 67-87.

[2] M. Asare and J. Sakoe, The effects of electronic banking on financial services in Ghana, *Res. J. Financ. Account.* 6(16) (2015), 147-155.

[3] J. N. Taiwo, M. E. Agwu, A. A. Babajide, T. C. Okafor and A. A. Isibor, Growth of bank frauds and the impact on the Nigerian banking industry, *Journal of Business Management and Economics* 4(12) (2016).

[4] Bank of Ghana, F. Banks and S. D. Institutions, BANK OF GHANA Banking Sector Report, *Corp. Gov. Dir.*, 2018.

[5] Avanti H. Vaidya and S. W. Mohod, Internet banking fraud detection using HMM and BLAST-SSAHA hybridization, *Int. J. Sci. Res.* 3(7) (2014), 574-579.

[6] L. Kovács and S. David, Fraud risk in electronic payment transactions, *Journal of Money Laundering Control* 19(2) (2016), 148-157. https://doi.org/10.1108/JMLC-09-2015-0039

[7] Z. Ghahramani, An introduction to hidden Markov models and Bayesian networks, *Int. J. Pattern Recognit. Artif. Intell.* 15(1) (2001), 9-42. https://doi.org/10.1142/S0218001401000836.

[8] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* 77(2) (1989), 257-286. https://doi.org/10.1109/5.18626

[9] A. Devi, Mobile banking: the revolution in digitalization of financial services with special reference to State Bank of India, *Int. Res. J. Manag. Sci. Technol.* 9(4) (2018), 49-58.

[10] S. Kovach and W. V. Ruggiero, Online banking fraud detection based on local and global behavior, *ICDS 2011, Fifth Int. Conf. Digit. Soc.*, 2011, pp. 166-171 [Online]. Available: https://www.thinkmind.org/articles/icds_2011_6_40_90006.pdf

[11] W. Wei, J. Li, L. Cao, Y. Ou and J. Chen, Effective detection of sophisticated online banking fraud on extremely imbalanced data, *World Wide Web* 16(4) (2013), 449-475. https://doi.org/10.1007/s11280-012-0178-0

[12] S. D. Avghad and M. S. Joshi, Securing online banking transaction using predictive approach of hidden Markov model, *Int. J. Comput. Appl.* 128(7) (2015), 14-17. https://doi.org/10.5120/ijca2015906603

[13] M. Carminati, R. Caron, F. Maggi, I. Epifani and S. Zanero, BankSealer: A decision support system for online banking fraud analysis and investigation, *Comput. Secur.* 53 (2015), 175-186. https://doi.org/10.1016/j.cose.2015.04.002

[14] X. Wang, H. Wu and Z. Yi, Research on bank anti-fraud model based on K-Means and hidden Markov model, *2018 3rd IEEE Int. Conf. Image Vis. Comput. (ICIVC),* 2018, pp. 780-784. https://doi.org/10.1109/ICIVC.2018.8492795

[15] I. Achituve, S. Kraus and J. Goldberger, Interpretable online banking fraud detection based on hierarchical attention mechanism, *IEEE Int. Work. Mach. Learn. Signal Process. (MLSP)*, 2019, pp. 1-6. https://doi.org/10.1109/MLSP.2019.8918896

[16] S. S. Mhamane and L. M. R. J. Lobo, Internet banking fraud detection using HMM, *2012 3rd Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT'12),* 2012, pp. 1-4. https://doi.org/10.1109/ICCCNT.2012.6395910

[17] A. Srivastava, A. Kundu, S. Sural and A. Majumdar, Credit card fraud detection using hidden Markov model, *IEEE Trans. Dependable Secur. Comput.* 5(1) (2008), 37-48. https://doi.org/10.1109/TDSC.2007.70228

[18] N. Khare and S. Y. Sait, Credit card fraud detection using machine learning models and collating machine learning models, *International Journal of Pure and Applied Mathematics* 118(20) (2018), 825-838.

[19] R. Wedge, J. M. Kanter, K. Veeramachaneni, S. M. Rubio and S. I. Perez, Solving the false positives problem in fraud prediction using automated feature engineering, *Lecture Notes in Computer Science*, vol. 11053, Springer, Cham, 2019, pp. 372-388. https://doi.org/10.1007/978-3-030-10997-4_23

[20] M. Malekpour, M. Khademi and B. Minae-Bidgoli, A hybrid data mining method for intrusion and fraud detection in e-banking systems, *J. Comput. Intell. Electron. Syst.* 3 (2014), 1-6. https://doi.org/10.1166/jcies.2014.1068

[21] L. Duan, L. Xu, F. Guo, J. Lee and B. Yan, A local-density based spatial clustering algorithm with noise, *Inf. Syst.* 32(7) (2007), 978-986.
https://doi.org/10.1016/j.is.2006.10.006