Intelligent Citizenship Identity through Family Pedigree Using Graph-Signature Based Random-Forest Model

*[1]Adedayo D. Adeniyi, [2]Semiu O. Oladejo and [1]Tiwalade M. Usman
[1]Department of Computer science, Kaduna Polytechnic, Kaduna, Nigeria
[2]Department of Mathematics, Gombe State University, Gombe, Nigeria
drdayo@kadunapolytechnic.edu.ng|abdsemiu@yahoo.com|tiwalade.usman@aun.edu.ng

Received: 31-JAN-2021; Reviewed: 28-MAR-2021; Accepted: 29-MAY-2021
http://dx.doi.org/10.46792/fuoyejet.v6i2.609

Abstract- There has been a global upsurge of interest in the topic of citizenship identity over the past decades, specifically in the world dominated by profound insecurity, inequalities, proliferation of identities, and rise of identity politics,engendered by capitalism. However finding effective solution to these problems has been rendered difficult. To alleviate these problems, this paper presents an analytical Machine learning model that suitably combined the graph signature with random forest techniques. This study presents the design and realization of a novel Intelligent Citizenship Identity through family pedigree using Graph Signature based random forest (GSB-RF) model. The study also showcases the development of a novel graph signature technique referred to as Canonical Code Signature(CCS) method. The CCS method is used at the pre-processing stage of the identification process to build signature for any given tuple. Performance comparisim between the present system and the baseline techniques which includes: the K-Nearest Neighbour and the traditional Random Forest shows that the present system outperformed the baseline method studied. The proposed system shows capability to perform continuous re-identification of Citizens based on their family pedigree with ability to select best sample with low computational complexity, high identification accuracy and speed. Our experimental result shows that the precision rate and identification quality of our system in most cases are equal to or greater than 70%. Therefore, the proposed Citizenship Identification machine is capable of providing usable, consistent, efficient, faster and accurate identification, to the users, security agents, government agents and institutions on-line, real-time and at any-time.

Keywords- Canonical code,Citizenship Identity, Family pedigree,Graph-Signature,Machine learning, Random-forest

1 INTRODUCTION

The recent surge of theoretical interest in citizenship identity has been a matter of pressing concern across the socio-political, economic and security spectrum.The apparent rise in the politics of identity, insecurity, inequality and proliferation of identities in recent years has brought to force the age-old tension in the heart of concerned individuals. Apparently the fragmented political identity and conflicting political loyalty and obligations associated with these conditions posed important socio-economic, political and security challenges to any nation (Purvis and Hunt, 1999).

This study presents the design and realization of an intelligent citizenship identity through family pedigree using Graph Signature Based Random Forest(GSB-RF) algorithm. It aims at determining the identity of a given citizen through their family lineage and to develop an online application that can be accessed by security agents, government agents and parastaters to identify an individual through their family pedigrees. The pedigree chart technique adopted will allow users to be able to view record of other family members from a single member of the family presented to the system.

The contribution of this work is in threefold. A novel analytical machine learning algorithm for the purpose of generating and mining the identity of an individual from a large population based database is proposed. Attention is specifically focused on improving the search efficiency and identification accuracy of the citizenship identification process by hybriding the graph signature with the random forest technique. This is to enable the designers and administrators of citizenship identification systems to have varieties of algorithms from which the best performing algorithm can be selected. The developed intelligent citizenship identification system through family pedigree, powered by Graph Signature Based Random Forest technique is computationally efficient, accurate and faster for scalable implementation. It is as well capable of handling a large dataset, overcome computational inaccuracy, search inefficient, long processing time and scalability challenges common to many traditional algorithms.

A novel graph based signature technique called Canonical Code Signature (CCS) model is proposed. This method is used to build signature for a given individual or node at the pre-processing stage of the identification process, before applying the proposed random forest model. This is achieved by generating the canonical sum of the graph connections according to distance from the root of the tree on a Depth-First-Search (DFS) hierarchy. This technique helps to build a unique identity of an individual for identification purposes. The application of the CCS technique will significantly improve the proposed system's results since the topology information given allows us to categorize the basic graph entities and subsequently form a basis for carrying out identification of an individual. A specific intelligent Citizenship identification system using an in-house developed experimental website is constructed. The website was

*Corresponding Author

Section B- ELECTRICAL/COMPUTER ENGINEERING & COMPUTING SCIENCES
Can be cited as:
Adeniyi A.D., Oladejo S.O. and Usman T.M. (2021): Intelligent Citizenship Identity Through Family Pedigree Using Graph-Signature Based Random-Forest Model, FUOYE Journal of Engineering and Technology (FUOYEJET), 6(2), 58-66. http://dx.doi.org/10.46792/fuoyejet.v6i2.609

developed with PHP programming language at the front-end and was hosted using XAMP/Apache HTTP server with MySQL database management system at the back-end. The developed system is a website that allows the identity managers and administrator to enter information relating to an individual, the system build her signature by generating a unique identification code for the individual, stores it in a data mart and generates a plastic Id-card for the individual. The information on this card can be used by security or government agent to establish the identity of an individual by entering her Id-number which serves as her signature.

