2013 Sixth International Conference on Developments in eSystems Engineering

Novel Approach for Generating the Key of Stream Cipher System Using Random Forest Data Mining Algorithm

Samaher Hussein Ali
Department of Information Network
Faculty of Information Technology, University of Babylon,
Hilla, Iraq (00964)
samaher@itnet.uobabylon.edu.iq

Abstract— **Key generator is part of the stream cipher system that is responsible for generating a long random sequence of binary bits key that used in ciphering and deciphering processes. Therefore, key generator is the heart of the stream cipher system. A system with traditional key generating techniques such as using the feedback shift register is vulnerable to cypher attacks which render it to be ineffective and insecure. A suggested novel method to overcome this problem is to use a Random Forest-Data Mining (RF-DM) algorithm. It incorporates using new types of linear and nonlinear functions to mix the plain text with the key in the encryption process and the cipher text with the reverse key in the decryption process. The method is successful in satisfying the secrecy goal of the stream cipher system because it uses two types of dual randomization to baffle the attacker or cryptanalysist. By mathematical logic and prove by experiments, we generate the key successfully by encrypting messages of different sizes . This proves the ablity to generated the unique key for each message based on its length without repeating the key in most cases. In addition, by combining the use of the random forest data mining technique as the randomizating principle in the ciperhing process makes it exteremely difficult for any attacker because the attacker does not have access or understanding of the not only the technique used but also the process to decipher. The result from our experiment shows that the longer the number of the m*essage size, the longer is the length of the generated key, which determines its strength and hence, its complexity to break by brute force.***

Keywords: *Random forest algorithm, data mining, stream cipher, key generator, linear and nonlinear functions, security, attacks, cryptanalysis*.

I. INTRODUCTION *(HEADING 1)*

Cryptography is the study of secret (crypto) writing (graphy). Cryptography is the art or science encompassing the principles and methods of transforming an intelligible message into one that is unintelligible, and then retransforming that message back to its original form. it consist of: (i) cipher an algorithm for transforming an intelligible message into one that is unintelligible by transposition and/or substitution methods, (ii) key some critical information used by the cipher, known only to the sender & receiver, (iii) encipher the process of converting plaintext to cipher text using a cipher and a key, and (iv) decipher the process of converting cipher text back into plaintext using a cipher and a key .

The main requirements of any cryptography system are:

1. The encryption and decryption transformation must be efficient for all keys.
2. The system must be easy to use.
3. The security of the system must depend only on the secrecy of the key and not on the secrecy of the algorithm of the encryption /decryption processes.
4. It should be computationally infeasible for a cryptanalyst to determine the deciphering transformation from intercepted ciphertext, even if the corresponding plaintext is known.
5. It should be computationally infeasible for a cryptanalysis to determine the plaintext from the interpreted ciphertext, in addition to providing confidentiality, cryptography is often asked to do other jobs.
6. It should be possible for the receiver of a message to authentication its origin.
7. It should be possible for the receiver of a message to verify its integrity that it has not been modified in transit.
8. A sender should not be able to nonrepudiation by falsely deny later that he/she did not sent the message

Encryption is the science used to maintain the secrecy and confidentiality of the information within a data stream. Encryption ensures that data can pass between communicating parties in a transparent unhindered manner and render the efforts of attcakers to be useless. The main objectives of encryption (access, confidentiality, integrity, authentication). Given the importance of secure information communications, especially with the growing use of the World Wide

978-1-4799-5264-9/14 $31.00 © 2014 IEEE
DOI 10.1109/DeSE.2013.54

Web and numerous business applications. Which impact on the development combination data mining algorithm (CDMA) science because of hackers and curious to penetrate some of the information for cryptography in a constant state of development.[1]

Many paper disuss the stream cipher system, [2] it has two drawbacks: First, it limits the possibility to detect errors when decrypting. Second, an attacker can insert controlled changes to parts of the cipher text and may achieve a wanted modification of the plaintext. [3] It proposed an efficient stream cipher algorithm which generates 23 random bits in each
round of processing by parallel random number generator and 115 bits of Initial Vector
remind some of problems in the stream cipher system, ***Firstly;*** the length of key in the LFSR or NLFSR determin by the number of satates that contain binary random valuses, ***secondly;*** the number of states in both LFSR or NLFSR have limited number of successful connection and ***thirdly;*** both registers can generation the key and given secrit results if the message is short or devided in to multi sub messages but the results are become easy broken with the long messages because the key is repeat more than one and the logic of encreption is known for all. Therefore, in this paper, we attempt to satisfying the secrecy goal of the stream cipher system because it uses two types of dual randomization to baffle the attacker or cryptanalysist. In addition, we hope to generate the unique key for each message based on its length without repeating the key in most cases.

The remained of this paper is organization as follow: section II presents the methodology and techniques used. Section III explains the novel methodology used in predicate. Each step in novel methodology has been explained and analyzed extensively. Section IV Illustrates the implementation of the methodology and the results. In addition, Section V explain the main assumption and limitation of this work. Section VI shows conclusions of this work together with some recommendations for future work in this field.

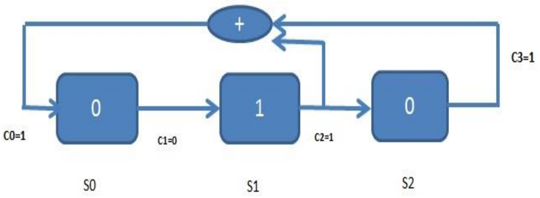
II. METHODOLOGIES AND TECHNIQUES USED

A stream cipher is a system in which the key is fed to an algorithm, which uses the key to generate a finite sequence. The algorithm is usually referred to as the sequence generator or key stream generator. Stream ciphers should possess the following characteristics [6]: (i) Easy to implement. (ii) High speed in generating key stream. (iii) Computationally secure and difficult to break it, even against known plaintext attack with a large key space, and the time period of the key bit string sequence should be greater than the message length. (iv) Key sequence consist of random characteristics to overcome efforts by attackers.

