Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Social Media Analytics

Anti-Fraud Schema System

Highlights

Fast Prototyping Networks

Identity-Based Cryptosystem

Discovering Thoughts, Inventing Future

Volume 16 Issue 4 Version 0.1

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 16 Issue 4 (Ver. 1.0)

Open Association of Research Society

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://globaljournals.us/terms-and-condition/menu-id-1463/

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin: 452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Global Journals Constitutional Editorial Board

~Integrated~

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Identity-Based Cryptosystem based on Tate Pairing

By Ramesh Ch, K Venugopal Rao & D Vasumathi

GNITS

Abstract - Tate Pairings on Elliptic curve Cryptography are important because they can be used to build efficient Identity-Based Cryptosystems, as well as their implementation essentially determines the efficiency of cryptosystems. In this work, we propose an identity-based encryption based on Tate Pairing on an elliptic curve. The scheme was chosen cipher text security in the random oracle model assuming a variant of computational problem Diffle-Hellman . This paper provides precise definitions to encryption schemes based on identity, it studies the construction of the underlying ground field, their extension to enhance the finite field arithmetic and presents a technique to accelerate the time feeding in Tate pairing algorithm.

Keywords: identity-based crytosystems, tate pair, elliptic curves and digital certificates.

GJCST-C Classification : D.4.6, E.3

Identity-Based Cryptosystem based on Tate Pairing

Ramesh Ch [α], K Venugopal Rao [σ] & D Vasumathi [ρ]

Abstract- Tate Pairings on Elliptic curve Cryptography are important because they can be used to build efficient Identity-Based Cryptosystems, as well as their implementation essentially determines the efficiency of cryptosystems. In this work, we propose an identity-based encryption based on Tate Pairing on an elliptic curve. The scheme was chosen cipher text security in the random oracle model assuming a variant of computational problem Diff Hellman. This paper provides precise definitions to encryption schemes based on identity, it studies the construction of the underlying ground field, their extension to enhance the finite field arithmetic and presents a technique to accelerate the time feeding in Tate pairing algorithm.

Keywords: identity-based crytosystems, tate pair, elliptic curves and digital certificates.

I. Introduction

The advent of asymmetric encryption represented a great advances in safety of computers, especially because it solved the problem of key exchange algorithms for symmetric encryption. But attacks have been taking the advantage of the fact that it does not have a guarantee on who and the true owner of a public key, so that a user can impersonate another easily by making use of a necessary mechanism of association between a public key and its owner.

To resolve this problem was created the mechanism of certified digital, that uses a hierarchical structure of certifying authorities, able to ensure properly the possession of a given public key. This mechanism works very well in open organizations such as the internet.

In 1984 a model-based cryptographic identities was proposed by Shamir [1]. This model was intended to prevent the use of Digital Certificates, using the identity of the user as its public key. This identity could be an address of e-mail, Social Security number, full name, or a combination is of these elements. The private key would be obtained through a trusted third party(TA - trust authoraty). With this, digital certificates would be necessary only in identification of this central authority, drastically reducing their use. A problem that exists in this idea is the knowledge of the private key by the central authority, needed a total expectations by the user, which requires a lot of care from practical and legal point of view.

On the other hand, does not need the entire infrastructure of hierarchical authorities for the management of the keys by making the model more simple and suitable for organizations where hierarchy and its limitations are well controlled.

Shamir developed a signature scheme based on identities, whose operation is similar to the RSA. He also speculated on the existence of a scheme that has a problem that has been solved in practice by the cryptosystem of Boneh and Franklim [2], whose safety has been rigorously demonstrated.

a) Signature Scheme Based on Identities of Shamir

The signature scheme of Shamir based on Identities and all other forms of encryption based on identities, being divided into four steps:

1. *Setup:* this step and held by authority of expectations to generate the global parameters of the system and the master key, which will underpin that only the TA can generate private keys.
2. *Generation of private key:* this algorithm receives as input the master key and the identity of a user, returning the associated private key.
3. *Signature:* given a private key and a message, the algorithm returns the signature.
4. *Checking:* given an identity, a message and a signature, the algorithm returns true if the signature of that message matches the identity supplied, and returns false if contradicts.

II. Introductory Concepts

a) Security

We will now define some important issues to determine the security of an algorithm based on an additive group, as is the case of elliptic curves encryption [4]:

- *Problem of discrete logarithm:* Given $Q = nP$, determine n.
- *Problem Computational Diffie-Hellman:* Three Data points P, aP, bP, determine abP.
- *Problem of decision Diffie-Hellman:* Four Data elements P, aP, bP and cP belonging to a group G, answer true if and only if $C \cong ab(Mod \#G)$.

Author α: Dept. of Computer Science, G.Narayanamma Institute of Technology and Science Hyderabad, India.
e-mail: chramesh@gmail.com
Author σ: Dept. of Computer Science, G.Narayanamma Institute of Technology and Science, Hyderabad, India.
Author ρ: Dept. of Computer Science, Jawaharlal Nehru Technological University, Hyderabad, India.

One of the first uses of pairings was made by Joux [5]. In this article he showed how the decision has to be taken to issue the Diffie-Hellman can be easy through the bilinear maps, thus managed to produce an application for key sharing among three parties in a single round.

b) Elliptic Curves

An elliptic curve E defined over a finite field F_p^m and a set of points P = (x, y) with x,y ϵ F_p^m such that $y^2 + a_1 xy+a_3y +a_2 = x^3 + a^2x^2 + a^4x + a^6$ (standard medium Weiers trass) for a_i ϵ F_p^m there, beyond the point at infinity, denoted by ∞.

Setting up an operation in an appropriate sum, the elliptic curve form an additive Abelian group with neutral element given by the point at infinity.

An operation widely used in elliptic curve cryptography and scalar multiplication, where a point P and coupled with it own times k to k ϵ Z. A point of order n such that an extent NP = ∞ and n the smallest positive integer this property.

III. Identity-based Encryption

The central idea of the public key cryptographic system based on Identity is very simple, because of the fact that the public key is a numeric value without explicit direction and which can be calculated from string of any significance?. In [1], it was proposed that the public key can be the user's identity, such as name , email address, social security number, cell phone number, IP address, serial number of electronic devices, etc.

Is the public key is predetermined (equal to the identity), and then calculate the secret key ? The answer to this question comes with the first model of security assumptions: there is a CA, with the following main responsibilities:

- Create and maintain safe custody of a secret master key S_{AC}
- Identify and record all users of the system
- Calculate the secret keys of the users
- Deliver the secret keys securely (with confidentiality and authenticity)

In 1984, Shamir described the model and algorithms for digital signature. It took almost two decades until efficient encryption algorithms were discovered and demonstrated for the identity -based model to create interest among researchers and industry.

For comparison, in Table 1, we see that the secret key is calculated according to the secret system of authority and the user's identity. For a convenient f, it is not feasible to recover the master key from the ID values. And just the authority is able to generate secret keys, so that secret itself is a guarantee that the use of ID will work in cryptographic operations involving the owners identity.

To encrypt a message to the owner ID or verify a signature ID, user ID using the identity over the public parameters of the system, They include the public key of the authority (see Figure 1).

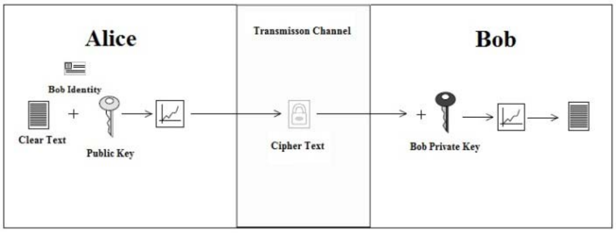
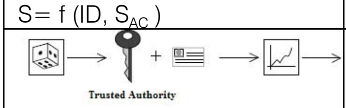


Figure 1: Encrypting the model based on the identity

Table 1: Attributes of cryptographic identity -based public key style

Secret key	Public key	Warranty
S= f (ID, S_{AC})	ID	S
		
Calculated by the authority and chosen by the user or shared with the user	Chosen by the user or shared with the user formatted for authority	

To decrypt a message to ID or to create a signature, the secret key ID is required.

a) Advantages

The identity -based model is attractive because it has many interesting advantages. The first is that the public key can in most cases be easily remembered by humans. Very different from the conventional public key, which is usually a binary string with hundreds or thousands of bits? The identity can be informed by the user to their partners and there is no requirement to maintain key directories.

To be able to view the saving processing time, storage costs and data transmissions, we will recall, for example, as It is generally a cryptographic operation with PCI. If Bob wants to encrypt a message to Alice, first of all, he must obtain the certificate that was issued to Alice (consulting a public directory or Alice itself). Bob needs check the validity period and the signature contained in the certificate. The signature verification is a process that sometimes runs the certification path of the certifying authorities involved in the hierarchy until they reach the root certification authority. If nothing goes wrong, Bob can save the Alice certificate for future use.

However, before each use, Bob need to consult a validation authority to verify that the certificate has not been revoked (often, a referral to a server that is online). Once the certificate is valid and not revoked, Bob extracts the public key of Alice, encrypts the message and transmits.

In identity-based model, just if the system parameters are authentic Bob can encrypt a message based on the identity of Alice and send (considering that identity withdrawal is treated as explained below).

A peculiarity of identity -based model is that the public key can be used before the secret key calculation. Thus, it is possible to encrypt a message for those who have not registered with the system authority or has secret key for decryption. In contrast to the model based on certificates, the user must first register and get the certificate, and then to receive an encrypted message under your public key.

b) *Disadvantages*

The first disadvantage, which is characteristic of identity-based systems is the custody of keys. As explained above, the system authority has the ability to generate secret keys of all users under their responsibility. This implies that the authority reaches to the level of confidence that defined in [10]. Consequently, you can decrypt any encrypted texts that have access (if you can identify the recipient's identity). You can also sign on behalf of any user and there is no irreversibility guarantee. Therefore, it is essential that the system of authority is reliable enough for eavesdropping of shares or counterfeiting as these are controllable.

Custody of property keys, referenced by key escrow in English texts is not always undesirable. Within a company , for example, if all sensitive documents and data are encrypted by the employee who created it , the board may have access to decryption in case of death or termination of the employee . When there is need for monitoring the content of encrypted e-mail, it can also be justifiable custody of keys. However, for most applications, custodial key is a disadvantage.

Another point unfavourable to identity -based model is the need for a secure channel for distribution of secret keys. If delivery occurs in networked and remote environment, it is necessary to ensure mutual authentication and delivery with secrecy.

Another concern that one must have in identity -based model is the possibility of identity revocation. If the secret key of a user is compromised, its identity should be repealed. Therefore, it is not recommended to simply use the number of CPF or mobile phone, for example, as a user identifier.

c) *Additional features*

As noted by [1], the identity -based model is ideal for groups of users, such as executives of a multinational company or branch of a bank, once the headquarters of these corporations can serve as system authority in all trust. Applications small scale, where the cost of deploying and maintaining an ICP are prohibitive, are candidates for the use of identity -based model. When the disadvantages cited above are not critical, the characteristics model allow interesting implementations.

Some examples of services with time availability confidential document that can be revealed to the press or to a particular group , only from certain date and time ; bids an auction that should be kept secret until the end of negotiations ; or view a film that should be enabled only within the rental period contracted.

The identity -based model has also been the subject of studies in search for alternatives to SSL / TLS, to Web applications, as shown in [7]. With the elimination of certificates the process of distributing public keys and access control will be simplified. Similarly, the model has been explored to provide security in a number of other application areas , such as grid computing and sensor networks (see for example [5] and [8]) and other applications.

IV. Pairings

A pairing and a pair of mapping linearly independent points of an elliptic curve elements of a finite field is not cyclic. We denote the pairing of two points P and Q e(P, Q). The properties listed below are very interesting for cryptographic applications, are present both in pairing as Weil pairing Tate:

- *Identity:* Pairing a pair of matched points and mapped to the neutral element of the underlying finite field
- *Bilinearidade:* Data three points P, Q, R, pairing P + Q and R and the multiplication of the P and R pairing by pairing Q and R. This property is the most important of all, because through it we get the following:
- $e(P,nQ) = e(P,Q)^n = e(nP,Q)$
- *Do not degeneration:* If P and Q are linearly independent, so their pairing and distinct from the neutral element of the underlying finite field.
- *Efficiency:* data any two points, its pairing can be calculated efficiently by a computer.

a) *Tate Pairing*

K is an integer such that F_q^k contains the n nth roots of unity. Pairing Tate and defined through the following mapping:

$$e : E[n] \times E/nE \rightarrow F_q^k/(F_q^k)^n$$

where E [n] are the points P of the curve such that nP = ∞.The Tate pairing can be calculated as e(P, Q) = g (D) where D and a divider point Q associated with a function whose rational divider n[P] - n [∞]. The Miller algorithm [Mil04] can be used to calculate the function g.

Menezes, Okamoto, and Vanstone [6] pairings used to perform a transformation of an elliptic curve points super singular to elements of a finite field generated by the unitary roots of unity. This transformation has allowed a large reduction in the difficulty of the discrete logarithm problem for these curves.

Sakai, Ohgishi and Kasahara [8] made possible the construction of a ciframento protocol based on identities using pairings, this solved the problem proposed by Shamir in his article.

V. Proposed Scheme

Now we can describe in detail the proposed scheme.

Configuration: Given k, the PKG singles groups of bilinear maps, G_1, G_2 and G_t, of prime order $p> 2^k$ generators $Q \in G_2$,$P=Ø(Q) \in G_1$,$g=e(P,Q) \in G_t$ Select s random belonging to Z^*_p a public key of $Q_{pub} = SQ \in G_2$ system summary cryptographic functions H_1, H_2 and H_3.

Generation of key pair: For an identity ID, the private key and $S_{ID} =\frac{1}{H1(ID)+S} Q \in G_2$.

Encryption: Given a message M , the identity of the sender ID_r and the identity of the recipient ID_d, random x is used belonging to Z^*_p to calculate

$$r=g^x ,C=M \oplus H_3 (r) \text{ and } h =H_2 (M,r).$$

It is estimated $S=(x + h) \varphi (S_{ID})$ and $T = x(H_T (ID_r)P + \varphi(Q_{pub)}$.

The cipher text and the triple (c, S, T).

Deciphering and verification: Given the triple (c, S, T) and the identity of the ID_R sender is calculated as

$$r = e(T, S_{IDd}), M = c \oplus H_3 (r) \text{ and } h = H_2 (M, r).$$

Accept message if $r = e(S, H\text{-}1 (ID_r)Q+Q_{pub})g^{-h}$, in which case the message M and signature (h, S) are returned.

VI. Review

This proposed scheme is interesting because their safety was demonstrated by Barreto semantically, in order to not be subject to attacks that occur when they are used some optimizations of Weil and Tate pairings. Also, please note that the simple junction of the features of this scheme and signature represents a gain of security.

But there is a problem that has not been discussed, which is the abrogation of the private key. This question this open and represents a major problem for the security of any key establishment protocol, because the User can and should change your private key regularly. The problem is in the fact that the private key calculation is deterministic, that is, given the master key sea identity ID, the algorithm always returns the same private key. As the public key and the very identity, the User can not change your identity to obtain a new private key, and needed some other solution. Other asymmetric encryption schemes do not have this problem because the public key is published and revoked with its corresponding private key.

VII. Conclusion

In this work it was possible to see that cryptosystems based on Identities are very interesting and represent an area of research that is growing. However the joint utilization of digital certificates and Identity-Based Protocols can be even more interesting as these two possible solutions to the problem of ensuring association between public key and its owner seem to be complementary.

References Références Referencias

1. A. Shamir, “Identity-based cryptosystems and signature schemes”, Advances in Cryptology - Proceedings of CRYPTO 84, Lecture Notes in Computer Science, 196 (1985), 47–53.
2. D. Boneh and M. Franklin, “Identity-based encryption from the Weil pairing”, Advances in Cryptology – CRYPTO 2001, Lecture Notes in Computer Science, 2139 (2001), 213–229. Full version: SIAM Journal on Computing, 32 (2003), 586–615.
3. K. Paterson and G. Price, “A comparison between traditional public key infrastructures and identity-based cryptography”, Information Security Technical Report, 8(3) (2003), 57–72.
4. W. Mao. Modern Cryptography -theory and practice. Prentice Hall, 2004.
5. A. Joux. A one round protocol for tripartite Diffie-Hellman. In W. Bosma, editor, Algorithmic Number Theory, IV-Symposium (ANTS IV), LNCS 1838, pages 385–394. Springer-Verlag, 2000.
6. A. J. Menezes, T. Okamoto, and S. A. Vanstone. Reducing elliptic curve logarithms to a finite field. In IEEE Trans. Info. Theory, number 39, pages 1636–1646, 1983.
7. L. Adleman and M. Huang, “Function field sieve methods for discrete logarithms over finite fields”, Information and Computation, 151 (1999), 5–16.
8. R. Sakai, K. Ohgishi, and M. Kasahara. Cryptosystem based on pairing. In Symposium on Cryptography and Information Security, Okinawa, Japan, January 2000.
9. O. Ahmadi, D. Hankerson and A. Menezes, “Software implementation of arithmetic in F3m”, International Workshop on Arithmetic of Finite Fields (WAIFI 2007), Lecture Notes in Computer Science 4547 (2007), 85–102.
10. ANSI X9.62, Public Key Cryptography for the Financial Services Industry: The Elliptic Curve Digital Signature Algorithm (ECDSA), American National Standards Institute, 1999.

11. A. Atkin and F. Morain, “Elliptic curves and primality proving”, Mathematics of Computation, 61 (1993), 29–68.
12. R. Balasubramanian and N. Koblitz, “The improbability that an elliptic curve has subexponential discrete log problem under the Menezes-Okamoto-Vanstone algorithm”, Journal of Cryptology, 11 (1998) 141–145.
13. P. Barreto, S. Galbraith, C. ´O h´Eigeartaigh, and M. Scott, “Efficient pairing computation on supersingular abelian varieties”, Designs, Codes and Cryptography, 42 (2007), 239–271.
14. P. Barreto, H. Kim, B. Lynn and M. Scott, “Efficient algorithms for pairing-based cryptosystems”, Advances in Cryptology – CRYPTO 2002, Lecture Notes in Computer Science, 2442 (2002), 354–368.
15. P. Barreto, B. Lynn and M. Scott, “Efficient implementation of pairing-based cryptosystems”, Journal of Cryptology, 17 (2004), 321–334.
16. P. Barreto and M. Naehrig, “Pairing-friendly elliptic curves of prime order”, Selected Areas in Cryptography – SAC 2005, Lecture Notes in Computer Science, 3897 (2006), 319–331.
17. B. den Boer, “Diffie-Hellman is as strong as discrete log for certain primes”, Advances in Cryptology – CRYPTO ’88, Lecture Notes in Computer Science, 403 (1996), 530–539.
18. A. Boldyreva, “Efficient threshold signatures, multisignatures and blind signatures based on the gap-Diffie-Hellman-group signature scheme”, Public Key Cryptography – PKC 2003, Lecture Notes in Computer Science, 2567 (2003), 31–46.
19. D. Boneh, X. Boyen and H. Shacham, “Short group signatures”, Advances in Cryptology – CRYPTO 2004, Lecture Notes in Computer Science, 3152 (2004), 41–55.
20. D. Boneh, G. Di Crescenzo, R. Ostrovsky and G. Persiano,“Public key encryption with keyword search”, Advances in Cryptology – EUROCRYPT 2004, Lecture Notes in Computer Science, 3027 (2004), 506–522.
21. D. Boneh and M. Franklin, “Identity-based encryption from the Weil pairing”, Advances in Cryptology – CRYPTO 2001, Lecture Notes in Computer Science, 2139 (2001), 213–229. Full version: SIAM Journal on Computing, 32 (2003), 586–615.
22. D. Boneh, C. Gentry, H. Shacham and B. Lynn, “Aggregate and verifiably encrypted signatures from bilinear maps”, Advances in Cryptology – EUROCRYPT 2004, Lecture Notes in Computer Science, 2656 (2003), 416–432.
23. D. Boneh, B. Lynn and H. Shacham, “Short signatures from the Weil pairing”, Advances in Cryptology – ASIACRYPT 2001, Lecture Notes in Computer Science, 2248 (2001), 514–532. Full version: Journal of Cryptology, 17 (2004), 297–319.
24. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Anti-Fraud Schema System for Identification and Prevention of Fraud Behaviors in E-Commerce Services

By Qinghong Yang, Wei Xing, Xiangquan Hu, & Yan Quan Liu

Southern Connecticut State University

Abstract - This study aims to determine the best practices and provide a model of the technical solutions that can effectively and systematically limit fraudulent transactions of online orders in e-commerce services, using the methods of analytical mining and case studies. Based on a process of fraud prevention and detection performed in the e-business Dangdang, Inc., a leading online retailer in China, twelve identifying features of fraudulent order data were extracted and compiled into a feature matrix. Logistic regression with this matrix was then used to build a model to judge if an order was fraudulent. The model was tested using various order data with machine learning techniques to meet the requirements of being effective, correct, adaptive, and persistent. Then an online detection and prevention schema was established and the hypothesis of so-called Behavior Pattern Change Assumption (BPCA) was proven.