The system search through the database, using the search state space tree with limited backtracking. This will assist the security and government agents to easily establish the identity of the individuals through their family lineage and to set priority for public security intervention to reduce crime and criminalities in the society at large. The classical algorithm, that is, the Graph signature based random forest (GSB-RF) algorithm that form the basis for the development of the graph based data mining and machine learning system is presented alongside the proposed graph based CCS signature technique.The performance evaluation of the proposed system is carried out against two baseline methods which are: the K-Nearest Neighbour and the traditional Random Forest. This is to justify the rationale for the selection of the signature based random forest identification model. The results of the conducted experiment shows that the present system outperformed the baseline method studied. The present system shows capability to perform continuous re-identification of Citizens based on their family pedigree with ability to select best sample with low computational complexity, high identification accuracy and speed, with precision rate and identification quality of equal to or greater than 70% in most cases. Finally, a systematic presentation of the experimental result is carried out and the developed system can be implemented online and in real-time basis.

2 Related Work

This section reviews related works germane to this work, highlighting major survey that covers prior works on the topic. More recent works that share similarities with our approach will also be discussed. The review is specifically organized into sub sections as follows.

2.1 Overview of Machine Learning Techniques

Machine learning is a branch of artificial intelligent system that concerns the design of algorithm that enables a system to learn from data rather than through explicit programming. (Hurwitz and Kirsch, 2018; Smola and Vishwanathan,2008; Lewu, 2017). The learning process involves making the machine to find statistical regularities or other patterns in the data, thus a number of machine learning algorithms almost similar to how human might approach a learning task. Machine learning task can be approached in a number of ways, which includes supervised learning, unsupervised learning, reinforcement learning, neural networksand deep learning (Hurwitz and Kirsch, 2018; Smola and Vishwanathan,2008; Lewu, 2017;Yara, Nadine, Nicholas and Mariette, 2019).

According to different scholars in machine learning field, there are different algorithms that can be used to approach different problems, these includes Bayesian, Decision tree, KNN, K-Means, SVM, ANN, SOM, Random forest etc. (Singh and Das, 2007; Hssina, Merbouna, Ezzikouri and Erritali, 2014; Adeniyi, Wei and Yongquan, 2015; Adeniyi, Wei and Yang 2018). A number of these algorithms have been studied; some show weakness in handling large data sets, inefficient rules when tested on new datasets, scalability problem, slow response, inability to handle noisy data sets, inaccuracy and computational complexity. The present system adopts a hybrid technique that combines the random forest method with graph signature technique to bring about a more scalable, faster, accurate, efficient system capable of handling large datasets with no computational complexity.

2.2 Overview of Graph Signatures

There are a number of known graph representation tools that address different areas of problems. Graph signature is a technique used in capturing the local topology information surrounding each graph node. A number of different graph signatures have been reported by different scholars in recent time, these includes: data signature reported by Wong Foote, Chine, Mackey and Perrine, (2006),Control flow graph-based signature by Saleh, Ratazzi and Xu (2017), the use of canonical code presented by Hsien, Hsu, Ti and Kuo (2014), weighted graphs node signature by Jouili and Tabbone (2009). etc. Most of these methods have been proved to be efficient in their different area of applications. However, a number of them are faced with one challenge or the other, ranging from low scanning speed, insecurity, and some of these techniques are subject to manipulation by intruders and potential false positive. The Present work share a similar design philosophy with some of these existing works but with a different design approach which serves to overcome some of the challenges of most of the existing methods..

2.3 Overview of Related Identity System

A person's identity can be described as a set of characteristics, physical, social and legal that fully describes and characterizes that individual as an active member of human society and can differentiate the individual from the rest of the populace.Nguyen (2003). In order to augment security in our society and in our everyday life, different scholars have proposed different approaches to citizenship identification and authentication, so as to provide proactive ways of anticipating security vulnerabilities before damage is done. (Nguyen, 2003; Hocking, Furnell, Clarke and Reynolds, 2013; Luis-Garcia, Alberola-Lopez, Aghzout and Ruiz-Alzola, 2003; Rovisco and Lunt, 2019.). A weak National identification system can lead to exclusion of citizens from social services programs, inability to verify identities can also lead to insecurity of lives and properties, fraudulent activities and more. Therefore,

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

there is a need for a robust identification system, to reduce cost of governance, increase security system and facilitate easy financial transactions Khan (2014). The goal of this work is to give a comprehensive study of the citizenship identity system. The work aimed to introduce a technology enabled citizenship identification system through family pedigree with the use of graph signature and random forest techniques.

3 Methodology

This section presents the distinct sets of activities, actions, tasks and frameworks for the design and implementation of the Intelligent citizenship Identification System, it also showcases the application of the new methodology to analyse the population census figures database of the Nigerian National population commission. It presents how an on-line, real-time Intelligent Citizenship Identification system is developed to assist government agents and security agents to identify an individual through their family pedigree and to assist security experts to be able to detect crime and criminals in the society so as to set priorities for security of lives and properties, in order to reduce crime and criminalities and its effect on individuals and the public in general. The graph signature based random forest algorithm developed will assist the designers and administrators of security and identity management system to have more varieties of algorithms from which they can select one with the best performance.

3.1. Experimental Design

The proposed Intelligent Citizenship Identification model is trained on Nigerian 2006 population census data, extracted from the archive of the Nigerian national Population Commission. The details of the extracted population data is kept anonymous and secured. The study subjects consist of cohort of extended families and their members from 25 local government areas of six states from different geo-political regions of the country. We studied the population records retrospectively and discovered that some families have up to forth generations, while most families' lineage stops at the third generations living. We assigned identity code to each level of the families generation based on their state and local government code. Each individual in the family is assigned a unique identification code which is linked to her family lineage code to form her signature. Test data were individually matched with records in the training set according to their family relationship. The identities of the individuals were determined using our proposed graph signature based random forest model. The identity of an individual is reveled based on similarities in their signature with respect to other individuals' signatures in the family network.