In general the following two examples explaine how the stream cipher system works by the tradition linear and nonlinear feedback shift register:

Example 1: Let the initial State of Sift Register is {0, 1, 0} and Feed Back coefficient 1011 find States, then cipher the massage GOOD

Sol:



Max Length of Key = $2^{n}-1$

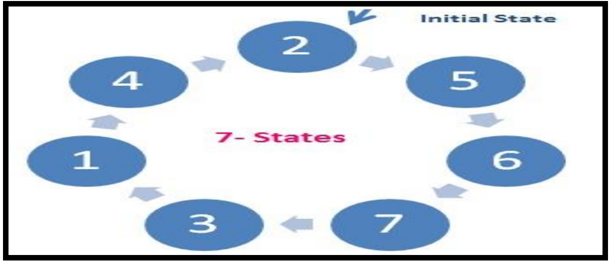
Number of Correct Connection = $\frac{\phi(2^{n}-1)}{n}$

Where; n number of states Max Length of Key

=(2^3) -1 = 7

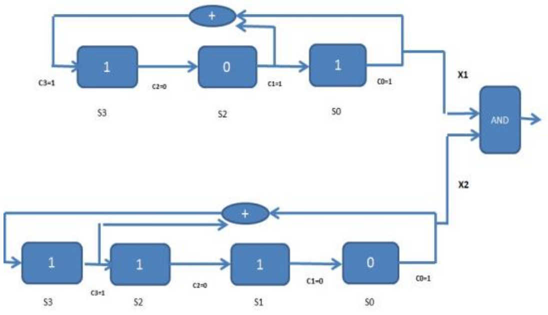
No	FB	S0	S1	S2	Output
		0	1	0	
1	1	1	0	1	0
2	1	1	1	0	1
3	0	1	1	1	0
4	0	0	1	1	1
5	1	0	0	1	1
6	0	1	0	0	1
7		0	1	0	0

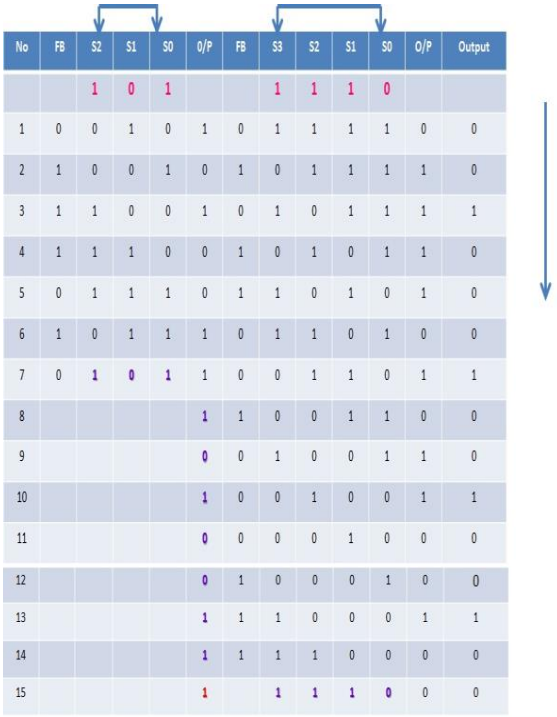
Key = 0101110



Message = GOOD
G = 7 = 0111
O = 15 =1111
O = 15 = 1111
D = 4 = 0100
Plain Text = 0111 1111 1111 0100
Key = 0101 1100 1011 1001

Example 2: Let the nonlinear FBSR consist of two registers the first have three states while the second have four states, initial State of the first Sift Register is {1, 0, 1} and Feed Back coefficient 1011 while, initial state of the second sift register is {1, 1, 1, 0} and Feed Back coefficient 11001 find the key generation using AND LOGIC:





No	FB	S2	S1	S0	O/P	FB	S3	S2	S1	S0	O/P	Output
		1	0	1			1	1	1	0		
1	0	0	1	0	1	0	1	1	1	1	0	0
2	1	0	0	1	0	1	0	1	1	1	1	0
3	1	1	0	0	1	0	1	0	1	1	1	1
4	1	1	1	0	0	1	0	1	0	1	1	0
5	0	1	1	1	0	1	1	0	1	0	1	0
6	1	0	1	1	1	0	1	1	0	1	0	0
7	0	1	0	1	1	0	0	1	1	0	1	1
8					1	1	0	0	1	1	0	0
9					0	0	1	0	0	1	1	0
10					1	0	0	1	0	0	1	1
11					0	0	0	0	1	0	0	0
12					0	1	0	0	0	1	0	0
13					1	1	1	0	0	0	1	1
14					1	1	1	1	0	0	0	0
15					1		1	1	1	0	0	0

Key = 00100 01001 00100

The main quations here are:

1. **Can stream cipher use block cipher?** Stream cipher can used as a block cipher with efficiently fixed small sized sub-messages to comply with real time applications. That is a consequence of the sort time interval available to process a sub-message, e.g. A5 of GSM cellular telephony.
2. **Can stream cipher use as secrit cipher method with along message?** Stream cipher can't used as a secrit cipher method with along message for many reason (i.e., The main weaknesses of stream ciphers) [6] and [10]: (i) No integrity. (ii) Known-plaintext attack is very dangerous. (iii) Key stream is ever repeated.
3. **Can stream cipher generate a unique key for each message?** No, because the leght of key was determined by the intial value in the states and number of correct conections.

