Dhinaharan Nagamalai
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

3rd International Conference on Artificial Intelligence and Applications (AI-2017) December 30~31, 2017, Chennai, India

AIRCC Publishing Corporation

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-78-6
DOI : 10.5121/csit.2017.71801 - 10.5121/csit.2017.71805

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 3rd International Conference on Artificial Intelligence and Applications (AI-2017) was held in Chennai, India, during December 30~31, 2017. The 3rd International Conference on Computer Science and Information Technology (CSTY-2017) and The 3rd International Conference on Signal and Image Processing (SIGI-2017) was collocated with The 3rd International Conference on Artificial Intelligence and Applications (AI-2017). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The AI-2017, CSTY-2017, SIGI-2017 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, AI-2017, CSTY-2017, SIGI-2017 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the AI-2017, CSTY-2017, SIGI-2017.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Natarajan Meghanathan

Organization

General Chair

David C. Wyld	Southeastern Louisisna University, USA
Jan Zizka	Mendel University in Brno, Czech Republic

Program Committee Members

Ahmad Rawashdeh	University of Central Missouri, United States
Ahmed Korichi	University of Ouargla, Algeria
AL-Shatnawi	Al al-Byte University, Jordan
Atallah M	Al al-Byte University, Jordan
Ayush Singhal	Contata Solutions, USA
Azeddine Chikh	University of Tlemcen, Algeria
Barhoumi Walid	SIIVA-LIMTIC Laboratory, ENICarthage, Tunisia
Carlo Sau	Universit degli Studi di Cagliari, Italy
Chaker LARABI	Universit de Poitiers , France
Chin-Chen Chang	Feng Chia University, Taiwan
Chuanzong Zhang	Aalborg University, Denmark
Claudio Gallicchio	University of Pisa, Italy.
Dabin Ding	University of Central Missouri, United States
Dac-Nhuong Le	Haiphong University, Vietnam
Dalel BOUSLIMI	Institut Mines- Telecom, France
Dongping Tian	Baoji University of Arts and Sciences, China
Duan Keqing	Wuhan Early Warning Academy, China
Elaheh Pourabbas	National Research Council, Italy
Emad Awada	Applied Science University, Jordan
Emad Eldin Mohamed	Canadian University Dubai, UAE
Fabio Gasparetti	Roma Tre University, Italy
Farida Bouarab-Dahmani	Mouloud Mammeri University of Tizi-Ouzou, Algeria.
Farzin Piltan	University of Ulsan, Korea.
Fatma Outay	Zayed University DXB, UAE
Fernando Zacarias Flores	Universidad Autonoma de Puebla, Mexico
Gammoudi Aymen	University of Tunis, Tunisia
Hacer Yalim Keles	Ankara University, Turkey
Hamid Alasadi	Basra University, Iraq
Hamzeh Khalili	Universitat Politecnica de Catalunya (UPC), Spain
Hanming Fang	Logistical Engineering University, China
Hari Krishna Garg	National University of Singapore, Singapore
Hassan Ugail	University of Bradford, UK
Hayet Mouss	Batna Univeristy, Algeria
Issac Niwas Swamidoss	Nanyang Technological University, Singapore
Jamal El Abbadi	Mohammadia V University Rabat, Morocco
Jiting XU	ebay, USA
John Tass	University of Patras, Greece

Jun Zhang	South China University of Technology, China
Jyoti Ohri	National Institute of Technology, India.
Klimis Ntalianis	Athens University of Applied Sciences, Greece
Kulwinder Singh Parmar	Punjab Technical University, India
Lark Kwon Choi	The University of Texas at Austin, USA
Lei Zhang	University of Surrey, UK
Mahdi Salarian	University of Illinois, USA
Mahmood Ali Mirza	DMS SVH College of Engineering, India
Manik Sharma	DAV University, India
Mike Turi	California State University-Fullerton, USA
Mohamad Badra	Zayed University, Dubai, UAE
Mohamedmaher Benismail	King saud University, Saudi Arabia
Mohammad alsarem	Taibah University, KSA
Mohammad Masdari	Islamic Azad University, Iran
Mohammad Rawashdeh	University of Central Missouri, United States
Mohammad Siraj	King Saud University, Saudi Arabia
Mostafa Ashry	Alexandria university, Egypt
mourchid mohammed Ibn	Tofail University Kenitra, Morocco
Necmettin	Erbakan University, Turkey
Neda Firoz	Ewing Christian College, India
Noura Taleb	Badji Mokhtar University, Algeria
Oleksii K.Tyshchenko	Kharkiv National University of Radio Electronics, Ukraine.
Ouafa Mah	Ouargla university, Algeria
Paulo Roberto Martins de Andrade	University of Regina, Canada
Prakash Duraisamy	University of Central Missouri, United States
Prantosh kumar Paul	Raiganj University, India
Prateek Agrawal	Lovely Professional University, India
Razieh malekhoseini	Islamic Azad University, Iran
Samy S. Abu Naser	Al-Azhar University, Palestine
Santosh Kumar Nanda	Eastern Academy of Science and Technology, India.
Shoeib Faraj	Institute of Higher Education of Miaad, Iran
Sitanath Biswas	Gandhi Institute for Technology, India
Taeghyun Kang	University of Central Missouri, United States
Temur Z. Kalanov	Institute of Electronics, Uzbekistan.
Wonjun Lee	The University of Texas at San Antonio, USA
Xuechao Li	Auburn University, USA
Zhao Peng	Huazhong University of Science and Technology, China

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

3rd International Conference on Artificial Intelligence and Applications (AI-2017)

3rd International Conference on Computer Science and Information Technology (CSTY-2017)

3rd International Conference on Signal and Image Processing (SIGI-2017)

A COMPARATIVE STUDY FOR ICA MULTIUNIT ALGORITHMS

Doru CONSTANTIN[1], Emilia CLIPICI[2] and Alina-Florentina ȘTEFAN[3]

[1,3]Department of Mathematics-Informatics, University of Pitesti,
Street Targu din Vale, No.1, Pitesti, Romania
[2]Department of Finance, Accounting and Economics, University of Pitesti,
Street Targu din Vale, No.1, Pitesti, Romania

ABSTRACT

We present the comparative study of convergence for multiunit algorithms based on negentropy function for estimating the independent components.

KEYWORDS

Independent Component Analysis (ICA),Blind Source Separation (BSS), Signal Processing, Negentropy function

1. INTRODUCTION

A fundamental problem in neural network research, as well as in many other disciplines, is finding a suitable representation of multivariate data, random vectors. For reasons of computational and conceptual simplicity, the representation is sought as a linear transformation of the original data. In other words, each component of the representation is a linear combination of the original variables. Well known linear transformation methods include principal component analysis, factor analysis, and projection pursuit. Independent component analysis is a recently developed method in which the goal is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible [9,7]. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation.

