# Classification of Substitution Ciphers using Neural Networks

*G.Sivagurunathan†, V.Rajendran††, and Dr.T.Purusothaman†††*

*†Research Scholar, Anna University, Coimbatore, Tamilnadu, India*
*††Research Scholar, Anna University, Coimbatore, Tamilnadu, India*
*†††Assistant Professor, Faculty of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamilnadu, 641013 India*

**Summary**
Most of the time of a cryptanalyst is spent on finding the cipher technique used for encryption rather than the finding the key/ plaintext of the received cipher text. The Strength of the classical substitution cipher's success lie on the variety of characters employed to represent a single character. More, the characters employed more the complexity. Thus, in order to reduce the work of the cryptanalyst, neural network based identification is done based on the features of the cipher methods. In this paper, classical substitution ciphers, namely, Playfair, Vigenère and Hill ciphers are considered. The features of the cipher methods under consideration were extracted and a backpropagation neural network was trained. The network was tested for random texts with random keys of various lengths. The cipher text size was fixed as 1Kb. The results so obtained were encouraging.

***Key words:***
*Cipher text, Classifier, Back propagation neural network, Playfair cipher, Hill Cipher, Vigenère cipher*

## 1. Introduction

Encryption is a primary method of protecting valuable electronic information. Encryption is a process to transform a piece of information into an incomprehensible form. The input to the transformation is called plaintext (or clear text) and the output from it is called cipher text (or cryptogram). The reverse process of transforming cipher text into plaintext is called decryption (or decipherment). The encryption and decryption algorithms are collectively called cryptographic algorithms (cryptographic systems or cryptosystems). Both encryption and decryption processes are controlled by a cryptographic key, or keys. In a symmetric (or shared-key) cryptosystem, encryption and decryption use the same (or essentially the same) key; in an asymmetric (or public-key) cryptosystem, encryption and decryption use two different keys: an encryption key and a (matching) decryption key, and the encryption key can be made public (and hence is also called public key) without causing the matching decryption key being discovered (and thus a decryption key in a public-key cryptosystem is also called a private key).

Various attacks on the cipher text are performed to identify the plaintext. The main problem of the cryptanalyst is to find the method employed and the encryption key used. Any encryption algorithm is breakable, but the real problem is that the cryptanalyst should be able to break the cipher text within a given time frame, because after that time frame the information so obtained may be useless. Most of the useful time of cryptanalyst is wasted in finding the method or encryption algorithm. So if it is possible to identify the encryption algorithm employed then the task of the cryptanalyst becomes easier. Various methods for identifying ciphers have been employed earlier. Identification of permutation, substitution and Vigenère ciphers was done using frequency analysis and classified depending upon a cost value [1].An attempt was made to identify block ciphers like DES Blowfish etc using pattern recognition methods[2].Other ciphers like stream cipher SEAL and Enhanced RC6 have been identified using neural networks [3].

### 1.1 Artificial neural networks

A neural network is a computational method inspired by studies of the brain and nervous systems in biological organisms. It is a Computing system made of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external input. Animals are able to react adaptively to changes in their external and internal environment, and they use their nervous system to perform these behaviours. An appropriate model/simulation of the nervous system should be able to produce similar responses and behaviours in artificial systems. The nervous system is build by relatively simple units, the neurons, so copying their behaviour and functionality should be the solution. Neurons work by processing information. They receive and provide

Manuscript received March 5, 2010
Manuscript revised March 20, 2010

information in form of spikes.

An artificial neural network is composed of many artificial neurons that are linked together according to specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.

An artificial neural network may contain an input layer, output layer and hidden layers ( If necessary). The hidden layer may be employed if linear classification is not possible. Each layer consists of several neurons.

A neuron is considered to be an adaptive element. Its weights are modifiable depending on the input signal it receives, its output value and the associated teacher response (if available). Thus the neuron will modify its weights based only on the input and/or output. One of the distinct strengths of neural networks is their ability to generalize. The network is said to generalize well when it sensibly interpolates input patterns that are new to the network. Assume that a network has been trained using the data x, x1 through x5, The figure 1 illustrates bad and good generalization examples at points that are new and are between the training points. Neural networks provide, in many cases, input-output mappings with good generalization capability.