To avoid the above problems and make the stream cipher system is efferent again; we must find new tool to generation the long and unique key for each message. This work, attempt to useful of data mining techniques in generation that key base on the same principle and architecture of steam cipher in encryption and decryption process.

Data Mining (DM) is not just a single method or single technique but rather a spectrum of different approaches, which searches for patterns and relationships of data [1]. But in this work, DM is used to search for long unique binary sequence that it represented the key of developed stream cipher system. The question is: ***Which technique from data mining techniques can satisfy this goal?*** In general the data mining techniques can perform the following tasks: clustering, classification, generation association rules and forcasting (i.e., predication). In this work, the main task is forcasting of the key(i.e., generation sequence of continuous binary values using as a key). Therefore, we deal with the regression random forest data mining algorithm.

We can define the Random Forest (RF) as collection of classifier that involves of multi Decision Trees (DTs). In general each data set divided in two subsets, the first take 2\3 of the total dataset called bagging while, the second subset takes the remind 1\3 of the total datasett called out-of bag. Non-Parametric Random Forests croups of trees increased from bagging data. ***In classification process***, the trees are jointed by widely polling with one poll for each tree completed all the trees in the forest. ***For regression***, forests are formed by dive number of polling for each group on the total number of polling for all tree, Figure 1 shows how the RF working. For more dtials about this tool see [7][8].

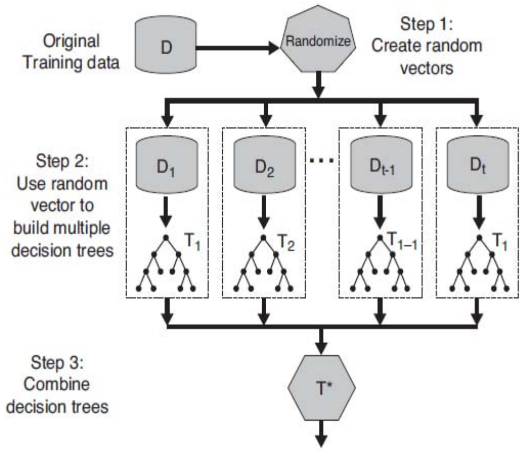


Fig 1: Random Forests [7]

As a result, random forests can given results more effientity than any other single classifier or regression tree [2].

In general, RFs deal with ***two kinds of randomness***. *First* in the process of selection a random sample of predictors while the other randomness is satsify by utilizing a random sample for increase the size of the each tree.

The main highlight points of random forests described down:

1. RF is consider one of the Machine Learning (ML) algorithms that provide the accurate results for many data bases, in addition, it constracts a most accurate classifier..
2. RF ecexiations usefully on huge data setes.
3. RF can treatments thousands of inter variables not including variable loss.
4. It provides predication of the all-important variables in the classification.
5. It generates an internal unbiased estimate of the generalization error as the forest building progresses.
6. RF can be considering as effective methodology solves the missing value problem, and upholds accuracy if a large amount of the values are missing [3].
7. In the case of, we have unbalance data sets need to classification, Rf is sutible tool because it provides balancing error.
8. The relation between the variables and the classification can be find through design the models that give information describe these relation.
9. “RF calculates nearness between pairs of cases that can be used in clustering, locating outliers or (by scaling) give interesting views of the data” [4]. These capabilities can be spread to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
10. RF can be used as a trial way for discovering variable interaction [5].

Tenfold Cross Validation (TCV) is one of the verification method used to ensure that the whole database is tested to determine the fregments will be used as training the dataset for subsequent use and testing the dataset.

In this paper, we attempt to solve the main drawbacks of the stream cipher system explained above by combination between the advantage of data mining represent by random forest and cryptography represented by the main principles of stream cipher system as explaining in the next section.

III. The Novel Methodology for Cipher System

The main idea of this research is to demonstrate or prove how we can find the novel methodology to generate a long unique binary sequence of bits using as a key of stream cipher system. (i.e., tool is satisfied the same purpose of feedback shift registers in stream cipher which solve the main weakness points of that system it). In addition, the novel methodologies satisfy the same purpose and it consider more robust and confidence compare with a ***“Linear Feedback Shift Register (LFSR)”*** or “***Nonlinear Feedback Shift Register (NLFSR).***

For three reasons the main three weakness points of LFSR and NLFSR are:

1. The length of key in the LFSR or NLFSR determin by the number of states that contain binary random valuses,
2. The number of states in both LFSR or NLFSR have limited number of successful connection; and
3. Both registers can generation key given secrit results if the message is short or devided in to multi sub messages but the results are become easy broken with the long messages because the key is repeat more than one and the logic of encreption is known for all.

To ensure the key used in the decipher process is true, we base on the tenfold cross validation procedure in that procedure separated the dataset into ten equal fragments. After that, a tree is constracted using 90 % of the total data (that called ‘training set’) and checking base on the remining 10% of the total data(that called’testing set’). Then the procedure continouse ten by builing another tree by the same method but it using the different rate of training and testing data sets

The following, Table (1): presents the main linear and nonlinear functions used to test the methodology. Followed by the main two procedures used in the encryption and decryption phases. In addition, Figure 1 shows the main architure of the novel methodology. Figure 2 presents the block diagram of the cipher process and Figure 3 explains the block diagram of decipher process.