2. NEGENTROPY FUNCTION FOR ONE-UNIT ALGORITHMS

The negentropy function is a measure of the nongaussianity and is defined based on the entropy function. The entropy function H of a random vector y with density function $p_y(\eta)$ have the expression:

$$H(y) = -\int p_y(\eta) \log p_y(\eta) \tag{1}$$

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 01– 09, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71801

A fundamental result of information theory is that a gaussian variable has the largest entropy among all random variables of equal variance [3,7]. This means that entropy could be used as a measure of nongaussianity.

To obtain a measure of nongaussianity that is zero for a gaussian variable and always nonnegative, one often uses a normalized version of differential entropy, called negentropy. Negentropy J is defined as follows:

$$J(y) = H(y_{gauss}) - H(y) \tag{2}$$

where y_{gauss} is a gaussian random variable of the same correlation (and covariance) matrix as y.

Negentropy approximations

There are some approximations of the negentropy function used in practical applications. The classic method of approximating negentropy is using higher-order cumulants:

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 \tag{3}$$

where y is assumed to be of zero mean and unit variance.

Another approximation is based on two nonquadratic functions G^1 and G^2 so that G^1 is odd and G^2 is even, and we obtain:

$$J(y) \approx k_1(E\{G^1(y)\})^2 + k_2(E\{G^2(y)\} - E\{G^2(\nu)\})^2), \tag{4}$$

where k_1 and k_2 are positive constants, ν is a gaussian variable of zero mean and unit variance and y is assumed to have zero mean and unit variance [6,7,9].

In the case where we use only one nonquadratic function G, the approximation becomes:

$$J(y) \approx [E\{G(y)\} - E\{G(\nu)\}]^2 \tag{5}$$

The gradient algorithm

Taking the gradient of the approximation of negentropy in (5)with respect to w and taking the normalization $E\{(w^T z)^2\} = \quad w^2 = 1$ we obtain:

$$\Delta w \propto \gamma E\{zg(w^T z)\} \tag{6}$$

$$w \leftarrow \frac{w}{w} \tag{7}$$

where $\gamma = E\{G(w^T z)\} - E\{G(\nu)\}$ and ν being a standardized gaussian random variable. For function g we may use:

$$g_1(y) = \tanh(a_1 y) \tag{8}$$

$$g_2(y) = y\exp(-\frac{y^2}{2}) \tag{9}$$

$$g_3(y) = y^3 \tag{10}$$

where $1 \le a_1 \le 2$ is a constant.

The algorithm for one independent component estimation

1. Data centering (make its mean zero).
2. Data preprocessing (whitening data) and obtain z.
3. Choose an initial value for w of unit norm and an initial value for γ.
4. Update scheme by

$$\Delta w \propto \gamma z g(w^T z),$$

where the function g is defined in (8), (9), (10).

5. Normalize the vector w by:

$$w \leftarrow \frac{w}{w}.$$

6. If the sign of γ is not known a priori, update

$$\Delta\gamma \propto (G(w^T z) - E\{G(v)\}) - \gamma.$$

7. If the algorithm not converged, go back to Step 4.

The fixed-point algorithm for ICA model estimation

From the gradient method in (6) we may establish the following fixed-point iteration:

$$w \leftarrow E\{zg(w^T z)\} \tag{11}$$

After rewriting the (11) relation we have:

$$w = E\{zg(w^T z)\} \Leftrightarrow (1+\alpha)w = E\{zg(w^T z)\} + \alpha w \tag{12}$$

According to the Lagrange conditions $E\{G(w^T z)\}$ under the constraint $E\{w^T z\} = \quad w^2 = 1$ are obtained at points where the gradient of the Lagrangian is zero:

$$E\{zg(w^T z)\} + \beta w = 0 \tag{13}$$

Now let us try to solve this equation by Newton's method, which is equivalent to finding the optima of the Lagrangian by Newton's method. Denoting the function on the left-hand side of (13) with F, we obtain its gradient:

$$\frac{\partial F}{\partial w} = E\{zz^T g'(w^T z)\} + \beta I \tag{14}$$

Apply a reasonable approximation:

$E\{zz^T g'(w^T z)\} \approx E\{zz^T\}E\{g'(w^T z)\} = E\{g'(w^T z)\}I$. Thus we obtain the following approximative Newton iteration:

$$w \leftarrow w - \frac{E\{zg(w^T z)\} + \beta w}{E\{g'(w^T z)\} + \beta} \tag{15}$$

This algorithm can be further simplified by multiplying both sides of (16) with $\beta + E\{g'(w^T z)\}$. This gives the following form:

$$w \leftarrow E\{zg(w^T z)\} - E\{g'(w^T z)\}w \tag{16}$$

This is the basic fixed-point iteration in FastICA.

The FastICA algorithm for estimating one independent component

1. Data centering.
2. Data preprocessing and obtain z.
3. Choose an initial value for vector w of unit norm.
4. Apply the updating rule:
$$w \leftarrow E\{zg(w^T z)\} - E\{g'(w^T z)\}w,$$
where function g is defined in (8), (9), (10).
5. Normalize the vector w:
$$w \leftarrow \frac{w}{w}.$$
6. If the algorithm not converge, come back to 4.

3. MULTI-UNIT ALGORITHMS FOR ICA MODEL ESTIMATIN

It is possible to find more independent components by running an one-unit algorithm many times and using different initial points but with the property like the vectors w_i corresponding to different independent components are orthogonal in the whitened space [6,7,9,13].

3.1. The IC's estimation by deflationary orthogonalization

For deflationary orthogonalization is using the GramSchmidt method. This means that we estimate the independent components one by one and alternate the following steps:

1. Set the desired number of ICs to estimate m and initialization $p = 1$.
2. Initialize w_p.
3. Do an iteration of a one-unit algorithm and obtain w_p.
4. Do orthogonalization transformation:

$$w_p \leftarrow w_p - \sum_{j=1}^{p-1} (w_p^T w_j) w_j \qquad (17)$$

5. Normalize the vector w_p:

$$w \leftarrow \frac{w}{w}.$$

6. if w_p has not converged back to step 3.
7. Set $p \leftarrow p+1$. If p is not greater than m back to step 2.

3.2. The IC's estimation by symmetric orthogonalization

In this case the vectors w_i are estimated in parallel, not estimated one by one. Thus the symmetric orthogonalization methods enable parallel computation of ICs. The general form of this algorithm is:

1. Set the desired number of ICs to estimate m.
2. Initialize $w_i, i = 1,...,m$.
3. Do an iteration of a one-unit algorithm on every w_i in parallel scheme.
4. Do a symmetric orthogonalization of the matrix $W = (w_1,...,w_m)^T$.
5. If w_p not converged back to step 3.

The symmetric orthogonalization of W can be accomplished by:

$$W \leftarrow (WW^T)^{-1/2} W \qquad (18)$$

The inverse square root $(WW^T)^{-1/2}$ is obtained from the eigenvalue decomposition of $WW^T = Ediag(d_1,...,d_m)E^T$:

$$(WW^T)^{-1/2} = Ediag(d_1^{-1/2},...,d_m^{-1/2})E^T \qquad (19)$$

A simpler alternative is the following iterative algorithm:

1. Calculate $W \leftarrow W/\ \ W$.
2. Calculate $W \leftarrow 3/2W - 1/2WW^TW$.
3. If the matrix WW^T is not close enough to identity matrix then go to step 2.