Keywords: e-commerce services, fraud behavior, determination, fraud prevention, case studies, logistic regression, machine learning.

GJCST-C Classification : K.4.4, H.2.1

Anti-Fraud Schema System for Identification and Prevention of Fraud Behaviors in E-Commerce Services

Qinghong Yang [α], Wei Xing [σ] , Xiangquan Hu, [ρ] , & Yan Quan Liu [ω]

Abstract- This study aims to determine the best practices and provide a model of the technical solutions that can effectively and systematically limit fraudulent transactions of online orders in e-commerce services, using the methods of analytical mining and case studies. Based on a process of fraud prevention and detection performed in the e-business Dangdang, Inc., a leading online retailer in China, twelve identifying features of fraudulent order data were extracted and compiled into a feature matrix. Logistic regression with this matrix was then used to build a model to judge if an order was fraudulent. The model was tested using various order data with machine learning techniques to meet the requirements of being effective, correct, adaptive, and persistent. Then an online detection and prevention schema was established and the hypothesis of so-called Behavior Pattern Change Assumption (BPCA) was proven. The results show the model can detect 94% of fraudulent orders. The Anti-fraud Schema System established for Dangdang is shown to be the best model for the determination and prevention of fraudulent behaviors in the e-commerce services.

Keywords: e-commerce services, fraud behavior, determination, fraud prevention, case studies, logistic regression, machine learning.

I. Introduction

Electronic commerce has enjoyed rapid growth in recent years [1], as more and more people have accepted online shopping. However, along with the growing number of transactions, there are a growing number of fraud activities. The temptation of economic gain and the difficulty of internet supervision have led to a great number of online fraud activities. Hackers can steal online accounts and use these accounts in criminal activities [2]. Prevention of fraud activities in order to provide a safe online shopping experience is a challenge for electronic commerce [3]. EBay is the leading e-commerce company around the globe, and every day thousands of customers trade through eBay. Therefore, eBay has hired experts from the National Aeronautics and Space Administration (NASA) of the U.S. to develop an anti-fraud model to detect and prevent fraud activities.

E-commerce started late in China, and few resources have been devoted to the anti-fraud field, so systematic anti-fraud solutions are especially scarce. As a leading business-to-consumer e-business in China founded in 1999, China Dangdang, Inc. offers products mainly in the categories of books, audios, digital devices, and household merchandise. Dangdang made an initial public offering (IPO) on the New York Stock Exchange in November 2010and had over 9,000,000 active customers. Because of its main business is online, Dangdang shows great interest in solving Internet-related fraud problems, especially those involving online orders directly affecting its customers, as its key strategy is to grow its e-business,

Beginning with are iew of past studies relevant to the present research, this paper provides detailed process descriptions of an anti-fraud model's development (Section 3), and discussions of its implementation with a real data mining process (Section 4). The results were concluded significantly that proven the best practice of fraud determination and detection in real situation of e-commerce order transactions.

II. Literature Review

A review by Hogan [4] summarized research on fraud behavior over the past decade. Prior to that researchers mainly focused on fraud in the areas of accounting, auditing and finance activities [5].The growth of online transactions led to a growth of fraud activities, and the lack of supervision online made it easy to commit fraud [3]. The online market has some unique features that attract fraud activities, namely information asymmetry, online transactions and the uncertainty of traders' identities and commodities [6].

Lou & Wang's research reveals that though many methods can be used to prevent fraud activities, e-commerce should use a systematic method to solve the problems uniformly [7]. Account information is collected and the information used to summarize the patterns of fraud to adapt to different situations [8]. Latch (1999) pointed out that using classification algorithms such as ID3 and C4.5 to detect fraud patterns and identify accounts with suspicious activity and then allowing humans to make the final judgment could work well in detecting fraud [9].

Author α σ ρ: School of Software, Beihang University, Beijing 100029, China.

Author ω: Southern Connecticut State University, New Haven, CT 06515, USA. e-mail: liuy1@southernct.edu

Detecting fraud behavior and managing fraud risk require the design and application of a fraud-detection model. Eining et al. find that auditors can manage different levels of fraud risk better and make unanimous auditing decisions by using an expert system [10]. Green and Choi use a neural network to detect fraud behavior and achieve satisfactory results [11].

Ohlson found that the identifying features of fraud activity could be used to alert sellers to fraud activities during financial transactions [12]. Lenard and Alam used logistic regression in detecting fraud activities [13]. This method has also been employed in several researcheslater on [14][15].

Maranzato [1] researches how to detect fraud activities in an e-commerce system. He then uses logistic regression to detect and identify features of credit fraud, and he also points out that logistic regression depends greatly on the data quality [16].

This work focuses on how to detect and prevent fraud activities in online transactions, especially before fraudulent orders are completed in the environment of e-commerce.

III. Research Design

This work combines analytical mining and case study using the real data of Dangdang's sale transactions to identify the common patterns of detecting and preventing fraud activities from occurring within online orders.

a) *Collecting and Processing Data*

Real customer order records collected from Dangdan's transaction logs, including initially identified fraudulent orders by the company's customer services officers, were processed in the following five phases:

Phase one: collect order data from Dangdang (01/01/2014-07/09/2014) and statistically analyze it to identify the features that can distinguish fraudulent orders from legitimate ones. Purposely, we used this method to find key features of fraud activities to and build a feature matrix system from machine learning.

Phase two: use the order data collected from the same period of time (01/01/2014-07/09/2014) to develop and train a logistic regression model to predict orders that are unusual.

Phase three: test the order data collected from Dangdang (07/01/2014-08/31/2014) with this logistic regression model to reveal how well the model works on the condition that the fraud ratio of orders is high.

Phase four: test the order data collected from Dangdang (10/07/2014-10/28/2014) with this logistic regression model again to reveal how well the model works on the condition that the fraud ratio of orders is low.

Phase five: employ this model in a non line environment with real customer records to assess the usefulness of the model.

b) *Making Behavior Pattern Change Assumption*

Unusual orders are associated with customer behaviors differing from normal ones. As empirical evidence in daily sales has accumulated, the researchers are convinced with such a set of hypothetical rules that may direct the fraud discovering process, as we so call Behavior Pattern Change Assumption (BPCA).

Rule 1, for most of the e-commerce user accounts, customer behaviors are consistent with shipping address, receiver name, receiver phone, payment habits, and so on, remaining steady. This is called 'steady behavior.' Sudden changes of some or all of these attributes may indicate fraudulent behavior.

Rule 2, when an order is confirmed as a fraudulent order, all the orders whose receiver address, receiver IP and so on are same as this order are considered suspicious. This is because one hacker may steal multiple accounts and make multiple orders; however, the IP and address may stay the same. It's like one fraudulent order infects the IP or address. This is known as the 'suspicion infection'.

Rule 3, hackers won't add their own money to an account but will just deplete the balance in the account or do other things that won't benefit the account but will deplete all possible resources from the account. They want to maximize their profit. This is called ''maximum rob".

c) *Underling Research Procedure*

An outline of the research procedure for this study consists of defining the case, analyzing the data, until extracting, evaluating and implementing the outcomes shown in Figure 3.

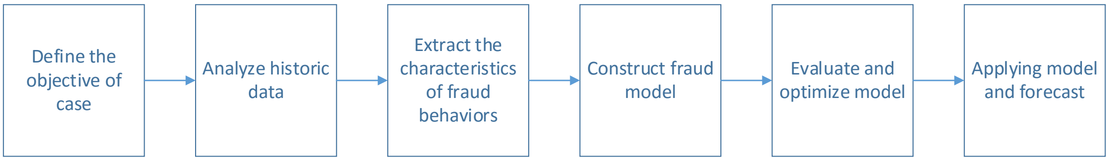

Define the objective of case
Analyze historic data
Extract the characteristics of fraud behaviors
Construct fraud model
Evaluate and optimize model
Applying model and forecast

Figure 3: Research process of detecting fraudulent orders

Analyzing order data: The data from orders that have been marked as fraudulent by customer service are analyzed.

Feature extraction: Custom service and technology experts brainstorm to extract some features that may distinguish fraudulent orders from legitimate ones and statistically test them.

Model construction: Use the statistical analysis and apply an algorithm to the data and establish the norm of the algorithm. Use order data to create a logistic regression model.

Test and optimize the model: Use test order data to test and assess the performance of the model, then continue optimizing the model.

Model application: Use the fraudulent order detection model to judge online transaction orders then assess the performance and economy value of the model.

IV. Construction of a Fraudulent Orders' Detection Model

In an attempt to create effective technical solutions that could systematically limit fraudulent transactions of Internet orders, a fraudulent order detection model based on the Behavior Pattern Change Assumption (BPCA) was developed, consisting of the following process.

a) *Determination of Fraudulent orders*

Fraudulent orders occur when a hacker steals a customer's account and uses the balance in the account to purchase goods for him/her self. Fraudulent orders are confirmed when customers call customer service to complain. Customer service staff will also call the customer to check if the customer or hacker places an unusual order, which usually is the primary method to determine whether the order is a "regular" or "fraud" order.

b) *Process of Analysis*

The core idea of fraudulent order detection is to compare normal orders with fraudulent orders to find identifying features that distinguish them. These features can then be used to judge if an order is fraud or not with BPCA.

There are three steps to extract the features of fraudulent orders:

Step 1: Customer service staffs locate fraudulent orders because of customer complaints.

Step 2: Statistical analysis of commonly used information such as the IP address of the order, receiver name, receiver address and receiver phone number.

Step 3: Compare normal orders with fraudulent orders to identify features that distinguish fraudulent orders from normal ones.

c) *Analysis of Source Data*

Labelled order data are provided by Dangdang customer service, and analyzing these data can verify BPCA at some level. When a hacker places a fraudulent order, the receiver name, receiver number and receiver address are different from the normally used information. Because hackers don't want to use real addresses, they may use some generic rough ones ending with 'county', 'block', 'corner' or 'street'. Six features that can distinguish fraudulent orders from normal ones therefore were identified.

We analyzed the six features using real data from Dangdang to determine their effectiveness in identifying fraudulent orders. The source data were from Dangdang's order data between January 1 to July 9 of 2014, in a total of 2075 fraudulent orders and 1513 stolen accounts.

Table 4.1: Items that can distinguish fraudulent orders

item ID	Definition	Frequency
rough_addr	Is the address rough?	71.1%
usually_city	Is the receiver city usually used?	42.5%
usually_tel	Is the receiver phone usually used?	35.5%
usually_name	Is the receiver name usually used?	34%
usually_email	Is the receiver email usually used?	32.6%
payment_ratio,0.05	Pay for the order using extra money instead of the money stored in the account, the ratio of extra money is more than 0.05	7.3%

Taking the feature rough_ addr as example, frequency in the table means that 71.1% of all orders that have rough_ addr are fraudulent orders. According to the results of these statistical tests, a basic idea of fraudulent orders emerges and this result will help in building a machine-learning model. These results also reveal some interesting facts. First is that rough_ addr is a very identifiable feature from which to detect fraudulent order. Usually_ city, usually_ tel, usually_ name, usually _email all show some potential to detect fraudulent orders. Payment_ ratio verifies that hackers just want to maximize their profit, which is the rule 3 of BPCA. These simple statistical results are useful at some level, to solve the fraudulent order detection problem and to further verify that BPCA, the machine learning algorithm is needed.

d) Fraudulent order Detection Model using Logistic Regression

Firstly, Choosing an appropriate algorithm. Logistic regression is a suitable algorithm for fraudulent order detection because it is not hard to apply, and the company's customer service staff can easily interpret the result.

Assuming some identification features as previously described. The identifying features and order data are input into the regression, and then the algorithm builds a model where each feature has a coefficient showing how much this feature can affect the result. The features with low coefficients (i.e. that don't significantly affect the results) are removed and the model is run again.

Formula 4-1 shows the result of the model. If the model returns a result of 1, then the order is fraud, less than that indicating otherwise.

$$Y=\begin{cases} 1 \text{ fraud order} \\ 0 \text{ normal order} \end{cases} \quad (4\text{-}1)$$

Using identification features as $x = (x\_1,x\_2,\ldots,x\_p)$, logistic regression can be represented as Formula 4-2.

$$Pf = P(Y=1|x) = \pi(x) = 1/(1+e^{(-g(x))}) \qquad (4\text{-}2)$$

$$\text{Where } g(x) = \beta\_0+\beta\_1 x\_1+\beta\_2 x\_2+\cdots+\beta\_p x\_p \quad (1\le p\le n)$$

Using maximum likelihood estimation, the coefficients β0, β1, βp can be obtained. According to Formula 4-2, an interested party can calculate the possibility of one order being a fraudulent order. When using a binary classifier, it becomes necessary to pick a threshold; if the possibility is greater than the threshold, the order is a fraudulent order, otherwise it may be innocent. The value of the threshold can be anywhere between 0 and 1.However, if the threshold is too low, the model would be unstable and if the threshold is too high, the recall rate would not be ideal. This paper chooses 0.75 as the threshold.

Second, conducting characteristic statistics and extraction. The most critical process to build the fraud detection model is to select identification features. The statistical information in Table 4.1 shows fraudulent orders always having features such as different receiver name, receiver address, receiver city, and receiver telephone number, which can ascertain fraudulent orders made by hackers going directly to their own addresses.

Features 1-14 in Table 4.2 are deduced by the concept "steady behavior" of BPCA. Receivers related information changes mean that this order might be a fraudulent order. Features 15-19 deduced by the concept "suspicion infection" of BPCA mean that if some receiver's phone numbers or receiver addresses have been complained about before, new orders that have the same receiver address and receiver phone number have the possibility of being fraudulent orders. Feature 20 is based on the statistical results shown in Table 4.1, if the receiver address is rough, there is a high possibility that the order is fraudulent order. Features 21-23 are based on "maximum rob" of BPCA, meaning that the hackers want to make the most profit possible out of the stolen account.

Table 4.2: Identification features of the logistic regression model

Feature ID	Feature name（x）	Explanation
1	name_dubious_count	Complaint number of this receiver name.
2	name_cust_dubious_count	How many customer IDs are related to this receiver name?
3	tel_home_dubious_count	Complaint number of this receiver telephone number.
4	tel_home_cust_dubious_count	How many customer IDs are related to this receiver telephone number?
5	tel_mobile_dubious_count	Complaint number of this receiver mobile phone number.
6	tel_mobile_cust_dubious_count	How many customer IDs are related to this receiver mobile phone number?
7	orderip_dubious_count	Complaint number of this receiver IP address
8	orderip_cust_dubious_count	How many customer IDs are related to this receiver IP address?
9	addr_dubious_count	Complaint number of this receiver address.
10	addr_cust_dubious_count	How many customer IDs are related to this receiver address?
11	permid_dubious_count	Complaint number of this receiver permid*
12	permid_cust_dubious_count	How many customer IDs are related to this receiver permid*
13	email_dubious_count	Complaint number of this receiver email.
14	email_cust_dubious_count	How many customer IDs are related to this receiver email?

15	name_frequency_count	How many orders does the customer make using this receiver name in history?
16	tel_home_frequency_count	How many orders does the customer make using this receiver telephone number in history?
17	tel_mobile_frequency_count	How many orders does the customer make using this receiver mobile number in history?
18	city_frequency_count	How many orders does the customer make using this receiver city in history?
19	addr_frequency_count	How many orders does the customer make using this receiver address in history?
20	rough_address	Is this address rough?
21	whole_price	The total price of the order.
22	Payment	How much money should the receiver pay when they receive this package?
23	payment_ratio	How much should be paid apart from using the account balance?
24	Intercept	The constant in logistic regression model.

*permid is used to identify customers. Whether or not the customer is logged in, Dangdang will save a permid on the device being used to browse Dangdang.

These identification features are used to compose an identification features matrix, and then training data is used to train the matrix. Some of the identification features may not be effective in detecting fraudulent orders, and some ineffective features are eliminated during the training process.

V. Application and Results

The implementation of the logistic regression discussed above helps develop a fraudulent order detection model and test its effectiveness using the real order data of Dangdang as the experimental subject.

a) Preparation and Preprocessing of Data

Preprocessing of data is a key problem in machine learning, because in most cases data is incomplete, noisy and incompatible. The result of a machine-learning algorithm depends greatly on the quality of data. Data preprocessing includes: data cleaning, data integration, data conversion and data reduction [17].

Because of the volume of data, a sample of the total order data has been used with a ratio of fraudulent orders versus normal orders of from 1:5 to 1:9.

Continuous numbers were assigned to discrete sections. For example, the total money was divided into sections [0,10), [10,50), [50,100), and >100. Discretization can be used when the focus is only on relative value instead of absolute value. The discretization formula used in this paper is ln (x+1)/ln2. Discretization is useful to describe nonlinear relationships and solve the hidden flaws in data [11].

b) Process and Application of Model

Based on Formulae 4-1 and 4-2, R programming language was used to create a logistic regression model and then train the model to obtain coefficients.

$$g(x) = β\_0+β\_1\ x\_1+β\_2\ x\_2+\cdots+β\_p\ x\_p(1\leq p\leq 23) \quad (5\text{-}1)$$

After training, the coefficients, namely β0，β1，…βp can be figured out as shown in Table 5.1.

Table 5.1: Coefficients of first training

Index	X (Feature name)	**B** (coefficient)
1	email_dubious_count	6.990
2	name_cust_dubious_count	0.619
3	rough_address	0.532
4	orderip_cust_dubious_count	0.168
5	permid_cust_dubious_count	0.166
6	addr_dubious_count	0.075
7	tel_mobile_dubious_count	0.062
8	tel_home_cust_dubious_count	0.049
9	addr_cust_dubious_count	0.003
10	whole_price	0.000
11	name_frequency_count	0.000
12	email_cust_dubious_count	0.000
13	Payment	-0.000
14	addr_frequency_count	-0.004
15	tel_mobile_frequency_count	-0.006
16	tel_home_dubious_count	-0.023

17	payment_ratio	-0.030
18	tel_mobile_cust_dubious_count	-0.031
19	permid_dubious_count	-0.056
20	orderip_dubious_count	-0.097
21	city_frequency_count	-0.341
22	name_dubious_count	-0.462
23	tel_home_frequency_count	-0.464
24	Intercept	-5.161

c) Optimization and Second Training of Model

After first training, the coefficients of the first model were obtained. Then, based on the analysis of these results, 13 features were deleted and 1 new feature added. Features were deleted based on three rules:

Rule 1: If one feature's coefficient is 2 magnitudes lower than the biggest coefficient, it should be deleted.