TABLE 1: THE MAIN LINEAR AND NON LINEAR FUNCTIONS

Name of Functions	# Variable	Functions
Linear	Two	F(Y)=P1+P2*Message+P3*Key
Quadratic	Two	F(Y)=P1+P2* Message +P3* Message ^2+P4* Key +P5* Key ^2+ P6* Message * Key
Product	Two	F(Y)=P1+P2* Message * Key
XOR	Two	F(Y)= Message Xor Key
AND	Two	F(Y)= Message And Key
OR	Two	F(Y)= Message OR Key

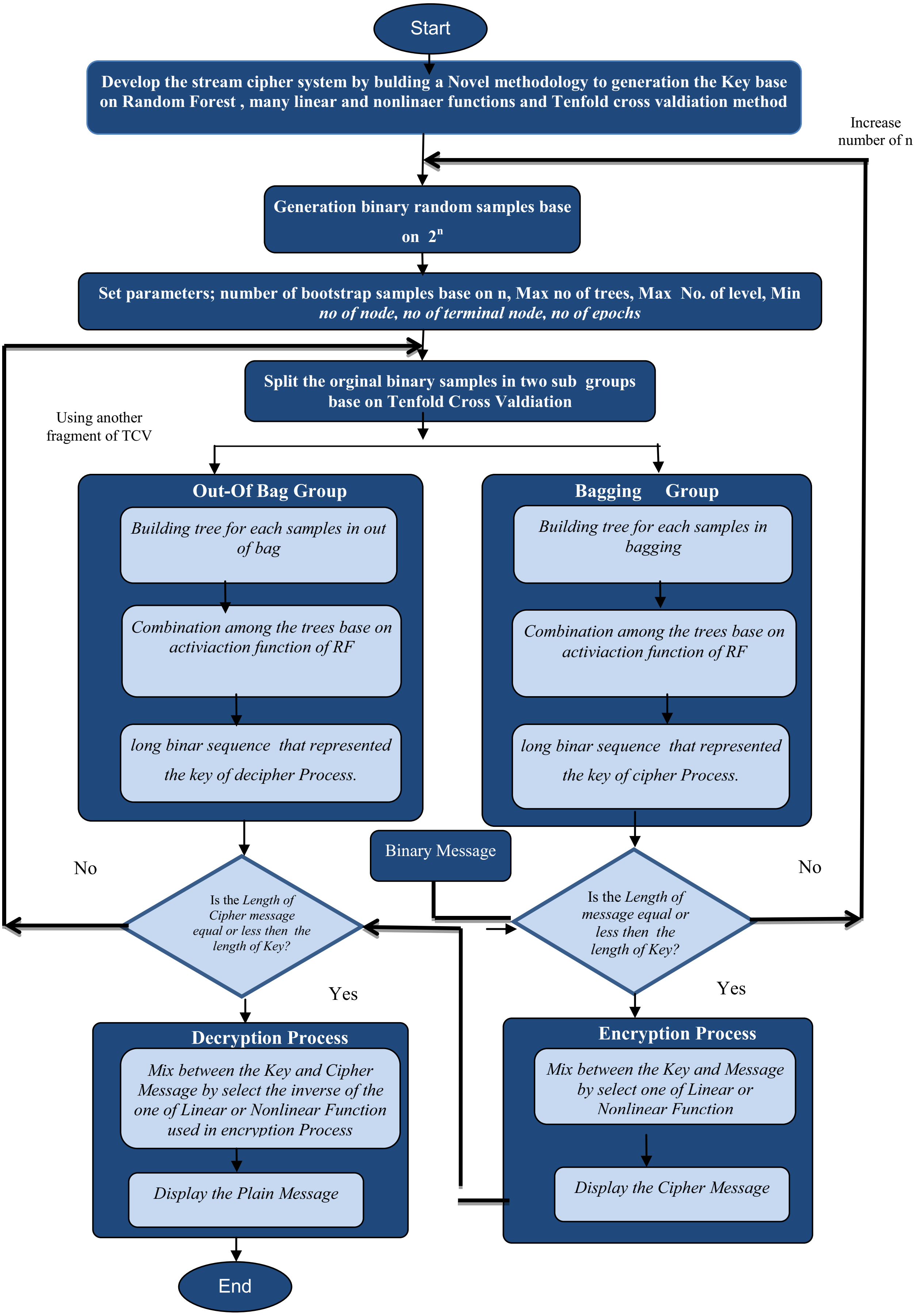

Start
Develop the stream cipher system by bulding a Novel methodology to generation the Key base on Random Forest , many linear and nonlinaer functions and Tenfold cross valdiation method
Increase number of n
Generation binary random samples base on 2^n
Set parameters; number of bootstrap samples base on n, Max no of trees, Max No. of level, Min no of node, no of terminal node, no of epochs
Split the orginal binary samples in two sub groups base on Tenfold Cross Valdiation
Using another fragment of TCV
Out-Of Bag Group
Building tree for each samples in out of bag
Combination among the trees base on activiaction function of RF
long binar sequence that represented the key of decipher Process.
Bagging Group
Building tree for each samples in bagging
Combination among the trees base on activiaction function of RF
long binar sequence that represented the key of cipher Process.
Binary Message
No
Is the Length of Cipher message equal or less then the length of Key?
Yes
No
Is the Length of message equal or less then the length of Key?
Yes
Decryption Process
Mix between the Key and Cipher Message by select the inverse of the one of Linear or Nonlinear Function used in encryption Process
Display the Plain Message
Encryption Process
Mix between the Key and Message by select one of Linear or Nonlinear Function
Display the Cipher Message
End

Figure 1: Archticture of the the novel approach for generating the key of stream cipher system using random forest data mining algorithm