4. EXPERIMENTAL RESULTS FOR CONVERGENCE OF THE MULTI-UNIT ALGORITHMS

By using the FastICA algorithm we can determine the components independent and was considered the estimate of the independent components problem of a mixture of signals. The original signals are obtained from the mixing matrix signals. For estimate de ICA model we have two multi-unit algorithms: the algorithm based on the deflationary orthogonalization and the algorithm based on the symmetric orthogonalization. In the experimentally applications we choose the following nonlinear functions for function g used in the algorithms:

1. default function $g(u) = u^3$.
2. function tanh $g(u) = tanh(u)$.
3. function gauss $g(u) = u * exp(-u^2/2)$.
4. function $g(u) = u^2$.

To compare convergence for the two types of approaches, by deflating and symmetrically transformation, using the four functions mentioned above, was considered for example the following mixing matrix form:

$$A = \begin{pmatrix} 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 \end{pmatrix} \qquad (20)$$

The application establish the seven independent components approximation of the original signals and the convergence is shown in the next table by average of the iterations number:

Table 1. The mean number of steps for convergence.

No. item	Function	Symmetric	Deflationary
1.	$g(u) = u^3$	83 steps	12-8-8-5-5-5-2
2.	$g(u) = \tanh(u)$	18 steps	16-14-14-10-5-4-2
3.	$g(u) = u * \exp(-u^2/2)$	16 steps	12-8-16-21-17- -
4.	$g(u) = u^2$	17 steps	14-13-16-26- - -

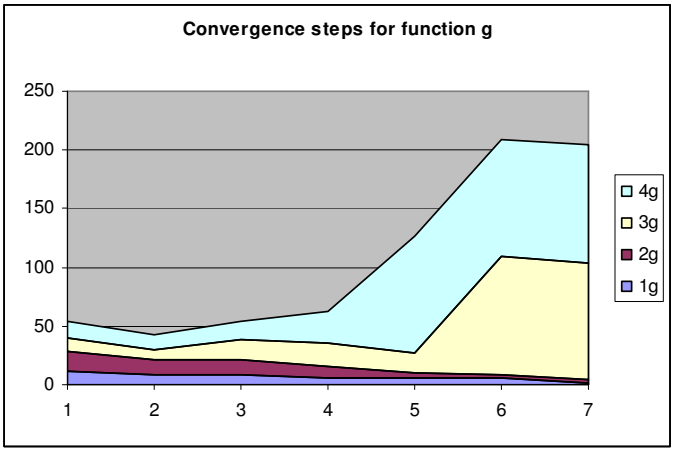


Figure 1. Convergence results for versions of function g

From above table that presents the number of steps of convergence multi-unit algorithms with symmetric and deflationary orthogonalization note that for the algorithm based on the symmetric orthogonalization the function of type 3, 4 and 1 produce a suitable results of convergence expressed through number of steps, and for the algorithm based on the deflationary orthogonalization the function of type 1 and 2 produce a suitable results of convergence. In case of IC's estimation by deflationary orthogonalization algorithm we note a high complexity to estimate the last two or three independent components for $g(u) = u * \exp(-u^2 / 2)$ and $g(u) = u^2$.

5. CONCLUSIONS

For estimating the independent components was used the negentropy function like a contrast function. By using the negentropy we may derive the updating rule for ICA estimation and obtain the general gradient one-unit algorithm, the fastica algorithm and the multi-unit algorithms based on the symmetric and deflationary orthogonalization. For the multi-unit algorithms based on the negentropy function and the symmetric and deflationary orthogonalization were established the experimental results that illustrating the performance of original signals recognition in terms of convergence.

REFERENCES

[1] C.M. Bishop, Neural Network for Pattern Recognition, Clarendon Press, 1995.

[2] A. Cichocki, R. Unbehauen, Neural Networks for Signal Processing and Optimization, Wiley, 1994.

[3] D., Constantin, L., State, A Comparative Analysis on a Class of Numerical Methods for Estimating the ICA Model, Proc. International Conference on Computers, Communications & Control, 2008.

[4] D., Constantin, L., State, A Version of the FastICA Algorithm Based on the Secant Method combined with Simple Iterations Method, ICISP 2008, Image and Signal Processing, LNCS, Springer, 2008.

[5] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 1991.

[6] R. Gonzalez, P. Wintz, Digital Image Processing, Addison-Wisley, 1987.

[7] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley-Sons, 2001.

[8] T.W. Lee, Independent Component Analysis - Theory and Applications, Kluwer, 1998.

[9] J.V. Stone, Independent Component Analysis: A Tutorial Introduction, Mit Press, 2004.

[10] K.I. Diamantaras, S.Y. Kung, Principal Component Neural Networks: Theory and Applications, Wiley, 1996.

[11] Daniela Dănciulescu, Mihaela Colhon, Gheorghe Grigoraş, A System of Knowledge Representation for Right Linear Grammars Generation, Broad Research in Artificial Intelligence and Neuroscience, vol 8(1), ISSN 2068-0473, E-ISSN 2067-3957, pp. 42-51, 2017.

[12] Cosmin Sabo, Petrică C. Pop, Honoriu Vălean, Daniela Dănciulescu, An Innovative Approach to Manage Heterogeneous Information Using Relational Database Systems, Proceedings of International Conference on Intelligent Systems Design and Applications, ISDA 2016, pp 1-10, 2016.

[13] Viorel Negru, Gheorghe Grigoraş, Daniela Dănciulescu, Natural Language Agreement in the Generation Mechanism based on Stratified Graphs, Proceedings of the 7th Balkan Conference in Informatics (BCI 2015), Craiova, Romania, 2015.

[14] Daniela Dănciulescu, Formal Languages Generation in Systems of Knowledge Representation based on Stratified Graphs, INFORMATICA 2015, vol. 26, no. 3, pp. 407-417, ISSN 0868-4952, 2015.

[15] Daniela Dănciulescu, Mihaela Colhon, Systems of knowledge representation based on stratified graphs. Application to Natural Language Generation, Carpathian Journal of Mathematics, 32(1), pp. 49-62, 2014.

[16] Daniela Dănciulescu, Mihaela Colhon, Splitting the structured paths in stratified graphs. Application in Natural Language Generation, Analele ştiinţifice ale Universităţii Ovidius Constanţa, Seria Matematică, vol. 22, no. 2, pp. 59-69, ISSN: 1224-1784, 2014.

[17] Daniela Dănciulescu, Nicolae Tăndăreanu, Splitting the structured paths in stratified graphs, Modeling and Development of Intelligent Systems, Proceedings of the 3th Int Conference on Modeling and Development of Intelligent Systems, ISSN 2067-3965, 2014.

[18] Daniela Dănciulescu, Systems Of Knowledge Representation Based On Stratified Graphs And Their Inference Process, 9th International Conference of Applied Mathematics, 2013.