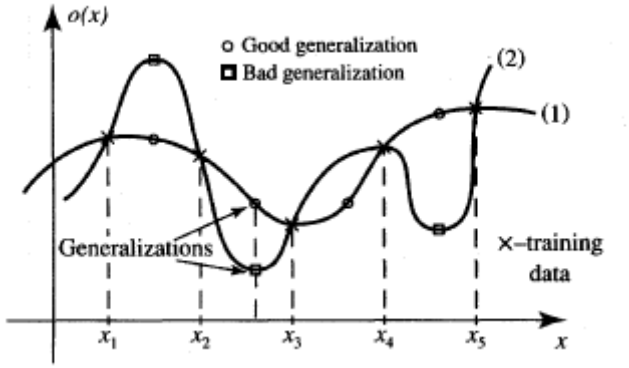


**Figure.1 Generalization of a neural network**

The enhanced standard Back propagation algorithm [4] can train any network as long as its weights, net input and transfer functions have derivative functions. Here the weights are adjusted according to gradient descent.

$$\Delta w = -k \frac{\partial E}{\partial w} \quad (1)$$

$$= \eta \frac{\partial E}{\partial w} \quad (2)$$

where η is the learning rate , Δ*w* is the weight change and E is the sum of squares of error

The problem with the standard gradient descent method is that it at times gets trapped into local minima, and hence variations were suggested. In Gradient descent algorithm with momentum, the weights are adjusted according to gradient descent. However some weightage is given to the previous weight change also.

$$\Delta w_n = \alpha \, \Delta w_{n-1} + \eta (1 - \alpha) \frac{\partial E}{\partial w} \quad (3)$$

where α is the smoothing factor for applying the momentum and η is the learning rate.

## 2.Classical Ciphers[5]

Most of the classical ciphers are substitution type. Each character may be represented by one (Caesar cipher) or many characters (Vigenère). Some of the ciphers are block ciphers, which converts one plaintext character block into one cipher character block. The description of the ciphers used for identification in this paper are given below

### 2.1 Playfair Ciphers

The Playfair cipher is a polygraphic cipher; it enciphers more than one letter at a time. The Playfair cipher enciphers digraphs – two-letter blocks. An attack by frequency analysis would involve analyzing the frequencies of the digraphs of plaintext. Complications also occur when digraph frequencies are considered because sometimes common plaintext digraphs are split between blocks.

The playfair cipher has a password length of 25 characters. All the characters are involved in the key matrix which is 5 X 5 Matrix. Both I and J share the same space or it may be any other lower frequency character which is left out in the matrix. The plaintext characters are encrypted into cipher text taking two characters at a time. The cipher text characters will be the intersection of row of the first plaintext character and the column of the second plaintext character and the second cipher text character will be the first plaintext character column and the second character's row. If the both plaintext characters are same then they are separated by a filler character, say "X'. So, no two adjacent cipher text characters will be the same. If the plaintext characters are in the same row, the next characters in the same row after the plaintext characters are taken as cipher characters. The same applies for plaintext characters in the same column.

The Playfair cipher is a simple example of a block cipher, since it takes two-letter blocks and encrypts them to two-letter blocks. A change of one letter of a plaintext pair will always change at least one letter, and usually both letters of the cipher text pair. However, blocks of two letters are too small to be secure, and frequency analysis, is usually successful.

The features of Playfair ciphers are

1. No two adjacent characters are identical
2. Only 25 alphabets will be used in the encryption process for the conversion of plaintext to cipher text.
3. The number of characters in the Plaintext is always even.
4. If 'AB' is encrypted as 'UH' then 'BA' will be encrypted as 'HU'.
5. The plaintext character will not be represented by itself in the cipher text.

### 2.2 Hill Ciphers

Hill ciphers are asymmetric ciphers where one key is used for encryption and a second key (the key inverse) is used for decryption. The Hill cipher is a cryptosystem that enciphers blocks. Any block size may be selected, but it might be difficult to find good keys for enciphering large blocks. The advantage of having large blocks is that change of one character in a plaintext block may change potentially all the characters in the corresponding cipher text block. Hill cipher uses invertible matrices for encryption. An invertible matrix of sufficient order is used as key. The number of characters to be converted in each block depends on the order of the matrix. As the order of the matrix increases the diffusion property of the cryptosystem increases, but it is very difficult to find invertible matrices of higher order.

Characters of the size of the order are selected and multiplied with the key matrix. The resulting matrix is operated on modulo 26 and the elements of this matrix now contain the cipher text. Hill cipher has more diffusion property which means that frequency statistics of letters, in a plaintext are diffused over several characters in the cipher text, and hence much more cipher text is needed to do a meaningful statistical attack.

The features of Hill Cipher are

1. Strong against Frequency analysis
2. All characters (A to Z) are employed for encryption and hence all the characters may be present in the cipher text.
3. Higher order keywords are very rare as it is difficult to find invertible matrices both of which contain integers only.
4. Higher diffusion property and increases with matrix order.