Rule 2: If one feature's coefficient is not logical, it should be deleted.

Rule 3: If one feature is considered not logical after discussion with experts, it should be deleted. Features that have the pattern "*** _cust_dubious_count" all have low coefficients and after discussion it was determined that these features are not very logical, so they were deleted (see Table 5.2).

Table 5.2 : Features that were deleted and why

X (Feature name)	β(coefficient)	Reason for deletion
Payment	-0.000	r1
addr_cust_dubious_count	0.003	r3
email_cust_dubious_count	0	r3
name_dubious_count	-0.462	r2
name_cust_dubious_count	0.619	r3
orderip_cust_dubious_count	0.168	r3
permid_cust_dubious_count	0.166	r3
tel_home_dubious_count	-0.023	r1
tel_home_cust_dubious_count	0.049	r3
tel_mobile_cust_dubious_count	-0.031	r3
name_frequency_count	0	r1
tel_home_frequency_count	-0.464	r3
tel_mobile_frequency_count	-0.006	r1

A new feature, phone_address was added. Phone_address is a complex feature that can be calculated by this rule: if neither the receiver mobile number nor the address of one order have ever been used in this account, then phone_address is the number of total history orders of this account, unless phone_address is 0. This feature was added based on the reasoning that the more orders a user has bought, the lower the likelihood of them changing receiver address and mobile number at the same time. The final logistic model is shown in Table 5.3.

Table 5.3 : Final features and coefficients of logistic model

Sequence	X (Feature name)	β(coefficient)
1	city_frequency_count	-0.405
2	addr_dubious_count	0.305
3	email_dubious_count	2.680
4	orderip_dubious_count	0.561
5	phone_address	0.887
6	tel_mobile_dubious_count	0.993
7	whole_price	0.338
8	addr_frequency_count	-1.050
9	payment_ratio	-0.200
10	Intercept	-1.395
11	permid_dubious_count	0.605
12	rough_address	0.406

By analyzing the final model, BPCA is testified and the following conclusions are reached.

The coefficients of city_frequency_count and addr_frequency_count are negative, which means if the receiver city and receiver address of an order have been used in this account many times, the order is less suspicious.

The coefficients of addr_dubious_count, email_dubious_count,tel_mobile_dubious_count,permid_dubious_count and orderip_dubious_count are all positive.

This result confirms the hypothesis of "suspicion infection": the new orders that have the same receiver address, mobile number and IP as previous fraudulent orders are considered suspicious.

The coefficient of rough_address is positive, which means that an order with a rough address is suspicious because those committing fraud do not want to supply their address.

The coefficient of par_rate is negative, which means that if the customer should pay extra money over the value of their account balance, the order is less suspicious.

The coefficient of whole_price is positive, which means that fraudulent orders tend to be greater in total price. However, the absolute value of the coefficient is one of the lowest ones, so this tendency is not very important.

The coefficient of phone_address is positive, which means that if one account has made many orders in history and now uses both a new receiver mobile number and a new address to make a new order, then the new order is suspicious.

The results shown in table 5.3 verify the BPCA.

d) Test and Performance of Model

Using the values (x1, x2, x4….) of the features of one order as input, the model calculated a possibility (Pf as shown in Formula 5-2) of this order being a fraudulent order. If the possibility is greater than the threshold (0.75), then this order is a fraudulent order, otherwise legitimate.

$$Pf = P(Y=1|x) = \pi(x) = 1/(1+e\,\hat{}\,(-g(x))\,) \qquad (5\text{-}2)$$

$$g(x) = \beta\_0+ \beta\_1\, x\_1+\beta\_2\, x\_2+\cdots+\beta\_p\, x\_p \qquad (1\le p\le 11)$$

The values of β0, β1, … βp are shown in Table 5.3. X1, x2…xp are provided by Dangdang's data system.

Four categories were defined: FF (judged as fraud, in fact is fraud), FC (judged as fraud, in fact is clear), CF (judged as clear, in fact is fraud), and CC (judged as clear, in fact is clear). Human check number is the number of orders that should be checked by customer service officers, recall rate is the rate of fraudulent orders that can be detected by the system, the calculation of these two values are shown below.

Human check number=FF+FC (5-3)

Recall rate=FF/(FF+CF) (5-4)

Two experiments were designed to test the effects of the fraud detection model in different conditions of time and situation.

Experiment 1 used Dangdang order data from 07/01/2014 to 08/31/2014 as the test data set. It detected 395 of a total of 417 fraudulent orders. The result is shown in Table 5.4 where it can be seen that the model works well in a situation where the rate of fraudulent orders is high.

Table 5.4 : Results of Experiment 1

Index	Value	Index	Value
Fraudulent order number	417	CF	22
Normal order number	346725	FC	896
Total order number	347142	CC	345829
Threshold	0.75	Recall	0.95
FF	395	Human check number	1291

Experiment 2 used Dangdang order data from 10/07/2014 to 10/28/2014 as the test data set. The model detected 48 of a total of 51 fraudulent orders. The result is shown in Table 5.5. Three fraudulent orders were missed. One of them was the first order of a new account, and the other two were a situation in which one customer used another customer's gift card, but the account using the gift card was not stolen.

Table 5.5 : Results of Experiment 2

Index	Value	Index	Value
Fraudulent order number	51	CF	3
Normal order number	270414	FC	1443
Total order number	270465	CC	268971
Threshold	0.75	Recall	0.94
FF	48	Human check number	1491

The fraudulent order rate in Experiment 2 is 0.000189, which was less than the rate 0.001201 of Experiment 1. It can be seen that the recall rate and human check number are close to that in Experiment 1.

e) Performance of the Model

The changes in numbers of fraudulent orders and amount of money stolen were analyzed to show the usefulness of the fraudulent order detection system. The anti-fraud system using this fraud detection model started to run on 06/24/2014. Figure 5.1 shows the change in fraudulent order numbers from 01/2014 to 01/2014. Figure 5.2 shows the change in amount of money being stolen from 01/2014 to 01/2014. Based on the information in Figures 5.1 and 5.2, it appears that after the system was implemented in 06/2014, the fraudulent order problem was controlled. Such a fraud detection system was not in place in Dangdang before, as shown in figure 5.1. Prior to April 2014 there were not many fraudulent orders found, because Dangdang had no way to control the situation. April 2014 could be a bench mark in China. That year a number of accounts were terribly leaked in China. Dangdang, as one of the biggest ecommerce entities, were attacked by hackers who used the leaked accounts

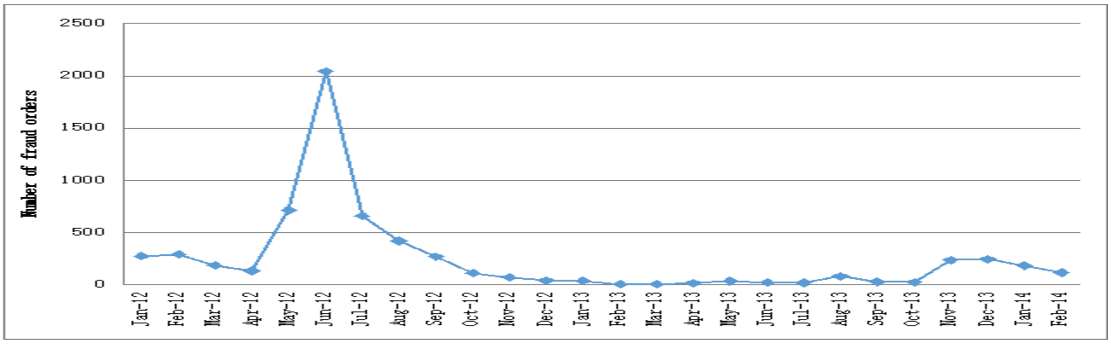


Figure 5.1 : Trend of fraudulent order number per month

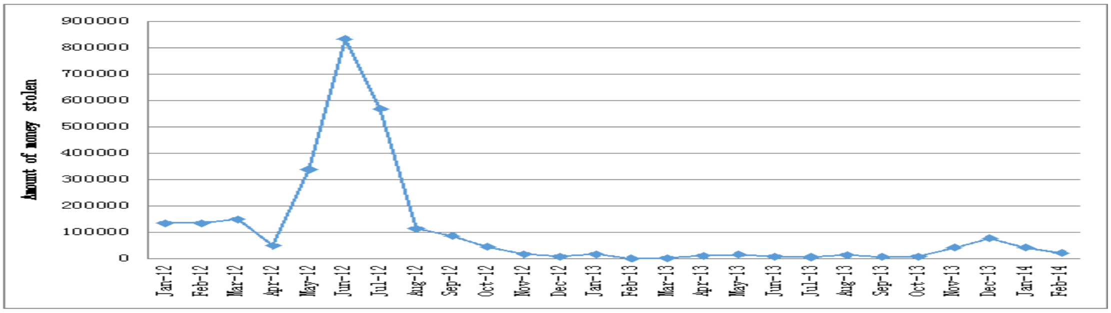


Figure 5.2 : Amount of money being stolen per month

94% of all fraudulent orders can be detected by this system. After the anti-fraud detection system was implemented, fewer complaints were received. Combining this system with human rechecking, both the number of fraudulent orders and the amount of money stolen were reduced.

VI. Conclusions

We found that trading fraud in online ecommerce can be detected and prevented. Order fraud is a key problem of trading online and is very harmful to companies. By focusing on the determination of these online order frauds, the real situation and data of Dangdang was used with these two steps to follow.

First, statistical analysis was carried out to determine the basic differences between fraudulent orders and normal ones, and then a logistic regression model was created. Finally, different experiments were designed to test the validity and effectiveness of the Fraudulent orders' Detection Model using twelve identifying features, which is the best model approved by the company in identification and prevention of fraud behaviors in e-commerce services.

a) Pattern of Detecting and Preventing Online Seller and Customer Fraudulent orders

First, statistical analysis of order data was carried out to provide a basic idea of the characteristics of fraudulent orders. Then, the features that can help in distinguishing fraudulent orders from normal ones were extracted and used to format a feature matrix. The feature matrix and logistic regression algorithm were used to build a fraudulent order detection model and carry out optimizations. Finally, the model's effectiveness was tested, and the model was used to detect real time orders and keep track of the performance of the model.

There would be no way to determine fraudulent orders without the implementation of this new developed model, which has been allowing enabled the company's customer service staff to successfully catch and free zesuspicious accounts. About 94% of fraudulent orders were detected in Dangdang in the past year with the model. It helped the company reduce fraudulent orders and therefore could be instructive to and implemented by similar electronic commerce entities.

b) Economical Significance of Detecting and Preventing Fraudulent order

Reducing the number of fraudulent orders can benefit both customers and companies. First, fewer fraudulent orders means fewer customers losing money and more customers enjoying their shopping experience. Second, fewer fraudulent orders means that companies will receive fewer complaints and customer satisfaction will be higher. The public image of companies is improved with fewer fraudulent orders.

VII. Future Work

This report describes innovative research on the role of features of online orders and accounts in monitoring online transaction activities. This work resulted in a model that can detect patterns of fraud activity in online transactions, and judges the likelihood of each transaction being fraudulent. When the likelihood of a transaction being a fraud is high, human checks are still required to make a final judgment. Therefore, human resources are needed in identifying fraud. There are a variety of fraud activities, so different detection processes are required. Therefore, the process and model of this work can in future be adjusted to be used in more situations.

Acknowledgements This research was supported by Dangdang (http://www.dangdang.com). Thanks to Fu Qiang, VP of Technology, Ju Qi, Technical director, and other staff of Dangdang for their support and help. Thanks to the technical help of Michael Wagner of Yale University Social Sciences Information Center. Thanks to Hu Saiquan of Yale University Economics and Management School, and Sam of Yale Computer Science School for discussion of research methods. Thanks to Ms. Arlene and BiBielefield, as well as Mr. Keven and Quan Wang for corrections to the paper.

References Références Referencias

1. Jans, M., N. Lybaert, and K. Vanhoof. 2009. A framework for internal fraud risk reduction at IT integrating business processes: The IFR framework. The International Journal of Digital Accounting Research 9: 1–29.
2. Kim, T.K., Lim, Y.J. & Nah, J.H. (2013) Analysis on fraud detection for internet service. International Journal of Security and Its Applications, 7 (6): 275-284.
3. Shim, S. & Lee, B. (2010) An economic model of optimal fraud control and the aftermarket for security services in online marketplaces. Electronic Commerce Research and Applications, 9(5): 435-445.
4. Hogan, C. E., Rezaee, Z., Riley, R.A. & Velury. U.K. (2008) Financial statement fraud: Insights from the academic literature. Auditing: A Journal of Practice & Theory, 27(2): 231–252.
5. Trompeter, G.M., Carpenter, T.D., Desai, N., Jones, K.L. & Riley Jr., R.J. (2013) A synthesis of fraud-related research. Auditing: A Journal of Practice & Theory, 32 (S1): 287-321.
6. Klein, B., & Leffler, K. B. (1981) The role of market forces in assuring contractual performance. The Journal of Political Economy, 89 (4): 615–641.
7. Lou, Y.I. & Wang, M.L.(2009) Fraud risk factor of the fraud triangle assessing the likelihood of fraudulent financial reporting. Journal of Business & Economics Research, 7(2): 61–78.
8. Michael & Adler (1971) pointed out that the sole concern of fraud risk factor studies was the consideration of fraudulent behavior or how to detect or to deter fraud.
9. Lach, J. (1999) Data mining digs in. American Demographics, 38-45.
10. [JW1] Eining, M.M., Jones, D.R., & Loebbecke, J.K. (1997) Reliance on decision aids: an examination of auditors' assessment of management fraud. Auditing: A Journal of Practice and Theory, 16 (2): 1-19.
11. Green, B.P. & Choi, J.H. (1997) Assessing the risk of management fraud through neural network

technology. Auditing: A Journal of Practice and Theory, 16 (1): 14–28.

12. Ohlson, J.A. (1980) Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18: 109–131.
13. Lenard, M.J. & Alam, P. (2009) An historical perspective on fraud detection: From bankruptcy models to most effective indicators of fraud in recent incidents. Journal of Forensic & Investigative Accounting, 1(1): 1-5.
14. Zhang, H., Lin, Z., & Hu, X. The effectiveness of the escrow model: an experimental framework for dynamic online environments. Journal of Organizational Computing and Electronic Commerce, 17 (2): 119–143.
15. Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013) Applied Logistic Regression. Wiley & Sons, New York.
16. Maranzato, R., Pereira, A., do Lago, A.P. & Neubert, M. (2010) Fraud detection in reputation systems in e-markets using logistic regression. Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre. Switzerland.
17. Lek, M., Anandarajah, B., Cerpa, N. & Jamieson, R. (2001) Data mining prototype for detecting e-commerce fraud. The 9th European Conference on Information Systems, Bled, Slovenia.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Social Media Analytics using Data Mining

By Hibatullah Alzahrani

Saudi Arabian Cultural Mission

Abstract - There is a rapid increase in the usage of social media in the most recent decade. Getting to social media platforms for example, Twitter, Facebook LinkedIn and Google+ via mediums like web and the web 2.0 has become the most convenient way for users. Individuals are turning out to be more inspired by and depending on such platforms for data, news and thoughts of different clients on various topics. The substantial dependence on these social platforms causes them to produce huge information described by three computational issues in particular; volume, velocity and dynamism. These issues frequently make informal organization information exceptionally complex to break down physically, bringing about the related utilization of computational method for dissecting them.

Keywords: *social network, data analysis, data mining, social media platform.*

GJCST-C Classification: *H.2.8, K.4.2*

SOCIALMEDIAANALYTICSUSINGDATAMINING

Strictly as per the compliance and regulations of:

Social Media Analytics using Data Mining

Hibatullah Alzahrani

Abstract- There is a rapid increase in the usage of social media in the most recent decade. Getting to social media platforms for example, Twitter, Facebook LinkedIn and Google+ via mediums like web and the web 2.0 has become the most convenient way for users. Individuals are turning out to be more inspired by and depending on such platforms for data, news and thoughts of different clients on various topics. The substantial dependence on these social platforms causes them to produce huge information described by three computational issues in particular; volume, velocity and dynamism. These issues frequently make informal organization information exceptionally complex to break down physically, bringing about the related utilization of computational method for dissecting them. Information mining gives an extensive variety of strategies for recognizing valuable information from huge datasets like patterns, examples and standards. Various data mining strategies are utilized for useful data recovery, factual displaying and machine learning. These systems generally do a sort of pre-processing of data, performs the data analysis and information. This study examines distinctive information mining procedures utilized as a part of mining different parts of the informal community over decades going from the chronicled systems to the forward model.

Keywords: social network, data analysis, data mining, social media platform.

I. Introduction

Data mining is an instrument which helps in finding different patterns in the dataset under analysis and connections inside the information. Mining of information finds concealed data from substantial data bases. Analysis of social media platforms has attracted much consideration in the form of chart information administration in research field. To guarantee important information mining results it's better to comprehend the information better. There are a few components which has made the analysis of information on social platforms pick up tremendous significance by data scientists. Couple of such variables incorporates the presence of enormous measure of information on these platforms, the representation of this data in dashboard forms as diagrams.

Data mining is an intelligent procedure inside which advancement is characterized by revelation through either programmed alternately manual strategies. Organizations can gain from their exchange information more about the conduct of their clients and in this way can enhance their business by making use of this information science can acquire from observational information, new bits of knowledge on exploration questions. Web use data can be broke down and used to advance data access. Along these lines information mining creates novel, unsuspected understandings of information.

II. Issues in Analysing Social

DATA Linkage Based - In linkage-based examination, analysis on the linkage conduct of the system with a specific end goal to decide essential hubs, groups, joins, and developing locales of the system is build. This analysis gives a decent outline of the wide development conduct of the network and it gets easy to gauge the current situation of the data flowing in these networks on the social media platform.

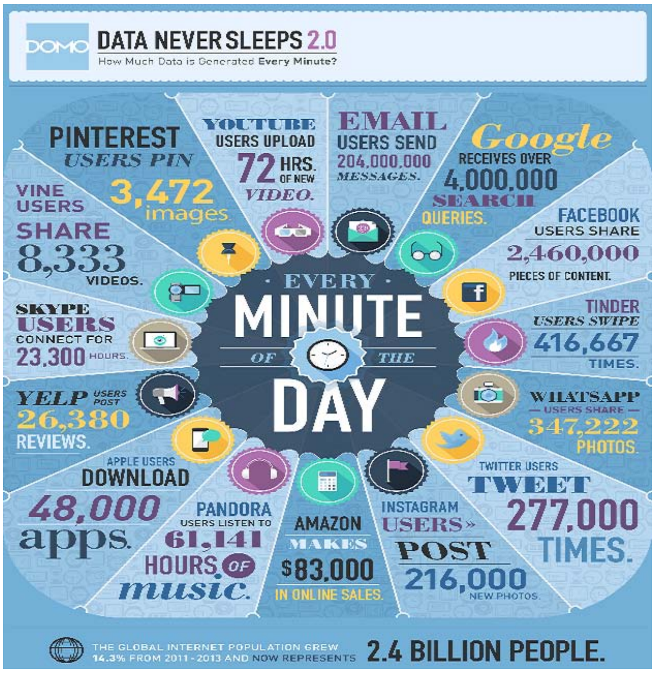

DOMO
DATA NEVER SLEEPS 2.0
How Much Data is Generated Every Minute?
PINTEREST USERS PIN 3,472 images.
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
VINE USERS SHARE 8,333 VIDEOS.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
SKYPE USERS CONNECT FOR 23,300 HOURS.
EVERY MINUTE OF THE DAY
TINDER USERS SWIPE 416,667 TIMES.
YELP USERS POST 26,380 REVIEWS.
WHATSAPP USERS SHARE 347,222 PHOTOS.
APPLE USERS DOWNLOAD 48,000 apps.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
AMAZON MAKES $83,000 IN ONLINE SALES.
INSTAGRAM USERS» POST 216,000 NEW PHOTOS.
TWITTER USERS TWEET 277,000 TIMES.
THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011-2013 AND NOW REPRESENTS 2.4 BILLION PEOPLE.