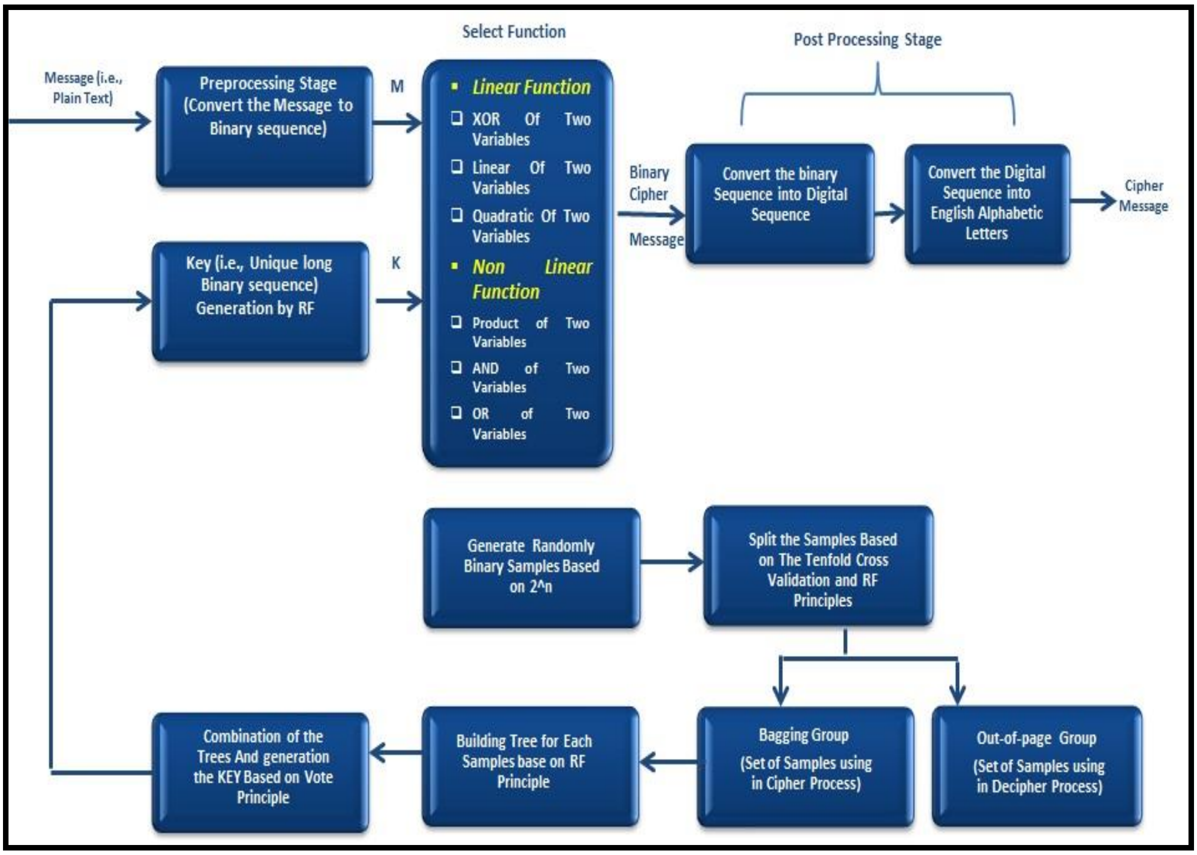

Select Function
Post Processing Stage
Message (i.e., Plain Text)
Preprocessing Stage (Convert the Message to Binary sequence)
M
▪ Linear Function
❑ XOR Of Two Variables
❑ Linear Of Two Variables
❑ Quadratic Of Two Variables
▪ Non Linear Function
❑ Product of Two Variables
❑ AND of Two Variables
❑ OR of Two Variables
Binary Cipher Message
Convert the binary Sequence into Digital Sequence
Convert the Digital Sequence into English Alphabetic Letters
Cipher Message
Key (i.e., Unique long Binary sequence) Generation by RF
K
Generate Randomly Binary Samples Based on 2^n
Split the Samples Based on The Tenfold Cross Validation and RF Principles
Combination of the Trees And generation the KEY Based on Vote Principle
Building Tree for Each Samples base on RF Principle
Bagging Group (Set of Samples using in Cipher Process)
Out-of-page Group (Set of Samples using in Decipher Process)