[19] Nicolae Tăndăreanu, Irina Tudor (Preda), Daniela Dănciulescu, Applications of stratified graphs in optimal planning, Proceedings of the 12th Conference on Artficial Intelligence and Digital Communications (AIDC), pp.7-24, 2012.

[20] W. Gardner, Introduction to Random Processes with Applications to Signal and Systems, Macmillan, 1986.

[21] M. Girolami, Self-Organising Neural Networks - Independent Component Analysis and Blind Source Separation, Springer Verlag, 1999.

[22] D.A., Popescu, N., Bold, O., Domsa, Generating assessment tests with restrictions using genetic algorithms, 12th IEEE International Conf. on Control & Automation (IEEE ICCA 2016), 1-3 June, (2016) Kathmandu, Nepall.

[23] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag New York, 2002.

[24] Oppenheim, R. Schafer, Discrete-Time Signal Processing, Prentince Hall, 1989.

[25] S. Russell, P. Norvig, Artificial Intelligence. A Modern Approach, Prentice Hall, 1995.

INTENTIONAL BLANK

CONVOLUTIONAL NEURAL NETWORK APPLIED TO THE IDENTIFICATION OF RESIDENTIAL EQUIPMENT IN NON-INTRUSIVE LOAD MONITORING SYSTEMS

Deyvison de Paiva Penha and Adriana Rosa Garcez Castro

Institute of Technology, Federal University of Para, Belém, Brazil

ABSTRACT

This paper presents the proposal of A new methodology for the identification of residential equipment in non-intrusive load monitoring systems that is based on a Convolutional Neural Network to classify equipment. The transient power signal data obtained at the time an equipment is connected in a residence is used as inputs to the system. The methodology was developed using data from a public database (REED) that presents data collected at a low frequency (1 Hz). The results obtained in the test database indicate that the proposed system is able to carry out the identification task, and presented satisfactory results when compared with the results already presented in the literature for the problem in question.

KEYWORDS

Convolutional Neural Networks, Identification of Residential Equipment, Non-Intrusive Load Monitoring, NILM System, Energy Conservation

1. INTRODUCTION

The reduction and rationalization of electricity consumption are increasingly becoming priorities, not only for residential consumers, but also for electric power companies and government. Considering this concern, which is worldwide, research in Non-Intrusive Load Monitoring (NILM) has been emphasizing. Research in this area began in 1992 with the presentation of the work of George W. Hart [1] and since then many works have been presented, focusing on the various stages of a NILM system.

A NILM system has as main objective to measure an aggregate load of a residence through a single sensor, placed in the central meter of the residence. From the aggregate load, measured over a period of time, it is possible, through specific software, to carry out an identification of the electric equipment in operation and obtain the individual consumption thereof, in addition to obtaining the operating hours of each equipment [1]. This information can be used by residential consumers to take actions aimed at reducing and rationalizing their consumption, thus ensuring greater energy efficiency. In addition to this main functionality of the NILM systems, it is also possible to highlight: the possibility of identifying the load profile of a residence; possibility of identifying non-standard behavior of loads; possibility of detection of power failures and thefts; possibility of the use of the information of the load disaggregated by the electric power concessionaires that can promote aid to their customers in the process of identification of waste

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 11– 21, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71802

during peak hours, thus helping to reduce consumption during these periods, offering for this incentive to consumers [2].

Considering the good results already presented by the academic community involving deep neural networks for the NILM problem, this paper presents the results obtained from the application of Convolutional Neural Networks for the problem of equipment identification. Here, unlike what we already have in the literature, a CNN network was developed to identify the type of equipment from the transient power signal data obtained at the moment an equipment is connected. The choice of the use of the transient power signal is due to the fact that each type of equipment presents different transient signal characteristics, depending on the generation mechanism, which is suitable for the development of classification systems. For the development and testing of the proposed system, the public database was used, and much used by researchers in the area, REDD (Reference Energy Disaggregation Dataset) [3]. This database has data of several equipments that were collected individually in 6 different residences at a frequency of 1 Hz. The system was developed to identify 7 equipments, these being classified as on / off loads, multilevel or variable.

2. NON-INTRUSIVE LOAD MONITORING SYSTEMS

The non-intrusive load monitoring aims to obtain a good approximation of the various electric devices in operation in a residence, using dedicated hardware and software [4]. The monitoring and identification of loads are performed based on the analysis of measurements of a single point of current and voltage of the aggregate load obtained through a meter outside the residence. Since each electrical equipment has its own profile of energy consumption called the electric signature, the developed algorithms try to identify such signatures in the aggregate load curve, thus indicating the periods of operation of the equipment and their respective energy consumption. The methodology of a NILM system is based on four main steps, as can be seen in Figure 1, which are the signal acquisition, event detection, characteristic extraction and equipment identification, as can be seen in Figure 1.

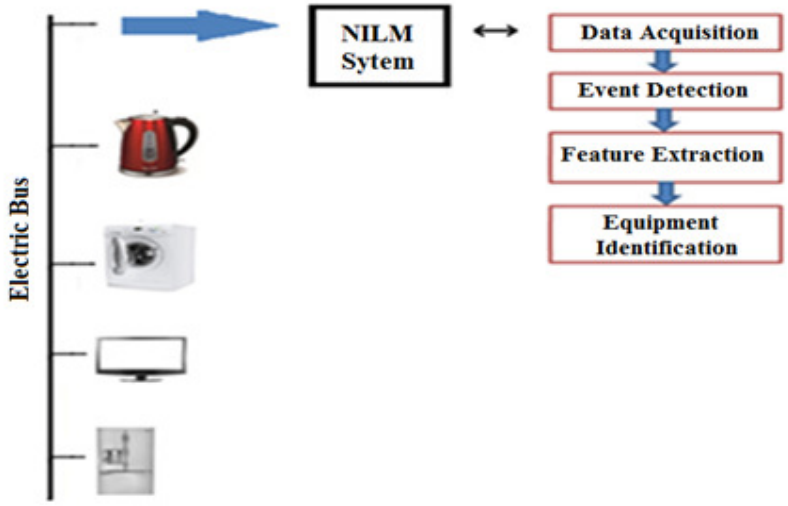


Figure 1 Residential electricity system with integrated NILM system

During the signal acquisition step, the aggregate load is measured through a single sensor on the main branch that is outside the residence. Figure 5 shows an example of the load measured over a period of 1 hour for one of the 7 equipment chosen (Refrigerator). For this stage we use the public database REDD (Reference Energy Disaggregation Data Set), being one of the most used in the

field of NILM systems research. REDD consists of data collected in six households, and contains aggregate electrical power data collected at the 1Hz frequency [3]. Table 1 shows the equipment per household that was measured in REDD.