Figure 1 : Source: https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/

Content Based - Numerous social media platforms for example, Flickr, Message Systems, and YouTube comprises of massive data which can be utilized keeping in mind the end goal to enhance the quality of the data analysis. For instance, a photo sharing site, for example, Flickr contains a huge measure of content and picture data as client labels and pictures. Also, blog systems, email systems and message sheets contain content substance which are connected to each other. Consolidating content-based examination with linkage-based investigation gives more successful results in a

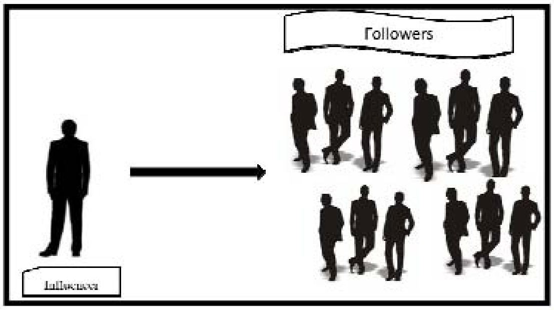
Author: e-mail: mesfer66@gmail.com

wide assortment of uses. For example groups which are planned with content substance are much wealthier as far as conveying data about the topical ability of the fundamental group.

III. Analysing Social Media Platform Data

a) Graph Theory

Graphs theory is most likely the principle strategy in the analysis of social media platforms in starting era of such platform [1]. The methodology is used on social media platforms data with a specific end goal to decide critical components of the system, for example, the hubs and connections (for instance influencers and the devotees). Influencers on these community have been recognized as clients that have sway on the exercises or feeling of different clients by method for followership or impact on choice made by different clients on the system. This hypothesis has ended up being exceptionally powerful on matter scale datasets [2]. This is on account of it is equipped for byepassing the working of a genuine visual representation of the information to run on information frameworks. Centrality measure was used to investigate the representation of force and impact that structures bunches cohesiveness on social media platforms [3]. Parameterized centrality metric is used to deal with the system structure and to rank hubs availability. Their work framed an expansion of a-centrality approach which measures the quantity of reduced ways that exist among hubs [4].

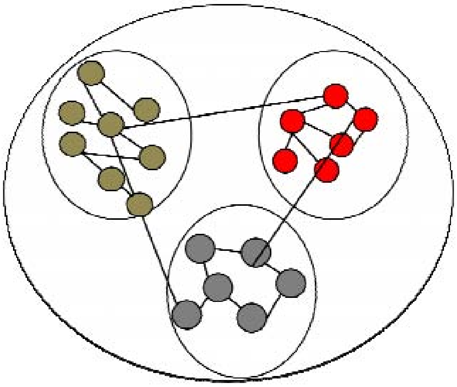


b) Community Detection

Community can be defined as small gathering inside a bigger system. This arrangement is most common qualities of informal community locales. Customers with comparative interest structure comes together and forms groups on social media platforms consequently showing solid sectional structure [5]. Groups on these platforms are similar to other groups in real business scenarios and are extremely perplexing in nature and hard to distinguish. Applying the proper instruments in identifying and comprehending the conduct of system groups is significant as this can be utilized to demonstrate the dynamism of the community they come from [6]. Various clustering principles have been given for the grouping strategies to distinguish groups on informal community with various leveled grouping being for the most part utilized. This procedure is a mix of numerous strategies used to gathering hubs in the system to uncover quality of individual gatherings which is then used to circulate the system into groups [7].Vertex grouping has a place with progressive grouping techniques, diagram vertices can be determined by including it in a vector space so that pairwise length between vertices can be measured. Basic similarity measures of various leveled bunching focus on number of regular system associations shared by two hubs [8].Two individuals on social system with a few common companions will probably be nearer than two individuals with less common companions on the system [9]. Clients in the same social system group regularly prescribe things and administrations to each other taking into account the experience on the things or administrations included. This is known as recommender framework

c) Recommender System

Taking into account the commonality between hubs in the social media platform hubs CF method which is known as collaborative filtering can be used which shapes one of the three classes of the recommender framework (RS), can be utilized to study affiliation among clients. Things can be prescribed to a client taking into account the rating of their common association [10]. Where CF's primary drawback is that of information sparsity, content-based (another RS technique) investigate the structures of the information to deliver suggestions. Be that as it may, the cross breed approaches normally propose suggestions by joining CF and data based proposals [11]. The analysis in proposed approach named EntreeC, a framework that pools learning based RS and CF to prescribe eateries. The work in enhanced CF calculation by utilizing an insatiable execution of various leveled agglomerative grouping to propose pending gatherings or diaries in which data scientists can present their work [12].



d) Semantic Web

The Semantic Web stage makes information sharing and re-use conceivable over various applications and group edges. Finding the evolvement of Semantic Web (SW) improves the information of the conspicuousness of Semantic Web People group and imagines the combination of the Semantic Web. There has been a lot of work done where this is utilized FOAF which is known as friend of a friend to investigate how nearby and worldwide group level gatherings create and develop in substantial scale of social media platforms on the Semantic Web [13]. The study uncovered the advanced layouts of social structures and conjectures future float. In the same way application model of Semantic Online analysis of social media platforms model makes the ontological field library of these platforms consolidating with the ordinary blueprint of the semantic web to achieve keen recovery of the Web administrations [14]. Besides others have enhanced the open-source Web-Harvest structure for the accumulation of online platforms information with a specific end goal to study structures of trust upgrade what's more, of online investigative affiliation [15]. Semantic Web is a moderately new territory in informal organization examination and exploration in the field is as yet developing.

IV. Conclusion

The ascent of social media platforms gives exceptionally solid impacts to the set of methods created for mining diagrams and social systems. Social media platforms are established in numerous sources of information and at various scales. In this scope data mining gives capable approach to execute and make utilization of database. In this paper we have quickly looked into the different information mining methods which are utilized for informal organization investigation also, its applications. It is very important to study social networks from the business perspective and hence by doing it successfully organizations can get insights into their current market landscape and can further leverage that knowledge into framing their plan of action and marketing strategies to improvise their positioning in the market and leapfrog the competition.

References Références Referencias

1. Borgatti, S. P., Everett, M. G.: A graph-theoretic perspective on centrality. Social networks 28, 466-484, 4, 2006.
2. Burt, R S.: Brokerage and closure: An introduction to social capital. Oxford University Press, 2005.
3. Ghosh, R., Lerman, K.: Parameterized centrality metric for network analysis. Physical Review E, 83(6), 066118, 2011.
4. Scott, J.: Social network analysis: developments, advances, and prospects. Social network analysis and mining, 1(1), 21-26, 2011.
5. Aggarwal, C.: An introduction to social network data analytics. Springer US, 2011.
6. Fortunato, S.: Community detection in graphs. Physics Reports, 486(3), 75-174, 2010.
7. Girvan, M., Newman, M. E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826, 2002
8. Newman, M.: Networks: An introduction. Oxford University Press, 2010.
9. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P. Community detection in socialmedia Data Mining and Knowledge Discovery, 24(3), 515-554, 2012.
10. Burke, R.: Hybrid recommender systems: Survey and Experiments. User Modelling and User-Adapted Interaction, 12(4): 331–370, 2002
11. Liu, F., Lee, H. J.: Use of social network information to enhance collaborative filtering performance. Expert Systems with Applications, 37, 4772-4778, 2010.
12. Pham, M. C., Cao, Y., Klamma, R., Jarke, M.: A clustering approach for collaborative filtering recommendation using social network analysis. J. UCS, 17(4), 583-604, 2011.
13. Murthy, D., Gross, A., Takata, A., Bond, S.: Evaluation and Development of Data Mining Tools for Social Network Analysis. In Mining Social Networks and Security Informatics (pp. 183-202). Springer Netherlands, 2013.
14. Ruan, X. H., Hu, X., Zhang, X.: Research on Application Model of Semantic Web-Based Social Network Analysis. In Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 2 (pp. 455-460). Springer Berlin Heidelberg, 2014
15. Zhou, L., Ding, L., & Finin, T.: How is the semantic web evolving? A dynamic social network perspective. Computers in Human Behaviour, 27(4), 1294-1302, 2011.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

MAGED: Metaheuristic Approach on Gene Expression Data: Predicting the Coronary Artery Disease and the Scope of Unstable Angina and Myocardial Infarction

By E.Neelima & M.S.Prasad Babu

GITAM University

Abstract - The Genetic risk prediction strategies found in practice for coronary artery disease are not significant to estimate the scope of adverse cardiovascular events such as unstable angina and myocardial infarction. Hence in regard to this objective, this manuscript contributed a metaheuristic approach to predict coro- nary artery disease and the scope of unstable angina and myocardial infarction. The proposed metaheuristic is built from the gene expression data of blood samples collected from patients with coronary artery disease diagnosed, unstable angina and Myocardial Infarction. The data also includes gene expression data collected from the blood samples taken from the people clinically proven as salubrious (healthy). The relation between genes and gene expressions are considered as the state of input to devise the metaheuristic.

Keywords: *micro array, coronary artery disease, unstable angina, myocardial infarction, gene expression data, gene expression profiling, metaheuristics, machine learning.*

GJCST-C Classification : *J.3, H.2.1*

MAGED: Metaheuristic Approach on Gene Expression Data: Predicting the Coronary Artery Disease and the Scope of Unstable Angina and Myocardial Infarction

E.Neelima α & M.S.Prasad Babu σ

Abstract - The Genetic risk prediction strategies found in practice for coronary artery disease are not significant to estimate the scope of adverse cardiovascular events such as unstable angina and myocardial infarction. Hence in regard to this objective, this manuscript contributed a metaheuristic approach to predict coro- nary artery disease and the scope of unstable angina and myocardial infarction. The proposed metaheuristic is built from the gene expression data of blood samples collected from patients with coronary artery disease diagnosed, unstable angina and Myocardial Infarction. The data also includes gene expression data collected from the blood samples taken from the people clinically proven as salubrious (healthy). The relation between genes and gene expressions are considered as the state of input to devise the metaheuristic. In order to find the confidence of the relation between gene and gene expression a bipartite graph is built between them. The experimental study evincing that the prediction performance of the proposed model is substantial that compared to other benchmarking models.

Keywords: micro array, coronary artery disease, unstable angina, myocardial infarction, gene expression data, gene expression profiling, metaheuristics, machine learning.

I. Introduction

Cardiovascular diseases are the critical reason of human deaths happening worldwide. The statistics indicating that this disease causes annually around 17.3 million deaths [1].The inadequate blood supply to the heart causes necrosis of myocardial tissue, which is clinically referred as Myocardial Infarction (MI).

The MI was claimed 7.6 million deaths among 58 million deaths worldwide in 2005 [2]. The advancements in clinical practices to diagnose and prevent MI are evinced to be not significant, since the count of human deaths due to MI is high that compared to the deaths caused by any other disease [1] [2].

The current diagnosis of MI is based on clinical symptoms including chest pain and difficulty to breath, ECG pattern variants, and potential drop and raise of blood floating in cardiac muscles (cardiac troponins also referred as cTns) [3]. Though the phenomenal advances in clinical diagnosis strategies found, still the substantial constraints are evinced in current clinical diagnosis strategies. The advances in hs-cTni assays [4] have evinced high detection of cardio vascular disease cases (Increased true positive rate) but significant normal cases have been labeled as cardio vascular prone (decreased true negative rate), which is a potential constraint. Another advanced approach of diagnose the cardio vascular disease diagnostic measure is the cardiac miRNAs as biomarkers [5]. The prediction outcomes of this model are trivial due to limited size and tissue specific expression. Hence it is obvious to have more significant and automated detection strategies, which are using the cardiac miRNAs as primary input [6]. The serum inflammatory markers such as BNP, CRP are also considered as cardiovascular biomarkers but the detection accuracy observed with slight improvement [7][8][9].

The acts such as clinical pathology and biology are the crucial to define cardiac biomarkers, which are expensive and less accurate. In contrast to this, the gene expression profiling quantifies the gene expressions formed by the large quantity of genes in order to identify biomarkers, which is analogous and concurrent across the multiple pathways. Hence the gene expression profiling is potential and feasible to quantify the biomarkers to diagnose cardio vascular diseases [10]. The biomarkers defined by Gene expression profiling are potential and those are not evinced by the pathology and biology based clinical processes.

The rest of manuscript describes the related work in section 2, the Metaheuristic Approach on Gene expression Data (MAGED) that followed by section 4, which elaborates the experimental study of the proposal. Finally the section 5 concludes the contribution of the manuscript.

II. Related Work

Gene expression analysis is a potential approach to discover profound biomarkers of cardio vascular diseases. The contemporary literature contains significant contributions in defining biomarkers through gene

Author α: Assistant Professor, Department of CSE GITAM University, Visakhapatnam, AP, INDIA. e-mail: eadha.neelima@gmail.com

Author σ: Professor, Department of CS & SE Andhra University, Visakhapatnam, AP, INDIA. e-mail: profmspbabu@gmail.com

expression analysis. Randi et al., [11] devised a gene expression analysis that conceded 482 genes associated to the composition of plaques found in arteries. Many of these genes were not considered for atherosclerosis in earlier diagnosis strategies. Archacki et al.,[12] proposed a gene expression profiling strategy that resulted 56 different genes for atherosclerosisprone and salubrious human coronary arteries. Among these 56, the 49 genes were not associated to coronary artery disease earlier. The model devised in [13] discovered set of genes those enables classification according to age and sex, which are having strong association with obstreperous CAD in the patients, who are not diagnosed as diabetic. The contributions in [14] and [15] profiled variant gene expressions to differentiate the cardio myopathies with influence of ischemic and non-ischemic conditions. Min KD et al., [16] contributed profiling and analysis of gene expressions to notice the divergent genes associated to congestive heart failure. Suresh R et al., [17] studied the salubrious and MI patients that discovered biomarkers and imbalanced pathways those significant evince the reappearance MI in patients effected once with MI.

Liew et al.,[18] defined sequence tags from gene expressions using microarray analysis that compares mRNA molecules found in cellular components of the blood with mRNA molecules found in9divergent human tissues comprising heart. The correlation observed from this comparison concluded that 84% of mRNA molecules were overlapped with mRNA molecules of heart and 80% were overlapped with mRNA molecules of other tissues. mRNA molecules of cellular components of the blood are costing minimal and feasible to access in order to substitute gene expression in other tissues.

The contributions found in contemporary literature are specific to discover the influential genes of Myocardial Infarction. None of these are capable to identify the given gene expression is prone to CAD under MI and UA or the expression is salubrious. This evinces the need of novel contributions to discover the state of a given gene expression is prone to CAD under MI and UA or salubrious. This helps to deploy the case based reasoning to treat the patients prone to CAD under MI and UA differently. In this regard this manuscript attempted to define metaheuristic approach on gene expression data (MAGED) to discover the state of a given gene expression is prone to CAD under MI and UA or salubrious. The MAGED is machine learning strategy that learns from the labeled gene expression data of Cardia Vascular Diseased, Unstable Angina, Myocardial Infarction and Salubrious cases.

III. Metaheuristic Approach on Gene Expression Data

The objective of the MAGED is to define a metaheuristic scale by the knowledge gained from the given gene expression data. In order to this the given gene expressions are partitioned into their respective categories of coronary artery disease (CAD), unstable angina (UA), Myocardial Infarction (MI) and salubrious (blood samples taken were diagnosed as healthy). The data also includes gene expression data collected from the blood samples taken from the people clinically proven as normal.

The genes involved in each gene expression are considered as features of the respective category. Since the gene expression contains dense number of genes and majority of them may be insignificant to respective category of the disease. Henceforth, the feature optimization process (see sec 3.1) will be carried out to eliminate these insignificant features. The gene range will be discretized further to compare two genes through equality by approximation (see sec 3.2). Afterwards the confidence of each feature towards all categories of gene expression data will be assessed (see sec 3.3) that follows the assessment of each gene expression confidence against the features of all categories (see sec 3.4). Further the confidence obtained for each feature and gene expression of respective category will be used as input to define the metaheuristic scales to estimate the scope of coronary artery disease, the unstable angina and myocardial infarction.

a) Feature Optimization

For each disease context considered, the gene expression dataset $D_i = \{e(i)_1, e(i)_2, ...e(i)_{|D_i|}\}$ of size $|D_i|$ will be considered for training towards defining metaheuristic scale. Each gene expression is representted by sequence of genes for the set of features selected of respective diseases context. This description binds to all datasets ofgene expressions representing coronary artery diseases, Unstable Angina, Myocardial Infarction.

Let $D_n = \{e(n)_1, e(n)_2, ..., e(n)_{|D_n|}\}$ be the set of gene expressions collected from the blood samples of salubrious cases. The sets $F_i = \{f(i)_1, f(i)_2, ..., f(i)_{|F_i|}\}$ and $F_n = \{f(n)_1, f(n)_2, ..., f(n)_{|F_n|}\}$ are feature sets of gene expressions represented by D_i and D_n respectively. The attribute set $G(i)_j = \{g(ij)_1, g(ij)_2, .\ g(ij)_{|G(i)_j|}\}$ be the set of genes as values observed for feature $f(i)_j$ of gene expressions represented by D_i. Similarly the attribute set $G(n)_j = \{g(nj)_1, g(nj)_2, .\ g(nj)_{|G(n)_j|}\}$ be the set of genes as values observed for feature $f(n)_j$ of gene expressions represented by D_n.

Since the gene expression is the combination of numerous count of genes, the size of feature set can lead to process complexity. In order to overcome the process complexity, the insignificant features should be identified and discarded. The feature $f(i)_j$ of F_i is said to be insignificant feature, if genes $G(i)_j$ of $f(i)_j$ are almost similar to the genes $G(n)_j$ of feature $f(n)_j$ of F_n .Hence to identify the insignificant features, we adopt hamming distance that applied on genes of each feature as vectors from each disease and normal cases. The hamming distance with 0 or less than the given threshold indicates that the respective feature is insignificant. The process of hamming distance is explored below:

i. *Hamming Distance*

The value of Hamming Distance obtained here is to denote the difference between genes assigned to same feature from gene expression data of diseased and normal cases. This is one of the significant strategy to assess the difference between to elements in coding theory.

The hamming distance between given vectors $CX = \{cx_1, cx_2, ..., cx_n\} \& CY = \{cy_1, cy_2, ..., cy_m\}$ of size n and m respectively will be measured as follows:

```
Let $CZ \leftarrow \phi$ // is a vector of size 0
$foreach \ \{i \exists i = 1, 2, 3, .....\max(n.m)\}$ Begin
$if \ (\{cx_i \exists cx_i \in CX\} - \{cy_i \exists cy_i \in CY\}) \equiv 0 \ then$
$CZ \leftarrow \{cx_i \exists cx_i \in CX\} - \{cy_i \exists cy_i \in CY\}$
Else
        $CZ \leftarrow 1$
End
```

$$hd_{CX \leftrightarrow CY} = \sum_{j=1}^{|CZ|} CZ\{i\}$$

// $hd_{CX \leftrightarrow CY}$ is the hamming distance between CX and CY , $CZ\{i\}$ is the i^{th} element of the vector CZ and $|CZ|$ is the size of the vector CZ

b) *Gene and Gene Expression Confidence Assessment*

Then these genes found for each optimal feature of respective gene expression data set and the gene expressions of that data set will be used further to assess the gene and gene expression confidence.