Figure 2: Architectural block diagram of the ciphering phase

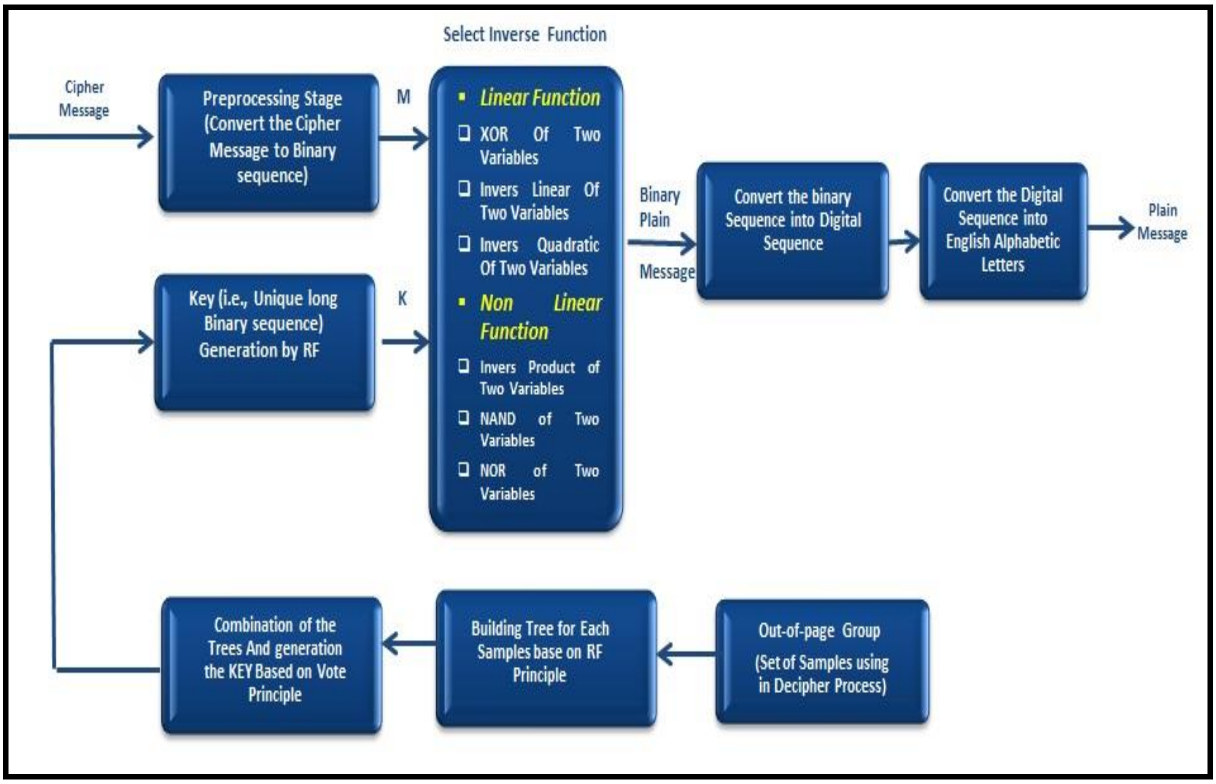

Select Inverse Function
Cipher Message
Preprocessing Stage (Convert the Cipher Message to Binary sequence)
M
▪ Linear Function
❑ XOR Of Two Variables
❑ Invers Linear Of Two Variables
❑ Invers Quadratic Of Two Variables
▪ Non Linear Function
❑ Invers Product of Two Variables
❑ NAND of Two Variables
❑ NOR of Two Variables
Binary Plain Message
Convert the binary Sequence into Digital Sequence
Convert the Digital Sequence into English Alphabetic Letters
Plain Message
Key (i.e., Unique long Binary sequence) Generation by RF
K
Combination of the Trees And generation the KEY Based on Vote Principle
Building Tree for Each Samples base on RF Principle
Out-of-page Group (Set of Samples using in Decipher Process)

Figure 3: Architectural block diagram of the deciphering phase

Pseudo Code of Encryption Process

```
Input: Plain Text ( i.e., plain message with any length)
Output: Cipher Text (i.e., cipher message)
• Step 1: Population initialization through generation binary random samples base on $2^n$
• Step 2: Set parameters; number of bootstrap samples base on n, Max no of trees, Max No. of level, Min
          no of node, no of terminal node, no of epochs
• Step 3: Split the orginal binary samples in two sub groups, first sub group called bagging and other sub group
          called out-of bag .
• Step 4: Pass the first sub group "bagging" that take 2 or 3 from the total number of binary samples to cipher
          phase.
• Step 5: Building tree for each samples in bagging sub group.
• Step 6: Combination among the trees base on activation function of RF. The result of combination is long binary
          sequence that represented the key.
• Step 7: Covert the plain text to binary text
• Step 8: If the length of binary sequence gererated by step 5 less than length of binary plain text:
      ▪ Goto step 1 and increase n.
      ▪ Else Goto step 9
• Step 9: Select one of the linear or nonlinear functions to mix between binary plain text and binary sequence
          generated by step 6:
              Ci=(ki linear function Pi)mod 26
                          or
              Ci=(ki nonlinear function Pi)mod 26
• Step 10: End Encryption Process
```

Pseud code of Decryption Process

```
Input: Cipher Text (i.e., cipher message)
Output: Plain Text ( i.e., plain message)
• Step 1: Pass the second sub group" out-of bag" that take 1/3 from the total number of binary samples to
          decipher phase.
• Step 2: Building tree for each samples in out-of bag sub group.
• Step 3: Combination among the trees base on activation function of RF. The result of combination is long
          binary sequence that represented the key.
• Step 4: Covert the cipheir text to binary text
• Step 5: If the length of binary sequence generated by step 3 less than length of binary cipher text:
      ▪ Repeat the key to become equal the length of cipher text
      ▪ Else use the key without repeat
```

- ***Step 6**: Select reverse of the linear or nonlinear function used in encryption phase to mix between binary*
    - *cipher text and binary sequence generated by step 3.*
- ***Step** 7: Covert the binary results to digit character bas on the total number of English alphabetic*
    - *Pi=(ki reverse linear function Ci)Mod 26*

      or
    - *Pi=(ki reverse nonlinear function Ci)Mod 26*
- ***Step 8**: Covert the digit to the character forward in English alphabetic. The result sequence of character that represent the plain text.*
- ***Step 9:** End Decryption Process*

IV. EXPERIMENTS AND RESULTS

As we explained previously, the generate of key in the stream cipher system using novel methodology " random forest data mining algorithm" consists of three phases:

- ***A preprocessing phase*** that building the binary random samples.
- ***An encryption phase*** that building the key base on RF from bagging sub group and selects one of the linear or nonlinear fuction to mix between the key and binary plain message.
- ***Final phase called decipher phase*** that building the revers key based on RF from out-of bag sub group and select one of the reverse linear or nonlinear function mapping of that used in the encryption phase to mix between the reverse key and binary cipher message) .

Each system phase depends on the previous. The main objective of the phases is to generate a key base new methodology. Figure 4 shows how the novel methodology can generation the uique binary sequence that it is length more than the length of the orginal message and this point prove the novel methodology provide the users of high level of security and it made the message can't be broken . This point is conclusion from all experiments that performance to test the novel methodology.

The following tables explain different experiments perform using the novel methodology on different plain taxes.

