Delving

Journal of Technology and Engineering Sciences

An Open Access Journal

ISSN 0975-5829

December 2020

Vol. 4 Issue 2

Articles

ACROPOLIS
Enlightening wisdom

Acropolis Institute of Technology & Research Indore

www.acropolis.in

Delving: Journal of Technology and Engineering Sciences

ISSN 0975-5829

Prof. (Dr) Kamal K Sethi
Editor in Chief
Professor & Head Computer Science and Engineering
Acropolis Institute of Technology & Research Indore India

Prof. Praveen Bhanodia
Editor
Associate Professor Computer Science and Engineering
Acropolis Institute of Technology & Research Indore India

Editorial Office
Acropolis Delving
Department of Computer Science and Engineering
Acropolis Institute of Technology & Research Indore
Bypass Road, Manglia Square, Manglia,
Indore, Madhya Pradesh – 453771
India

Cryptosystems in Asymmetric Cryptography for Securing Data at various Level

Sartaj Singh[1*], Ashok Sharma[2]

[1,2]School of Computer Science and Engineering, Lovely Professional University
Phagwara, India
sartaj.singh@lpu.co.in, drashoksharma@hotmail.com

Abstract

With upcoming threats in digital world, we need to work continuously in the area of security in all aspects form hardware to software as well in data modeling. Rise in social media activities and hunger of data by various entities leads to cybercrime and more attack on privacy and security of persons. Cryptography has always been employed to avoid access to important data by using many processes. Symmetric key and Asymmetric key cryptography have been in used for keeping data secrets in rest as well in Transmission mode. Various cryptosystem has been evolved from time to time to make the data more secure. In this research article, we are studying various cryptosystem in asymmetric cryptography and their application with usefulness and much emphasis is given to Elliptic curve cryptography involving algebraic mathematics.

Keywords: *Cryptography, Symmetric key cryptography, Asymmetric key cryptography*

I. INTRODUCTION

Cryptographic algorithm has been proven a key aspect in cryptosystem for maintaining authentication and confidential messages across the networks. Encryption and Decryption are the primary necessity for privacy security on the internet [18]. Creating secret keys S, is important to encrypt and decrypt the message. The number of bits contained by message is the size of key. Therefore, the key size must always be greater than the bits used by message. A cipher text, can be decrypted back into original message only by using the correct key K. A comprehensive key search, which may be called 'brute-force' also, is the rudimentary technique used for identifying the concept K. Many a time, it has been seen that a weakness of the key schedule of the cipher happens to better the efficiency of a comprehensive key-search-attack. With a sea-change in the improvement of technology, the computing performance of technology, the computing performance always tend to make the comprehensive key-search-attack a better practice against keys having a fixed length. At the time of the designing of DES, it was deemed very secure, as compared to a comprehensive key-search, apart from being a less costly hardware.

It is also possible that comprehensive key-search may be employed on desktop work stations or personal computer. Whereas comprehensive search of DES 56 – bit key

*Corresponding Author
Sartaj Singh
Research Scholar, School of Computer Science
Lovely Professional University, Phagwara, Punjab, India. ✉*Email:* sartaj.singh@lpu.co.in

space is likely to require a thousand years on the best available computer of the day, the development of internet made the utilization of many gadgets in a speedy search divided by the key space. The search is executed by partitioning it and dividing small portions to large scale. As a result, some specifically designed ultra-modern computer was need of the hour. In such computer a DES key was actually split into 22 hours during the month of January, 1999. The existing rate of growth in computer power is about 80-bit key, which is likely to offer a reasonable amount of security for say about 12-15 years more. It seems impossible that 128-bit keys, which are employed in International Data Encryption Algorithm (IDEA), will be separated by the comprehensive findings in the near future. Same may be the case with the forth coming AES. In this research, to enhance the security and to reduce the size of M in RSA, Huffman Compression technique and DES are considered as proof of concept. The security factor is decided by RSA algorithm, with the actual use of RSA cryptosystem [2]

RSA modulus is basically the result of 2 large primes; and the primes tend to become larger. As a result, an attack will necessitate far greater time-span to factor it. A number with bigger prime factors with specified properties is going to make it easier to factor. For instance, this will happen, To enhance the security and for compression, several authors have proposed different methods which are illustrated in a few publications. This review highlights some important techniques of the said algorithms that has been carried out already. They are presented in the following subsections.

II. REVIEW OF COMMUNICATION TRANSMISSION

According to the Shannon's coding theorem, $\log_b P$ is the optimal length of code for a symbol. Here b is the number of symbols used to render output codes, while p is used for the input symbol likelihood. But there is a shortcoming of arithmetic coding in it. The update operation is slow; and so is the model look up. In this method, at least one multiplication per event is needed for fully precise form of arithmetic coding; whereas in some cases of implementations up to two divisions and two multiplications are needed, per even. Both, Lempel-Ziv coding as well as, Huffman coding are much faster, so far so, speed is concerned. It is because this method is represented straightway in the data structure. There is one more drawback of arithmetic coding – arithmetic codes. Resist poorly the occurring errors, when these are used with adaptive models. In a coded file, a single bit error, consequently makes the decoder’s internal state show to be in error. As such, it makes the rest of the encoded file incorrect. Actually, this drawback is found almost in all the adaptive codes. Lempel-Ziv codes and adaptive codes of Huffman are no exception [3].

Huffman coding shows many useful and nice properties. It is widely used in many applications. However, there are also some significant limitations in it. The main drawback in this code is that any error in the coded sequence of bits is likely to propagate at the time of decoding. This issue is faced in many codes of varied length. The boundaries between code words cannot be ascertained beforehand, unlike fixed length codes. it can be found only while the process of decoding is in progress. The error moves into the succeeding codeword if the wrongly decoded sequence of bits does not occur at the end on the boundary of correct code word [4].

Tarek M Mahmoud presented a novel and efficient technique. It is a method comprising encryption for the security of short message services in the operating system of

Symbian. This method is employed for the safe and secure sending of SMS from one mobile phone to the other. RSA based technique of encryption is also employed to dwindle the possibility of spying and eaves dropping. But here again, the problem arises that encryption enhances the length of text message. As such, the bandwidth is not utilized. Clemens Guhmann and Stephan Rein offered the method of data compression in mobiles. It offers less complicated mathematics coding compression of the script transmitted through cell phone. Biham and Seberry bought in yet another method known as 'rolling arrays. This method comprises rotations and permutations. But this method is also not devoid of shortcomings. Its limitation is that the total 256 keystream KS does not depend on the M, which is to be encrypted[5][6].

Fenwick suggests the use of prevailing predefined variable length codes, as well as universal. They can also show satisfactory compression. It offers a pretty essential and better density of maximum files. The higher the compression fraction, the more effectual the algorithm is. But it does not mention about the security of the message while transmitting from one to the other. Data compression methodologies for loss free data presented by Porwal S et. al gives a comparative performance of Huffman and arithmetic encoding is greater and better as compared to that of Huffman's encoding. The time used and channel bandwidth is lowered and it is far better than Huffman encoding. When compared with Huffman coding, the compression speed is actual low in encoding[6][7].

Sreelaja and Pai suggest another method for the creation of KS, known as Ant Colony Optimization (ACO). This method is employed for the distribution of the characters for encryption in the M. An also Artificial Ant does not find their counterparts. At the same time, the encryption time grows higher. It is so because the spectacle deposition depends on difficulty. It does not re-create real ants' act. As per Imad Khaled Salah et. at analysis RSA cryptosystem cannot be broken by any attack algorithm in an effective and efficient way. Most attack seen the result of system's misuse or wrong choice of parameters. Majid Bakhtiari and Mohd. Aizaini Maarof created an effective and efficient stream cipher algorithm to produce 115 bits in a single round of the process randomly. At the same time, it enhanced the processes' resistance, in comparison to algebraic and correlation attacks of Berlekamp-Massey. However, some computers may be unable to create random bits effectively and efficiently[8][9]

Iwan Handoyo Putro, Petrus Santoso suggested yet another novel technique for data compression. It is an arithmetic encoding-based technique. This method highlights the drawbacks of the length of the message. It showed a way-out for transmitting a message with more than 180 characters. Text compression, as well as, superfast searching has been shown by Khurana U and Koul. They demonstrated a better and efficient technique which provided higher compression ratios and speedier search through the text. In this way, if necessitates the development of mathematical models. In addition, increasing the speed of public key cryptographic algorithms and efficient implementation is also mandatory[10][11]

Various codes have been proposed in the literature for data compression and they are categorised as fixed length and variable length codes. Variable length codes use some statistical method in contrast with lengthy and fixed codes. Short codes are given to groups of signs or symbols with a upper probability of existence in the variable length code. Longer codes are assigned to lower probability symbols or group of symbols.

Individuals who design and implement lengthy codes will tackle these two problems, namely assigning unambiguously decodable codes and assigning codes with the very less average width. Huffman code [12] is one of the length codes of the function. It is a compression of loss-free data to represent a character in some other form. Using the known RSA public-key algorithm, the compressed code will be sent for encryption.

A symmetric block encryption algorithm translates a static block length of M data into similar sized blocks of C data, while a stream cipher that works on smaller M units usually transforms bits or bytes. The block's flow ciphers are highly rapid than ciphers. The problem facing most stream ciphers is to produce one random bit in each process round as the output flow of the cryptosystem raises the risk of algebraic similarity with these cryptosystems [13]. The one-time pad of Vernam is one of the stream ciphers that uses a randomly generated bits series. As the whole Ks is random, if he / she sees the P, even an adversary with infinite computational tools might guess the P. While the one-time pad of Vernam is perfectly safe, it is too hard to remember and store a K because the size of K is always taken as the size of M, so it is at least practical [14].

Abdul-Kader, Hadhoud [15], in their study proved Minaam that DES fares better than 3DES. Muthumanickam T [16] proposed that the Rijndael‟s S-Boxes are the dominant element of the round function in terms of required logic resources. Each Rijndael round requires sixteen copies of the S-Boxes, each of which is an $8bit \times 8bit$ look-uptable, requiring more hardware resources. In [17] Mijanur Rahaman, Md. Masudul Islam proposed the technique based on quantum computing which changes standard communication and processing in cloud servers.Moreover authors also highlighted some benefits and issues related to the quantum computing progress. Erdem S S Yanik Ko T [18] proposed a computational method in a finite field GF(2N) by integrating it in a immense ring Rp. They proved that the multiplication operation is a product of convolution and the rearrangement of bits is the squaring operation. Multiplication operation in Rp has complexity N+1, which is approximately is doubly efficient than optimal normal basis multiplication (ONB).

In [19] Longa P and Miri A described an innovative methodology to create the composite operations of the form $dP+Q$ by applying the special addition with identical z-coordinate to the setting of generic scalar multiplications over prime fields. They showed that their methods offer the lowest costs, given by $1I+(9L)M+(3L+5)Sand$ $1I+(9L)M+(2L+6)S$, when using only one inversion. Symmetric-key algorithm possess many limitations. A novel approach proposed by Gopinath Ganapathy and Mani K in generating the K from the Ks for stream ciphers using PPT (Primitive Pythagorean Triples) [20] has been proved that this method is used to reduce the number of keys to be stored and distributed. The requirement of storage space is minimized and the P is not taken as such. Instead, a code is generated based on Huffman code and also a mutation process is employed at various levels of the Huffman tree of each character. It is proved that the mutated code of each character is used for encryption with Ks, the corresponding ciphertext obtained from the encryption process is not easily predictable.

Md. Rubaiyat Hasan [21] presented a Huffman based LZW data compression encoding method to transmit a digital image from source to destination. The study concluded the proposed technique resulted better speed of transmission with less time period. Rajan et.al [22] used secure RSA technique for the implementation of encryption and decryption of file. It shows that Modified RSA Encryption Algorithm (MREA) can be

used to encrypt files and transmit itto other end for decryption. It is suitable for small sized files as large sized file take more time for processing. Monisha Sharma et al.[23], proposed a model based on partition and scanning patterns to encrypt or decrypt the image [24].