3.2 Data Collection

A sample of 18,427 anonymous individuals was randomly selected from the Nigerian 2006 population census data obtained from the Nigerian Population Commission (NPC) archive based on their family lineage from 25 local government areas of six states from different geopolitical regions of the country. The raw population data extracted were cleansed in order to eliminate irrelevant or noisy entries and a database was developed.

3.3 The Proposed Graph Signature Model

A tree T is a set of nodes storing elements in a Parent-Child relationship with the following properties. If T is non-empty, it consists of a specific node popularly referred to as the root of T, which has no parent. Each node V of T differs from the root has a unique parent node W; Every node with parent W is a child to W.

A tree T can either be empty or consists of a node r, referred to as the root of T, and a possible empty sets of tree in which the roots are the children of r. Two nodes that are children of the same parents are called siblings. An external node V are nodes that has no children, an internal nodes V are nodes with one or more children.
A node U is an ancestor of a node V, if $U = V$ or U is an ancestor of the parent of V. i.e. Node V is a descendant of node U. An edge of a tree T is a pair of nodes(U, V) such that U is the parent of V or vice versa.Goodrich, Tamassia and Mount (2011). A path of a tree T is a sequence of nodes in which any two consecutive nodes form an edge eg. Great-Grand-Father/Grand-Father/Father/Son.

A tree is a connected graph without any cycle, it is a connected acyclic graphGoodrich, Tamassia and Mount (2011). According to Balakrishnan and Ranganathan (2012), a family of a sub tree must satisfy the Helly property.
Let $r = \{T_i : i \in I\}$ be a family of sub trees of a tree T.
Suppose

$$\forall i, j \in J \subset I, T_i \cap T_j = \phi \qquad (1)$$

$$\text{To show that } \bigcap_{j \in \mathrm{J}} T_j \neq \phi \qquad (2)$$

If some tree $T_j \in r$, $i \in r$, is a single vertex tree
$\{V\}$ (that is, K_i), then clearly,

$$\bigcap_{j \in \mathrm{J}} T_j \neq \{V\} \qquad (3)$$

Assume each tree $T_i \cap T, i \in J$ has a minimum of two vertices. If this is true for all trees with at most n vertices:
Let T be a tree with $(n + 1)$ vertices
Let V_0 be an end vertex of T and U_0 be V_0's unique neighbor in T
Let $T_i' = T - V_0$, $i \in J$ and $T_i' = T - V_0$.

By induction hypothesis, the result is true for the tree T_i\`.
Also, $\boldsymbol{T_i'} \cap \boldsymbol{T_j'} \neq 0, \forall \boldsymbol{i}, \boldsymbol{j} \in \boldsymbol{J}$.

If T_i and T_j have a vertex $U(\neq V_0)$ in common, then T_i and T_j have U_0 also in common, so it is for , $\boldsymbol{T_i'}$and $\boldsymbol{T_j'}$, Hence by induction hypothesis $\bigcap_{j \in \mathrm{J}} Tj\grave{} \neq \phi$. Hence equation (2) is proofed.

3.4 Graph Signature

This work aims to enhance the search effectiveness of our data mart, through designing of a smart ranking function by generating a graph signature for each tuple. The signatures are then ranked in ascending order. The tuple will then be classified into a family through the nearest Top-K class that the given tuple belong. A signature for a

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

particular tuple by generating a canonical code for the given tuple was built. The canonical code can be defined by any feasible representation of the given tuple.Hsien, Hsu, Ti and Kuo (2014). In this case the canonical code is derived by concatenating some feasible features of the given node or tuple; which includes the state code(ST_{id}), local government code(LG_{id}),Family code(FM_{id}),Family generation level code(FG_{id}), and the individual node's personal code(PS_{id}). The Family generation level code is derived by concatenating the computed factorial from the first generation to the generation in which the individual tuple belong in the family lineage, using the expression:

$$FG_{id} = n! + (n-1)! + (n-2)! + \ldots\ldots.. + 1. \qquad (4)$$

For instance, if a tuple belongs to the third generation, its Family generation level code ($FG_{id)}$ will be:

$$3! + (3-1)! + (3-2)!$$
$$= 3! + 2! + 1$$

The FG_{id} = 6 + 2 + 1, = 621, where "+" is string concatenation. Therefore the signature of a given node V which belong to the n^{th} generation in the family lineage will be:

$$Sig(v) = ST_{id} + LG_{id} + FM_{id} + FG_{id} + PS_{id}. \quad (5)$$

where "+" is string concatenation.

FM_{id} and PS_{id} are unique numbers generated randomly by the computer.

3.5 The Proposed Random Forest Model

Breima (2001) is widely believed to have first come up with the idea of random forest in the early 2000. His work which combines several randomized decision trees and aggregate their pre-directions by averaging, has been considered to be successful as a general purpose classification and regression method.Biau and Scornet (2015). The method has been proved versatile enough to be applied to large scale bbb problems and are easily adapted to various ad-hoc learning task. However, mathematical analysis of the entire algorithm is difficult. A forest is a graph without a circle; a tree is a connected forest. The term random forest is a generic expression for amassing random decision tree, no matter how the trees are obtained (Biau and Scornet, 2015).