In order to this, initially the gene pairs will be defined such that each pair contains two genes and each gene representing different feature of the same dataset. Then we assess the associativity support of each gene pair. The associativity support can be described as the ratio of gene expressions contains that pair against the total number of gene expressions in respective dataset. The process of assessing associativity support of each gene pair is described in following section (see sec 3.2.1).

ii. *Assessing gene pair correlation*

Let P_i be the set and contains all possible unique gene pairs from respective dataset D_i . The possible unique gene pairs will found as follows:

For each gene expression $e(i)_j$ of respective dataset D_i , find all possible unique pairs of genes and add to P_i . Then correlation of each pair $\{p_j \exists p_j \in P_i\}$ as follows.

Let $\{g_k \exists g_k \in p_j\}$ and $\{g_l \exists g_l \in p_j\}$ be the two genes paired as $\{p_j \exists p_j \in P_i\}$, then the correlation $s(p_j)$ of the pair p_j is

$$s(p_i) = \frac{\sum_{v=1}^{|D_i|} \{1 \exists \{g_k, g_l\} \subseteq e(i)_v\}}{|D_i|}$$

//The ratio of number of gene expressions contain both genesagainst total number of genes

The correlation of each pair of genes found in gene expressions of each respective gene expression data set of coronary artery disease, unstable angina, myocardial infarction and normal cases should be estimated using the process explored in sec 3.2.1.

iii. *Assessing Gene and Gene Expression Confidence*

In order to assess the confidence of genes and gene expressions of respective gene expression dataset D_i , a mutual relation graph will be formed between gene expressions and genes of respective D_i . There will be an edge between a gene and gene expression if and only if the selected gene exists in that gene expression. Then each edge between gene and gene expression is weighted as follows.

```
        $\forall_{j=1}^{|G(i)|} \{g_j \exists g_j \in G(i)\}$ Begin
        $\forall_{k=1}^{|D_l|} \{e(l)_k \exists g_j \in e(l)_k\}$ Begin
                        $w_{g_j} = 0$
        $\forall_{m=1}^{|e(l)_k|} \left\{g_m \exists g_m \in e(l)_k \wedge g_j \neq g_m\right\}$ Begin
        $p_m = \{g_j, g_m\}$
                        $w(g_j) += s(p_m)$
                    End
                        $w_{g_j \leftrightarrow e(l)_k} = \frac{w(g_j)}{|e(l)_k| - 1}$
End
End
```

The weights obtained for edges between genes and gene expressions in mutual graph are further used to assess the gene and gene expression confidence towards respective CAD (coronary artery disease), UA (unstable angina), MI (myocardial infarction) and Normal datasets.

Further we measure the each feature confidence towards gene expression dataset D_i as follow

```
            $\forall_{j=1}^{|G(i)|} \{g_j \exists G(i) \ni g_j\}$ Begin
$c_{g_j \Rightarrow D_i} = \sum_{k=1}^{|D_i|} \{w(g_j) \exists e(i)_k \ni g_j \wedge D_i \ni e(i)_k\}$
```

//aggregating the weight of gene g_j towards each gene expression $e(i)_k$ of respective dataset D_i and the same is considered as the respective gene confidence towards dataset D_i

End

Similarly each respective gene expression confidence towards gene expression dataset D_i is measured as follows

$$\overset{|D_i|}{\underset{j=1}{\forall}}\{e(i)_j \exists D_i \ni e(i)_j\} \text{ Begin}$$

$$c_{e(i)_j \Rightarrow D_i} = \sum_{k=1}^{|G(i)|}\{w(g_k) \otimes c_{g_k \Rightarrow D_i} \exists e(i)_j \ni g_k \wedge D_i \ni e(i)_j\}$$

//The sum of product of each gene weight and the respective gene confidence, such that the gene exists in selective gene expression is the confidence of that gene expression

End

The confidence of genes and gene expressions of each respective gene expression data set of CAD, UA, MI and salubrious cases should be estimated using the process explored in sec 3.2.2.

c) Defining metaheuristics to CAD, UA, MI and Salubrious scope

Further the confidence of gene expressions of gene expression datasets D_{CAD} , D_{UA} , D_{MI} and

$$m_{CAD} = \frac{\sum_{i=1}^{|D_{CAD}|}\{c_{e(CAD)_i \Rightarrow D_{CAD}} \exists D_{CAD} \ni e(CAD)_i\}}{|D_{CAD}|} \quad \text{//Aggregate}$$

mean of the respective gene expressions confidence of coronary artery disease gene expression dataset D_{CAD}

In order to identify the lower and upper bounds of m_{CAD} , the mean absolute distance of D_{CAD} is assessed as follows

$$m_{CAD}^{ad} = \frac{\sum_{i=1}^{|D_{CAD}|}\sqrt{\left(m_{CAD} - c_{e(CAD)_i \Rightarrow D_{CAD}}\right)^2}}{|D_{CAD}|}$$

Then the lower and upper bounds of m_{CAD} is assessed as

$ml_{CAD} = m_{CAD} - m_{CAD}^{ad}$ // lower bound of m_{CAD}

$mu_{CAD} = m_{CAD} + m_{CAD}^{ad}$ // upper bound of m_{CAD}

Similarly metaheuristics for UA (unstable angina), MI (Myocardial infarction) and salubrious (healthy) scope

$$m_{UA} = \frac{\sum_{i=1}^{|D_{UA}|}\{c_{e(UA)_i \Rightarrow D_{UA}} \exists D_{UA} \ni e(UA)_i\}}{|D_{UA}|} \quad \text{//Aggregate}$$

mean of the respective gene expressions confidence of Unstable Angina gene expression dataset D_{UA}

The mean absolute distance of D_{UA} is

$$m_{UA}^{ad} = \frac{\sum_{i=1}^{|D_{UA}|}\sqrt{\left(m_{UA} - c_{e(UA)_i \Rightarrow D_{UA}}\right)^2}}{|D_{UA}|}$$

Then the lower and upper bounds of m_{UA} is assessed as

$ml_{UA} = m_{UA} - m_{UA}^{ad}$ // lower bound of m_{UA}

$mu_{UA} = m_{UA} + m_{UA}^{ad}$ // upper bound of m_{UA}

$$m_{MI} = \frac{\sum_{i=1}^{|D_{MI}|}\{c_{e(MI)_i \Rightarrow D_{MI}} \exists D_{MI} \ni e(MI)_i\}}{|D_{MI}|} \quad \text{//Aggregate}$$

mean of the respective gene expressions confidence of myocardial Infarction gene expression dataset D_{MI}

The mean absolute distance of D_{MI} is

$$m_{MI}^{ad} = \frac{\sum_{i=1}^{|D_{MI}|}\sqrt{\left(m_{MI} - c_{e(MI)_i \Rightarrow D_{MI}}\right)^2}}{|D_{MI}|}$$

Then the lower and upper bounds of m_{MI} is assessed as

$ml_{MI} = m_{MI} - m_{MI}^{ad}$ // lower bound of m_{MI}

$mu_{MI} = m_{MI} + m_{MI}^{ad}$ // upper bound of m_{MI}

$$m_N = \frac{\sum_{i=1}^{|D_N|}\{c_{e(N)_i \Rightarrow D_N} \exists D_N \ni e(N)_i\}}{D_N} \quad \text{//Aggregate mean of}$$

the respective gene expressions confidence of salubrious gene expression dataset D_N

The mean absolute distance of D_N is

$$m_N^{ad} = \frac{\sum_{i=1}^{|D_N|}\sqrt{\left(m_N - c_{e(N)_i \Rightarrow D_N}\right)^2}}{|D_N|}$$

Then the lower and upper bounds of m_N is assessed as

$ml_N = m_N - m_N^{ad}$ // lower bound of m_N

$mu_N = m_N + m_N^{ad}$ // upper bound of m_N

d) Predicting the state of gene expression

The metaheuristics devised (see section 3.3) will be used further to assess the CAD, UA and MI scope of a given gene expression e . The confidence of given gene expression

$$c_{e \Rightarrow CAD} = \frac{\sum_{i=1}^{|G(D_{CAD})|}\left\{c_{g_i \Rightarrow CAD} \otimes w(g_i) \exists g_i \in G(D_{CAD}) \wedge e \ni g_i\right\}}{\sum_{j=1}^{|G(D_{CAD})|}\left\{c_{g_j \Rightarrow CAD} \otimes w(g_j) \exists g_j \in G(D_{CAD})\right\}}$$

// the aggregate of product of each gene confidence and weight of that exists in $G(D_{CAD})$ and e , which divides by the aggregate of confidence of all genes exists in $G\ D(\ _{CAD}\ )$

Further the confidence of e towards D_{UA}, D_{MI} and D_N assessed as :

$$c_{e\Rightarrow UA} = \frac{\sum_{i=1}^{|G(UA)|}\left\{c_{g_i\Rightarrow UA}\otimes w(g_i)\exists g_i \in G(UA)\wedge e \ni g_i\right\}}{\sum_{j=1}^{|G(UA)|}\left\{c_{g_j\Rightarrow UA}\otimes w(g_j)\exists g_j \in G(UA)\right\}}$$

// the aggregate of product of each gene confidence and weight of that exists in $G(D_{UA})$ and e, which divides by the aggregate of confidence of all genes exists in $G(D_{UA})$.

$$c_{e\Rightarrow MI} = \frac{\sum_{i=1}^{|G(MI)|}\left\{c_{g_i\Rightarrow MI}\otimes w(g_i)\exists g_i \in G(MI)\wedge e \ni g_i\right\}}{\sum_{j=1}^{|G(MI)|}\left\{c_{g_j\Rightarrow MI}\otimes w(g_j)\exists g_j \in G(MI)\right\}}$$

// the aggregate of product of each gene confidence and weight of that exists in $G(D_{UA})$ and e, which divides by the aggregate of confidence of all genes exists in $G(D_{UA})$.

$$c_{e\Rightarrow N} = \frac{\sum_{i=1}^{|G(N)|}\left\{c_{g_i\Rightarrow N}\otimes w(g_i)\exists g_i \in G(N)\wedge e \ni g_i\right\}}{\sum_{j=1}^{|G(N)|}\left\{c_{g_j\Rightarrow N}\otimes w(g_j)\exists g_j \in G(N)\right\}}$$

// the aggregate of product of each gene confidence and weight of that exists in $G(D_{UA})$ and e, which divides by the aggregate of confidence of all genes exists in $G(D_{UA})$.

Then these confidence values of gene expression e with respect to CAD, UA, MI and N will be used to estimate the given expression state is salubrious, prone to coronary artery disease, Unstable Angina or Myocardial Infarction according to the following conditions.

$$(c_{e\Rightarrow CAD}\ge mu_{CAD})\vee(c_{e\Rightarrow UA}\ge mu_{UA})\vee(c_{e\Rightarrow MI}\ge mu_{MI})$$

Coronary Artery Disease Confirmed (highly prone to either of three disease conditions)

$$(c_{e\Rightarrow CAD}\ge m_{CAD})\wedge(c_{e\Rightarrow UA}\ge ml_{UA})\wedge(c_{e\Rightarrow MI}\ge ml_{MI})$$

Coronary Artery Disease Confirmed (prone to CAD and either or both of the UA and MI)

$$\text{If}\ \left(\begin{array}{l}(c_{e\Rightarrow CAD}\ge ml_{CAD})\wedge\\(c_{e\Rightarrow UA}\ge ml_{UA})\wedge\\(c_{e\Rightarrow MI}\ge ml_{MI})\wedge\\(c_{e\Rightarrow N}< m_{MI})\end{array}\right)$$

Then Prone to Coronary Artery Disease

$$\text{if}\ \left(\begin{array}{l}(c_{e\Rightarrow CAD}< ml_{CAD})\wedge\\(c_{e\Rightarrow UA}< ml_{UA})\wedge\\(c_{e\Rightarrow MI}< ml_{MI})\wedge\\(c_{e\Rightarrow N}> m_{MI})\end{array}\right)$$

Then Salubrious state Confirmed

$$\text{if}\ \left(\begin{array}{l}(c_{e\Rightarrow CAD}< m_{CAD})\wedge\\(c_{e\Rightarrow UA}< m_{UA})\wedge\\(c_{e\Rightarrow MI}< m_{MI})\wedge\\(c_{e\Rightarrow N}\ge mu_{MI})\end{array}\right)$$

Then Prone to Salubrious state

IV. Experimental Study

The experimental study was carried out on a set of gene expressions taken from multiple benchmark datasets [19]. The number of gene expressions used are 1114, which are the combination of coronary artery Disease (286 expressions), Unstable Angina (275 expressions), Myocardial Infarction (277 expressions) and salubrious condition (276 expressions). The gene expressions of respective category are considered as separate datasets labeled as D_{CAD}, D_{UA}, D_{MI} and D_N. Each dataset D_{CAD}, D_{UA}, D_{MI} and D_N partitioned into test and training sets. The 75% of gene expressions of each dataset are considered as training set and rest 25% of gene expressions considered as test set.

The metaheuristics obtained from the given training set were explored in table 1.

Table 1: The metaheuristics obtained from training data

Training Set	834 (CAD:214, UA:206,MI:207, N:207)
m_{CAD}	0.582474187
m^{ad}_{CAD}	0.095593654
ml_{CAD}	0.486880533
mu_{CAD}	0.678067841
m_{UA}	0.615957277
m^{ad}_{UA}	0.103864099
ml_{UA}	0.512093178
mu_{UA}	0.719821376
m_{MI}	0.646638853

m_{MI}^{ad}	0.099722167
ml_{MI}	0.546916686
mu_{MI}	0.74636102
m_{N}	0.631593026
m_{MI}^{ad}	0.068999373
ml_{N}	0.562593653
mu_{N}	0.700592398

Table 2 : The prediction statistics of the SDS

Test Set	280 (CAD:72, UA:69,MI:70, N:69)
True Positives	197
True Negatives	54
False Positives	15
False Negative	14
CAD, UA and MI gene expression Prediction Value (positive prediction value, PPV)	0.929245283
Salubrious gene expression Prediction Value (Negative Prediction value, NPV)	0.794117647
Detection Accuracy	0.896428571
AD, UA and MI gene expression prediction Rate (True Positive Rate)	0.933649289
Salubrious gene expression Prediction rate (True Negative Rate)	0.782608696

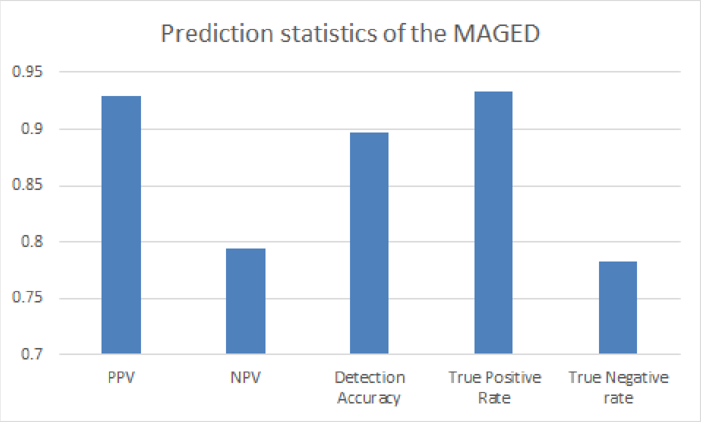


Figure 1: The prediction statistics observed for MAGED

The 280 (CAD: 72, UA: 69, MI: 70, N: 69) gene expressions were used to assess the prediction accuracy of the proposed MAGED. The MAGED assessed the given input gene expressions such that 197gene expressions are true positives (the detection of CAD, UA and MI gene expressions are true), 15gene expressions are false positive (falsely detected as CAD, UA or MI), 54gene expressions are true negatives (detecting gene expressions as salubrious is true) and 14gene expressions are false negative (detecting gene expressions as salubrious is false). Hence the CAD, UA or MI gene expression prediction value (also known as precision or positive prediction value) is 0.93, Salubrious Gene Expression prediction value is 0.79, the CAD, UA and MIgene expression detection rate (also known as sensitivity) is 0.93, the salubrious gene expression detection rate (also known as specificity) is 0.782 and the overall success rate (also known as accuracy, which is the ratio between true prediction of all types of gene expressions and all given number of gene expressions) is 0.90. These statistics indicating that the MAGED is find to significant to identify the CAD, UA and MI prone gene expressions with success percentage of 93% (since sensitivity is 0.93), but the detection of salubrious cases, the success rate is 78% (since specificity is 0.782). The computer aided medical diagnosis should be more robust to deliver high sensitivity at the cost of specificity. Hence the Model MAGED is scalable and robust to predict the CAD, UA and MI prone gene expressions.The prediction statistics observed from the experimental study of the MAGED are visualized in fig1.

V. Conclusion

This paper introduced a learning model that device heuristics to scale the given patient record is disease prone or normal. The proposed learning model delivers two heuristics called Scale to Diseased health Scope and Scale to Normal Health Scope. In contrast to the existing benchmarking models, these heuristics are further used as scales to assess the given patient record is disease prone or normal. The medical records labeled as diseased and normal are used to device the heuristics $sdhs$ and $snhs$ respectively. In order to this all unique values of all the attributes are considered as features, and then the influence weight of these features towards their respective datasets. The influence weights further will be used to assess the influence weight of the each record in dataset. From these influence weights of the records of respective dataset will be used to assess the proposed heuristics. The experimental results are optimistic and concluding the prediction accuracy and robustness. This work can be extended to identify the impact of feature correlation towards minimizing the process and computational complexity of the learning process.

References Références Referencias

1. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. Circ. 2015; 131:29–32.
2. Mendis, S., Thygesen, K., Kuulasmaa, K., Giampaoli, S., Mähönen, M., Blackett, K. N., & Lisheng, L. (2011). World Health Organization definition of myocardial infarction: 2008–09 revision. International journal of epidemiology,40(1), 139-146.
3. Thygesen K, Alpert JS, White HD. Universal definition of myocardial infarction. Europ Heart J. 2007; 28:2525–2538.
4. Eggers KM, Lind L, Venge P, Lindahl B. Will the universal definition of myocardial infarction criteria result in an over diagnosis of myocardial infarction?. The Amer J of Card. 2009; 103:588–591.
5. Wang Z, Luo X, Lu Y, Yang B. miRNAs at the heart of the matter. J of Mol Med. 2008; 86:771–783.
6. de Planell-Saguer M, Rodicio MC. Detection methods for microRNAs in clinic practice. Clin Biochem. 2013; 46:869–878. doi: 10.1016/j.clinbio chem.2013.02.017 PMID: 23499588.
7. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engström G, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. The J of the Amer Med Assoc. 2009; 302:49–57.
8. Shah T, Casas JP, Cooper JA, Tzoulaki I, Sofat R, McCormack V, et al. Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts. Inter J of Epid. 2009; 38: 217–231.
9. Wilson PWF, Pencina M, Jacques P, Selhub J, D'Agostino R, O'Donnell CJ. C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ: Card Qual and Outc. 2008; 2: 92–97.
10. Pedrotty DM, Morley MP, Cappola TP. Transcriptomic biomarkers of cardiovascular disease. Prog in Card Dis. 2012; 55: 64–69.
11. Randi AM, Biguzzi E, Falciani F, Merlini P, Blakemore S, Bramucci E, et al. Identification of differentially expressed genes in coronary atherosclerotic plaques from patients with stable or unstable angina by cDNA array analysis. J of Throm and Haem. 2003; 1: 829–835.
12. Archacki S, Angheloiu G, Tian XL, Tan FL, DiPaola N, Shen GQ, et al. Identification of new genes differentially expressed in coronary artery disease by expression profiling. Phys Genom. 2003; 15: 65–74.
13. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in nondiabetic patients. BMC Med Genom. 2011; 4: 4–26.
14. Kittleson MM, Ye SQ, Irizarry RA, Minhas KM, Edness G, Conte JV, et al. Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. Circ. 2004; 110: 3444–3451.
15. Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, et al. Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. Phys Genom. 2005; 21: 299–307.
16. Min KD, Asakura M, Liao Y, Nakamaru K, Okazaki H, Takahashi T, et al. Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. Bioch and Biophy Res Comm. 2010; 393: 55–60.
17. Suresh R, Li X, Chiriac A, Goel K, Terzic A, Perez-Terzic C, et al. Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction. J of Mol and Cell Card. 2014; 74: 13–21.
18. Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. The J. of Lab and Clin Med. 2006; 147: 126–[132].
19. https://www.ebi.ac.uk/ega/datasets

This page is intentionally left blank

Global Journals Inc.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Wildfire Predictions: Determining Reliable Models using Fused Dataset

By Hariharan Naganathan, Sudarshan P Seshasayee, Jonghoon Kim, Wai K Chong & Jui-Sheng Chou

Arizona State University, United States

Abstract- Wildfires are a major environmental hazard that causes fatalities greater than structural fire and other disasters. Computerized models have increased the possibilities of predictions that enhanced the firefighting capabilities in U.S. While predictive models are faster and accurate, it is still important to identify the right model for the data type analyzed. The paper aims at understanding the reliability of three predictive methods using fused dataset. Performances of these methods (Support Vector Machine, K-Nearest Neighbors, and decision tree models) are evaluated using binary and multiclass classifications that predict wildfire occurrence and its severity. Data extracted from meteorological database, and U.S fire database are utilized to understand the accuracy of these models that enhances the discussion on using right model for dataset based on their size. The findings of the paper include SVM as the best optimum models for binary and multiclass classifications on the selected fused dataset.