### 2.3 Vigenère ciphers

The Vigenère Cipher, proposed by Blaise de Vigenere from the court of Henry III of France in the sixteenth century, is a progressive poly alphabetic substitution method. The set of related mono alphabetic substitution rules makes use of 26 Caesar ciphers with shifts of 0 to 25.[6]The table used for encryption can be created for a simple alphabet from A to E, which can be extended to all letters from A to Z .Each row in a table can be created by a simple shift of the previous row. Thus a vigenere cipher of password one can be considered as a Caesar cipher as this involves only one shift of the alphabets and thus forming a Caesar cipher

**Table 1 – vigenere table for alphabet A to E**

| Plain text / key | **A** | **B** | **C** | **D** | **E** |
|---|---|---|---|---|---|
| **A** | A | B | C | D | E |
| **B** | B | C | D | E | A |
| **C** | C | D | E | A | B |
| **D** | D | E | A | B | C |
| **E** | E | A | B | C | D |

To derive the cipher text using the table, for each letter in the plaintext, one finds the intersection of the row given by the corresponding keyword letter and the column given by the plaintext letter itself. It can be modeled mathematically as, **C = (P+K) % 26** Where C is the cipher text letter and P, K are plain text and key letters.
Decipherment of an encrypted message is equally straightforward This time one uses the keyword letter to pick a row of the table and then traces down the row to the column containing the cipher text letter. The index of that column is the plaintext letter. It can be modeled mathematically as**, P = (C-K) % 26**

The Features of Vigenère Ciphers are

1. Since this method employs 26 Caesar ciphers, each character may be represented in 26 different ways and each character may be used to represent 26 characters.
2. Kasiski's technique for finding the length of the keyword was based on measuring the distance between repeated bi grams in the cipher text and can be used to find the length of keyword.
3. If the keylength is 'n' then every character at $(n+m)^{th}$ character ( m < n)will follow the same column and hence each can be treated as individual Caesar ciphers. Thus we will have 'n' Caesar ciphers.

## 3. Experiment and Results

### 3.1 Training the network

The features of the ciphers under consideration were extracted and a back propagation neural network was trained. 900 samples of 1 Kb were taken for training (300 for each classical substitution cipher). Back propagation network so designed had 3 layers including a hidden

layer. Tansigmoidal and logsigmoidal learning rules were adopted.

Some of the features that were considered are

1. The number of alphabets in the cipher text.
2. The number of adjacent duplicates
3. The distance between the digram duplicates etc

These features were selected because, Playfair cipher consists of 25 alphabets only and hence it may form a feature that can be used for its identification. Similarly, in the vigenere cipher, a character will be represented by the same another character whenever the password character is same. Hence the adjacent duplicates can be used as a measure to find the password length and it is used as a feature.

The network converged when the validation check reached 3000. This network was simulated and tested with MATLAB.

### 3.2 Testing the network

The network was tested for different type of input ciphers and the various conditions that were considered are

1. Different Password length and same Plaintext
2. Same Password and Different Cipher texts
3. Different Password length and Different Plaintext

The plaintexts were taken from general English books and password lengths varying from 2 to 25 were used to train. Each plaintext was encrypted with all the three encryption methods .Each cipher was tested for at least 1000 text files and the results are tabulated below

#### 3.2.1 Different Password length and same Plaintext

**Table 2 Identification of Ciphers for different Password Lengths**

| SNo | Input Cipher | No of Files | Classified as | | | Percentage of identification |
|---|---|---|---|---|---|---|
| | | | Playfair | Vigenere | Hill | |
| 1 | Playfair Cipher | 2000 | 2000 | 0 | 0 | 100 |
| 2 | Vigenère Cipher | 2000 | 1 | 1740 | 259 | 87 |
| 3 | Hill Cipher | 2000 | 27 | 283 | 1690 | 84.5 |

The same plaintext files of 1Kb were used for the classification of the three ciphers under consideration, but the password length was varied from 2 to 26 for Playfair ciphers. All the encrypted files were correctly classified as Playfair cipher. For the same set of plaintext files, Vigenère cipher was used for encryption and its password length was varied from 2 to 26. Some of files were misclassified as Hill cipher and very few as Playfair. The password length of Hill cipher was varied from 2 to 20 and the same set of files of 1Kb were used for encryption.

#### 3.2.1 Same Password length and Different Plaintexts

Plaintexts of 1Kb length were encrypted using passwords of different length every password length was tested with at least 500 files for identification. All the passwords were randomly generated and the plaintexts were taken from general English textbooks. Password lengths varying from 2 to 25 were tested for Playfair and vigenere ciphers .Since higher password lengths of hill cipher are rare, hill cipher was tested with password lengths varying from 2 to 20.