TABLE 2: RESULT OF SAMPLE 2^N , N=2

Plain Message	Computer
Function	XOR
Key	00001110000
Cipher Message	nkhgfs
Revers Function	XOR
Retrieve message	Computer

TABLE 3: RESULT OF SAMPLE 2^N , N=3

Plain Message	UNIVERSITY BABYLON
Function	XOR
Key	0010110001100110110100010000000
Cipher Message	RcbYGRXZAVD AEJYDPA
Revers Function	XOR
Retrieve Message	UNIVERSITY BABYLON

TABLE 4: RESULT OF SAMPLE 2^N , N=3

Plain Message	DEPARTMENT
Function	AND

Key	00000110100010100101111101011
	10
	00010100000100010100001110 10
	01110011001101110011110000 01
	0001101011100110010010 11

Cipher Message	AAFARCAAIQ
Revers Function	NAND
Retrieve Message	DEPARTMENT

TABLE 5: RESULT OF SAMPLE 2^N , N=4

Plain Message	SOFTWARE
Function	XOR
Key	1010010110100011100 11001011100010111 1010
Cipher Message	GYUKEcae
Revers Function	XOR
Retrieve message	SOFTWARE

TABLE 6: RESULT OF SAMPLE 2N , N=5

Plain Message	STAGE FOUR
Function	XOR
Key	010010100010011111 000101011000110101010 0010101010011100
Cipher Message	bbTaO dUAb
Revers Function	XOR
Retrieve Message	STAGE FOUR

TABLE 7: RESULT OF SAMPLE 2^N N=5

Plain Message	SOFTWARE
Function	OR
Key	1111001110101100111 000001001100010110100
Cipher Message	dOXeXGVU
Revers Function	NOR
Retrieve Message	SOFTWARE

TABLE 8: RESULT OF SAMPLE 2^N , N=6

Plain Message	INFRMATION TECHNOLOGY
Function	XOR

Key	11100 01000 01111 11001 11000
	01001 01100 11110 01100 11000
	0001100010 11110 10110 00011 0
	0110 00111 00000 10111 11111 0

Cipher Message	UFKXJFMNEWOYNSBBIdARR
Revers Function	XOR
Retrieve Message	INFRMATION TECHNOLOGY

TABLE 9: RESULT OF SAMPLE 2N , N=7

Plain Message	COMPUTER AND DATA SECYRITY
Function	XOR
Key	0001101110101001111101111010 0010110010001110001111100001 1001011100001101
Cipher Message	BAYRKCIACdOGCODcUEMV OJJWQc
Revers Function	XOR
Retrieve Message	COMPUTER AND DATA SECYRITY

TABLE 10: RESULT OF SAMPLE 2^N , N=8

Plain Message	HELLO AHMED IAM ALI
Function	XOR
Key	0101000100101100011110011O 0110101111100111101000111O 0111101001001110000111000 1 0100010111111100111011101 1 0101110011011100100100010 1 1101100111101001011111100 0 0110
Cipher Message	NAcMcbXdRHQABYCeCUR
Revers Function	XOR
Retrieve message	HELLO AHMED IAM ALI

TABLE 11: RESULT OF SAMPLE 2^N , N=9

Plain Message	HELLO MOHAMMAD WHERE ARE YOU GOING
Function	XOR
Key	1101111100010101101111101100 0111011010111001110111110000 0101000000110011111000101100 0100001100100101111101101111 0011101101110001101010011111 1000001011010101011000110000 011111110010011011
Cipher Message	bYBQVZZAPMYBQGcDCYLM bMTZSBERMYOOZ
Revers Function	XOR
Retrieve Message	HELLO MOHAMMAD WHERE ARE YOU GOING

V. ASSUMPTIONS AND LIMITATIONS

The main assumption of this paper is design the new key generator using data mining techniques to solve the problem of a stream cipher system that it become passive at this time because it suffer from many drabacks as explained in section 2. This new methodology not handel the problem of stream cipher system but also it attempts to test new functions from linear and nonlinear such as “product, quadratic and linear add to that xor, and , or ” to mix between the orginal message and the key and the invers of that function to mix between the key and cipher message to generate the orginal plain message.

We can summarize the main benefits of the anovel methodology by the following points:

1. Conceptual framework and architecture for the design and evaluation of security system for virous length of messages
2. Test Linear and non Linear functions of two variables and It given a well results and speed of execution.
3. Complete and successful prototype design
4. Implementation of prototype on hardware
5. Risk factors to be resolved
6. Prove the accuracy and the speed of the suggest key generator in solve the steam cipher problem

In addition, RF is consider as one of the statistical tools that proves good performance in many fields but by experiments we find the combination between RF as key generator and steam cipher principle to design delevelop secruit system lead to increase the sapace complexity and the reduce the time complexity.

V1. CONCLUSION

Key generator is part of the stream cipher system that is responsible for generating a long random sequence of binary bits key that used in ciphering and deciphering processes. In that system, we used LFSR or NFSR as a tool to generate the key and mix between the original message and the key in cipher phase or mix between the cipher message and key in the decipher phase but this system becomes "passive/out-of-fashion" at recent year for many reasons:

Firstly; the length of key in the LFSR or NLFSR determin by the number of satates that contain binary random valuses, ***secondly;*** the number of states in both LFSR or NLFSR have limited number of successful connection and ***thirdly;*** both registers can generation the key and given secrit results if the message is short or devided in to multi sub messages but the results are become easy broken with the long messages because the key is repeat more than one and the logic of encreption is known for all.

Therefore, this research present a novel methodology to generate a long unique binary sequence of bits using as a key of stream cipher system. And it test new functions to mix between plain message and key or cipher message and key in cipher and decipher process respectively. In addition it consider more robust and confidence compare with a LFSR or NLFSR.