III. REVIEW OF PROCESS IN ASYMMETRIC KEY CRYPTOSYSTEM

Unlike symmetric key cryptography, asymmetric key cryptosystem used two keys private and public and this process is similar to locking and unlocking padlocks with keys as given in the figure 1.

Figure1.Asymmetric Key Cryptosystem Message Locking and Unlocking Process

Clearly, public key shared by receiver is required for encryption and receiver's private key is required for decryption at other end. The Basic principle of communication among sender and receiver is as shown in figure 2.

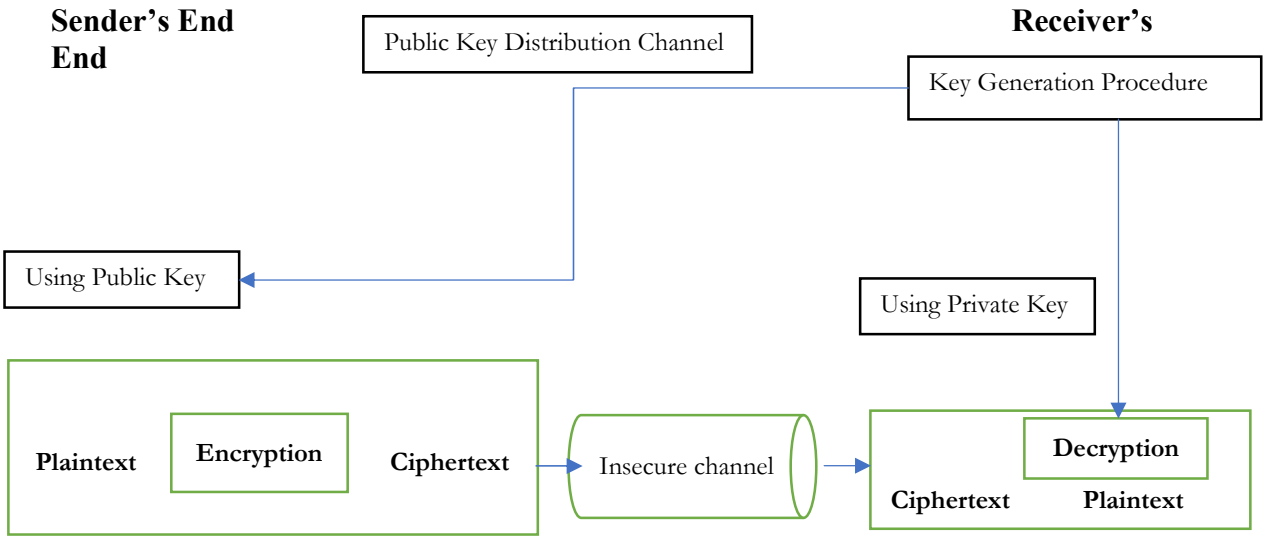


Figure 2. Communication Process in Asymmetric Key Cryptosystem

3.1 One-Way and Trap Door Function

One-way function satisfies the below listed properties
f is easy to compute means given x, y = f(x) can be easily computed
$f^{-1}$is difficult to compute means given y, makes invisible to calculate x- f^{-1} (y)a
and trapdoor function has additional property that gives y, a trapdoor (secret), x can be computed easily.

Mostly the communication process uses Trapdoor one -way function defined as under

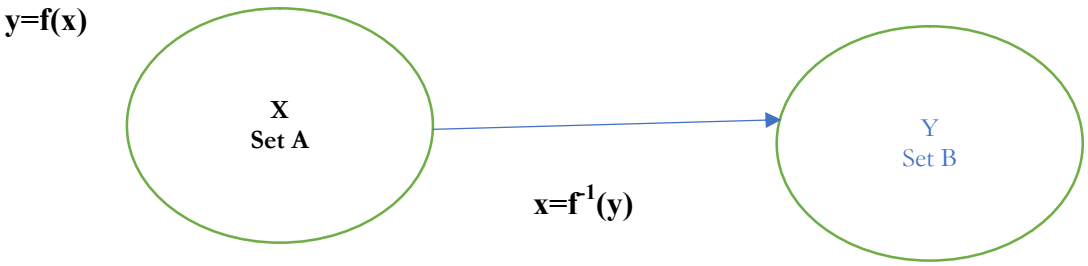


Figure 3.Simple Process of Mapping between Two Sets

A one way function where in addition to one way, given y a trapdoor(secret), x can be computed.

3.2 Knapsack Cryptosystem

Suppose we are having two set of n tuples,$x = [x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_n]$ and $y = [y_1, y_2, .., y_n]$ where first tuple is a predefined set and second tuple in which y_i is 1 or 0 defines elements of x are to be dropped in Knapsack the sum of elements dropped in Knapsack is given as:

$Sum = KS(y, x) = x_1y_1 + x_2y_2 + x_3y_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + x_ny_n$

Given y and x, we can find K_{sum} but $Y = inverse\_k_{sum}(K_{sum}, x)$ is difficult to find.But in case of super increasing tuple in which each element (except x_i) is greater than equal to sum of all previous element than K_{sum} and $inverse\_k_{sum}$ can be calculated easilyKnapsack Cryptosystem was an idea of Hellman and Merkle Communication Process in Asymmetric Key Cryptosystem and process of communication is depicted below in figure 4.

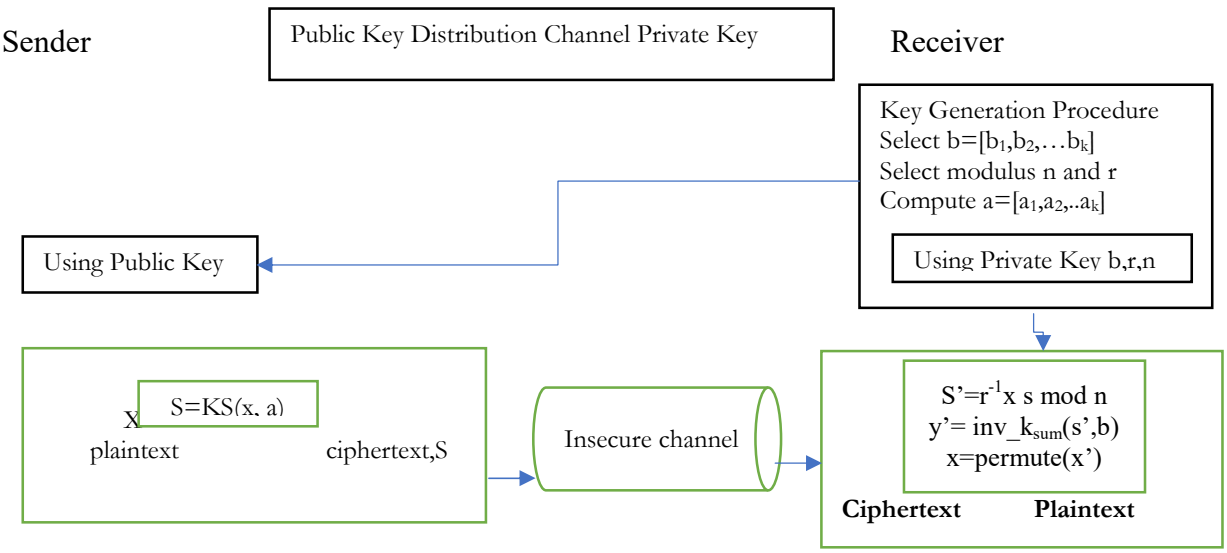


Figure 4 Secret communication with Knapsack Cryptosystem

3.2.1 Secret Communication with Knapsack Proceed as Under

Key generation in Knapsack
Create a super increasing K- tuple m = [m_1, m_2, ……m_k].
Select a modulus n, such that n>m_1+m_2+……+m_k.
Select any random integer r that is prime with n and 1 ≤ r ≤ n-1
Create temporary k-tuple L= [L_1, L_2, ……L_k] in which L_i = r x m_i mod n
Select a permutation of K objects and find new tuple O =permut (L)
Now public key is k-tuple O and he private key is n,r, and k-tuple m

3.2.2 Encryption of Message by Sender takes Place as Under

Sender converts their message to k-tuple
y = [y_1, y_2, ……, y_k] where y_i is 0 or 1 and y_i is simple text (plain text)
Sender uses knapsum routine to complete KS and deliver value of S(KS) as ciphertext.

3.2.3 Decryption of message by receiver takes place as under: -

Sender computes *S' = r^{-1}X KS mod n and then Sender uses inverse_KS to create y'.* Then sender permutes x' to find x. the tuple x is as Plaintext.

3.3 RSA Cryptosystem

RSA cryptosystem is more complex and is based on 2 exponents i.e. pk(public key) and pr(private key) and say PT (original text or plaintext)and CT (encrypted text or cipher text) and for generation of CT from PT sender applies $CT=(PT)^{pk}$ mod r.Reverse at them and use $PT=(CT)^{pr}$ mod n to decipher the ciphertext sent by sender . The mod r is a big number for process of key generation. Since in RSA, encryption &decryption is modular exponentiation and is feasible in polynomial time using fast exponentiation algorithm and also mod logarithm is a very hard for which there is no polynomial algorithm yet exist. So, it clearly indicates sender can encrypt in polynomial time (pk is public), receiver can decrypt in polynomial time but third person cannot decrypt because calculation of $(pk)^{th}$root of ciphertext CT using modular arithmetic is not feasible. The process of complexity related to RSA working is given in figure 5.

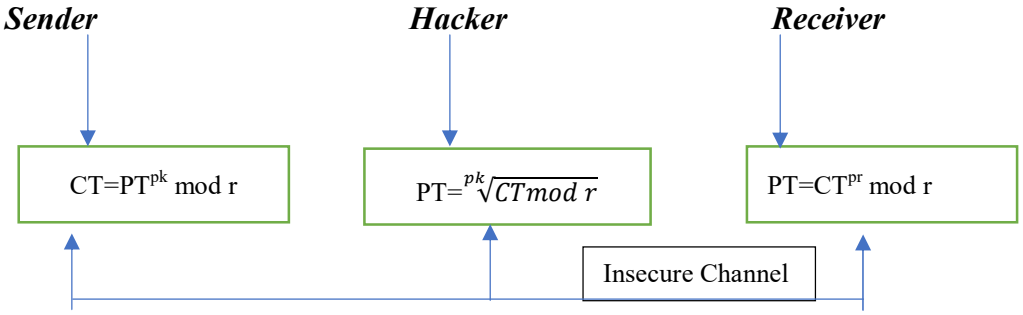


Figure 5 Complexity of Operation in RSA Cryptosystem

One communication in RSA takes places as under:

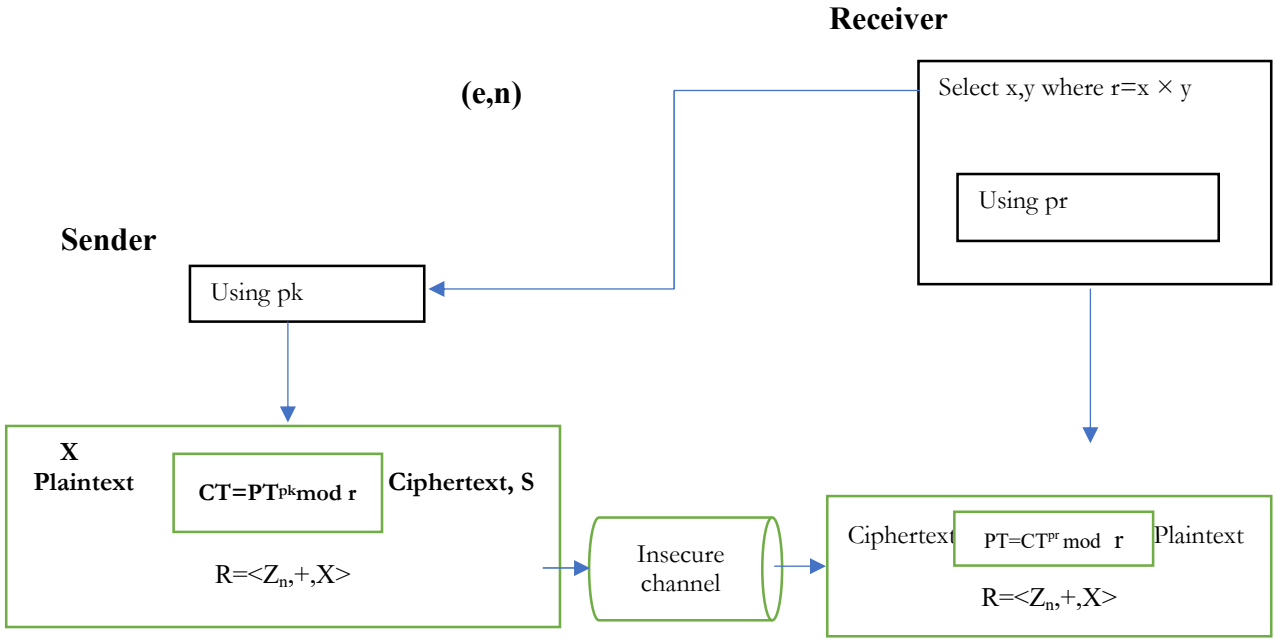


Figure 6. Process of Communication of RSA Cryptosystem

RSA uses public ring R = $<Z_n, +, X,>$ with two operations and a private group G = $<Z_{Ø(n)}*, x>$ and key generation takes place as under

3.3.1 Key Generation Algorithm

RSA_Keygen
Take any two large prime numbers x & y such that x≠y and
rx×y. ←
Ø(r) ←(x-1) × (y-1)
Choose e such that 1<pk< Ø(r) andpk is coprime to Ø(pk)
pr ← pk^{-1} mod Ø(r)
public_key ←(pk, r)
private_key ← pr
return public_key and private_key

3.3.2 Encryption and Decryption in RSA

RSA_encryption(PT,pk,r) // PT is plaintext
CT ← Fast_exp (PT, pk,r) // compute (PT^n mod n)
Return CT
RSA_decryption (CT,pr,r) // CT is Ciphertext
PT ← Fast_exp (CT,r, r)
Return PT

3.4 Rabin Cryptosystem based communication

Robin cryptosystem is based on exponentiation quadratic congruence where value of pk and pr are fixed as 2 & ½ respectively. Encryption is CT = PT^2 (mod n) and

decryption is PT = $CT^{½}$ (mod n) where n is public key and private key is tuple (p,r).Process of Robin crypto is depicted in figure 7.

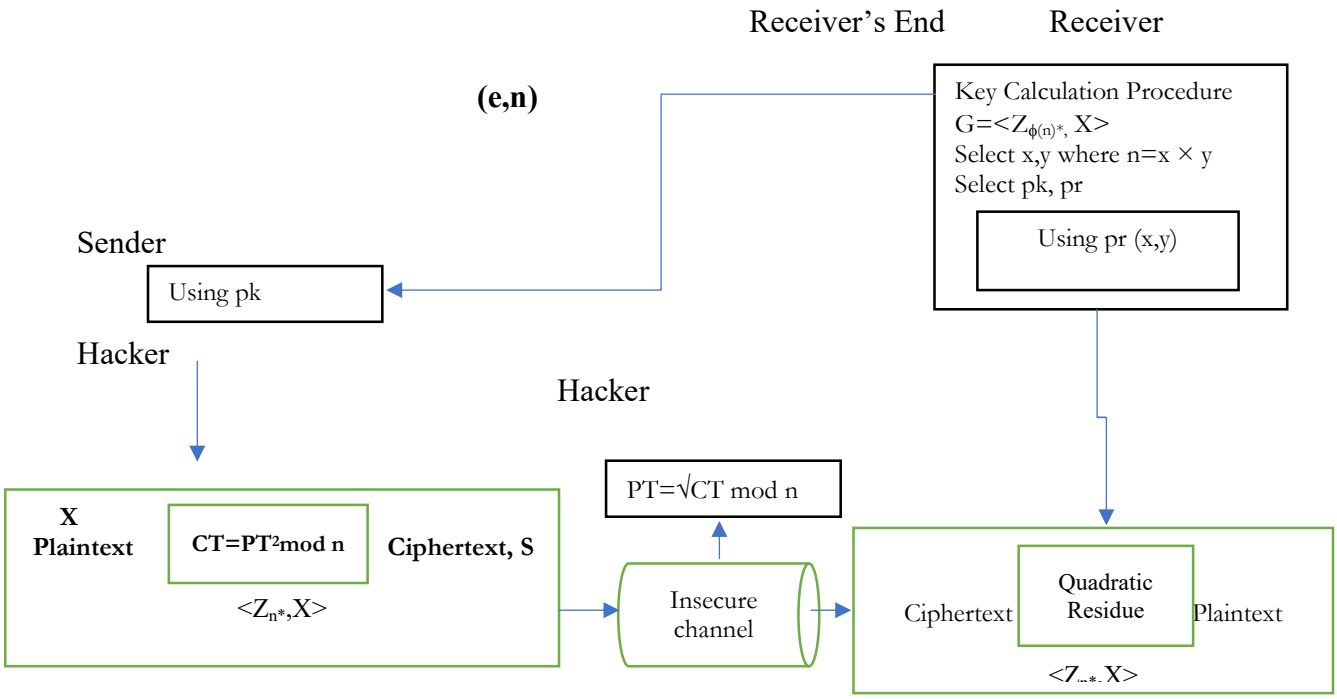


Figure 7 Process of Communication of Rabin Cryptosystem

3.4.1 Key_ gen in Robin cryptosystem

Select two large prime numbers, p and q in form of 4K+3 such that x≠y
n ⟵ x×y
Public_key ⟵ n
Private _key ⟵ (y, n)
Return private and public key

3.4.2 Encryption algorithm is as under

Rabin_ Encryption (n, PT)
CT ⟵ PT^2 mod n
Return CT

3.4.3 Decryption algorithm is as under

Rabin_Decryption (x,y,c)
s_1 ⟵ +$(CT^{(x+1)/4})$ mod x
s_2 ⟵ -$(CT^{(x+1)/4})$ mod x
t_1 ⟵ +$(CT^{(y+1)/4})$ mod y
t_2 ⟵ -$(CT^{(y+1)/4})$ mod y
m_1 ⟵ Chinese _remainder Thm (s_1, t_1, x, y)
m_2 ⟵ Chinese _remainder Thm (s_1, t_1, x, y)
m_3 ⟵ Chinese _remainder Thm (s_2, t_1, x, y)
m_4 ⟵ Chinese_remainder Thm (s_2, t_2, x, y)

Return m_1, m_2, m_3, m_4

Non-Deterministic nature of Rabin cryptosystem is very important for security of messages across the network as receiver can pick any one of the messages among four as final message but security of Rabin cryptosystem depends upon length of x and y only and is as secure as RSA Cryptosystem.

3.5 Elliptic Curves Cryptosystem

RSA is secure cryptosystem but it comes with issue of size of keys to secure the cipher. Elliptic curves cryptosystem offers same security with less key length based in elliptic curves. Elliptic curves are cubic equation in two variables and general form of equation is as Under

$$y^2+b_1\,xy+b_2y = x^3+a_1x^2+a_2x+a_3$$

And in case of real numbers: $y^2 = x^3+ax+b$

3.5.1 Elliptic curves

Elliptic curve considered as a basic building block for a better trapdoor. Elliptic curve is a cryptographic algorithm proposed in 1985 and is constructed on esoteric mathematics branch.

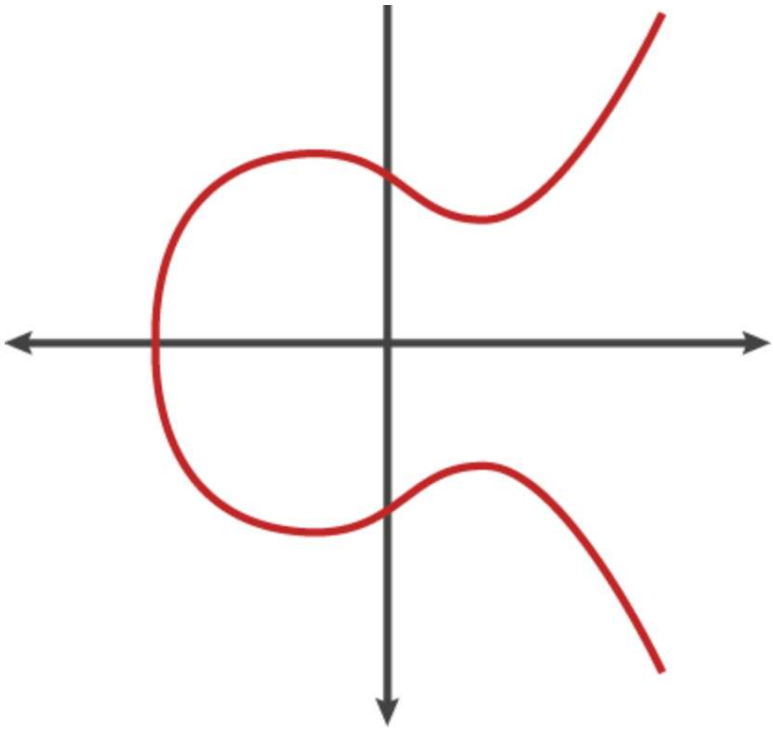


Figure 8 Elliptic curve for cryptography

Elliptic curves can be represented in many ways, however specifically it is represented as the points set which satisfy an equation in two variable in which one of the variable is

of degree two and other is of degree three. Elliptic curve have some attributes which make it suitable for cryptography as represented in figure 8. One of the attributes of fig 8 includes horizontal symmetry. The curve remains same on any point contemplate over x-axis. The other engrossing attribute includes every non-vertical lines will intersect the curve not more than 3 places.

To impose this condition one needs to limit numbers to a static range as in the case of RSA. The best method for this restriction is to use static whole numbers instead of allowing any numeric value plot on the curve. For formula formation of elliptic curve i.e. $y^2 = x^3 + ax + b$, one need to roll atop numbers when maximum number hits. The elliptic curve is termed as prime curve when maximum hit number is prime and is proved as magnificent cryptographic attributes.Figure 9, shows the elliptic curve for the equation $y^2 = x^3 - x + 1$ plotted for every number.

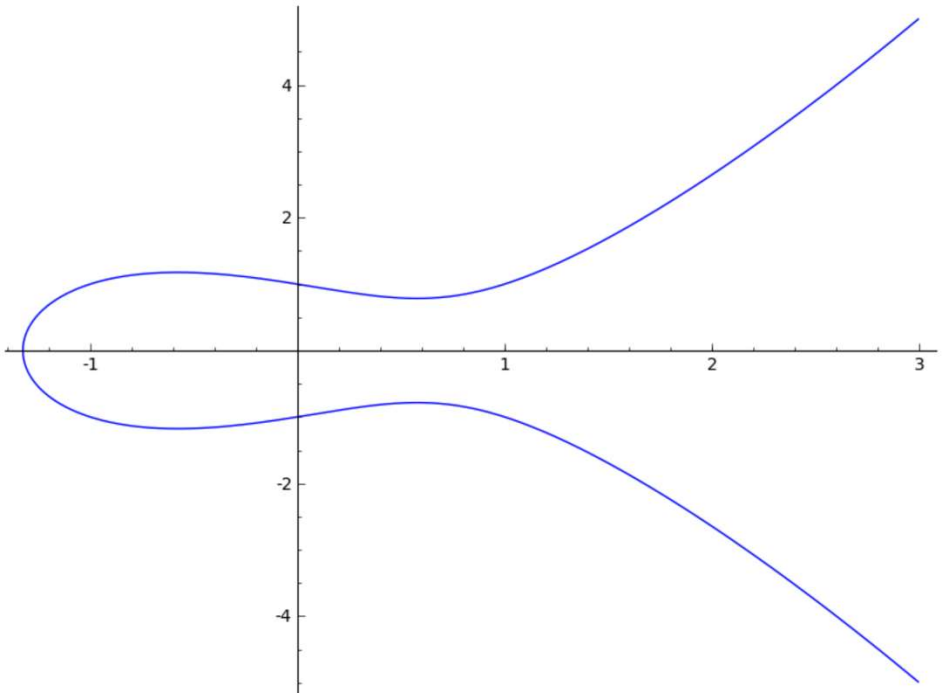


Figure 9: Elliptical curve for equation $y^2 = x^3 - x + 1$

By using this depiction, messages can be represented as point values on the curve. The message can be set on x-coordinate and by using equation we will get the value of y which results in point on the curve. And the following points are obtained
(70,6), (76,48), , (82,6), (69,22)

There are three major requirements of elliptic curve cryptosystem i.e. a maximum prime number, curve equation, and a public curve point. The private key is represented as a number say pr and public key represented as a dotted point pr times. The process of computation of private key from public key is termed as Elliptic Curve Discrete Logarithm (ECDL) function. This is termed as a Trapdoor function. ECDL is considered as a hard problem supporting elliptic curve cryptography. Many researchers working since past three decades to solve this problem but still no specific solution found by them.