In this study, attention is focused on designing and realization of a novel machine learning algorithm for the purpose of generating and mining the identity of an individual from a large population, based on their family lineage. Specifically, effort is focused on overcoming the challenges of the traditional random forest algorithm to improve its overall performance and usability in solving a number of problems. These challenges are overcome by introducing a graph signature version of the algorithm. The proposed graph signature based random forest has shown capability in overcoming the difficult mathematical analysis challenges, computational complexity, scalability challenges, inefficient search and processing time. The Intelligent Identity system has shown potential to provide a usable, efficient, faster and accurate identification to the users consistently. The present identity system powered by Graph-Signature Based Random Forest (GSB-RF) model as well provides a theoretical justification for other machine learning and identity systems. Performance evaluation of our work shows that the GSB-RF identification model has a high degree of accuracy and speed in a very large database when compared with the baseline methods.

3.6 Overall Architecture of the Entire GSB-RF Citizeship Identification System

After the data collection, the extracted data were cleansed a database was created and the cleansed data was stored in the database, after the database creation, the proposed Canonical Code Graph Signature (CCGS) technique was used to generate signature for each node that represents each individuals in each of the family trees in the forest. Each node's signature was stored in a datamart created and implemented using MySQL DBMS application. When a new individual is to be identified, her identity code(Signature) is given, the GSB-RF search through the data mart, select signatures at random, the signatures are then ranked in order of closeness to the given/test signature. The signature will then be classified into the family of the nearest Top-K class that the given signature belongs using the proposed GSB-RF model. The result of the experiment was stored in a data mart developed and implemented with MySQL DBMS software. An in-house program was developed using PHP software with XAMP/Apache HTTP server as host to implement the proposed GSB-RF algorithm online and in real-time basis. The overall architecture of the entire Intelligent citizenship identity system is shown in Figure 1.

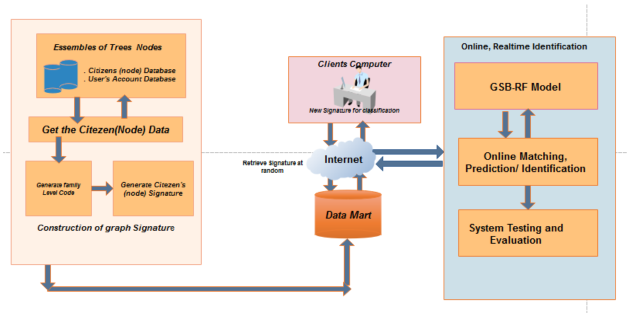
Essembles of Trees Nodes
. Citizens (node) Database
. User's Account Database
Get the Citezen(Node) Data
Generate family Level Code
Generate Citezen's (node) Signature
Construction of graph Signature
Clients Computer
New Signature for classification
Retrieve Signature at random
Internet
Data Mart
Online, Realtime Identification
GSB-RF Model
Online Matching, Prediction/ Identification
System Testing and Evaluation

Fig. 1:The overall architecture of the GSB-RF intelligent Citizenship identity system process flow

3.7 The Working of (GSB-RF) Model

A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(\mathbf{x}, \Theta_k), k=1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**. Breiman(2001) proposed a weighted forests in which the final prediction is the average of the individual tree results, the method was improved by incorporating tree-level weight to emphasize more accurate tree in prediction.

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

Given an ensemble of classifiers $h1(\mathbf{x}),h2(\mathbf{x}), \ldots ,hK (\mathbf{x})$, and with the training set drawn at random from the distribution of the random vector Y,**X**, define the margin function as *mg*(**X**,*Y*) ◎◎*avk I*(*hk* (**X**)◎*Y*)◎max*j*≠*y avk I*(*hk* (**X**)◎◎*j*).Where*I*(.•) is the indicator function. The margin measures the extent to whichthe average number of votes at **X**,*Y* for the right class exceeds the average votefor any other class. The larger the margin, the more confidence in theclassification.Growing an ensemble of trees and letting them vote for the most popular class have resulted in significant improvement in classification accuracy. Growing the tree ensembles often requires the generation of random vectors to govern the growth of each tree in the ensemble (Breiman, 2001). Originally the traditional random forest is an offline algorithmwhich required that the whole data set be given from the beginning. However, in this work, we proposed an online forest algorithmcalled signature based forest, in which the training data are generated over time and incorporated into the model. We first construct a graph signature for the given node; store it in a data mart while deferring classification until request is made to identify a tuple.

In our experiment, given an ensemble of family trees as shown in Figure 2., each tree consists of nodes representing each family member. Suppose our training database is restricted to the node described by the features ST_{id}, LG_{id}, FM_{id}, $FG_{id}$and PS_{id}. where, ST_{id} is the State identification code, LG_{id} is the Local government identification code, FM_{id} is the Family Identification code, FG_{id} is the family generation level code and PS_{id}., is the Personal identification code. The signature for a particular node can be derived by generating the canonical code for the node using equation 3.5, i.e. $ST_{id} + LG_{id} + FM_{id} + FG_{id} + PS_{id}$, where "+" is string concatenation. The family generation level code (FG_{id}) is derived using equation 3.4 i.e. $FG_{id}= n! + (n-1)! + (n-2)! + \ldots\ldots.+1$, FM_{id} and PS_{id} are unique numbers generated randomly by the computer system. The signature for each of the family tree's node in the forest are generated and stored in the data mart.