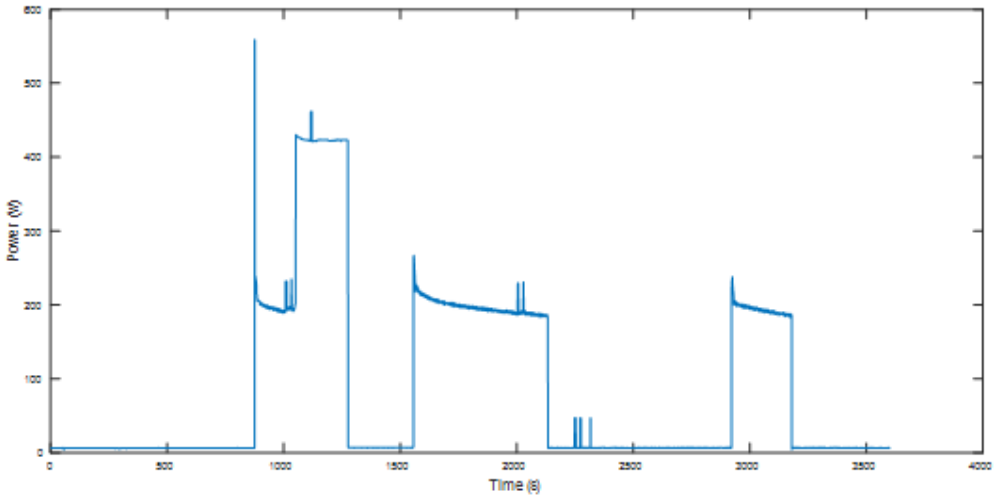


Figure 2 Load example measured over an hour

Table 1. Description of the houses and devices used in the evaluation in REDD data set [3].

House	Device Categories
1	Electronics, Lighting, Refrigerator, Disposal, Dishwasher, Furnace, Washer Dryer, Smoke Alarms, Bathroom GFI, Kitchen Outlets, Microwave
2	Lighting, Refrigerator, Dishwasher, Washer Dryer, Bathroom GFI, Kitchen Outlets, Oven, Microwave, Electric Heat, Stove
3	Electronics, Lighting, Refrigerator, Disposal, Dishwasher, Furnace, Washer Dryer, Bathroom GFI, Kitchen Outlets, Microwave, Electric Heat, Outdoor Outlets
4	Lighting, Dishwasher, Furnace, Washer Dryer, Smoke Alarms, Bathroom GFI, Kitchen Outlets, Stove, Disposal, Air Conditioning
5	Lighting, Refrigerator, Disposal, Dishwasher, Washer Dryer, Kitchen Outlets, Microwave, Stove
6	Lights, refrigerator, crazy washer, heater, clothes dryer, bathroom equipment, cooking utensils, cooker, electronic, air conditioning.

Still in the first stage, 7 electric appliances were chosen for the development of the planned system that was based on a convolutional neural network for the identification of the equipment These were as follows: a microwave, oven, stove, a dishwasher, an air conditioning, a washer/dryer and a refrigerator. The chosen equipment can be regarded as comprising the machines that consume most energy in a household. According to Batra [5], priority should be given to identifying the equipment that uses most energy in the dwellings because these appliances have the most significant features in the aggregate load and thus other appliances that consume less can be regarded as the only noisy items in the total aggregate load.

In the event detection stage, the on / off moments of equipment in a residence (event) are detected from the aggregate signal. In order to detect abrupt changes in the signal, a methodology was

used based on an analysis window that scans the whole measured aggregate load, and it is possible to identify the occurrence of an event when the difference between the final average) and the initial mean (left margin mean) of the window reach a predetermined threshold value, as can be seen in Figure 2. For each detected event, the first twelve transient samples were separated to form the training database of the system. The choice of the number of samples to be used as input to the system was based on the evaluation, for all equipment, of the number of samples sufficient to characterize a complete transient.

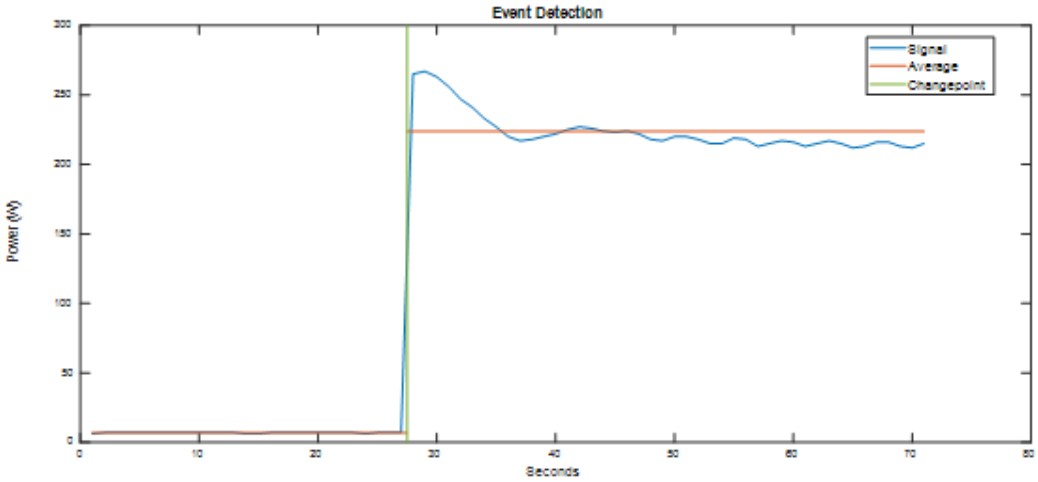


Figure 3 Event Detection Through Windowing

With the detected events, the third stage, of characteristic extraction or electric signatures, takes place. Electrical signatures represent a set of characteristics of voltage, current or power for a given equipment, and can be divided into macroscopic and microscopic. The macroscopic (low frequency) characteristics must be obtained from a sample period of up to one sample per cycle (1 Hz), which is the focus of this work. In the fourth and last step, from the characteristics / signatures extracted, we have the identification of each equipment for each detected event. Methods for identifying equipment used in NILM systems may be of the supervised or unsupervised type.

2.1. Previous work on approach NILM systems

In [6] the authors point out the main supervised techniques to solve NILM problems, such as Artificial Neural Networks (ANN), Supporting Vector Machines (SVM), Naive Bayes Classifier and K-Nearest Neighbor (KNN). Recently the researchers have turned their attention to the use of Deep Neural Networks to the problem of equipment identification. In [7] the authors apply 3 types of deep neural networks, a recurrent neural network based on Long Short Term Memory Units, a self-encoder neural network and a convolutional neural network, to predict the start and end time of an event of an equipment, as well as to predict the average demand of each device. In [8] the author sought to make an analysis of the various methods of deep learning to improve the performance of a NILM system. In [9], the authors used convolutional neural networks for the task of load disaggregation, promoting the individual identification of equipment loads based on the time series of the aggregate load. In [10], it is shown that CNN networks can also be used in the NILM context for equipment classification based on the VI path of an equipment.