IV. CONCLUSION

Various existing works related to encoding, compression and encryption has been studied in addition to attacks on various cryptosystem and enhancement of security, generation of keystream for symmetric key encryption algorithms and increasing the transmission speed of M have been studied. Still elliptic curve cryptography is very good candidate among all and in future we can think of hybrid approaches for more security of message in motion and rest as well.

REFERENCES

[1] Whitfield Diffie and Martin E. Hellman, "Exhaustive cryptanalysis of the NBS Data Encryption Standard", Computer Magazine, pp. 74-84, 1977.
[2] Lenstra A K and Verheul E R, "Selecting Cryptographic Key Sizes", The 2000 International Workshop on Practice and Theory in Public Key Cryptography (PKC2000), Melbourne, Australia, 2000.
[3] Shannon C E, "Certain Results in Coding Theory for Noisy Channels", Information and control, Vol. 1, pp. 6–25, 19574
[4] Tarek M Mahmoud, Bahgat A. Abdel-latef, Awny A. Ahmed, "Hybrid Compression Encryption Technique for Securing SMS", IJCSS, Vol. 3. Issue 6, 2010.
[5] Stephan Rein, Clemens Guhmann, Frank H. P. Fitzek, "Low-Complexity Compression of Short Messages", IEEE, Data Compression Conference, 2006.
[6] Biham E, Seberry J, Py (Roo), "A fast and secure stream cipher", Research Online: 2005.
[7]Fenwick P, "Burrows Wheeler Compression with Variable Length Integer Codes", Software–Practice and Experience, Vol. 32, No. 13, pp. 1307–1316, Nov. 2002.
[8] Porwal S, Chaudhary Y, Joshi J, Jain M, "Data Compression Methodologies for Lossless Data and Comparison between Algorithms", International Journal of Engineering Science and Innovative Technology (IJESIT), Vol. 2, Issue 2, Mar. 2013.
[9] Sreelajaa N K and Pai GAV, "Stream cipher for binary image encryption using Ant Colony Optimization based key generation", Journal of Applied Soft Computing, Vol. 12, pp. 2879–95, 2012.
[10]Imad Khaled, Salah, Abdullah Darwish and Saleh Oqeilli, "Mathematical Attacks on RSA Cryptosystem", Journal of Computer Science, Vol. 2, No. 8, pp:665-671, 2006.
[11] Majid Bakhtiari and Mohd Aizaini Maarof, "An Efficient Stream Cipher Algorithm for Data Encryption", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 3, No. 1, May 2011.
[12] Iwan Handoyo Putro, Petrus Santoso and Maya Basoeki, "A Short Text Compression Scheme based on Arithmetic Coding", 2007.
[13] Data_Compression available at website: http://en.wikipedia.org/wiki/ Data_Compression.
[14] Meier W and Staffelbach O, "Nonlinearity Criteria for Cryptographic Functions, Advances in Cryptology", EUROCRYPT '89, J-J Quisquater and J V andewalle Editors, Springer Berlin / Heidelberg, pp: 549-562, 1990.
[15] Charles P fleeger and Shari Lawrence P fleeger, Security in computing, Fourth Edition, Prentice Hall of India Pvt Ltd., New Delhi, 2007.

[16] Diaa Salama Abdul Minaam, Hatem M. Abdual-Kader, and Mohiy Mohamed Hadhoud, "Evaluating the Effects of Symmetric Cryptography Algorithms on Power Consumption for Different Data Types", IJ Network Security, Vol. 11 (2), 2010.
[17] Muthumanickam T, "Performance Analysis of Cryptographic VLSI Data", IRACST–International Journal of Computer Networks and Wireless Communications (IJCNWC), Vol. 2, No. 1, 2012.
[18] Mijanur Rahaman and Md. Masudul Islam, "An Overview on Quantum Computing as a Service (QCaaS): Probability or Possibility", International Journal of Mathematical Sciences and Computing (IJMSC), Vol. 2, No. 1, pp. 16-22, 2016.
[19] Erdem S S, Yanik T Ko C and C K., "Fast Finite Field Multiplication. In: C.K. Ko, c (ed.) Cryptographic Engineering", Springer, 2009.
[20] Longa P and Miri A, "New Composite Operations and Precomputation Scheme for Elliptic Curve Cryptosystems over Prime Fields. In: PKC 2008", LNCS, Vol. 4939, pp. 229-247, Springer, Heidelberg, 2008.
[21] Gopinath Ganapathy and Mani K, "Maximization of Speed in Elliptic Curve Cryptography Using Fuzzy Modular Arithmetic over a Microcontroller base Environment", Lecture Notes in Engineering and Computer Science, World Congress on Engineering and Computer Science (WCECS), IAENG, San Francisco,USA, Vol. 1, pp. 328-332, Oct. 2009.
[22] Rubaiyat Hasan M D, "Data Compression using Huffman based LZW Encoding Technique", International Journal of Scientific & Engineering Research, Vol. 2, No. 11, pp. 1-7, Nov. 2011.
[23] Rajan S Jamgekar and Geeta Shantanu Joshi, "File Encryption and Decryption Using Secure RSA", International Journal of Emerging Science and Engineering (IJESE), Vol. 1, Issue 4, Feb. 2013.
[24] Monisha Sharma, Chandrashekhar Kamargaonkar and Amit Gupta, "A Novel Approach of Image Encryption and Decryption by using partition and Scanning Pattern", International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 7, Sep. 2012. (IJARCET), Vol. 7, Issue 1, Jan. 2018.

Stock Market Prediction Using LSTM Techniques in Machine Learning

Shanu Gaur[1]*, Keshav Kori[2], Madhuri Nigam[3]
[1, 2] Computer Science and Engineering, Chhatrapati Shivaji Institute of Technology, Durg, Chhatisgarh, India
[3]Computer Science and Engineering, Sage University, Indore
shanu.nine@gmail.com, keshav.kori@csitdurg.in, mnigam1012@gmail.com

Abstract

One of the most multifaceted Artificial Intelligence Machine Learning issues is the offer worth forecast. It relies upon an assortment of elements that influence gracefully and request. This paper dissects various techniques for determining the future stock cost and gives a model utilizing a pre-constructed model that is adjusted to the Indian financial exchange. Costs of stocks are portrayed by time-arrangement information, and neural organizations are prepared to take in the examples from patterns in the current information. In this exploration, we additionally look at between aftereffects of a straightforward single LSTM model, a stacked LSTM model adjusted to the Indian market by utilizing the Indian BANK stock value informational index, and a mixture model utilizing both stacked auto-encoders and estimation investigation.

Keywords: - *Artificial Intelligence, Machine Learning, LSTM*

I. INTRODUCTION

The stock market is basically an aggregation of various buyers and sellers of stock. A stock (also known as shares more commonly) in general represents ownership claims on business by a particular individual or a group of people. The attempt [3] to determine the future value of the stock market is known as a stock market prediction. The prediction is expected to be robust, accurate and efficient. The system must work according to the real-life scenarios and should be well suited to real-world settings. The system is also expected to take into account all the variables that might affect the stock's value and performance. There are various methods and ways of implementing the prediction system like Fundamental Analysis, Technical Analysis, Machine Learning, Market Mimicry, and Time series aspect structuring. With the advancement of the digital era, the prediction has moved up into the technological realm. The most prominent and [3] promising technique involves the use of Artificial Neural Networks, Recurrent Neural Networks that is basically the implementation of machine learning. Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again.

*Corresponding Author
Shanu Gaur,
Research Scholar,
Computer Science and Engineering, Chhatrapati Shivaji Institute of Technology, Durg, Chhatisgarh, India
✉*Email: shanu.nine@gmail.com*

Traditional methods of prediction in machine learning use algorithms like Backward Propagation, also known as Back propagation errors. Lately, many researchers are using more of ensemble learning techniques. It would use low price and time [3] lags to predict future highs while another network would use lagged highs to predict future highs. These predictions were used to form stock prices. [1]

Stock market price prediction for short time windows appears to be a random process. The stock price movement over a long period of time usually develops a linear curve. People tend to buy those stocks whose prices are expected to rise in the near future. The uncertainty in the stock market refrain people from investing in stocks. Thus, there is a need to accurately predict the stock market which can be used in a real-life scenario..

II. LITERATURE SURVEY

2.1 Survey of Stock Market Prediction Using Machine Learning Approach

The stock market prediction has become an increasingly important issue in the present time. One of the methods employed is technical analysis, but such methods do not always yield accurate results. So it is important to develop methods for a more accurate prediction. Generally, investments are made using predictions that are obtained from the stock price after considering all the factors that might affect it. The technique that was employed in this instance was a regression. Since financial stock marks generate enormous amounts of data at any given time a great volume of data needs to undergo analysis before a prediction can be made. Each of the techniques listed under regression hasits own advantages and limitations over its other counterparts. One of the noteworthy techniques that were mentioned was linear regression. The way linear regression models work is that they are often fitted using the least squares approach, but they may alternatively be also be fitted in other ways, such as by diminishing the "lack of fit" in some other norm, or by diminishing a handicapped version of the least squares loss function. Conversely, the least squares approach can be utilized to fit nonlinear models. [1]

2.2 Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests.

The use of machine learning and artificial intelligence techniques to predict the prices of the stock is an increasing trend. More and more researchers invest their time every day in coming up with ways to arrive at techniques that can further improve the accuracy of the stock prediction model. Due to the vast number of options available, there can be n number of ways on how to predict the price of the stock, but all methods don’t work the same way. The output varies for each technique even if the same data set is being applied. In the cited paper the stock price prediction has been carried out by using the random forest algorithm is being used to predict the price of the stock using financial ratios form the previous quarter. This is just one wayof looking at the problem by approaching it using a predictive model, using the random forest to predict the future price of the stock from historical data. However, there are always other factors that influence the price of the stock, such as sentiments of the investor, public opinion about the company, news from various outlets, and even events that cause the entire stock market to fluctuate. By using the financial ratio along with a model that can effectively

analyze sentiments the accuracy of the stock price prediction model can be increased. [2]

2.3 Stock Market Prediction via Multi-Source Multiple Instance Learning

Accurately predicting the stock market is a challenging task, but the modern web has proved to be a very useful tool in making this task easier. Due to the interconnected format of data, it is easy to extract certain sentiments thus making it easier to establish relationships between various variable and roughly scope out a pattern of investment. Investment pattern from various firms show sign of similarity, and the key to successfully predicting the stock market is to exploit these same consistencies between the data sets. The way stock market information can be predicted successfully is by using more than just technical historical data, and using other methods like the use of sentiment analyzer to derive an important connection between people's emotions and how they are influenced by investment in specific stocks. One more important segment of the prediction process was the extraction of important events from web news to see how it affected stock prices. [3]

2.4 Stock Market Prediction: Using Historical Data Analysis

The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the stock market plays an important role in business and finance. The technical and fundamental analysis is done by sentimental analysis process. Social media data has a high impact due to its increased usage, and it can [6].be helpful in predicting the trend of the stock market. Technical analysis is done [6] using by applying machine learning algorithms on historical data of stock prices. The method usually involves gathering various social media data, news to extract sentiments expressed by individuals. Other data like previous year stock prices are also considered. The relationship between various data points is considered, and a prediction is made on these data points. The model was able to make predictions about future stock values.