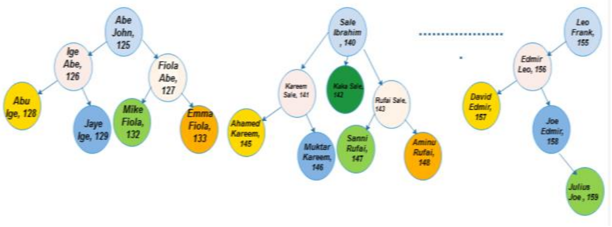

Fig. 2: An ensemble of family trees

In graph signature forest, we sort each nodes signature in ascending order. The identity of a given node is determined by comparing the node's signature with the list of the sorted node's signatures. The given node will then be classified into a family through the Top-K nearest tree class in the forest. The identity of a given individual is determined by the nearest Top-K signature to the test node. The GSB-RF assigned to the test node the topmost signature of its K-Closest training signature class. In this experiment, the value of K is taken to be 1 i.e. the K-1 rule is adopted. The algorithm for the proposed GSB-RF model is shown in Figure 3.

```
Input: n labeled graph $G_1$------ $G_n$.Output: node signatures database Sig($V_n$,$U_n$), Sig ($V_i$ ……$V_{nm}$)
Begin
    For K= 1 to m do
        For each node $V_i$ in $G_k$ do
            Get the feasible features of node $V_i$ ie. $ST_{id}$, $LG_{id}$, $FM_{id}$, $FG_{id}$and $PS_{id}$
            Compute $FG_{id}$ using the expression. $FG_{id}= n! + (n-1)! + (n-2)! + \ldots\ldots.+1$
            Compute the signature of node $V_{ik}$ using the expression:
             Sig ($V_{i,k}$) = $ST_{id} + LG_{id} + FM_{id} + FG_{id} + PS_{id}$, where "+" is string concatenation.
             Store Sig ($V_{i,k}$) and the feasible features in to the database/datamart.
        End for
    End for
    Input a random Sig of unknown classification( that is, given an individual signature /ID)
     Do until(Top-K, closest signature found)
    Retrieve available signature ($Sig_i$) at random without replacement from the data mart
    Let $Sig_i$ be a member of Top-K closest signature
    Else if ($Sig_i$ is closer to Sig than any previous Signature) Then
    Swap $Sig_i$ for the previous closest signature i.e add $Sig_i$ to the top of member of the K-closest
    signature
    End if
    Increment I by 1
    End do
    Classify Sig in the family of the Top-K signature
    End.
```

Fig. 3:. Algorithm Listing for the GSB-RF model

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

3.8 Application of the GSB-RF Model to Determine the Identity of an Individual Citizen

A population database which represent an ensemble of family tree was considered in figure 2. Each family member in the ensemble was represented as a node described by the following features, STid, LGid, FMid, FGidand PSid. To determine the identity of each node or individual citizen in the populace, the signature of each individuals has to be generated and stored in a data mart.

Given a particular citizen with Jaye Ige as name, EK as state of origin and MB as the local Government of origin with 36 and 05 as code for the state and local government respectively, and with personal ID, (PS_{id}) of 0133, family code 125 and belong to the third generation in the family lineage.We build the individual's signature by first deriving her family/Generation -ID i.e. (FG_{id}) by applying equation (4) i.e. FG_{id}= n! + (n-1)! + (n-2)! + ……..+1 = 3!+2!+1, = 6+2+1, = 621, where "+" is string concatenation.

Now, we can generate the signature for the given node (citizen) by generating the canonical code for the node using equation 3.5i.e.Sig ($V_{i,k}$) = ST_{id} + LG_{id} + FM_{id} + FG_{id} + $PS_{id,}$ where "+" is string concatenation. Therefore: Sig (Jaye Ige) = 36 + 05 + 125 + 621 + 0133, = 36051256210133. The signature of node/citizen with Jaye Ige as name will be:36051256210133, which can be represented on her plastic ID-card as EK-MB-125-621-0133. The whole process above will be repeated for all the given nodes of all family tress in the forest during registration and stored in a data mart.

Given a particular citizen **X_1's** identity card number/signature represented by the code: EK-MB-125-621-0133, which can be translated as: 36051256210133. Assuming the family of **X_1** is unknown. To determine the family of citizen **X_1** the GSB-RF model retrieves each stored signatures at random and sort them in order of closeness to the given signature. Table 1 shows the node's signatures sorted by closeness to the node **X_1**

Table 1. Data showing node's signatures sorted by closeness to node X_1

Node's Name	Family code	Closeness to node X_1
X_4	125	36051256220137
X_8	125	36051256210142
X_7	178	36051373210152
X_9	288	36081374210177
X_{10}	234	35051375210152
X_3	246	35051376210155

The GSB-RF approach, simply picks the node with the closest signature to node **X_1**, (i.e. the one from the top of the list) and use its family label with family code"125" to classify/identify the family that node **X_1** belongs. In the case the GSB-RF model predict **X_1** to belong to the family of node **X_4**.

4 System Evaluation and Analysis of Result

In this section, the novelties of our application were detailed with respect to the results of our experiment and related works. We evaluated the performance of the proposed system by applying the result of the conducted experiment. The quality of our citizenship identification system powered by GSB-RF model was evaluated by presenting the analysis of the experimental results.