This work differs from the other works by the fact that it possesses a single variable as input (transient power signal), while several authors already mentioned as [6] use current harmonics, current waveform, active and reactive power. In the context of the deep neural networks cited in

[7-10], this study performed better than the evaluation metrics used. This is due to the difficulty of other methods in classifying multi-state appliances, such as the dishwasher and the washing machine. In addition, the CNN already presented in the literature need to transform the transient signal of each equipment in an image (spectrogram), to extract the characteristics of the image through the intensity of the colors and finally to make the classification, while in our approach we use directness the power signal, causing our CNN to interpret these values as being the color intensity in an image.

2.2. Evaluating NILM Algorithms

In order to evaluate the performance of the proposed system, some evaluation metrics have been used that are generally used to evaluate equipment identification systems in the context of NILM systems:

Confusion Matrix: Allows an effective measure of the classification model, presenting the number of correct classifications versus classifications predicted for each class, on a set of examples [9]. The main diagonal presents for each class the correct classification number and the percentage that this number represents within the complete number of data of the class.

Accuracy: presents the percentage of positive and negative samples correctly classified on the sum of positive and negative samples.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

Being True positive (TP), the number of times an equipment is correctly classified as ON; True Negative (TN), the number of times an equipment is correctly classified as OFF; False Positive (FP) The number of times an equipment is incorrectly classified as ON and False Negative (FN) is the number of times an equipment is incorrectly classified as OFF.

Sensitivity: percentage of positive samples correctly classified on the total of positive samples.

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{Positive} \tag{2}$$

Precision: percentage of positive samples correctly classified on the total of samples classified as positive.

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

F-score: It is a weighted average of precision and sensitivity

$$\mathrm{F\text{-}score} = \frac{2 \times (\mathrm{Prec} \times \mathrm{Sens})}{(\mathrm{Prec} + \mathrm{Sens})} \tag{4}$$

3. CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network (CNN) can be considered as a variant of the neural network Perceptron of Multiple Layers (MLP). Instead of using fully-connected hidden layers, such as

MLP, the architecture of a CNN is based on the alternation of convolution layers - the layer that names the network; and pooling layers. Each layer will have a set of filters, also known as kernel, that will be responsible for extracting local features from an input. With this, we can create several convolution and pooling maps, containing several specific characteristics like borders, colour intensity, contours and shapes. Each feature map will have a shared set of weights, which decreases the computational complexity of the network [11]. Finally, we have the layer responsible for the classification process, which have the fully connected layer, which connects all the neurons of the layer before it to the output neurons, as shown in Figure 3.

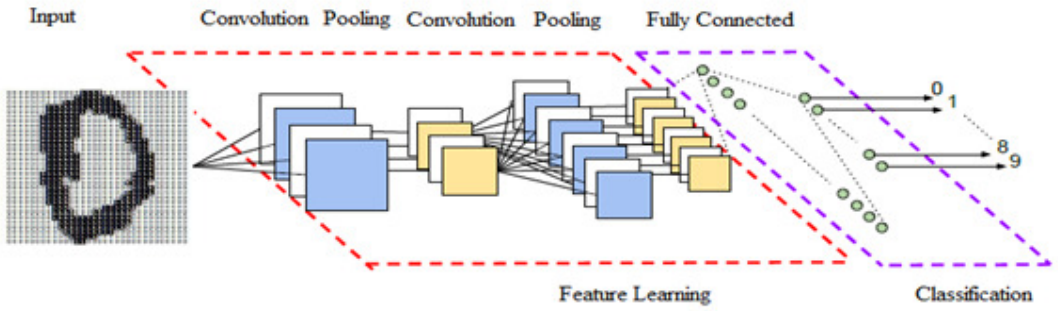


Figure 4 Illustration of the architecture of a CNN [11]

For this approach, which is focused on the classification of equipment through the behavior of its power transients, an architecture based on three layers of convolution followed by pooling was used. Between each convolution and pooling layer normalization is applied in the filter sets (batches), which serves to accelerate network formation and reduce sensitivity for initialization. In addition, we used the non-linear activation function (ReLU) which is simply the identity function for positive values. After the 3 layers of convolution and pooling a fully connected layer is used, followed by the Softmax function. This architecture, derived from a reduction in the convolutional network GoogLeNet [12] (that has five layers of convolution always followed by a pooling), is represented in Figure 4, containing specifications such as: the number of filters in each layer, the size of the stride and the configuration of the output layer.

The convolution layer consists of neurons that are responsible for extracting different sub-region resources from the input images [13]. These areas are derived from the filters used in this layer, being able to extract specific characteristics of the input. In this layer we specify the amount of filters, their sizes, in addition to the stride, which defines the size of the neighbourhood that each layer's neuron will process. [11]

The Pooling layer follows the convolutional layer reducing the number of connections to the following layers, being Max-Pooling in our work. A Max-Pooling layer returns the maximum values obtained in its filters. This layer does not perform any learning, but reduces the number of parameters to be learned in the following layers. [13,11]

The fully connected layer connects all the neurons of the anterior layer with the output neurons, which represent the classes to be classified. This layer combines all the characteristics (local information) learned in previous layers, sweeping the input to identify the highest standards. For our classification problem, it will combine the characteristics of the transients to classify the equipment. At the output of the classification layer, the Softmax activation function is applied which is responsible for performing the multi-class classification (for example: object recognition). [13,11]

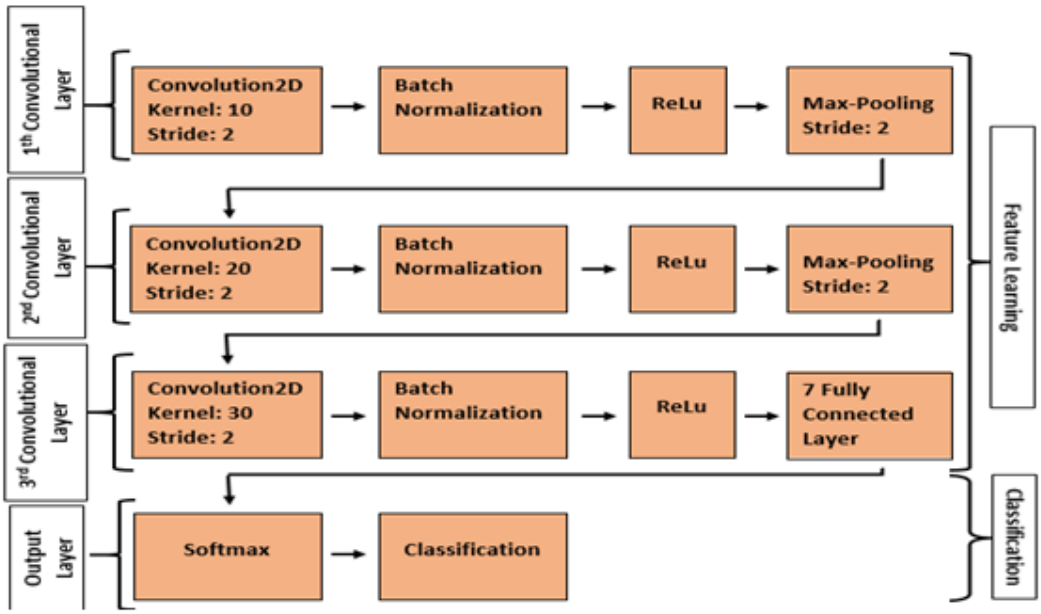


Figure 5 CNN architecture developed for the proposal

3.1. CNN Training

The database for developing the identification system had 448 patterns, when all the 7 appliances were taken into account. Each pattern has 7 transient samples for a particular appliance, thus forming a bidimensional matrix (7x448). The data were divided into training, validation and test categories, comprising approximately 60%, 20% and 20% respectively for the total number of patterns. Table 2 shows the arrangement of the data in greater detail.