2.5 A Survey on Stock Market Prediction Using SVM

The recent studies provide a well-grounded proof that most of the predictive regression models are inefficient in out of sample predictability test. The reason for this inefficiency was parameter instability and model uncertainty. The studies also concluded the traditional strategies that promise to solve this problem. Support vector machine commonly known as SVM provides with the kernel, decision function, and sparsity of the solution. It is used to learn polynomial radial basis function and the multi-layer perceptron classifier. It is a training algorithm for classification and regression, which works on a larger dataset. There are many algorithms in the market but SVM provides with better efficiency and accuracy. The correlation analysis between SVM and stock market indicates strong interconnection between the stock prices and the market index.

2.6 Predicting Stock Price Direction Using Support Vector Machines

Financial organizations and merchants have made different exclusive models to attempt and beat the market for themselves or their customers, yet once in a while has anybody accomplished reliably higher-than-normal degrees of profitability. Nevertheless, the

challenge of stock forecasting is so engaging in light of the fact that the improvement of only a couple of rate focuses can build benefit by a large number of dollars for these organizations. [6]

III. METHODOLOGIES

3.1 Classification

Classification is an instance of supervised learning where a set is analyzed and categorized based on a common attribute. From the values or the data are given, classification draws some conclusion from the observed value. If more than one input is given then classification will try to predict one or more outcomes for the same. A few classifiers that are used here for the stock market prediction includes the random forest classifier, SVM classifier.

3.1.1 Random Forest Classifier

Random forest classifier is a type of ensemble classifier and also a supervised algorithm. It basically creates a set of decision trees that yields some result. The basic approach of random class classifier is to take the decisionaggregate of random subset decision tress and yield a final class or result based on the votes of the random subset of decision trees.

Parameters

The parameters included in the random forest classifier are n_estimators which is total number of decision trees, and other hyper parameters like oob-score to determine the generalization accuracy of the random forest, max_features which includes the number of features for best-split. Min_weight_fraction_leaf is the minimum weighted fraction of the sum total of weights of all the input samples required to be at a leaf node. Samples have equal weight when sample weight is not provided.

3.1.2 SVM classifier

SVM classifier is a type of discriminative classifier. The SVM uses supervised learning i.e. a labeled training data. The output are hyper planes which categorizes the new dataset. They are supervised learning models that uses associated learning algorithm for classification and as well as regression.

Parameters

The tuning parameters of SVM classifier are kernel parameter, gamma parameter and regularization parameter.

- Kernels can be categorized as a linear and polynomial kernel calculates the prediction line. In linear kernels prediction for a new input is calculated by the dot product between the input and the support vector.
- C parameter is known as the regularization parameter; it determines whether the accuracy of model is increases or decreases. The default value of c=10.Lower regularization value leads to misclassification.

- Gamma parameter measures the influence of a single training on the model. Low values signifies far from the plausible margin and high values signifies closeness from the plausible margin.

IV. SYSTEM ARCHITECTURE

Kaggle is an online community for data analysis and predictive modeling. It also contains dataset of different fields, which is contributed by data miners. Various data scientist competes to create the best models for predicting and depicting the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a few companies.

The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features. These features are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection.

The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model. The splitting is done in a way that training data maintain a higher proportion than the test data.

The random forest algorithm utilizes a collection of random decision trees to analyze the data. In layman terms, from the total number of decision trees in the forest, a cluster of the decisiontrees looks for specific attributes in the data. This is known as data splitting. In this case, since the end goal of our proposed system is to predict the price of the stock by analyzing its historical data.

V. MODULE IDENTIFICATION

The various modules of the project would be divided into the segments as described.

5.1 Data Collection

Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we will be

analyzing the Kaggle dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

5.2 Pre Processing

Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

5.3 Training the Machine

Training the machine is similar to feeding the data to the algorithm to touch up the test data. The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyper parameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyper parameter values**.**

Finally, we will calculate a cross-validated score, for individual sets of hyper parameters. Then, we select the best hyper parameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

5.4 Data Scoring

The process of applying a predictive model to a set of data is referred to as scoring the data. The technique used to process the dataset is the Random Forest Algorithm. Random forest involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

VI. PROPOSED SYSTEM

As debated overhead stock market forecast is a huge subject and has a lot parts on which we can investigation upon, but one object all models have in common is their check on correctness of how well the models practical can fit to a given dataset and is it identical the results and forecasting correctly or not. Still each model has a few effects in common, they all need a list of companies of any stock exchange to forecast upon the three basic situations of market buy, hold, and sell and to do this the stock market data for each company against their tickers was stored in machine (to avoid larger accessing

time) and data manipulations were performed in order to prepare the dataset for additional machine learning classifiers which will ultimately forecast the marks and deliver the output.

6.1 Feasibility Study

To plaid the practicability of the overhead model the given productivity will be plaid and coordinated alongside the graph of the definite company for that period of time and distinguish the patterns. As a future Scope in our project we will further use quantopiana online platform for emerging trading approaches and back testing them, we will use it to advance a plan on quantopian and back test it to check the possibility of the tactic.

VII. RESULT AND DISCUSSION

As it can be grasped in the figure given underneath, one side it demonstrations the forecast counter spread of the company future prices, and additional figure demonstrations the graph of the company at that particular time of year in terms to the forecast and it can be detected that much of the outcomes are precise. As it can be perceived that the data spread is habitually saying buy the stock, it can be incorrect on the hold condition because the teaching data will never be perfectly stable ever, so supposedly if the model forecast buy then this would be 1722 correct out of 4527 which is still good and actually a better score than it attained, and it still is getting the above accuracy mark of 33% which is decent in a stock market analysis. Many situations will static be there which machines can miss out, supposedly this has circumstances to buy, sell, hold and sometimes the model can be penalised, say the model predictable a 2% rise in the following seven days, but the growth only went up to 1.5% and departed 2% the next day, then the model will forecast (buy, hold) rendering to the 1.5% rise in the seven days and give the predictable spread. A model can also be penalised if supposedly the growth went 2% up and then suddenly falls 2% short the next day, this sort of outcomes in real trading would be thoughtful and same goes for the classical of it turns out to be highly precise. Now observing at the spread and the graph of the company notice around the era of 2017 the company was growing in the market so therefore there were actually more buys, which rapidly fallen in 2018, but the data we mined was till 31, December 2017 and it displays that at the starting of the year it had lot of buys, hence 1722 out of 4527 which speedily was sold just in a tiny time hence a lot of sells more than the holds, giving 1424 out 4527, the model may not be perfectly accurate but has a very close range of decisions which can be accepted in real trading or using algorithms to trade.

VIII. CONCLUSION AND FUTURE ENHANCEMENT

Hereby, it can be proposed that no trading algorithm can be 100% effective, not only 100%, it will typically never be close to 70% but to attain even an accuracy of 40% or 35% is still good sufficient to get a good forecast spread. Although extreme attained

accurateness was 39%, it was still able to closely forecast the predictable outcome and have coordinated against the company graph. To make our expectation more efficient, it can be done by including bulky data sets that have millions of entries and could train the machine more powerfully. Different activities of stocks can lead to diverse raises or lows in the forecast price, use these movements to magistrate whether a company should be traded in or not. No training Data can ever be stable, hence there are always some unevenness which can be seen in the above data spread, but to still forecast close to an consequence will also lead to a good approach if it has greater than 33% accuracy. While, developing a strategy trader should always think to always have nominal imbalance while still being above 33%accurate.

It can also be determined that in a stock market, there is probable that some companies might not be associated at all, and mostly can be associated to each other, and can help justice movements of stock accordingly, we can scale affairs and see how much in percentages they are correlated.
Including gigantic data sets, to increase more effectiveness, and in data set if had nan values in tables, because of two simple reasons either a specific company wasn't opened during that time of year, or he data is not readily obtainable, in both the cases replace the null values with 0 , which is somewhat that trader might want to change while developing a trading tactic.
Furthermore, there can be back testing of the trading strategy, using zip line and quantopian a python platform for testing trading strategies and can see how well can a model fit into some random data of stock, and can the model from this random data of stock develop relations and correlations, and predict on terms of change.

REFERENCES

[1]. Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.
[2]. Loke.K.S. "Impact Of Financial Ratios And Technical Analysis On Stock Price Prediction Using Random Forests", IEEE, 2017.
[3]. Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, "Stock Market Prediction via Multi-Source Multiple Instance Learning." IEEE 2018.
[4]. VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. "Stock Market Prediction: Using Historical Data Analysis", IJARCSSE 2017.
[5]. SachinSampatPatil, Prof. KailashPatidar, Asst. Prof. Megha Jain, "A Survey on Stock Market Prediction Using SVM", IJCTET 2016. https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde.pdf
[6]. Hakob GRIGORYAN, "A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)", DSJ 2016.
[7]. RautSushrut Deepak, ShindeIshaUday, Dr. D. Malathi, "Machine Learning Approach In Stock Market Prediction", IJPAM 2017.
[8]. Pei-Yuan Zhou , Keith C.C. Chan, *Member, IEEE,* and Carol XiaojuanOu, "Corporate Communication Network and Stock Price Movements: Insights From Data Mining", IEEE 2018.

Comparison Analysis and Implementation of Prediction of Heart Disease

Pooja Yadav[1*], Kranti Jain[2]
Computer Science and Engineering, CSIT Durg, Chhattisgarh, India
poojayadav2409@gmail.com, kranti.jain@csitdurg.in

Abstract

In the current decade Heart disease is most common for all from 14-60 years of age. Now a days heart disease having many types of constraint to predict the information like hyper tension, Blood pressure increase, Blockage of Nervous system. So that the percentage of blockage must identified which surgery is good for the health of patient. In this paper we are trying to find the better method to predict and we also used algorithms for prediction. We use various algorithm for the prediction of blockage of nerves like Naïve Bayes, algorithm is analyzed on dataset based on hazard factors. Here we also try to use decision trees and grouping of algorithms for the calculation of heart disease based on the features. The results shown the comparison of various four algorithms of data mining which compare their accuracy of the result which one is better through graph..

Keywords: *Decision tree, Data mining, Heart Disease Prediction, Naïve Bayes, K-means, Machine learning.*

I. INTRODUCTION

The main topic is prediction using machine learning technics. Machine learning is widely used now a days in many business applications like e commerce and many more. Prediction is one of area where this machine learning used, our topic is about prediction of heart disease by processing patient's dataset and a data of patients to whom we need to predict the chance of occurrence of a heart disease. Data mining practices that are valuable in heart disease forecast with the assistance of dissimilar data mining tools that are accessible. If the heart doesn't function properly.The major challenge that the Healthcare industry faces now-a-days is superiority of facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service.Even though heart disease is acknowledged as the supreme chronic sort of disease in the world, it can be most avoidable one also at the same time. A healthy way of life (main prevention) and timely analysis. Heart expert's create a good and huge record of patient's database and store them. There is always a need to find out the Heart data set, such as Blood clotting, nerves damage, how much percentage blocking must be the challenging part of people. Additionally, there are situations when silent heart attack is preferred: for example, during an operation, seviour attack must be held in some times. It is hard for most people who are not familiar with a heart failure will take up certain parameters without an interpreter.