Keywords: support vector machines, k-nearest neighbor, k-fold cross-validation, decision tree stumps, forest fire, binary and multiclass classifiers.

GJCST-C Classification: H.2.8

Wildfire Predictions: Determining Reliable Models using Fused Dataset

Hariharan Naganathan [α], Sudarshan P Seshasayee [σ], Jonghoon Kim [ρ], Wai K Chong [ω] & Jui-Sheng Chou [¥]

Abstract- Wildfires are a major environmental hazard that causes fatalities greater than structural fire and other disasters. Computerized models have increased the possibilities of predictions that enhanced the firefighting capabilities in U.S. While predictive models are faster and accurate, it is still important to identify the right model for the data type analyzed. The paper aims at understanding the reliability of three predictive methods using fused dataset. Performances of these methods (Support Vector Machine, K-Nearest Neighbors, and decision tree models) are evaluated using binary and multiclass classifications that predict wildfire occurrence and its severity. Data extracted from meteorological database, and U.S fire database are utilized to understand the accuracy of these models that enhances the discussion on using right model for dataset based on their size. The findings of the paper include SVM as the best optimum models for binary and multiclass classifications on the selected fused dataset.

Keywords: support vector machines, k-nearest neighbor, k-fold cross-validation, decision tree stumps, forest fire, binary and multiclass classifiers.

I. Introduction

Wildfires are a major environmental hazard and a real world problem that affects human, wildlife and create damages to the economy. According to United States Department of Agriculture (USDA), fatalities caused by the wildfires are greater than structural fire and other disasters. Over 90% of the wildfires were caused by humans while others by a volcanic eruption and lightning. Data mining techniques have increased the possibilities of predicting forest fires that enhanced the firefighting capabilities in U.S. The National Interagency Fire Center (NIFC) provides daily information on wildfire events using various intelligence and predictive methods.

While data mining techniques are faster and accurate, it is essential to identify the right method for the database chosen.

Forest fire causes devastation to vegetation and building structures in the areas that it affects. Fire shapes the landscape and influences the bio-geo-chemical cycles (e.g. the ecological carbon cycle). Some technological advancements in fire-fighting are yet to balance the costs invested since hectares of forests are still destroyed every year. Efficient probabilistic models to detecting systems, real-time reporting and forecasting the trend have been developed using fire databases. Brillinger et al. developed a model based on location, elevation above sea level and fire and non-fire days. Despite numerous similar models with additional factors (weather and topography) were developed, USDA claims that the average burnt area is 7.3 million acres every year (2014).

An accidental small forest fire can lead to heavy loss of precious natural reserves on protected lands (Iyer, T, Paramesh, Murthy, & U, 2011). Forest fires are fueled by high temperature, strong wind, lack of precipitation, lightning, human negligence (e.g. cigarette and campfire), and arsons. A combination of these factors would make forest fire uncontrollably causing casualties and economic losses. The highly populated western part of United States (California and Oregon) are highly impacted by these factors according to USDA database. The Federal and State governments have developed many strategies to control forest fires, e.g. the National Cohesive Wild land Fire Management Strategy, Quadrennial Fire Review, and the National Fire Plan (Park, National, & Fire, 2003; Tania Schoennagel & Nelson, 2011). Also, they also provide daily forest fire information by associating meteorological conditions (e.g. lightning and lack of moisture) with potential fire hazards and thus isolate human-driven factors. The United State Department of Agriculture (USDA) forest fire services also conducts research on fire hazards to understand wild land fires, focusing its impact on the ecosystem.

II. Predictive Analytics

Predictive analytic is becoming a popular trend in predicting extreme events and disasters. Federal and state government, industrial activists (e.g. IBM), and non-profit agencies (e.g. Borealforest.org) conduct

Author α: Graduate Student, SSEBE, Arizona State University, Tempe, Arizona, 85287. e-mail: hnaganat@asu.edu

Author σ: Graduate Student, ECEE, Arizona State University, Tempe, Arizona, United States, 85287. e-mail: prash250491@gmail.com

Author ρ: Assistant Professor, Oklahoma State University, CMT, Stillwater, Oklahoma, United States, 74078. e-mail: jongkim@okstate.edu

Author ω: Associate Professor, SSEBE, Arizona State University, Tempe, Arizona, United States, 85287. e-mail: ochong@asu.edu

Author ¥: Professor, Construction Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, e-mail: jschou@mail.ntust.edu.tw

research to develop a generic model for predicting forest fires. Hierarchical information is a significant tool that connects factors and helps understand the start and growth of a forest fire. Such information helps fire managers make critical short and long-term decisions before the beginning of and during a wildfire. In addition to prediction, firefighting and fire restoration are also a part of wildfire mitigation. According to Burned Area Emergency Response (BAER), proper restoration and adaptation procedures after forest fires are a necessary and handy system to have. The active fire mapping program by the National Interagency Fire Center (NIFC) includes the location, severity, the type, burnt area, and the contaminant status of the wildfire region. It also specifies the causes of the fire that helps fire managers make a decision. The Wildfire Assessment System (WFAS) is a mapping tool that provides information on fuel and fire hazards. Also, the Federal government has a comprehensive fire prevention and prediction system that predicts, forecasts and contains information on forest fires through a national database on wildfires.

Predictive models integrating meteorological data from different weather stations (local sensors) and fire database still need improvement since it can possess lower predictive accuracy for larger fires. The accuracy also depends on the size of the database and its features. The motivation of this paper is to enhance the predictability of forest fires using predictive analytics to manage it effectively. The primary focus of this article is to develop prediction forecast models from spatial data, identify the areas prone to wildfires from previous meteorological and fire data using both binary and multi-class classifiers. While this is not a new approach, the applications have yet been fully tested to predict forest fire.

III. Research Objective

The objective is to understand the reliability of three techniques (that uses a dimensionality-reduced dataset) in predicting forest fires using USDA data. These techniques have been proven to provide insights for decision makers and computer scientists. The paper proposes a comparative study of the three techniques to analyze and predict forest fires using data from California, Idaho, Oregon, Nevada, and New Mexico. These states were selected due to the severity and frequency of occurrences between 2004 and 2014. The authors used three different predictive techniques in this paper to identify which one has greater accuracy with small-scale data.

Also, the data collection process involves feature extraction, and dimensionality reduction, to make the dataset more comprehensive. The paper is organized into sections that include objectives, a review of various fire predictions using support vector machine (SVM), K-nearest neighbors (KNN) and decision tree, addressing the gaps, research methodology, and discovery.

IV. Relevant Work

The section details on various models developed from previous studies, data mining techniques used in the models and finally addressing the gaps.

Climatic change is portrayed to be one of the reasons for wildfires at tropical regions (Over peck, Rind, & Goldberg, 1990). It is still a debate because fire is a set of complex set of interactions. According to National Oceanic and Atmospheric Administration (NOAA), 32 groups of scientists from around the world investigate 28 individual extreme events in 2014 and broke out various factors that led to the extreme events, including the degree to which natural variability and human-induced climate change played a role. The report added that the overall probability of California wildfires has increased (2,500 acres) due to human-induced climate change (EPA, 2014). Hence, fires not only impact carbon sequestration by forests but emit greenhouse gasses and releases most carbon as CO2, which potentially affect the climate. It has some potential positive feedback since greenhouse-gas-driven climate warming may increase fire activity.

Machine-learning models were frequently used to predict forest fires in different countries and states (Alonso-Betanzos et al., 2003; Bisquert, Caselles, Snchez, & Caselles, 2012; Cheng & Wang, 2008; Dale et al., 2001; De Groot et al., 2013; Flannigan, Stocks, & Wotton, 2000; French et al., 2008; Gavin et al., 2007; Martins Fernandes, 2001; Service & Mountain, 2002). Most of them relied on general models for both large and small database predictions.

After a preliminary review of related work on predictive systems used (on forest fire), regression models such as SVM with other metrics are found to be the most frequently used models (Cortez & Morais, 2007). Similarly, Cortez and Morais (2007) subsequently used a k-fold cross validation on the model with Root Main Square Error (RMSE). The neural network is an alternative model utilized on large data sets (Breiman, 2001). Breiman (2001) also utilized back propagation with controlled layers of data that serve the purpose of predictions. Additionally, the use of data mining techniques was used to extract through sensor networks (Safi & Bouroumi, 2013). Iyer et. al. (2011) utilized Waikato Environment for Knowledge Analysis (WEKA) as an interface to implement decision tree analysis and study the behavior of algorithms conditions.

SVM is an effective classification technique that supports kernel mapping of the data points to a higher dimensional space for small dataset (Cortez & Morais, 2007). SVM could be used with convex optimization method to determine the decision boundary to split dataset (Chang & Lin, 2011). Data mining techniques have been applied to identify the best model for predicting fire occurrence and spread (Cortez & Morais, 2007). The time dependence of the forest area burned in a given year is inherently chaotic, and the predictions become less accurate as time increases (Malarz, Kaczanowska, & Kulakowski, 2002). The features extracted from the predicted class through data mining allows machine learning algorithms to perform the function of data transformation (Iyer et al., 2011).

Viegas et al., (1999) examined five different methods of forest fire prediction and determined that the Canadian and modified Nesterov methods yielded the best overall performance. The K-Nearest Neighbor (KNN) method had also proven to be timely, cost-efficient, and accurate when applied in the Nordic countries and the United States (Finley, Ek, Bai, & Bauer, 2005). KNN is a non-parametric method used in regression analysis and the classification of data. The principle behind KNN is to determine, amongst the training data set, the points closest to the new point and predict the labels (Service & Mountain, 2002). Finley et al. (2005) utilized KNN approach that reduced the duration of the real-time mapping of USDA data set. Also, several other researchers utilized KNN to improve the prediction accuracies from data collected from remote sensors (Franco-Lopez, Ek, & Bauer, 2001; R. E. Mc Roberts, Magnussen, Tomppo, & Chirici, 2011; R. Mc Roberts, Nelson, & Wendt, 2002).

Two of the features of the decision tree are that it neglects the linearity of parameters or is independent of the meteorological, temporal and spatial data. It is not affected by missing values or outliers, as it splits the data on ranges rather than absolute values. It does not require the transformation or scaling of parameters like regression analysis. Also, the decision trees implicitly perform feature selection. Decision tree modeling has its origins in artificial intelligence research where the aim was to produce a system that could identify existing patterns and recognize similar class membership (Ofren & Harvey, 1996). Sensor nodes collect measured data and send to their respective cluster nodes that collaboratively process the data by constructing a neural network (Yu, Wang, & Meng, 2005). This process is expensive when compared to other methods since it involves installation of sensor systems. Service & Mountain (2002) included linear models (LMs), generalized additive models (GAMs), classification and regression trees (CARTs), multivariate adaptive regression splines (MARS), and artificial neural networks (ANNs) to identify which suits better for predicting forest fires. The comparative study concluded that the model's accuracy changes with the real time and assumed datasets.

Though there were different techniques and models developed, the paper compares three different techniques with same datasets for both binary and multiclass classification to determine the accuracy percentage of each technique. The following section in this paper explains the research methods and results obtained from the analysis.

V. Research Methodology

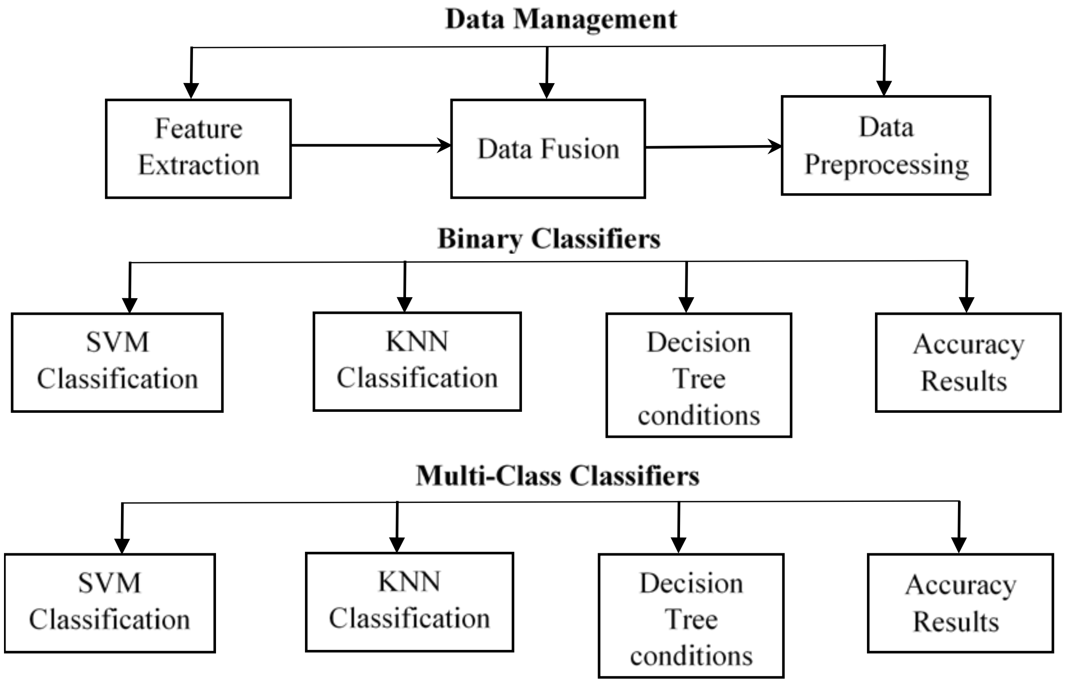
This paper utilizes three different data mining techniques, KNN, SVM, and decision tree models to identify the accuracy of each technique on a small database. The data collected (feature extracted) for this research are from two different reliable sources: 1) the US meteorological department (climate data such as maximum and minimum temperature, humidity, precipitation and snowfall); and 2) the US forest fire database (Burnt area, severity, latitude, longitude). The feature extraction is a prime factor that contributes towards machine learning. The data collected are fused using Python programming language and is cleaned, processed, and integrated into the models.

The primary intention of this paper is to utilize data fusion technique and identify the regions of severity using three different prediction methods.These results are compared with UCI repository data set to prove that the models in this paper perform better. The UCI dataset consists of Fire weather index, which serves as the core parameters towards detection of forest fires. The paper utilizes this information to derive the probability of occurrence of a forest fire and plot a performance curve. While predominantly, most machine-learning problems involve feature extraction as its defining factor, the model is assumed to behave like a black box. This paper aims at modifying the model at its root and fit them according to the dataset and its characteristics.

a) Feature extraction

The primary task of feature extraction is to understand the interpretations of the dataset. The output label needs to be clearly stated that helps in correlating and analyzing the data features. It can be done using the Fisher's information that provides a way of measuring the extent of how much one feature is dependent on another within the dataset. It provides the amount of information a feature has towards the prediction of the output label. The dataset is analyzed for its ability to undergo dimensionality reduction that helps to understand the output visually. The paper tests the hypothesis of predicting forest fires using meteorological data (interchangeably used with Climate Data) and fire data from the Monitoring Trends in Burnt Severity (MTBS) data source.

The algorithm and data extraction are learned at the University of California, Irvine machine learning repository that has data sets of forest fires from Portugal. The 517 samples from the UCI repository contains features from the Fire Weather Index such as FFMC and DMC. These serve as major contributing factors, which are derived from Fisher's information for predicting forest fires.



b) Data Fusion

The feature extracted data need to be fused together with specific date and region for all ten years. It is validated through the online metadata for US climate and MTBS data. In the Geospatial domain, we obtain localized points which on daily cycle records meteorological data. Additionally, the MTBS department also records the occurrence of forest fires. Using 'Beautiful Soup' library, a Python script is written that extracts data over a span of 10 years from 2004-2014. It is then fused with metadata that maps the occurrence of forest fire on a particular day with its respective climate data. It provides features such as Precipitation, Temperature (maximum and minimum), Burnt Area, Latitude, and Longitude of fire occurrence. If there is a date match with an occurrence of a fire, the dataset is integrated with its own forest fire affiliated data. If there is no burnt area, it is marked with a zero. It results in a wide separation between burnt severities and magnifies the confidence of prediction. While both datasets provide a binary label that allows us to predict if a forest fire occurred on a particular date given the meteorological data, the fused data also provides us with the liability to provide for the severity of the fire.

c) Data Preprocessing

Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, redundant information, noisy and unreliable data. While the dataset includes 21,000 samples from five states and seven different features with a small dimensionality, there is a need to look for false positives in the data and has to be omitted. Another python script is written that checks for such anomalies. It occurs because of the dataset during extraction, parses data at (0,0) latitude and longitude when there is no fire data against that date. Thus, it needs to be cleaned up or omitted to analyze in certain models.

Furthermore, this simplifies the search space a level further by consolidating valid samples. The first part is to infer the occurrence of forest fire whereas the second part is to identify the severity of the occurrences using MTBS reference table. It is performed using binary and multi-class classification while the former predicts the occurrence, the latter identifies the severity. The burnt severity is branched into five categories, namely: Very Small, Small, Medium, Large, and Very Large. Subsequently, these modes are separately passed through 3 models used for the classification of the data to derive meaningful results from the output.

d) Binary Classifiers

The process of Binary classification includes training, testing and validating data to determine the occurrence of wildfires from 21000 samples. These classification procedures are implemented in all three models respectively. Initially, a set of data is used to train the machine when the expected output is given to learning the pattern. Later, the data is tested to study the behavior of the machine and finally, the accuracy percentage is determined from each of the techniques by validation.

e) Multiclass classifiers

After training the machine to learn the prediction of burnt area from the sample provided by various features, the process of training and testing repeats with three different models for multi-class classifiers. The training includes severity data initially and then at the testing instance, the models are run to predict the right severity and validated later with real-time data to determine the accuracy percentage.

VI. Model Validation

The section validates three different models and explains the varied approaches used by the authors to improve the accuracy of prediction models. Support vector machines, K-Nearest Neighbors, and Decision tree stumps are trained and tested with modified algorithms to improve the accuracy.