Table 2. Data Organization.

N°	Equipment	Trai	Valid.	Test	Total
1	Refrigerator	60	14	15	89
2	Microwave	65	14	15	94
3	Stove	67	14	14	95
4	Oven	60	14	13	87
5	Dishwasher	61	12	15	88
6	Air conditioning	55	12	13	80
7	Washer / Dryer	80	15	18	113
#	Overall	448	95	103	646

The approach involves the direct use of 7 samples of power supply transient signals of the appliances as an entry to the CNN, without the need for the application of signal processing to images such as spectrogram [14], or binary images [15]. All that was necessary to achieve this was to re-size the entry of the training matrix to 4D, and thus take on the dimensions of 1x7x1x448, and in this way the CNN can interpret the data as a numerical 4-D matrix (an agglomeration of colored images). While the first three dimensions refer to height, width and channels, the last dimension must index the individual images, or rather, index the transients.

4. RESULTS

Table 3 shows the results obtained for the test data, after the training of the projected CNN network. The result are given in the form of metrics: sensitivity (Sens), precision (Prec) and F-score (F). The 3 assessment metrics used in this study can assist us in measuring the performance of the CNN from another perspective. Thus, for example, there are the Oven, Air-Conditioner and Washing-machine which were classified in a precise way, since they had a low amount of FP. However, they did not have the same level of performance for sensitivity, which measures the capacity of the system to predict correctly in the cases that really have it (TP). For this reason, the F-score is used to harmonize the two assessment metrics already mentioned and make a better comparison between the appliances by means of the F-score metric. Hence, it can seen from the analysis in Column F that the Air-conditioner and Oven had a score above 90%, which demonstrates that the model shown had an excellent performance.

Table 3. Results for Test Data

N	Equipment	Sens.	Prec.	F
1	Refrigerator	0.8667	0.6842	0.7647
2	Microwave	0.9333	0.7000	0.8000
3	Stove	0.6429	0.7500	0.6923
4	Oven	0.8462	1.0000	0.9167
5	Dishwasher	0.6667	0.7692	0.7143
6	Air conditioning	0.9231	1.0000	0.9600
7	Washer / Dryer	0.8889	1.0000	0.9412
#	Overall	0.8239	0.8433	0.8270

Table 4, in turn, shows the results obtained in the training, testing and validation simulations where the metrics used were accuracy and the F-score. On the basis of these results, it can be noted that although we are confronted with a complex classificatory problem, involving multi-stage types of equipment, the CNN on average, had a general rate of accuracy of 82.43% and an F-score of 82.46%, which are very promising results.

Table 4. Performance Results

Simulation	Acc.	F
Training	0.8795	0.8785
Validation	0.7684	0.7685
Test	0.8252	0.8270
Geral	0.8243	0.8246

Figure 5 shows the confusion matrix obtained for the test data which thus allows a broader view of the performance of our algorithm, as well as providing a detailed account of the results obtained in Table 3. The 6 appliances are defined as follows: Refrigerator (1), Microwave (2), Stove (3), Oven (4), Dishwasher (5), Air Conditioning (6) and washer/dryer (7). Each matrix column represents the categories of appliances predicted by the CNN, while the lines represent the real categories. The number of checks for each class can be found on the main diagonal of the

matrix. Thus, it can be inferred that the appliances that have FN values, had a reduction of sensitivity, such as the Stove and Dishwasher. The Refrigerator had the worst rate of precision owing to the fact that this appliance had had a high FP value, with 6 FP values and 13 TP values. However, the Washer-Dryer and Oven did not have any FP values, and attained a 100% precision rate.

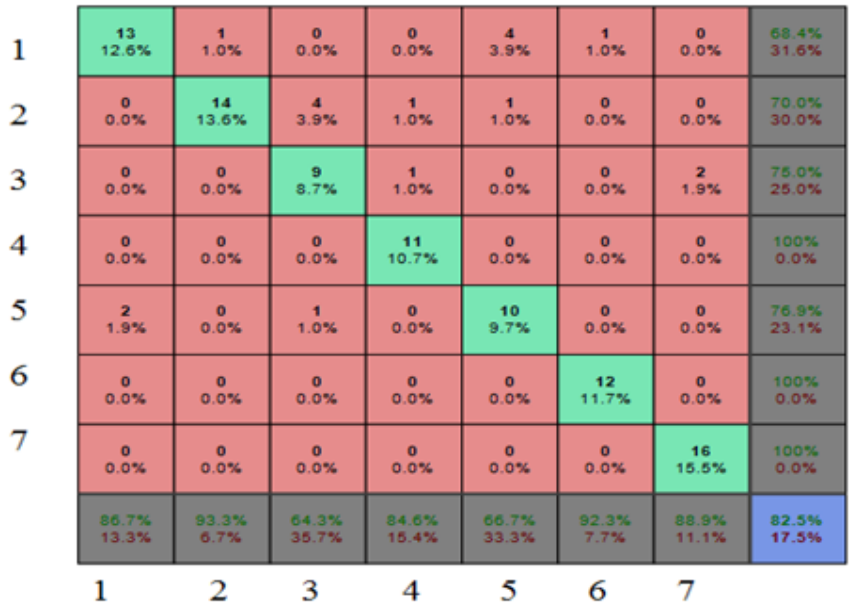


Figure 6 Confusion Matrix for Test Data

4.1. COMPARISON WITH STATE OF THE ART

In this section, we compare our results with some state-of-the-art NALM algorithms, proposed for low sampling rates and active power measurements. Table 5 presents the results of some systems already developed to identify equipment in NILM systems using as input the power transient measurements for low frequency. A direct comparison of results should be carried out with caution since for all the presented systems one has the database used for different training, test and validation and equipment and number of equipment also identified different.

Table 5. Comparison between systems presented in the literature

Authors	**Technique**	**Nº of Appliance categories**	**Sens**	**Prec**	**F**	**Acc**
This Study	CNN	7	0.82	0.84	0.82	0.82
Kelly [7]	Autoencoder	5	0.80	0.58	0.55	0.91
Kelly [7]	LSTM	5	0.69	0.39	0.39	0.68
WONG [16]	PDT	6	0.77	0.76	0.73	-----
Zhao [17]	GSP	8	0.51	0.89	0.64	0.77

[7] Uses long short-term memory; [16] Uses Particle-based Distribution Truncation (PDT) and [17] Uses Graph Signal Processing (GSP).