*Corresponding Author
Pooja Yadav
M Tech Scholar, Chatrapati Shivaji Institute of Technology Durg, Chhattisgarh, India.
✉Email:poojayadav2409@gmail.com

Thus, software that transcribes certain points of blockage must have find out and get

prediction of cover-up interactive training for people to learn a parameter. Heart Disease has become an important research field with the current focus on interactive feature selection and the Prediction analysis.

II. LITERATURESURVEY

In [2] Mohammed Abdul Khaleel has given a paper in the Survey of Techniques for the mining of information on Medical Data for Finding Frequent Diseases locally. This paper centers around dismember data mining systems that are required for restorative data mining especially to discover locally-visit sicknesses, for instance, heart ailments, lung threat, chest infection and so on. In-arrangement mining is the route toward removing data for finding latent models which Vembandasamy et al. performed work, to dissect and recognize coronary illness. In this, the calculation utilized was the Naive Bayes calculation. In the Naïve Bayes calculation, they utilized the Bayes hypothesis. Consequently Naive Bayes has the exceptionally ground-breaking to make suspicions freely. The pre-owned informational collection is acquired from diabetic examination establishments of Chennai, Tamilnadu which is a main foundation. There are in excess of 500 patients in the dataset. The instrument utilized is Weka and characterization is executed by utilizing 70% of Percentage Split. The precision offered by Naive Bayes is 86.419%.

In [3]. Costas Sideris, Nabil Alshurafa, HaikKalantarian, and Mo-HammadPourhomayoun have given a paper named Remote Health Monitoring Outcome Success forecast utilizing First Month and Baseline Intervention Data. RHS frameworks are viable in sparing expenses and lessening sickness. In this paper, they depict an up- evaluated RHM structure, Wanda-CVD that is cellphone-based and proposed to give far off educating and social assistance to individuals. CVD balancing activity measures are seen as a fundamental concentration by social protection relationship around the globe.

In [4]. L.Sathish Kumar and A. Padmapriya have given a paper named Prediction for similitudes of ailment by utilizing the ID3 calculation on TV and cell phone. This paper gives a modified and hid approach to manage perceiving plans that are concealed of coronary sickness. The given system use data mining strategies, for instance, the ID3 calculation. This proposed technique helps the individuals not exclusively to think about the maladies yet it can likewise assist with lessening the passing rate and tally of sickness influenced individuals.

In [5]. M.A.NisharaBanu and B.Gomathy have given a paper named Disease Predicting framework utilizing information mining strategies. In this paper, they talk about MAFIA (Maximal Frequent Item set calculation) and K-Means bunching. As arrangement is significant for the expectation of malady. The characterization dependent on MAFIA and K-Means brings about mistake.

In [6]. Wiharto and HariKusnanto have given a paper named Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. In this paper, they show a desire structure for heart disease using Learning vector Quantization neural framework computation The neural framework in this system recognizes 13 clinical incorporates as data and predicts that there is a proximity or nonattendance of coronary sickness in the patient, close by different execution measures.

In [7]. D.R. Patil and Jayshril S. Sonawane have given a paper named Prediction of Heart Disease Using Learning Vector Quantization Algorithm. In this paper they display a desire structure for heart contamination using Learning vector Quantization neural framework estimation The neural framework in this system recognizes 13 clinical incorporates as data and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various executionmeasures.

III. METHODOLOGY

3.1. Data Preprocessing

Cleaning: Data that we need to handle won't be spotless that is it might contain clamor or it might contain values missing of we measure we can't get great outcomes so to acquire great and impeccable outcomes we have to dispose of this, the cycle to take out this is information cleaning. We will fill missing qualities and can eliminate clamor by utilizing a few procedures like loading up with the most widely recognized an incentive in

Change: This includes changing information design starting with one structure then onto the next that is making them generally justifiable by doing standardization, smoothing, and speculation, accumulation strategies on information.

Coordination: Data that we need not cycle may not be from a solitary source here and there it very well may be from various sources we don't incorporate them it might be an issue while preparing so reconciliation is one of a significant stage in information pre-handling and various issues are considered here to incorporate.

Decrease: When we chip away at information it might be unpredictable and it might be hard to see here and there so to make them reasonable to the framework we will diminish them to the necessary arrangement so we can accomplish great outcomes.

To do this we have many machine learning algorithms out of which we the more widely used methods are Naïve Bayes classification technic and decision tree construction, in this decision tree construction we have many algorithms one which we took for this ID3 algorithm. The ID3 algorithm is one of the old algorithms which is used for building decision trees in the process of building a decision tree it handles missing values and removes outliers [2]. So we can build this decision tree even the data is not cleaned well. Decision tree constructs classification or regression models as a structure that is similar to a tree. It separates a dataset into fewer and fewer sub-sets while in the meantime a related decision tree is incrementally created. The last outcome is a tree with a choice point and leaf point [8]. A choice node has a minimum of 2 branches. Leaf nodes speak to a grouping or choice. The highest choice hub in a tree which compares to the best indicator called root point. Choice trees can deal with both all out and numerical information.

ID3 is an algorithm that is used to build decision trees [2]. ID3 has some features like removing outliers, handling missing values, and but their major disadvantage is to over-

fitting. And it's not so easy to implement as that of the Naïve Bayes algorithm.

Step 1: If all occasions in X are certain, then make YES node and end. On the off chance that all cases in X are negative, make a NO node and end. Generally select an element, B with qualities U1... Un, and make a choice node.

Step 2: Partition the preparation occasions in X into subsets X1, X2… Xn as indicated by the estimations of U.

Step 3: apply the calculation recursively to each of the sets Ai.

3.2 Naïve-Bayes Classification:

The Naïve-Bayesian classifier relies upon Bayes' speculation with autonomy suppositions among attributes [7-13]. A Naïve- Bayesian output is definitely not hard to run, with no entrapped repetitive parameter estimation which makes it particularly

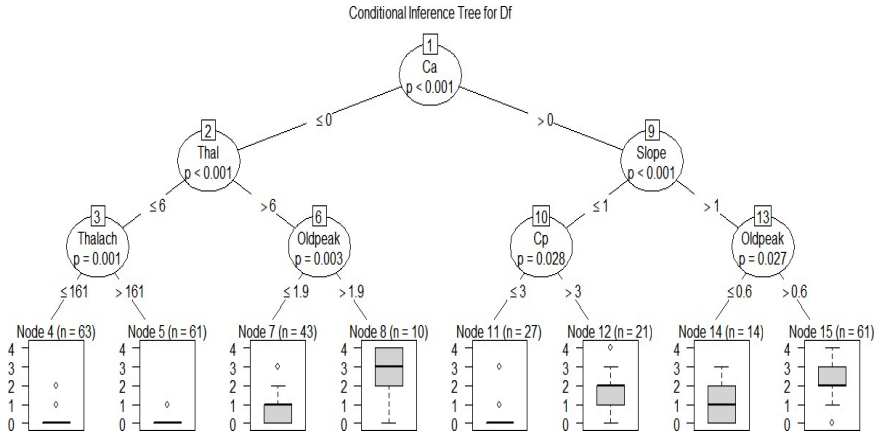
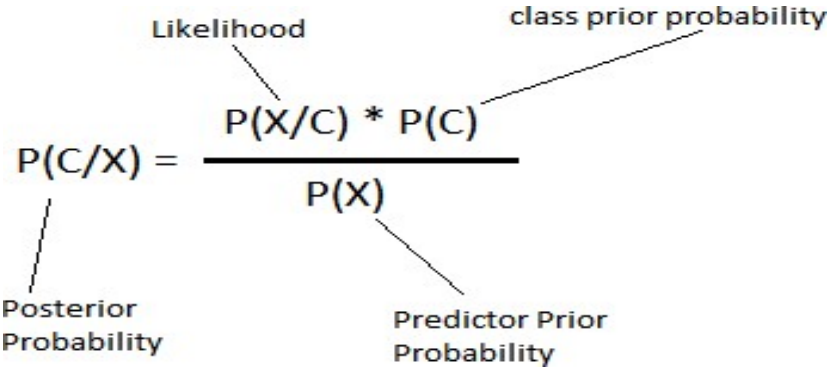


Figure 1 Demonstration of Naïve Bayes classification

supportive for broad datasets in spite of its effortlessness, the Naive Bayesian classifier generally completes its job shockingly good and is broadly used in light of the fact that it frequently outflanks high order techniques which are complex. The Naïve Bayes treats every variable as independent which helps it to predict even if variables don't



have proper relation [1].

- P(c|x) is the posterior probability of class (target) giv- en predictor(attribute)

- P(c) is the prior probability ofclass.
- P(x|c) is the likelihood which is the probability of predictor givenclass.
- P(x) is the prior probability ofpredictor.

3.3 K-Means

k-means clustering is one of the clustering technic used to cluster datasets based on nearest-neighbor here the data is clustered in k clusters based on a similarity between them we are also filled missing values of data using this k-means[6]. Once we clustered the data every dataset will come into any one of the clusters by using these clusters if we have missing values in the dataset we can fill those values as this is categorized into groups. Now as this missing values are all cleared we can apply different prediction technics on this for an example we can apply now as we know that for a dataset to be used for prediction in Naïve Bayes need to be pre-processed we can use this data for prediction in Naïve Bayes[1]. By different combinations of using these algorithms, we can achieve good accuracy.

We reviewed different papers on heart disease prediction out of all prediction technics and methods what everyone using when it comes to prediction is Naïve Bayes and decision trees we have different methods one of which that we used here is the ID3 algorithm. We took a medical data of heart disease patients from the UCI machine learning repository one of the popular repositories to get data for machine learning experiments it contains a record of nearly 300 patients we performed both this Naïve Bayes and ID3 technics on this training data using R tool. In the R tool, we used some 3rd party libraries like e1071 for implementing Naïve Bayes and part to construct a decision tree. In the data set that we took for implementing this contains variables

Thus preventing Heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the Heart diseases, and the predictions made are quite accurate.

The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and predictions are made with different algorithms KNN, Decision Tree, Random Forest, SVM, Logistic Regression etc. This is the Jupyter notebook code and dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'

I've used a variety of Machine Learning algorithms, implemented in Python, to predict the presence of heart disease in a patient. This is a classification problem, with input features as a variety of parameters, and the target variable as a binary variable, predicting whether heart disease is present or not.

3.4 Problem Statement

Previous research studies has examined the application of machine learning techniques for the prediction and classification of Heart disease. However, these studies focus on the particular impacts of specific machine learning techniques and not on the optimization of these techniques using optimised methods. In addition, few researchers attempt to use hybrid optimization methods for an optimized classification of machine learning. The most proposed studies in the literature exploit optimized techniques such as Particle Swarm Optimization and Ant Colony Optimization with a specific ML technique such as SVM, KNN or Random Forest.