By experiments, we find the key generatoing by a novel e methodology scessful encryption messages different in size (i.e., the length of message or *in other word*; the total number of letters in that messages) and prove ablity to generated the unique key for each message based on it' length without repeating the key in most cases. In addition, the system building base on the randomization principles of random forest data mining that mad process of the broking that system is very diffecult for any attacker because the attacker can't access of the out of bag package to using in decipher process

Future work, we can applying new functions such as Linear of (one, and three variables), cubic of one variable, Log of one variable in encryption and decryption phase base on the same architure of the that research.

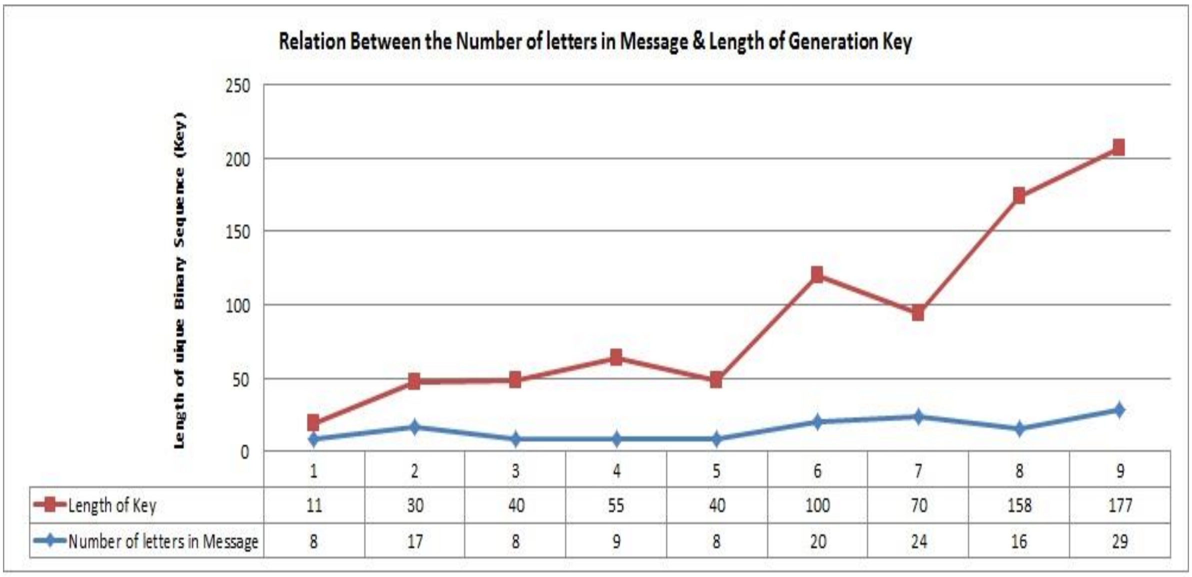


Figure 4: Relation between number of letters in message and the length of generated key

References

[1] Mitra S., "Data Mining: Multimedia, Soft Computing, and Bioinformatics," John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

[2] M. Robshaw. Stream ciphers. *Technical Report TR – 701.* RSALabs, July 1995

[3] Majid B. and Mohd A., **“**An Efficient Stream Cipher Algorithm for Data Encryption”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011

[4] David S. Siroky., “Navigating Random Forests and related advances in algorithmic modeling”, Statistics Surveys , Vol. 3 , PP 147–163 , ISSN: 1935-7516, DOI: 10.1214/07-SS033, 2009.

[5] Eibe Frank, ,"Book Data Mining, Practical Machine Learning Tools and Techniques" Department of Computer Science University of Waikato ,2nd ,558page,2005

[6] Jyun-Cheng Wang and David., “ Data mining techniques for customer relationship management” , Taiwan, ROC 2005

[7] Naphaporn Sirikulviriya and Sukree Sinthupinyo, ” Integration of Rules from a Random Forest “, , Chulalongkorn University, Bangkok, Thailand 2011

[8] Berbain, C. , Gilbert, H. and Patarin, J. , “QUAD: A multivariate stream cipher with provable security”, Journal of Symbolic Computation, 2009.

Available at: http://libhub.sempertool.dk.tiger.sempertool.dk/gmt/ivsl/elsevier/07477171_2009_44_12_1703-1723/S0747-7171(08)00183-1

[7] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, " Introduction to Data Mining **"**, University of Minnesota,USA,2006.

[8] Samaher Hussein Ali“Boook A New Trend Of Knowledge Discovery Toward Intelligent Data Analysis” , Germany, 2013. ISBN:978-3-659-40188-6.

[9] Samaher Hussein Ali, "Miner for OACCR: Case of Medical Data Analysis in Knowledge Discovery", Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on Digital Object Identifier: 10.1109/SETIT.2012.6482043 Publication Year: 2012 , Page(s): 962 - 975 .

[10] Sattar B. Sadkhan, Dr. Nidaa A. Abbas, Najat Shareef, “Digital Stream Cipher Generator Based on Multiplicative Cycle Group. TatraCrypt 2012.

[11] Samaher Hussein Ali,” A New Trend of Knowledge Discovery in Database using Data Mining Techniques” **IEEE** ,The 6th International Conference on Networked Computing and Advanced Information Management (NCM 2010), Kora, 2010.

[12] Hadzic, Fedja , Tan, Henry, Dillon and Tharam S., “Mning Unordered Embedded Subtrees Using TMG Candidate Generation Authors” **,** International Conference on Web Intelligence and Intelligent Agent Technology, 2008

Available at :

http://libhub.sempertool.dk/ivslAuth?url=http://libhub.sempertool.dk/gmt/ivsl/ieee/_2008_1__285292/4740404/4740405/4740462/5/10.1109/WIIAT.2008.403