VII. Svm Validation

Support vector machines (SVM) are learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. A set of training samples, each marked as belonging to one of two categories (0 or 1); an SVM training algorithm builds a model and make a not-probabilistic classifier. The samples are mapped so that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mappedinto that same space and predicted to belong to a category based on which side of the decision boundary they fall on in the domain space. The principle behind this model is to maximize the distance between the two classes that are positive and negative classes.

a) Modified Approach

The open source machine-learning library LIBSVM implements the algorithm for kernel SVM. SVM requires data to be represented as a vector of real numbers. The most trivial approach is to define simply the training and testing data and pass it to the SVM model. It provides the desired output regarding the input data. However, this paper aims at modifying the black box SVM model and analyzing it on the fused dataset. The first step was transforming the data into numerical data and then to the format for the LIBSVM package. While choosing a model for the SVM, several parameters are taken into accounts such as the penalty parameter, C, and the kernel parameters. We found that the model worked best when the soft margin constant C was kept at 100. The smaller value of C will tend to ignore the points close to the boundary and causes false results. Kernel parameters also have a significant effect on the SVM model. As our feature set is small, we chose the RBF kernel as it non-linearly maps data into a higher dimensional space and handles non-linear relationships between class labels and features. The degree of the polynomial controls the flexibility of the classifier. We found that the 5- degree polynomial works best as it has a greater curvature. The nu-SVM model sets a lower and upper boundary on the number of data points that lie on the wrong side of the hyperplane and is advantageous for controlling the number of support vectors.

b) Results

The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate. Figure 2 shows the area under the curve for the ROC on the SVM model. The true positive rate resembles the burnt area in the spatial domain, whereas the false positive rate identifies the non-burnt area in the spatial domain.

Table 1 : States and their predicted results using SVM

State	Date	Latitude	Longitude	Burnt Area
Nevada	04-25-2007	36.647	-116.435	330706
Idaho	06-15-2004	44.154	-115.566	9862
Oregon	07-20-2010	38.469	-112.473	42956
New Mexico	04-19-2008	37.623	-78.422	807
California	07-13-2010	36.215	-121.447	934

The above table randomly picks up tuples from each state of the test data and validates it against the MTBS metadata. It checks if the given forest fire occurred. It also crosses checks against its respective meteorological dataset.

Additionally, on analyzing the output as derived from MATLAB provides us with an accuracy of 75.67% using the SVM model with an RBF kernel over the given data set. The Mean square error obtained by implementing a Support Vector Regression model after taking a log(x+1) on the data set gives us 2.3117. It turns out to map onto the burnt area in a given spatial domain given its coordinates.

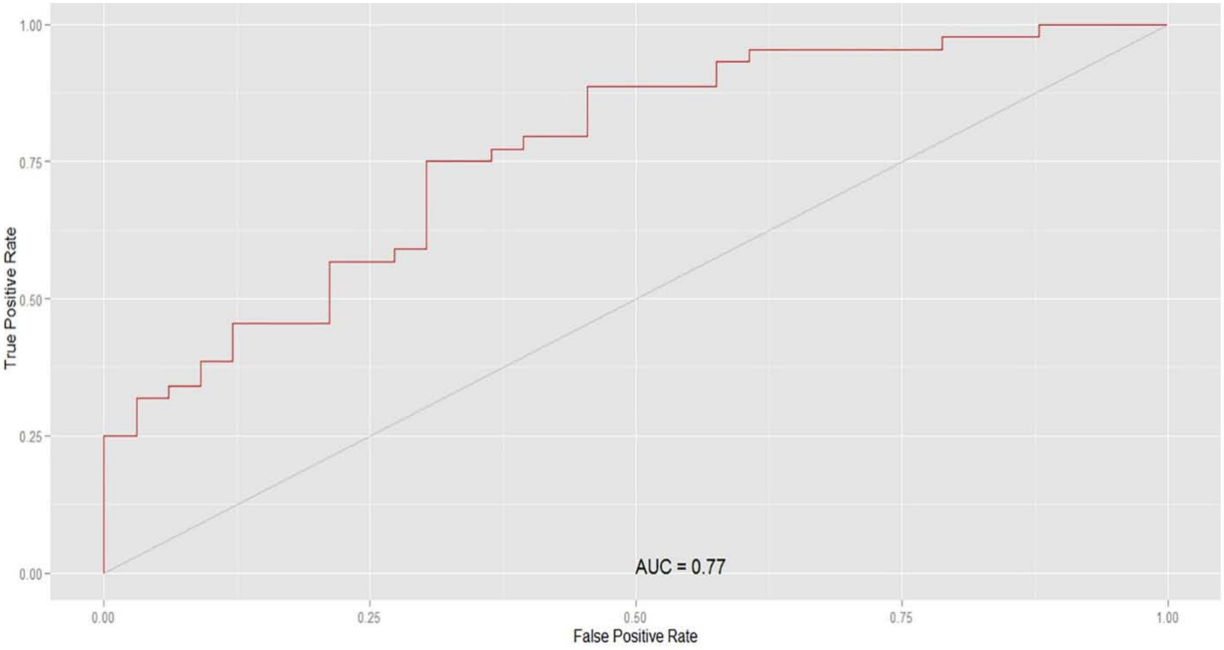


Fig. 2: ROC curve on the SVM model

Thus, the machine is trained with binary classifiers algorithm on an SVM model, and the accuracy is close to 65%. Similarly, the procedure is repeated with one more feature that is the severity type of burnt area, and the multiclass classification algorithms are run on SVM model. The accuracy percentage is around 42 %, which proves to below. It is because the SVM models are used for binary classifiers and not multiclass classification (Chang & Lin, 2011).

VIII. KNN Validation

Initially, a random set of points k is chosen. This k is the same number as neighbors and finds all the points in the training set that are closest. The weighted average of these points then moves k to a new place to balance the centroid in a spatial domain. Figure 3 shows the cells that depict the neighbors.

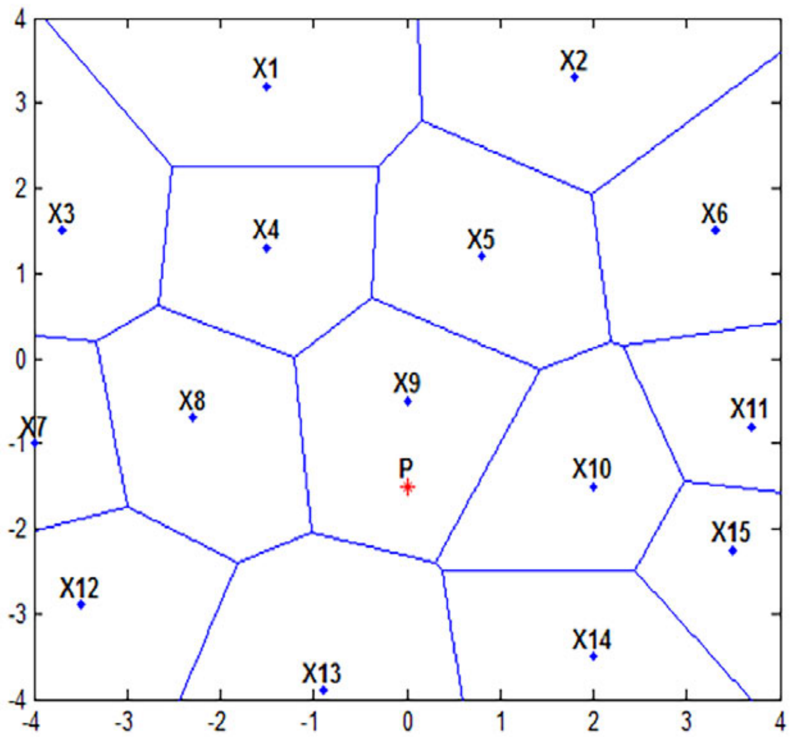


Fig. 3: Depiction of KNN using cell formations

Initially, k = 2, there would be {xj,yj} values where j ≠ size (D) closer to one of the k points. As we add another point to accommodate this phenomenon, the accuracy is accounted by the correctness of predicting the sample point in its respective polygon. The forest fire data occurs close to one another according to the feature space. Additionally, the features are localized to a spatial domain. Thus, if a model needs to predict the occurrence of forest fire based on meteorological data over a constrained area of land, its confidence is magnified if predicted correctly within the neighborhood of the previous occurrence with similar data. KNN does this exactly.

a) Modified Approach

Again manually altering the black box model, the author not only defined the model behavior but also increased the confidence by repeating the experiment several times. Each time the experiment is repeated, the number of neighbors is altered, and the behavior change of the pattern is observed and recorded.

Two different approaches tackle the model. First, the data set is separated into training and testing modules. The MATLAB code then produces an expected error from the training set. It is then matched against its test error or exact error, and the percentage of accuracy is derived using squared Euclidean distance. It is repeated several times to obtain a weighted average to test the validity of the code and the model. To elucidate further on this, we run a KNN model with up to 50 neighbors. With each new neighbor, an expected error is obtained on that models' neighbors' index. The test set is then applied to our trained model. The true error obtained here is compared to the expected error, and its accuracy is validated.

The second approach verifies the trained model and runs the k-fold cross validation on it. By this, the cross validation losses are obtained from each incremented neighbor. The index of which is then matched with the model that provides the least error. It provides us with an expected error per epoch. This, in turn, returns a minimum error of these neighbors. If the error obtained through cross-validation is lower than the expected error, the index at which the KNN flags optimum is incorrect and vice versa. This way the KNN model is used with both binary and multi-class classifiers.

b) Results

The KNN models are trained with UCI data primarily and then trained with the fused dataset. It is done to compare the accuracy and also to make machine optimize the pattern of output required.

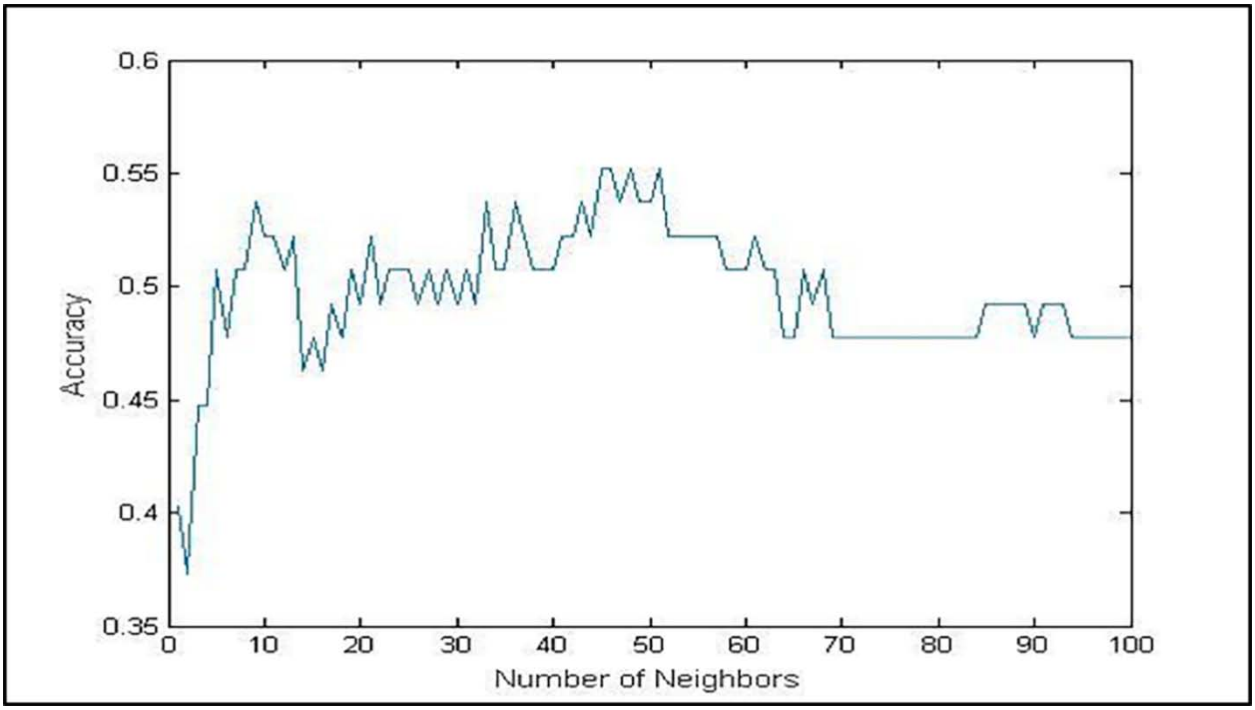


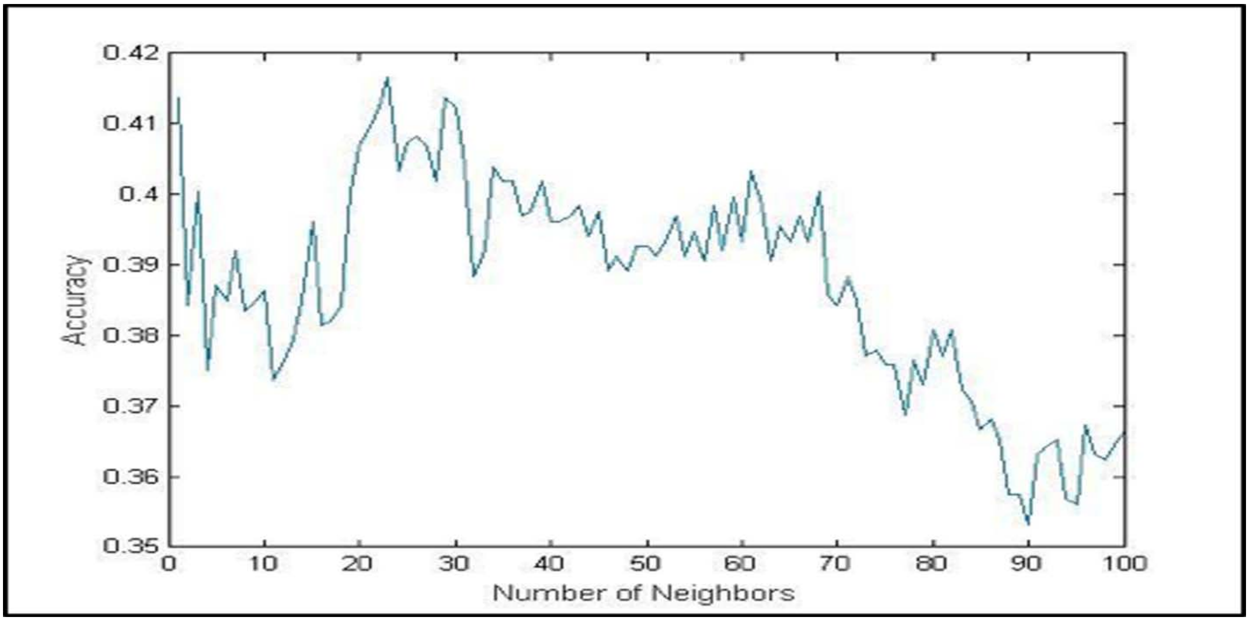
Fig. 4: Accuracy model of UCI dataset

The dimensionality reduction is made using MATLAB to depict visually seven features into two features namely the x-y plot. Similar to the SVM model, the KNN also picks up random tuples from the test data set and validates the error index against its corresponding neighbor. The accuracy is correspondingly determined with the confidence of prediction.

Table 2: States and their predicted results using KNN

State	Date	Latitude	Longitude	Burnt Area
Nevada	04-02-2007	39.014	-116.867	6662
Idaho	06-13-2004	45.153	-114.903	538167
Oregon	01-11-2010	28.903	-82.194	450
New Mexico	07-23-2009	65.625	-143.671	42649
California	10-21-2007	33.181	-116.430	197990

The accuracy percentage for UCI dataset is 53 % for binary and 40% for multiclass whereas the accuracy percentage of the fused dataset is close to 55% in binary and 44% in multi-class.



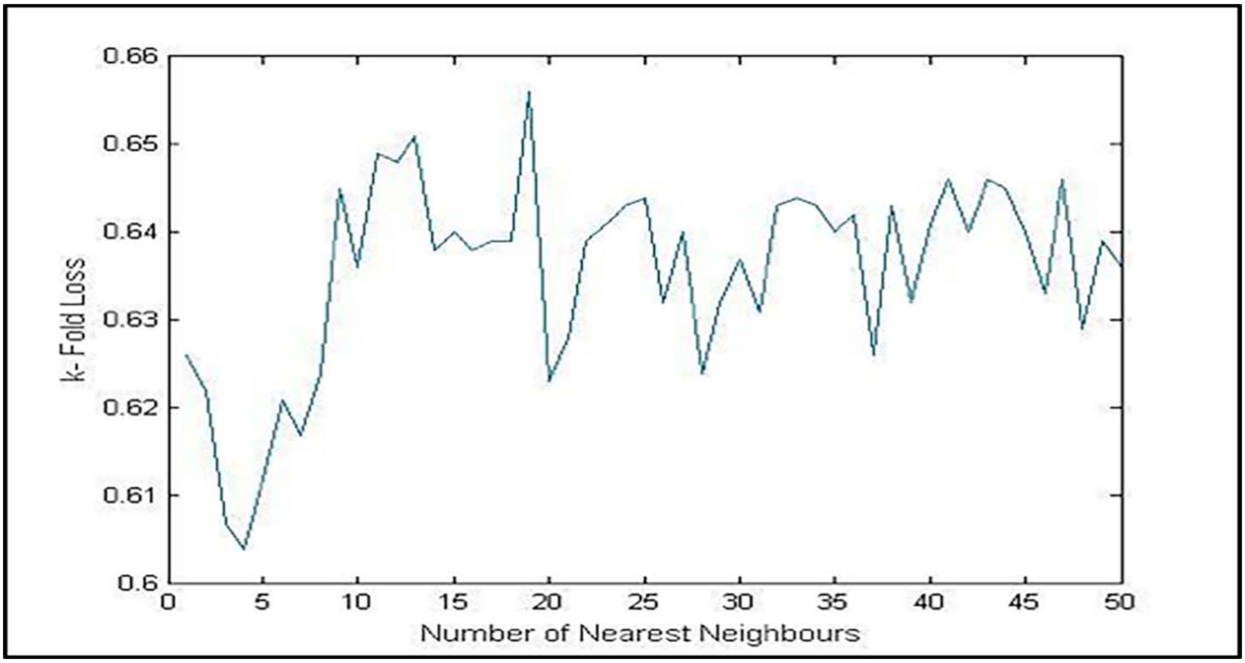


Fig. 5 and 6: Accuracy and Cross-validation model of fused dataset

Thus, KNN model under binary classifiers looks lesser than the SVM model for the small dataset. Figure 4 and 5 shows the accuracy and cross-validation model graphs of UCI and fused dataset.

IX. Decision Tree Validation

After the Nearest Neighbor approach to classification/regression, perhaps the second most intuitive model is Decision Trees. There are many possible trees can be used to organize (i.e., classify) the dataset. It is also feasible to get the same classifier with two very different trees. Tree classification becomes complex with lots of features. A tree that splits the data into all pure leaves is considered consistent with the data. It is always possible when no two samples have different outcomes but identical features. The hierarchy of the architecture leafs out in a manner where every level is a feature. The decision is made on a binary basis. Intuitively, the complexity of the tree increases the variance on the classification boundary.