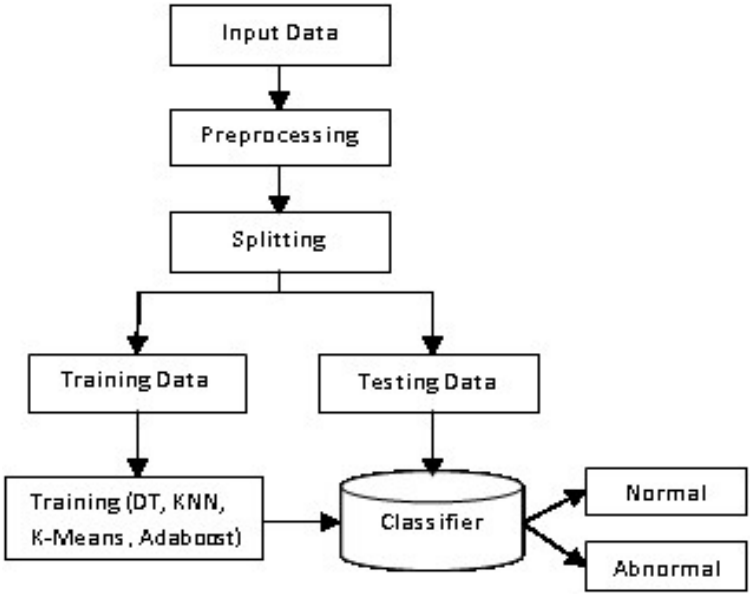


Figure 2 Problem Statement diagram

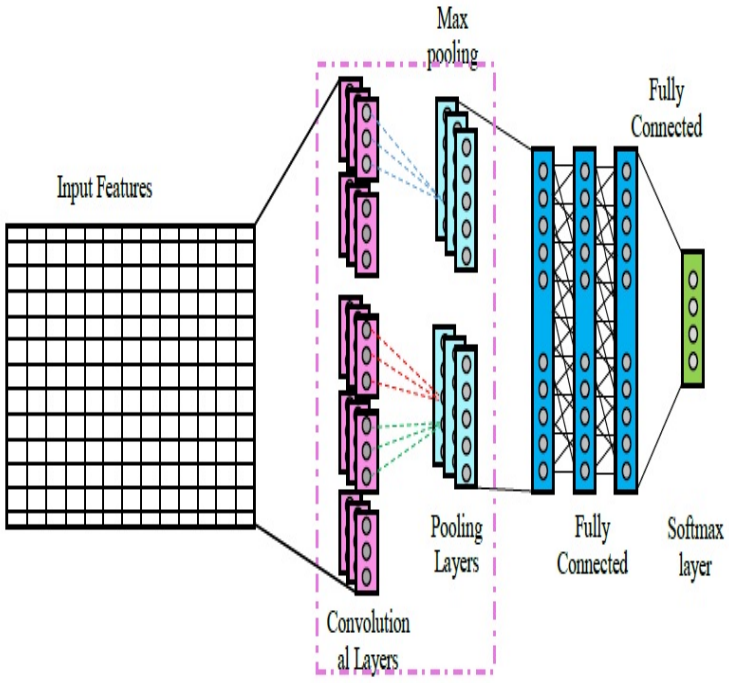


Figure 3 Structure of Convolution Neural Network

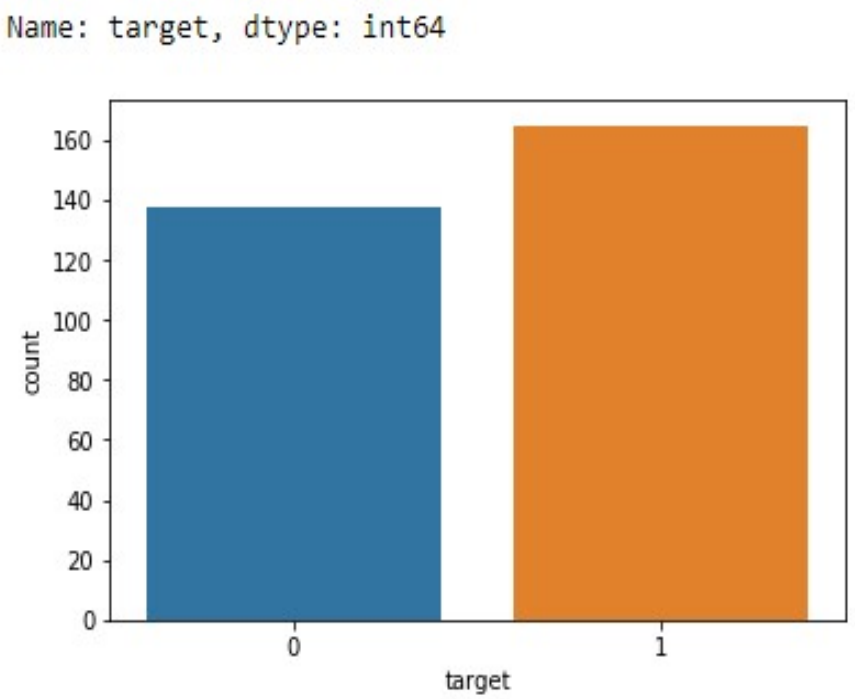


Figure 4 Analysis of Target Element

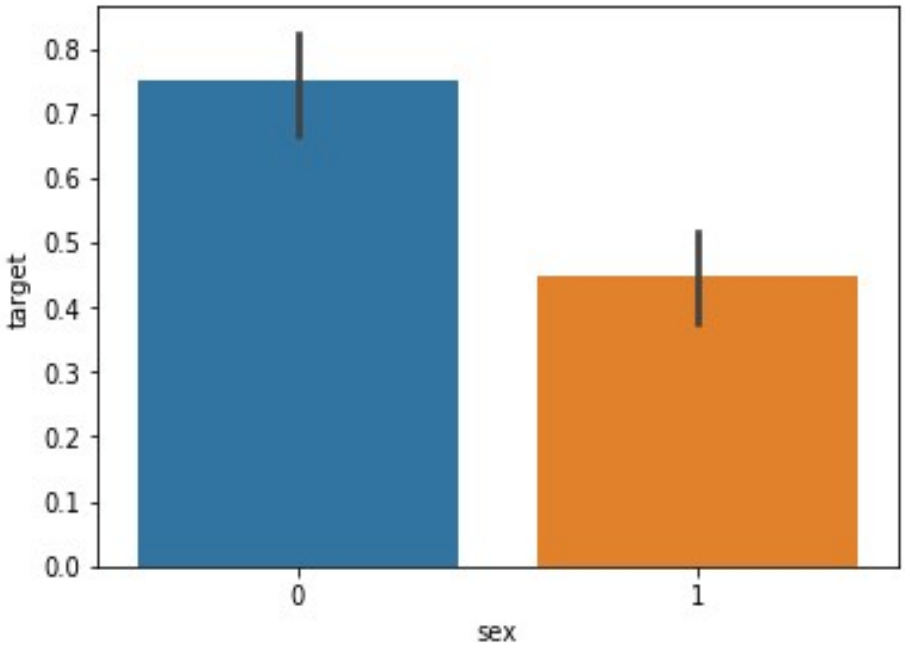


Figure 5 Analysis of the feature

```
In [7]: dataset.describe()
```

Out[7]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	3
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

Figure 6 Dataset Used

Performance Evaluation

$Accuracy = TP+TN/TP+FP+TN+FN$
$Precision = TP/\ TP+FP$
$Recall = TP/TP+FN$
$F\text{-}Measure = 2\times\ Precision\ \times\ RecallPrecision+Recall$

IV CONCLUSION

In this Paper we should find out during small datasets in some other cases most of time decision trees direct us to a solution which is not accurate, but when we look at Naïve Bayes results we are getting more accurate results with probabilities of all other possibilities but due to guidance to only one solution decision trees may miss lead. Finally we can say by this experiment that Naïve Bayes is more accurate if the input data is cleaned and well maintained even though ID3 can clean it self it cannot give accurate results every time, and in this same way Naïve Bayes also will not give accurate results every time we need to consider results of different algorithms and by all its results if a prediction is made it will be accurate. But we can use Naïve Bayes consider variables as individual we can use combination of algorithms like Naïve Bayes and K-means to get accuracy.

REFERENCES

[1] Sonam Nikhar, A.M. Karandikar "Prediction of Heart Disease Us- ing Machine Learning Algorithms" in International Journal of Ad- vanced Engineering, Management and Science (IJAEMS) June- 2016 vol-2
[2] Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal
[3] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, HaikKalantarian, Majid

Sarrafzadeh "Remote Health Monitoring Out- come Success Prediction using Baseline and First Month Interven- tion Data" in IEEE Journal of Biomedical and Health Informatics

[4] PonrathiAthilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Med- ical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march- 2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Pa- tients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[7] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dhavachelvan, P., Venkatachalapathy, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Alapatt, B.P., Kavitha, A., Amudhavel, J., "A novel encryption algorithm for end to end secured fiber optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amudhavel, J., Inbavalli, P., Bhuvaneswari, B., Anandaraj, B., Vengattaraman, T., Premkumar, K., "An effective analysis on har- mony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.

[10] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Bhuvaneswari Aadharshini, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journal of Advanced Re- search in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] Amudhavel, J., Kodeeshwari, C., Premkumar, K., Jaiganesh, S., Rajaguru, D., Vengattatraman, T., Haripriya, R., "Comprehensive analysis on information dissemination protocols in vehicular ad hoc networks", (2015) International Journal of Applied Engineering Re- search, 10 (3), pp. 2058-2061.

[12] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges, opportunities and analysis on security based cloud resource virtualization", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1458-1463.

[13] Amudhavel, J., Ilamathi, R., Moganarangan, N., Ravishankar, V., Baskaran, R., Premkumar, K., "Performance analysis in cloud au- diting: An analysis of the state-of-the-art", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2043-204.

Accelerated Testing for Durability of Reinforced Concrete

Sathish Kumar Sharma[1*], Shweta Kaushik[2]
[1*]Civil Engineering Department, Acropolis Institute of Technology and Research, Indore, India
[2]M.V.S.R Engineering College, Nadergul, Hyderabad-501510, Telangana
satishsharma@acropolis.in, shwetakaushik34@gmail.com

Abstract

A variety of cementing & supplementary materials are in use in concrete in India. Aggressive environments around deteriorate such cement concretes with passage of time. The paper focuses on results of accelerated testing of cement concrete materials in lab. and at sites from the standpoint of durability.
Key-Words: Durability, Deterioration, Simulation, Accelerated testing, Performance supplementary cementing materials.

I. INTRODUCTION

In present day concrete, a number of special cements find application. Supplementary Cementing Materials like Micro - silica & Fly Ash are also in use in High Performance Concrete. Such cement concretes have to encounter aggressive environments of chlorides, sulphates, nitrates etc. Interaction between special cement concretes & various aggressive environments needs to be understood from the view point of strength as well as durability. The paper discusses accelerated testing of cement concretes.

II. ACCELERATED TESTING

Deterioration of cement concretes is a time dependent phenomenon. One needs to know the long time performance in a short span of time to adopt appropriate measures for combating the aggressive environments which the cement concrete in question would be subjected to.

A. Methodology / approach

In order to accelerate the testing process, the following measures are adopted:

1. Cast test specimens of small size.
2. Increase the concentration of test chemicals (representing the aggressive environments).
3. Simulate the deterioration process such that accelerated results can be obtained.
4. Performance of special cement concretes should be compared with conventional cement concrete (made of Ordinary Portland cement or Portland Pozzolana).

*Corresponding Author
Dr. Satish Kumar Sharma
Professor and Head, Civil Engineering, Acropolis Institute of Technology and Research Indore. India
✉*Email:satishsharma@acropolis.in*

Results obtained in respect of parameters tested would yield comparative results of performance of various cement concretes in aggressive environments.

III. PERFORMANCE PARAMETERS

Performance parameters studied for assessing various cement concretes are in respect of:

1. Compressive strength
2. Durability

IV. EXPERIMENTAL ANALYSIS

For accelerated testing of various concretes, the following types of cement concretes (M20 grade) were chosen:

1. Ordinary Portland Cement (OPC) 33 grade
2. Portland Pozzolana Cement (PPC)
3. Sulphate Resisting Cement (SRC)
4. Portland Blast Furnace Slag Cement (PBFS)
5. OPC with replacement of 20% cement by 27.5% fly ash as per CBRI formula (OPF) [1].
6. SRC with replacement of cement by fly ash as per CBRI formula (SRF).
7. OPC with a waterproofing admixture (algae - proof) (OPCA).

Aggressive environments chosen for accelerated testing of cement concretes were:

1. 1.4% concentration of Chlorides in water (2 types of chlorides of Sodium & Magnesium) was used for testing.
2. 2.4% concentration of Sulphates (2 types of sulphates of Sodium & Magnesium) was used for testing.
3. Combination of 4% chlorides + 4% sulphates in water.
4. 4.4% Nitrate ($NaNO_3$) solution in water.
5. Sea water.