**Table 3 Identification of Ciphers for different Plaintext and same password length**

| S.No | Input cipher | No of Files | Classified as | | | Percentage of identification |
|---|---|---|---|---|---|---|
| | | | Playfair | Vigenere | Hill | |
| 1 | Playfair Cipher | 12000 | 12000 | 0 | 0 | 100 |
| 2 | Vigenère Cipher | 12000 | 0 | 9500 | 2500 | 79.2 |
| 3 | Hill Cipher | 10000 | 0 | 1565 | 8435 | 84.3 |

The same plaintext files of 1Kb were used and password length varied from 2 to 26 for Playfair ciphers. All the files were correctly classified as Playfair. For Vigenere cipher, password length was varied from 2 to 26 and the same files were used for encryption none of vigenere encrypted files got misclassified as Playfair cipher but few files was misclassified as hill cipher. The password length of Hill cipher was from 2 to 20.As with Vigenere, none of the hill cipher encrypted files was misclassified as Playfair cipher, but files were misclassified as vigenere cipher.

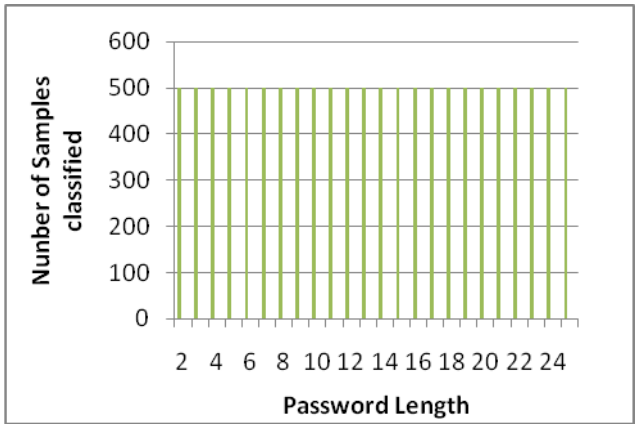


**Figure 3 Results of Playfair for different Plaintext encrypted with different password lengths**

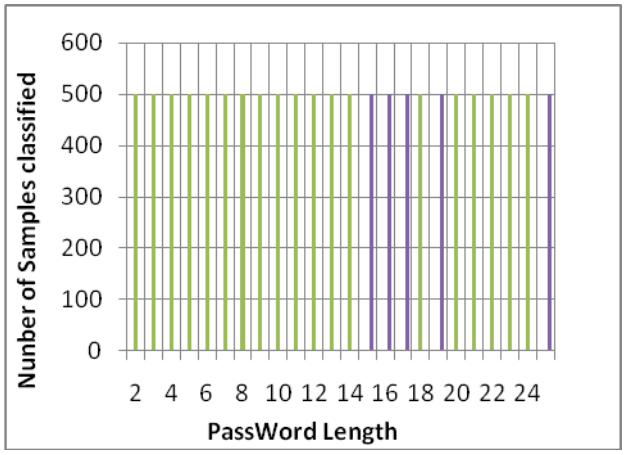


**Figure 4 Results of Vigenère for different Plaintext encrypted with different password lengths**

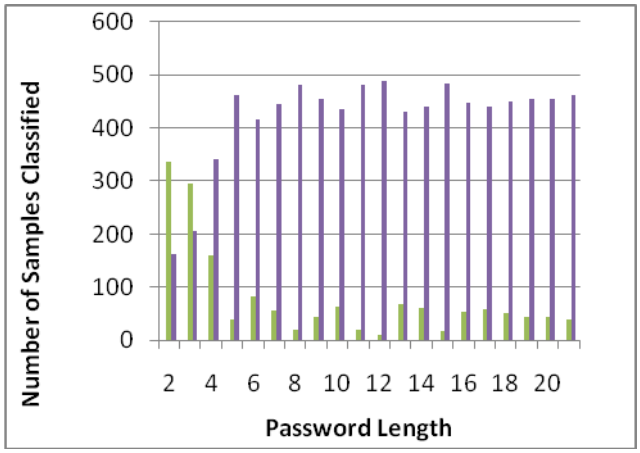


**Figure 5 Results of Hill Cipher for different Plaintext encrypted with different password lengths**

It can be seen that Playfair cipher was identified 100% correctly for all password lengths from 2 to 26. But some files which were encrypted by Vigenere cipher was misclassified as Hill cipher when password length was 15,16,17,19 and 25 Similarly, misclassifications of Hill cipher occurred the most when the password length was 2 and 3 and it was also observed that as the password length was increased the number of misclassicfications got reduced.