4.1 Realization of the Intelligent Citizenship Identity System Using PHP Programming Language

To implement the present GSB-RF model, an in-house application using the PHP programming language was developed with XAMP/Apache HTTP server as hosting server and MySQL database for data mart creation. The system accepts data relating to an individual, build her signature by generating a unique identification code for the individual, stores it in a data mart. Whenever this signature is entered into the system for the purpose of establishing the identity of the individual, the system searches through the database, the GSB-RF algorithm is applied. A family with the topmost distance to the test signature will be predicted as the family the individual belongs. The online implementation of the of the system can be done by simply substituting the data mart with the real user's URL server database on the server. The code can then be incorporated as part of the National Identity registration website code and can be triggered when necessary by the users.

4.2 Performance Evaluation of the Proposed GSB-RF Model

Performance evaluation allows the measurement of accuracy, efficiency and effectiveness of a model. This can be used to measure its ability to correctly predict new data set in real life analysis, rather than the training data set in which the model was trained (Idowu *et al.*, 2015; Nsofor, 2006).

4.2.1. Dataset for Evaluation

Evaluating a predictive system are usually based on appropriate data sets, Melville and Sindhwani, (2010). To this effect, the present work adopted the internal data sets, we applied the extracted Nigerian 2006 population census data from the Nigerian population commission archive as described in section 3.1.1. The extracted records were divided into five parts each, four parts were used as the training sets and the remaining one part was used as testing set. The families of the testing set was considered unknown, while the families of the training sets was considered known and are used to determine the families of the unknown.

4.2.2 Evaluation Metrics

In this study, an offline method of evaluation saw adopted and an F-Measure evaluation technique was adopted to evaluate the predictive quality of the proposed GSB-RF model. The F-measure according to Rijsbergen (1979), is the harmonic means of precision and recall.

$$\text{Precision (P)} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \qquad \text{Recall (R)} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

$$\text{F-measure (F)} = 2.\frac{\text{Precision.recall}}{precision+recall}$$

Since recall and precision are weighted, it is referred to as the F1-Measure. Figure 4 shows our experimental result in F1-Measure using our data sets.

The proposed GSB-RF predictive algorithm and the baselime methods which are the Traditional Random forest and the K-Nearest Neighbour techniques was run for over sixty times each using the developed PHP web application, the results of which was used for the purpose of evaluation of the present system. The results was recorded for the identified/relevant prediction, not identified/irrelevant prediction in all the cases, False negative(fn) i.e. the actual positive but identified as negative True positive(tp) i.e. the actual positive and identified as positive False positive (fp) i.e. the actual negative but identified as positive, and True negative (tn) i.e. the actual negative and identified as negative.The precision rate and the the F1-Measure are computed, the run-time of our system and the baseline methods waw also recorded at different length of identification in order to determine the execution speed of the proposed GSB-RF algorithm. The result of the experiment is shown in Figures 4 and 5.

4.3 Discussion

The result of this experiment shows an excellent performance of the proposed intelligent citizenship identity system powered by the Graph Signature Based Random Forest (GSB-RF) model. Performance comparison between the current system and two other base line methods that is, the traditional Random Forest(RF) and the K- nearest neighbour (KNN) was conducted using our data sets. The results shows a significant difference between the performance of the three algorithms studied. But in all cases the present Graph Signature Based Random Forest, Intelligent Citizenship Identity model shows a higher degree of accuracy in comparison with the baseline method as shown in Figure 4. It was established through the experimental results that in general, the F1-Measure values increases rapidly up to apeak point after which it goes down slightly in all cases. The quick increase indicates that precision is nearly steady while recall increases; the slight reduction implies that the precision decreases while recall is almost constant.

The experiment was conducted for about 60 different length of prediction, the GSB-RF predictive system has over 77% F1-Measure value for all sample identification/prediction length between 1 and 15 and thereafter maintain about 72% F1-Measure at longer prediction/identification length as shown in figure 4. Lowest F1-Measure are deteated from the K-NN when used on our experimental data set as shown in figure 4. The K-NN performed poorly compared to other methods. The traditional RF performed well at shoter length of prediction, but the performance rapidly decreases at long length of identification as shown in figure 4, therefore leads to a limited number of useful identification.

In order to further substantiate the excellent performance of the present GSB-RF Citizenship identification system, the run time of each algorithms was recorded at different length of identification so as to determine the speed of execution of each algorithm under the same experimental settings. The comparison of the run-time of the present GSB-RF with baseline methods i.e the traditional RF and the K-NN shows that the GSB-RF executes faster than the baseline methods as shown in figure 5. It was established that the run time of traditional RF and the K-NN methods increases rapidly as the length of identification increases while that of the present GSB-RF run time was stable initially at low length of identification, then increases a little at length greather than 25 ,then becomes stable as the length of identification increases.

The result indicates that the application of the GSB-RF model can lead to a more accurate and efficient citizenship identification process.The experimental results therefore shows that the present GSB-RF model remains accurate, efficient and appropriate predictive technique for this study. The low runtime of the proposed model makes it to drastically reduce computational loads. The application of the GSB-RF technique has in no small measure makes our system to achieve excellent results when it comes to a large data set and longer length of predictions and identifications. Finally, the GSB-RF prediction/identification model is capable of producing an accurate, scalable, efficient, faster and consistent prediction and identifications, especially in domain with large volume of data sets and at a very difficult predictive task like the citizenship registration and identification systems.