Accelerated tests chosen for evaluation of cement concretes were:

1. Testing of cement concrete cubes for compressive strength.
2. Testing of 40*40*160 mm prisms reinforced with steel bars for deterioration due to corrosion.
3. Testing of 150*150*750 mm reinforced concrete beams for compressive strength & deterioration due to corrosion & to study effect of cover to reinforcement.
4. Non - destructive tests using rebound hammer & ultrasonic pulse velocity for time dependent changes of strength parameters.
5. Salt spray test.
6. Polarization studies in respect of corrosion in specimens.
7. Electrolytic testing for corrosion.

Field studies were conducted in respect of strength parameters in:

1. Marine environment at Madras Harbour & Mandapam.
2. Chemical environment at Ammonium Chloride Plant in North Chennai.

Various tests as listed above were conducted in a total time span of 4 years.
After elaborate lab. & field studies as outlined above, performance indices were evolved for various special cement concretes in aggressive environments.

It was established that:

1. SRF performed much better than other cement concretes in environment of sulphates.
2. OPF performed better than other cement concretes in environments of chlorides and nitrates.
3. PPC & PBFS performed better than OPC & SRC in any aggressive environment in general.
4. SRC performed better than OPC in sulphates & performance was poor against chlorides.
5. OPCA performed better than OPC in general

After detailed investigations, performance indices were arrived at & are reported in tables 1 and 2.

Table 1. Ratings of performance of concretes in lab. Studies (1 to 7)

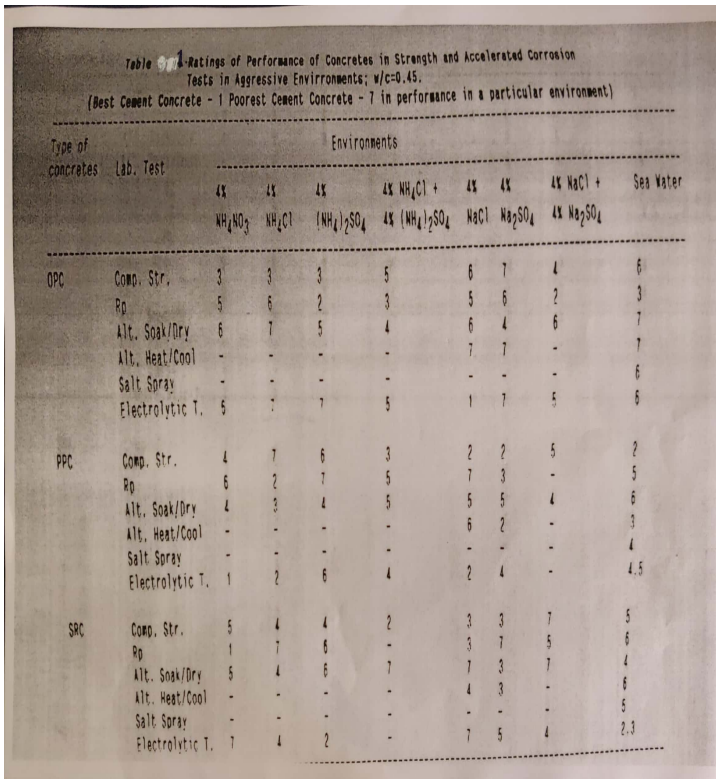


Table 1 - Ratings of Performance of Concretes in Strength and Accelerated Corrosion Tests in Aggressive Envirronments; w/c=0.45.
(Best Cement Concrete - 1 Poorest Cement Concrete - 7 in performance in a particular environment)

Type of concretes	Lab. Test	Environments							
		4% NH_4NO_3	4% NH_4Cl	4% $(NH_4)_2SO_4$	4% NH_4Cl + 4% $(NH_4)_2SO_4$	4% NaCl	4% Na_2SO_4	4% NaCl + 4% Na_2SO_4	Sea Water
OPC	Comp. Str.	3	3	3	5	6	7	4	6
	Rp	5	6	2	3	5	6	2	3
	Alt. Soak/Dry	6	7	5	4	6	4	6	7
	Alt. Heat/Cool	-	-	-	-	7	7	-	7
	Salt Spray	-	-	-	-	-	-	-	6
	Electrolytic T.	5	7	7	5	1	7	5	6
PPC	Comp. Str.	4	7	6	3	2	2	5	2
	Rp	6	2	7	5	7	3	-	5
	Alt. Soak/Dry	4	3	4	5	5	5	4	6
	Alt. Heat/Cool	-	-	-	-	6	2	-	3
	Salt Spray	-	-	-	-	-	-	-	4
	Electrolytic T.	1	2	6	4	2	4	-	4.5
SRC	Comp. Str.	5	4	4	2	3	3	7	5
	Rp	1	7	6	-	3	7	5	6
	Alt. Soak/Dry	5	4	6	7	7	3	7	4
	Alt. Heat/Cool	-	-	-	-	4	3	-	6
	Salt Spray	-	-	-	-	-	-	-	5
	Electrolytic T.	7	4	2	-	7	5	4	2.3

Table 2. Relative performance indices in field studies (with OPC in Intertidal zone at Mandapam as 100)

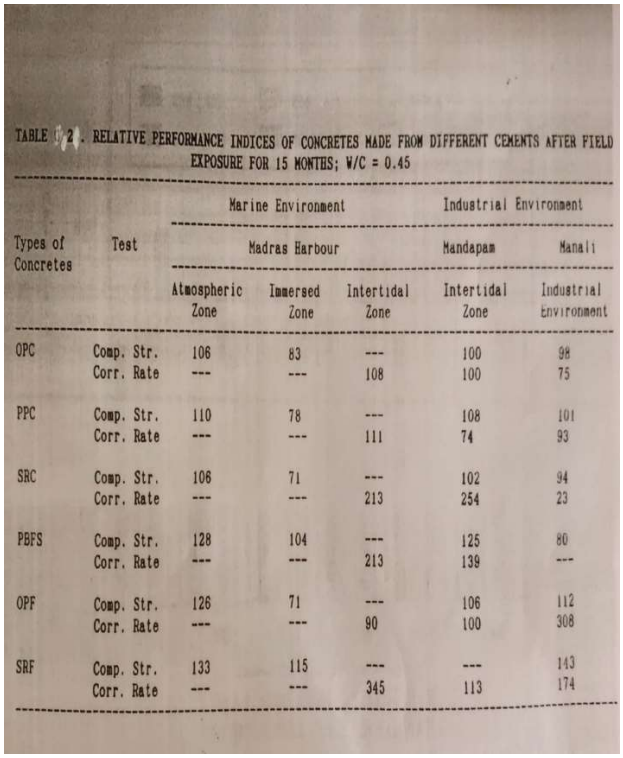


TABLE 2. RELATIVE PERFORMANCE INDICES OF CONCRETES MADE FROM DIFFERENT CEMENTS AFTER FIELD EXPOSURE FOR 15 MONTHS; W/C = 0.45

Types of Concretes	Test	Marine Environment			Industrial Environment	
		Madras Harbour			Mandapam	Manali
		Atmospheric Zone	Immersed Zone	Intertidal Zone	Intertidal Zone	Industrial Environment
OPC	Comp. Str.	106	83	---	100	98
	Corr. Rate	---	---	108	100	75
PPC	Comp. Str.	110	78	---	108	101
	Corr. Rate	---	---	111	74	93
SRC	Comp. Str.	106	71	---	102	94
	Corr. Rate	---	---	213	254	23
PBFS	Comp. Str.	128	104	---	125	80
	Corr. Rate	---	---	213	139	---
OPF	Comp. Str.	126	71	---	106	112
	Corr. Rate	---	---	90	100	308
SRF	Comp. Str.	133	115	---	---	143
	Corr. Rate	---	---	345	113	174

V. RESULTS AND DISCUSSION

Table 1 presents relative performance of various types of cement concretes in environments of chlorides, sulphates, nitrates and sea water. Various Cement concrete were tested in respect of polarization, compressive strength, polarization resistance, resistance in alternate soaking and drying, alternate heating and cooling, salt spray test and electrolytic test. The best relative performance is numbered as 1 followed by 2, 3….up to 7 in decreasing order. Table 2 represents performance indices of concretes made from 6 different cement concretes under field exposure in marine environment in Chennai at Madras Harbour at Mandapam. Specimens were exposed to atmospheric, intertidal and immersed zones of sea water. Specimens were also exposed to Ammonium Chloride at Manali, North Chennai.

Compressive strength and corrosion rates were evaluated. Relative performance was assigned performance indices of 100 and above so as to give relative idea of performance. Laboratory and field studies indicated the following observations:

Special formulations named as SRF& OPF performed better than other cement concretes as brought out above. This happened due to two - fold action of fly ash:

1. Strength gain because of formation of Calcium silicates.
2. Reduction of permeability which enhanced durability parameters.

VI.CONCLUSION

Based on accelerated testing of specimens in lab and at site, the following conclusions are arrived at:

1. Performance of SRF was seen to be the best against sulphates.
2. Performance of OPF was seen to be better against chlorides & other aggressive environments in general.
3. PPC & PBFS performed better than OPC.
4. OPCA performed better than OPC.

AKCNOWLEDGMENTS

Authors are thankful to Er. Nisha Mandele & Ms. Jayshree Solanki for setting the paper right in the prescribed format.

REFERENCES

[1] Rehsi, S.S. et. al., *"Proportioning Concrete Mix Containing Fly Ash"*, Proceedings of National Workshop on Utilization of Fly Ash,pp. B 171 - 174, Roorkee, May 1988.
[2] Kalyanasundaram, P., "Corrosion *of Steel in Reinforced Concrete"*, Material Testing lab., IIT Madras, pp. 1- 51, Dec. 1970.
[3] Eglington, M.S., *"Concrete & its Chemical Behaviour"*, Thomas Telford, pp. 6-29, U.K, 1987.
[4]Animtay,E.et.al.,*" Early Chloride Corrosion of Reinforced Concrete – A Test Report"*, Title No. 70 – 55, American Concrete Institute Journal ,pp. 606 – 611, September 1973.
[5] Gjorv, O.E. et. al., *"Diffusion of Chloride Ions from Sea Water into Concrete,"* Cement & Concrete Research, vol. 9, pp. 229 – 238, 1979.
[6] Crane, A.P., *"Corrosion of Reinforcement in Concrete Construction," Ellis Horward Limited"*, pp. 143 – 149, U.K, 1983.
[7]Harrison, W.H, *"Results of Exposure Test on Various Types of Concrete Blocks,"* Building Research Establishment, CP 761/74, U.K, August 1974.
[8] Kalousek, G.L.,et. al.,"Concrete for Long-time service in Sulphate Environment," Cement & Concrete Research, vol. 2, pp. 79 – 89, USA, 1972.
[9] Biczok, I., *"Concrete Corrosion, Concrete Protection,"* Akademia Kiado, Budapest, pp. 220 – 221, Hungary, 1972.
[10]Neville, A.E., "Properties of Concrete," 4th ed., ELBS, U.K, 1997.
[11] Sidhu, D.S., "A Study on the Corrosion Potential of Steel Embedded in Concrete made of Different Cements," Ph.D thesis, IIT Delhi, 1987.
[12] I.S 456, "Code of Practice for Plain and Reinforced Cement Concrete", B.I.S Appendix, New Delhi, 2000.
[13]Treadaway K.W.J., "Durability of Steel in Concrete," Deptt. Of Environment, Building Research Establishment, pp. 1.6, U.K.
[14]Sharma, Satish Kumar,"Performance of Concrete & Reinforced Concrete Materials in Aggressive Environments", Ph.D thesis, IIT Madras, June 1993.
[15] Klieger,P, "Durability Studies at the PCA," International Conference on Durability of Building Materials & Components", ASTM, STP 691, pp. 282 – 300, USA, 1980

AUTHORS

Dr. Satish Kumar Sharma, Professor and Head, Civil Engineering Department,
Acropolis Institute of Technology & Research, Manglia By-pass Road, Indore, India

Mrs. Shweta Kaushik, Assistant Professor,
M.V.S.R. Engineering College, Hyderabad, India