### 3.2.3 Different Password length and Different Plaintexts

Different plaintexts of 1Kb length were encrypted using password lengths varying from 2 to 26 for Playfair and Vigenère ciphers and 2 to 20 for Hill ciphers. These texts were different from the texts that were used for training and for other testing purposes. The passwords were randomly generated of required size/order.

**Table 4 Identification of Ciphers for different Plaintext and different password length**

| S. No | Input Cipher | No of Files | Classified as | | | Percentage of identification |
|---|---|---|---|---|---|---|
| | | | Play fair | Vigenere | Hill | |
| 1 | Playfair Cipher | 400 | 400 | 0 | 0 | 100 |
| 2 | Vigenère Cipher | 400 | 0 | 276 | 124 | 69 |
| 3 | Hill Cipher | 400 | 1 | 58 | 341 | 85.25 |

## 4. Conclusion

Identification of classical substitution ciphers like Playfair cipher, Vigenère cipher and Hill cipher was attempted using neural networks. Some of the features like adjacent duplicates and their frequency of repetition were used to identify the encryption method. It was seen that Playfair cipher was identified always irrespective of password length and plaintext. The Vigenère cipher was identified around 70 to 80 % and most of its misclassifications were as Hill Cipher. This happened when the password length was high, because as the password length increased, the number of repetitive adjacent duplicate characters was reduced. If it is possible to obtain lengthy cipher texts greater than 1Kb then the percentage of identification will be higher. Hill cipher was identified for all the cases successfully but failed when the password length was too low (2 and 3). This is because the diffusion property of the Hill cipher was low and hence the features of the resulting Hill

cipher were similar to that of Vigenère cipher. Hence most of the misclassifications of Hill cipher were Vigenère cipher. If the password length was increased then success rate of identifying Hill cipher was also increased.

## References


[1] Pooja Maheswari "Classification of ciphers", Indian Institute of Technology, Kanpur , 2001

[2] Sreenivasulu Nagireddy "A Pattern recognition approach to block cipher identification" Indian Institute of Technology, Chennai , 2008

[3] B.Chandra and P.Paul Varghese, "Application of cascade correlation Neural Network for cipher system Identification", World Academy of Science, engineering and technology 262007

[4] D. E. Rumelhart; G. E. Hinton and R. J. Williams; "Learning internal representations by error propagation", Parallel Data Processing, Vol.1, Chapter 8, the M.I.T. Press, Cambridge, MA, 1986, pp. 318-362.

[5] William Stallings, "Cryptography and Network Security Principles and Practices", Prentice Hall. 2006.

[6] Reinhard Wobst, "Cryptology Unlocked", John Wiley and sons ,2007


**G.Sivagurunathan,** completed his B.E in Electrical & Electronics Engineering in the year 1989 at Bharathiyar University, Coimbatore. He completed his M.E (Computer Science & Engineering) in Bharathiyar University, Coimbatore in the year 2001. He is currently pursuing his PhD in Anna University, Coimbatore. His research interests include Cryptanalysis and Digital image processing. He is a life member of ISTE and member of ACM.

**Mr.V.Rajendran** completed his B.E degree in the year 1986 at Madurai Kamaraj University, Madurai. He then completed his M.E degree in the year 2001 at Bharathiyar University, Coimbatore. Presently he is working as a senior lecturer at Government Polytechnic College for women, Coimbatore. Now he is pursuing his research work at Anna University, Coimbatore in cryptanalysis.

**Dr.T.Purusothaman**, completed his BE degree in the year 1988 under Madras University and completed his ME (CSE) degree at Government College of Technology, Coimbatore in the year2002. He did his Ph.D. under Anna University in 2006. He has 21 years of teaching experience. He is currently working as Assistant Professor (RD) in the Department of computer science and Engineering at Government College of Technology, Coimbatore. His research areas include Network Security and cryptography, Distributed operating systems, Mobile computing, Advanced Genetic Algorithm and Grid Computing. He has published ten Research articles in reputed Indian and International Journals and fifty Research articles in national and international conferences. He is currently supervising 20 Ph.D., Research Scholars. He is a life member of ISTE. He is the co-investigator of the project titled "Optimization of search time for evaluation of ciphers using Genetic Algorithm based cryptanalysis" sponsored by Ministry of Communications & Information Technology, New Delhi. He has presented a paper in International conference and attended a Faculty Development Program at Canada.