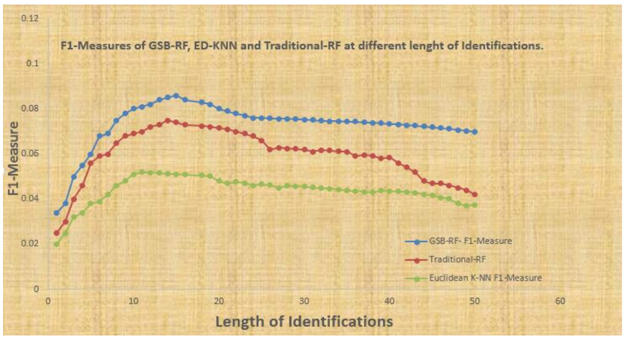

Fig. 4: F1-Measures of GSB-RF, ED-KNN and Traditional-RF at different lenght of predictions/ identifications

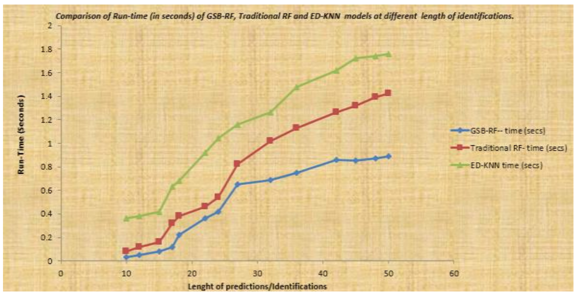

Fig. 5: Run-time (in seconds) of GSB-RF, Traditional RF and ED-KNN models at different length of Predictions identifications.

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

5CONCLUSION AND RECOMMENDATIONS

5.1 SUMMARY

This work is designed mainly to develop a simple, straightforward, robust, accurate, faster, flexible, scalable, transparent, easy to understand, easy to implement and consistentcitizenship identification system. This is achieved through the development of a novel predictive algorithm called Graph Signature Based Random Forest (GSB-RF). This study demonstrated that the proposed GSB-RF method can be effectively used to perform citizenship registration and identification through their family pedigree.The present designed system is a proof-of-concept, prototype of idea for using Graph Signature Based Random Forest technique to model Intelligent Citizenship identity through family pedigree.

This is aimed at assisting the security and government agents to easily establish the identity of any individual citizens through their family lineage and to assist them in detecting crime and criminals in the society, so as to set priority for security of lives and properties. The system also aimed at assisting the designer and administrators of citizenship identity and population management systems to have more varieties of algorithms from which the best performing algorithm can be selected and to improve their websites, especially the Nigerian National Population Commission website.

This is achieved by building signatures for any given individuals during registration through their family lineage. The signature is built by generating a canonical code for the individual which serves as basis for carrying out the identification of the individual. To achieve this, raw citizens bio data were collected from the archive of the Nigerian Population Commission, stored it in the system, the proposed canonical code graph signature technique was used to generate signature for each individual and store it in the data mart created. When a new individual citizen signature is given from her ID-card for identification purpose, the GSB-RF search through the data mart, select signatures at random, the selected signatures are then ranked according to their closeness to the given/test signature. The given signature will then be classified in to the family of the nearest Top-K signature in the list.

An in-house PHP program was developed to implement the GSB-RF predictor on an experimental website. The findings of the experimental study can now be used by the designers and administrators of the Nigerian population commission to plan the upgrade and improvement of their website. The rationale behind the selection of the GSB-RF predictor was achieved. The results shows that the GSB-RF Citizenship Identification engine can produce a more accurate, usable, efficient, faster and useful citizenship identification to the users, security agents, government agents and institutions online, real-time consistently.

5.2 CONCLUSION

This work focused on the application of a novel approach to the design and realization of an Intelligent Citizenship identity through family pedigree in a practical way; through the use of an analytical machine learning algorithm called Graph Signature Based Random Forest (GSB-RF) model. The system is capable of accepting information relating to an individual citizen, builds her signature by generating a unique identification code for the individual using the proposed canonical code signature (CCS) technique, stores it in a data mart and generates a plastic ID-Card for the individual.

The information on the ID-Card can be used by any security or government agent to establish the identity of the individual by entering her ID-Number, which serves as her signature to the system. The system applied the GSB-RF algorithm to search through the data mart, select signatures at random, the selected signatures are then ranked according to closeness to the given test signature. The test signature is then classified into the family of the nearest Top-K signature to the given signature. The result of our experiment shows that the Intelligent Identity engine powered by the GSB-RF model is capable of handling a large data sets, overcome computational inaccuracy, search inefficiency, long processing time and scalability challenges. The experimental result also shows that the precision rate and identification quality of the proposed identity engine in most cases are equal to or greater than 70%. Thus, the present system is capable of providing usable, faster, accurate, efficient identification and continuous re-identification of citizens consistently, online, real-time and at any time.

5.3 RECOMMENDATIONS FOR FUTURE WORK

The researchers are of the opinion that this study could be advanced further by introducing a biometric authentication technique to the identity system alongside the proposed Graph signature method. This is in order to make the system less vulnerable to technical weakness and potential misuse or abuse.