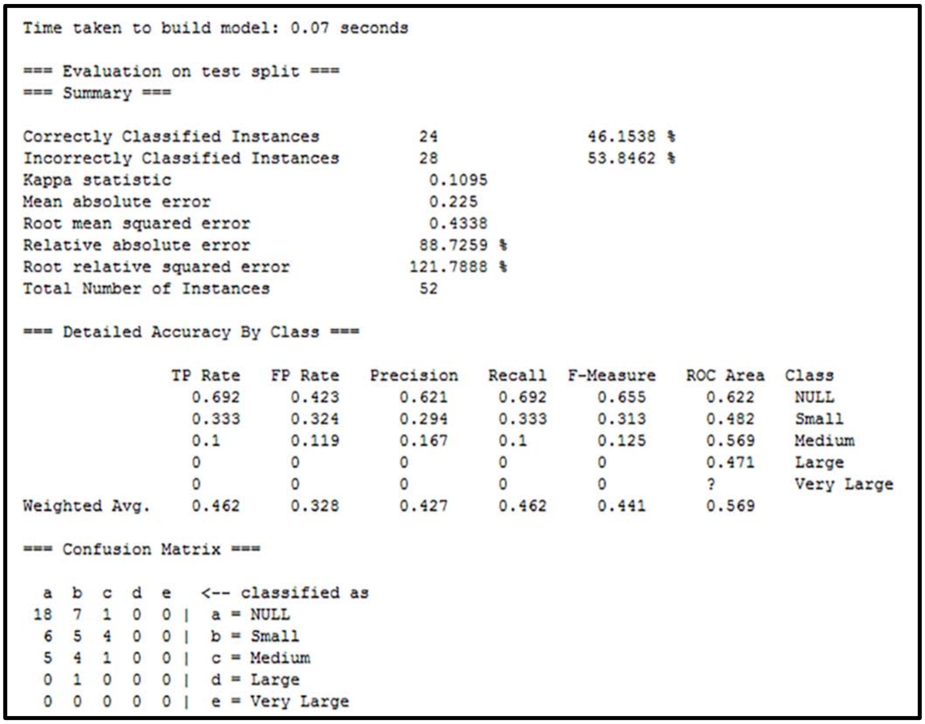
a) Modified Approach

The data is separated into testing and training. Using the C4.5 Decision Tree classifier, WEKA produced results that proved that the fused dataset had more accuracy than the 517 sample set. It can be reasoned merely due to some instances (21421 instances of data) than the 517 dataset. The smaller data set could overfit the model. The other reason is due to our better feature selection of spatial data (latitude) and meteorological data; the output has a higher attribute ranking.

Based on the C4.5 classifier model, the UCI 517 dataset could predict correctly at 46.15 % while the fused dataset could achieve 50%. With reduced error pruning, the rate could be increased roughly by 1%. The classifier is right in predicting the small fires. It achieves good accuracy with Prediction, Recall and ROC area. From the output file, it predicts better based on the features for a lower severity. Particularly, the area under ROC curve outputs the fused dataset at a value of 0.77 in most classes and with a weighted average of 0.636. In contrast, the weighted ROC curve area for UCI dataset is 0.569.

b) Results

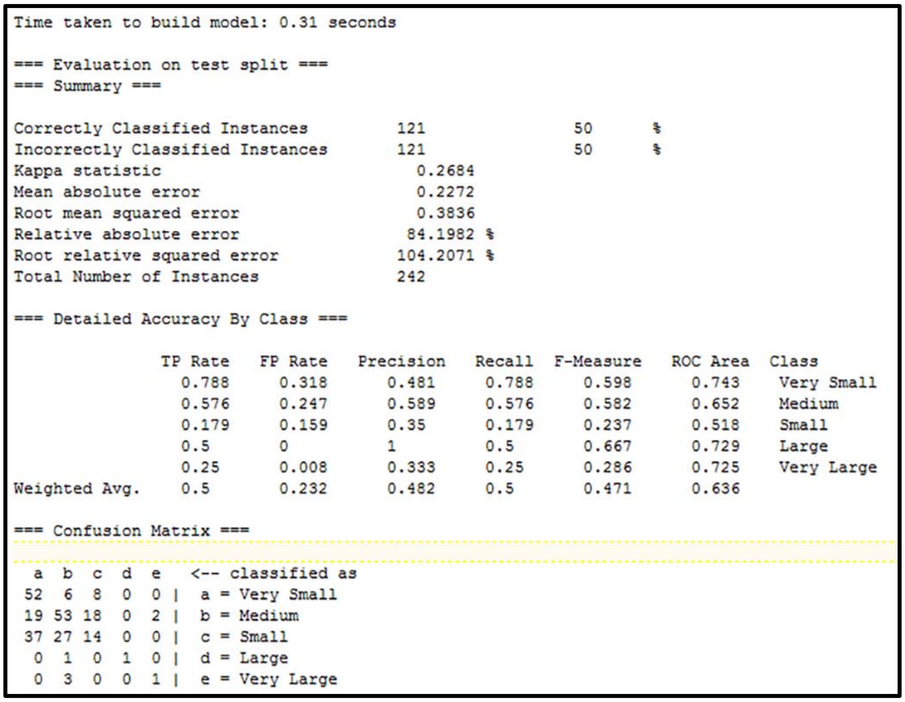
The classifier is developed using WEKA tool that serves best on controlling attributes, enhance visualization and preprocessing data, and availability of a variety of decision tree algorithms. Open-source workbench called WEKA is a useful tool to quantify and validate results, which can be edited and modified. WEKA can handle numeric attributes well, so we use the same values for the weather data from the UCI repository datasets. The class variable has to be a nominal one, to allow WEKA. As WEKA uses kappa stats for evaluating the training sets, a standard score of > 60 % means training set is correlated, using C4.5 simulations. C4.5 is the popular decision tree algorithm, and the WEKA employs the J48 that is an open-source Java implementation of C4.5. The C4.5 or J48 is an improved version of original ID3 that has additional support to handle continuous features in the data and a better bottom-up pruning methodology. The C4.5 automatically handles the pruning (to manage the overfitting) by default.



```
Time taken to build model: 0.07 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          24               46.1538 %
Incorrectly Classified Instances        28               53.8462 %
Kappa statistic                          0.1095
Mean absolute error                      0.225
Root mean squared error                  0.4338
Relative absolute error                 88.7259 %
Root relative squared error            121.7888 %
Total Number of Instances               52

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.692     0.423     0.621      0.692     0.655      0.622    NULL
                 0.333     0.324     0.294      0.333     0.313      0.482    Small
                 0.1       0.119     0.167      0.1       0.125      0.569    Medium
                 0         0         0          0         0          0.471    Large
                 0         0         0          0         0          ?        Very Large
Weighted Avg.    0.462     0.328     0.427      0.462     0.441      0.569

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 18  7  1  0  0 |  a = NULL
  6  5  4  0  0 |  b = Small
  5  4  1  0  0 |  c = Medium
  0  1  0  0  0 |  d = Large
  0  0  0  0  0 |  e = Very Large
```

Fig. 7: Decision Tree output on C 4.5 Algorithm on UCI dataset (Source: WEKA)

The class attribute of the burnt area that needs to be classified under supervised learning is a multiclass attribute that is based on the size of the burnt area.



```
Time taken to build model: 0.31 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         121               50      %
Incorrectly Classified Instances       121               50      %
Kappa statistic                          0.2684
Mean absolute error                      0.2272
Root mean squared error                  0.3836
Relative absolute error                 84.1982 %
Root relative squared error            104.2071 %
Total Number of Instances              242

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.788     0.318     0.481     0.788     0.598      0.743    Very Small
                 0.576     0.247     0.589     0.576     0.582      0.652    Medium
                 0.179     0.159     0.35      0.179     0.237      0.518    Small
                 0.5       0         1         0.5       0.667      0.729    Large
                 0.25      0.008     0.333     0.25      0.286      0.725    Very Large
Weighted Avg.    0.5       0.232     0.482     0.5       0.471      0.636

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 52  6  8  0  0 |  a = Very Small
 19 53 18  0  2 |  b = Medium
 37 27 14  0  0 |  c = Small
  0  1  0  1  0 |  d = Large
  0  3  0  0  1 |  e = Very Large
```

Fig. 8: Decision Tree output on C 4.5 Algorithm on fused dataset (Source: WEKA)

The accuracy percentage from binary classifiers is close to 57 % and percentage from multi-class classifiers is around 42 %. We employed the different algorithms for the Decision trees that could better suit the meteorological, spatial, and temporal data that are continuous.

X. K-Means Clustering

K-means clustering approach failed to deliver any useful results in this paper. The segregated dataset into five different classes to see the clustering based on the states were chosen and their burnt severity type respectively. This model changes its center after every iteration due to the highly localized data. Thus, it is unable to draw a conclusion on a stable centroid that distinctly separates the classes. Figure 8 depicts the clustering of burnt severity of five classes.

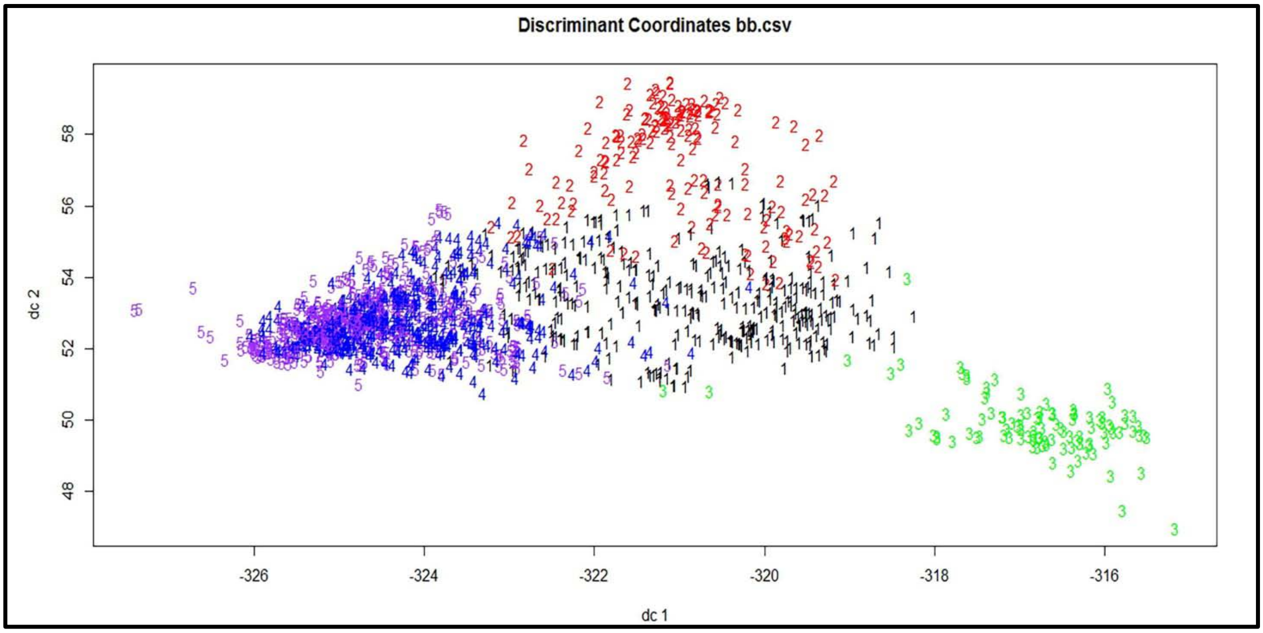


Fig. 9: K-Means Clustering plotted using the Burnt Area Severity

Due to this unlikely occurrence of overlapping data, no classifier can accurately suggest a stable or correct output. Hence, the clustering is omitted for this small-scale dataset.

XI. Conclusion

There are many research on forest fire predictions. There have been very fewer approaches to identify the accuracy of these models for both binary and multi-classifiers. The data fused is used to predict the occurrence by training the machine using latitude, longitude, temperature, humidity, burnt area, burnt area severity, precipitation, and snowfall. The purpose of this paper is to arrive at a model that predicts accurately in a small dataset on both binary classifiers and multi-class classifiers.

The validity of the model will be tested based on supervised learning of structured data. The research is chosen, as there is a need to have different models for different sizes of data. The actual experiment results will tell the suitable method and throw some light on the nature of the problem. Table 3 details on accuracy percentages of both binary and multiclass classifiers of three predictive techniques.

Table 3: Accuracy on various models

Model	Accuracy
SVM	Binary: 65%
	Multiclass: 42%
Decision Tree	Binary: 57%
	Multiclass: 42%
KNN	Binary: 55%
	Multiclass: 44%

From the table 3, it is evident that many parameters come into play while considering models on a small database. With respect to the database, SVM behaves as the optimal model to implement a binary classification and KNN for multiclass classification. The future focus is to improve the algorithms and add satellite images to extract more features and improve the accuracy of machine learning models. The research team also focuses on visualizing data and study of hypothesis over such small dimensionality using Inference and graphical models.

References Références Referencias

1. Alonso-Betanzos, A., Fontenla-Romero, O., Guijarro-Berdiñas, B., Hernández-Pereira, E., Paz Andrade, M. I., Jiménez, E., Carballas, T. (2003). An intelligent system for forest fire risk prediction and firefighting management in Galicia. Expert Systems with Applications, 25(4), 545–554. doi:10.1016/ S0957-4174(03)00095-2
2. Bisquert, M., Caselles, E., Snchez, J. M., &Caselles, V. (2012). Application of artificial neural networks and logistic regression to the prediction of forest fire danger in Galicia using MODIS data. International Journal of Wildland Fire, 21(8), 1025–1029. doi:10. 1071/WF11105
3. Breiman, L. (2001). Random Forrest. Machine Learning, 1–33. doi:10.1023/A:1010933404324
4. Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2, 27:1–27:27. doi:10.1145/1961189.1961199
5. Cheng, T., & Wang, J. (2008). Integrated spatiotemporal data mining for forest fire prediction. Transactions in GIS, 12(5), 591–611. doi:10.1111/j. 1467-9671.2008.01117.x
6. Cortez, P., & Morais, A. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. In New Trends in Artificial Intelligence (pp. 512–523).
7. Dale, V. H., Joyce, L. a, Mcnulty, S., Neilson, R. P., Ayres, M. P., Flannigan, M. D., P, M. (2001). Climate change and forest disturbances. Bioscience, 51(9), 723–734 ST – Climate change and forest disturbances. doi:10.1641/0006-3568(2001)051[07 23:CCAFD]2.0.CO;2
8. De Groot, W. J., Cantin, A. S., Flannigan, M. D., Soja, A. J., Gowman, L. M., & Newbery, A. (2013). A comparison of Canadian and Russian boreal forest fire regimes. Forest Ecology and Management, 294, 23–34. doi: 10.1016/j.foreco.2012.07.033
9. Finley, A., Ek, A., Bai, Y., & Bauer, M. (2005). K-Nearest Neighbor Estimation of Forest Attributes: Improving Mapping Efficiency BT - Proceedings of the fifth annual forest inventory and analysis symposium, 61–68.
10. Flannigan, M. D., Stocks, B. J. J., & Wotton, B. M. M. (2000). Climate Change and Forest Fires. The Science of the Total Environment. doi:10.1016/ S0048-9697(00)00524-6
11. Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbor's method. Remote Sensing of Environment, 77(3), 251–274. doi:10.1016/S0034-4257(01)00209-7
12. French, N., Kasischke, E., Hall, R., Murphy, K., Verbyla, D., Hoy, E., & Allen, J. (2008). Using

Landsat data to assess fire and burn severity in the North American boreal forest region: an overview and summary of results. International Journal of Wildland Fire, 17(4), 443–462. doi:10.1071/WF0800

13. Gavin, D. G., Hallett, D. J., Feng, S. H., Lertzman, K. P., Prichard, S. J., Brown, K. J., Peterson, D. L. (2007). Forest fire and climate change in western North America: Insights from sediment charcoal records. Frontiers in Ecology and the Environment. doi:10.1890/1540-9295(2007)5[499:FFACCI]2.0. CO;2
14. Iyer, V., T, S. S. I., Paramesh, N., Murthy, G. R., & U, M. B. S. (2011). Machine Learning and Dataming Algorithms for Predicting Accidental Small Forest Fires. Weather, (c), 116–121.
15. Malarz, K., Kaczanowska, S., & Kulakowski, K. (2002). Are Forest Fires Predictable? 13. doi:10. 1142/S0129183102003760
16. Martins Fernandes, P. A. (2001). Fire spread prediction in shrub fuels in Portugal. Forest Ecology and Management, 144(1-3), 67–74. doi:10.1016/ S0378-1127(00)00363-7
17. McRoberts, R. E. (2008). Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. Remote Sensing of Environment, 112(5), 2212–2221. doi:10.1016/j.rse.2007.07.025
18. McRoberts, R. E., Magnussen, S., Tomppo, E. O., & Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sensing of Environment, 115(12), 3165–3174. doi:10.1016/ j.rse.2011.07.002
19. McRoberts, R. E., Tomppo, E. O., Finley, A. O., & Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sensing of Environment, 111(4), 466–480. doi: 10. 1016/j.rse.2007.04.002
20. McRoberts, R., Nelson, M., & Wendt, D. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sensing of Environment, 82, 457–468. doi:10.1016/S0034-4257(02)00064-0
21. Ofren, R. S., & Harvey, E. (1996). A Multivariate Decision Tree Analysis of Biophysical Factors in Tropical Forest Fire Occurence. Integrated Tools Proceedings, 221–227.
22. Park, N., National, S., & Fire, I. (2003). Fire Monitoring Handbook. Program, 285. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search & q=intitle:Fire+Monitoring+Handbook #1
23. Safi, Y., & Bouroumi, A. (2013). Prediction of Forest Fires Using Artificial Neural Networks Description of the proposed method Artificial neural networks, 7(6), 271–286.
24. Schoennagel, T., & Nelson, C. R. (2011). Restoration relevance of recent National Fire Plan treatments in forests of the western United States. Frontiers in Ecology and the Environment, 9(5), 271–277. doi:10.1890/090199
25. Schoennagel, T., Nelson, C. R., Theobald, D. M., Carnwath, G. C., & Chapman, T. B. (2009). Implementation of National Fire Plan treatments near the wildland–urban interface in the western United States. Proceedings of the National Academy of Sciences of the United States of America, doi 10.1073. doi:10.1073/pnas.090099110
26. Schoennagel, T., Nelson, C. R., Theobald, D. M., Carnwath, G. C., & Chapman, T. B. (2009). Implementation of National Fire Plan treatments near the wildland-urban interface in the western United States. Proceedings of the National Academy of Sciences of the United States of America, 106(26), 10706–10711. doi:10.1073/pnas. 0900991106
27. Service, U. S. F., & Mountain, R. (2002). Comparing Five Modelling Techniques. Ecological Modelling.
28. Viegas, D. X., Bovio, G., Ferreira, A., Nosenzo, A., & Sol, B. (1999). Comparative study of various methods of fire danger evaluation in southern Europe. International Journal of Wildland Fire, 9(4), 235. doi:10.1071/WF00015

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

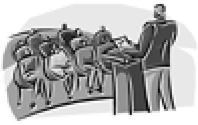
Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards “FARSC” title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

> The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.



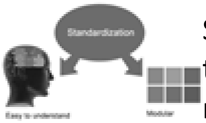
Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSC can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSC, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.



The FARSC will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSC member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A. Once you are designated as FARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSC members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSC member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSC is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSC can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSC member can decide its price and we can help in making the right decision.

The FARSC member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

The ' MARSC ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.
The “MARSC” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSC or William Walldroff, M.S., MARSC.

MARSC accrediting is an honor. It authenticates your research activities. After becoming MARSC, you can add 'MARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSC designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSC, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Auxiliary Memberships

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

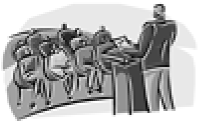
Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.
The author fees of such paper may be waived off up to 40%.



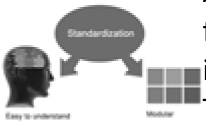
The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.



After nomination of your institution as “Institutional Fellow” and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

> "
>
> - In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
> - In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
> - In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone. "

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, “Abstract” word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript’s Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications.

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350