REFERENCES

Adeniyi D.A, Wei Z & Yongquan Y (2016) Automated web usage data mining and recommendation system using K-Nearest Neighbor(KNN) classification method. Journal of Applied Computing and Informatic. Vol.12, pp. 90–108, https ://doi.org/10.1016/j.aci.2014.10.001.

Adeniyi D.A, Wei Z & Yang Y, (2018), Risk Factors Analysis and Death Prediction in Some Life-Threatening Ailments Using Chi-Square Case-Based Reasoning (χ2 CBR) Model. Interdisciplinary Sciences: Computational Life Sciences. Volume 10, Issue 4, pp 854–874,https://doi.org/10.1007/s12539-018-0283-6.

Balakrishnan,R. & Ranganathan,K.(2012).A Textbook of Graph Theory.Springer. ISBN 978-1-4614-4529-6

Biau G, & Scornet E (2015). A random forest Guide Tour.,arxiv:1511.05741v\[maths.ST] 2015, pp. 1–42

Breiman L. (2001) Random forests. Machine Learning, vol. 45, pp. 5–32

Goodrich M.T., Tamassia R. & Mount M.D. (2011). Data Structures and Algorithms in C++, Second Edition. John Wiley & Sons Inc.

Hocking C.G., Furnell S.M., Clarke N.L., & Reynolds P.L. (2013). Co-Operative users identity verification using an Authentication Aura. Journal of Computers & Security.Vol. 39, pp. 486-502. http://dx.doi.org/10.1016/jcose.2013.09.011.

http://dx.doi.org/10.46792/fuoyejet.v6i2.609

Hssina B, Merbouna A, Ezzikouri H, & Erritali M. (2014). A Comparative study of decision tree ID3 and C4.5. International Journal of Advanced Computer Science Applications. Special issue on Advance in Vehicular Ad Hoc Networking and Applications. https://doi.org/10.14569/Speci alIssue.2014.04020 3

Hsien S., Hsu C., Ti Y., & Kuo C. (2014). Reducing the bottleneck of graph based data mining by improving the efficiency of labeled graph isomorphism testing. Journal of Data and Knowledge Engineering. pp. 17-33. http://dx.doi.org/10.1016ij.datak.2014.02.003

Hurwitz J., & Kirsch D.(2018). Machine learning for dummies. IBM Limited Edition. John Wiley & Sons Inc

Idowu, P.A., Williams, K.O. Balogun, J.A. & Oluwaranti, A.I. (2015). Breast Cancer Risk Prediction using Data mining classification Techniques. Transactions on Networks and Communications (TNC), Society for Science and Education, UK. Vol. 3, Issue 2, ISSN: 2054-7420. Doi: 10.14738/tnc.32.662.

Jouli S. &Tabbone S. (2009). Graph matching based on node signatures. 7th IAPR-TC-15 workshop on Graph-based representation in pattern recognition. Venise, Italy,Springer, 5534. pp. 154-163.

Khan G.A. (2014). Building robust identification system. Social protection and labor: Learning forum 2014. National Database and Registration Authority (NADRA), Ministry of interior, government of Pakistan.

Lewu N.D.(2017). Machine learning made easy with R. ISBN-13: 978-1546483755

Luis-Garcia R. Alberola-Lopez C., Aghzout O. & Ruiz-Alzola J. (2003). Biomertic identification systems. Journal of Signal Processing, 83, pp. 2539-2557. DOI: 10.1016j.sigpro.2003.08.001.

Melville, P., & Sindhwani, V. (2010). Recommender System. IBM T.J, Watson Research centre, pp. 1-18.

Nguyen T.(2003). National identification system. M.Sc. Thesis, Massachusetts Institute of Technology.

Nsofor, G. C. (2006). A comprehensive Analysis of predictive data mining techniques. M.Sc. Thesis, The University of Tennessee, Knoxville.

Purvis T., & Hunt A.(1999). Identity versus Citizenship: Transformations in the Discourses and Practices of Citizenship.*Social and Legal Studies*; SAGE Publications8; 457, DOI: 10.1177/a010358.

Rijsbergen, C.V., (1979). Information retrieval. London; Boston, Butter-worth, 2nd edition. ISBN: 0-408-70929-4.

Rovisco M., & Lunt P. (2019). Introduction: Performance and Citizenship. International Journal of Cultural Studies. Vol 22(5), pp. 615-629. Doi.org/10.1177/1367877919849965.

Saleh M., Ratazzi E.P., & Xu S. (2017). A control flow graph-based signature for packer identification. IEEE Milcom 2017, Track 3-Cyber security and Trusted computing. Pp. 683-688.

Singh A, & Das K.K. (2007). Application of data mining Technique in Bioinformatics. Dissertation, Department of Computer Scienceand engineering, National Institute of Technology, Rourkela.

Smola A., & Vishwanathan S.V.N.(2008). Introduction to machine Learning. Cambridge University press, Cambridge, UK. ISBN: 0 52182 583 0.

Wong P.C., Foote H., Chine G., Mackey P., & Perrine K.(2006). Graph signature for Visual analytics. IEEE Transactions on visualization and computer graphics. Vol , 2 no 6. pp. 1399-1413.

Yara R., Nadine H, Nicholas M., & Mariette A.(2019). Deep belief networks and cortical algorithms: A comparative study for supervised classification. Applied Computing and Informatics Vol.15, pp. 81–93.

http://dx.doi.org/10.46792/fuoyejet.v6i2.609