5. CONCLUSIONS

In this article, we describe how to apply CNNs to the recognition of technical equipment in an innovative manner. From the results obtained, the efficiency of the proposed system is clearly evident, where a weighted average of precision and sensitivity was obtained that was higher than 75%; and with an average degree of accuracy of 82%. The results obtained can be regarded as satisfactory when compared with the results of the identification systems already shown in the literature and also when account is taken of the complexity of the system put forward which was designed to identify loads in a multilevel or variable state.

One point that should be stressed with regard to the direct use of the power supply transient signal as an entry to the identification system, is that it speeds up the system. This means that it is a system that can achieve good results in classification by using data where the measured power is of a low frequency. This is beneficial since the use of low frequencies is common in available low–cost measuring devices which are currently being used for the development of NILM systems.

REFERENCES

[1] HART, George William. Nonintrusive appliance load monitoring. Proceedings of the IEEE, v. 80, n. 12, p. 1870-1891, 1992.

[2] FIGUEIREDO, Marisa. Contributions to Electrical Energy Disaggregation in a Smart Home. 2014. Tese de Doutorado. APA. Disponível em: <www:http://hdl.handle.net/10316/24256>. Acessado em: novembro de 2017.

[3] KOLTER, J. Zico; JOHNSON, Matthew J. REDD: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA. 2011. p. 59-62.

[4] KATO, Takekazu et al. Appliance Recognition from Electric Current Signals for Information-Energy Integrated Network in Home Environments. ICOST, v. 9, p. 150-157, 2009.

[5] BATRA, Nipun et al. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. arXiv preprint arXiv:1408.6595, 2014.

[6] WONG, Yung Fei et al. Recent approaches to non-intrusive load monitoring techniques in residential settings. In: Computational Intelligence Applications In Smart Grid (CIASG), 2013 IEEE Symposium on. IEEE, 2013. p. 73-79.

[7] KELLY, Jack; KNOTTENBELT, William. Neural nilm: Deep neural networks applied to energy disaggregation. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments. ACM, 2015. p. 55-64.

[8] DO NASCIMENTO, Pedro Paulo Marques. Applications of Deep Learning Techniques on NILM. 2016. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

[9] Wan He and Ying Chai. An Empirical Study on Energy Disaggregation via Deep Learning, in Advances in Intelligent Systems Research, volume 133, 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016), pp338-341, 2016

[10] DE BAETS, Leen et al. Appliance classification using VI trajectories and convolutional neural networks. Energy and Buildings, v. 158, p. 32-36, 2018.

[11] VARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In: Proceedings of the XXIX Conference on Graphics, Patterns and Images. 2016. p. 1-4.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in CVPR 2015, 2015.

[13] HIJAZI, Samer; KUMAR, Rishi; ROWEN, Chris. Using convolutional neural networks for image recognition. Tech. Rep., 2015. [Online]. Available: http://ip. cadence. com/uploads/901/cnn-wp-pdf.

[14] ABDEL-HAMID, Ossama et al. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, v. 22, n. 10, p. 1533-1545, 2014.

[15] Atabay, H.A.: Binary shape classification using convolutional neural networks. IIOAB J. 7(5), 332–336 (2016)

[16] WONG, Yung Fei; DRUMMOND, T.; ŞEKERCIOĞLU, Y. A. Real-time load disaggregation algorithm using particle-based distribution truncation with state occupancy model. Electronics Letters, v. 50, n. 9, p. 697-699, 2014.

[17] ZHAO, Bochao; STANKOVIC, Lina; STANKOVIC, Vladimir. On a training-less solution for non-intrusive appliance load monitoring using graph signal processing. IEEE Access, v. 4, p. 1784-1799, 2016.

AUTHORS

B. Sc. Deyvison de Paiva Penha is a Master Student in the Electrical Engineering Graduate Program in Federal University of Pará. He received his bachelor degree in 2009 at the Federal University of Pará.

Prof. Dr. Adriana Rosa Garcez Castro has a Master's degree in Electrical Engineering from the Federal University of Pará in 1995 and a PhD in Electrical Engineering from the Faculty of Engineering of the University of Porto in 2004. She is currently a Professor at the Federal University of Pará. His areas of interest are: Control of Electronic Processes and Computational Intelligence applied to Energy Systems.

INTENTIONAL BLANK

HPPS: HEART PROBLEM PREDICTION SYSTEM USING MACHINE LEARNING

Nimai Chand Das Adhikari[1], Arpana Alka[1] and Rajat Garg[2]

[1]Department of Mathematics, Indian Institute of Space Science and Technology, Thiruvananthapuram, India
[2]Department of Biotechnology, National Institute of Technology, Jalandhar, India

ABSTRACT

Heart is the most important organ of a human body. It circulates oxygen and other vital nutrients through blood to different parts of the body and helps in the metabolic activities. Apart from this it also helps in removal of the metabolic wastes. Thus, even minor problems in heart can affect the whole organism. Researchers are diverting a lot of data analysis work for assisting the doctors to predict the heart problem. So, an analysis of the data related to different health problems and its functioning can help in predicting with a certain probability for the wellness of this organ. In this paper we have analysed the different prescribed data of 1094 patients from different parts of India. Using this data, we have built a model which gets trained using this data and tries to predict whether a new out-of-sample data has a probability of having any heart attack or not. This model can help in decision making along with the doctor to treat the patient well and creating a transparency between the doctor and the patient. In the validation set of the data, it's not only the accuracy that the model has to take care, rather the True Positive Rate and False-Negative Rate along with the AUC-ROC helps in building/fixing the algorithm inside the model.

KEYWORDS

Heart Attack, Computation, Machine Learning, Data Analysis, Recommendation Systems, Neural Networks, Data Mining, Visualization, Artificial Intelligence

1. INTRODUCTION

The mortality rate in India and abroad is mainly due to heart attack. This calls for a vital check of the organ periodically for the wellness of all human beings. From the below figure of the heart, any major heart problem occurs when there is a blockage in the major arteries that carries the oxygenated blood [1]. The blockage causes huge pressure on the organ to pump the required amount of pure blood to the other parts of the body. The health care industry has huge amount of data that can be utilized to find the different patterns related to the heart problems with a probabilistic score. Here, we have collected the data from a survey of around 1000 patients from different parts of India and found out the correlation among the different risk factors that we have gathered.

The risk factors that has been taken as an input. in this survey are Family History, Smoking, Hypertension, Dyslipidemia, Fasting Glucose, Obesity, Life Style, CABG and High Serum in blood. Apart from the mentioned risk-factors, we have the demographic details as well. The most

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 23– 37, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71803

important thing that each diagnosis should prevent is the exposure to a normal human body to the CT Scan radioactive rays [2][3]. The CCTA (Coronary computed tomography angiography) is an imaging test for the heart to find out the places for the plaques build up in the blood vessels. This has an increased prone to the cancer for the human body exposed to high radiation [4]. Plaque is majorly built up due to the circulating substances in blood like fat, cholesterol and calcium, whose deposit in the inner side of blood vessel can effect the normal blood flow and can result in excessive pressure on the heart pump.

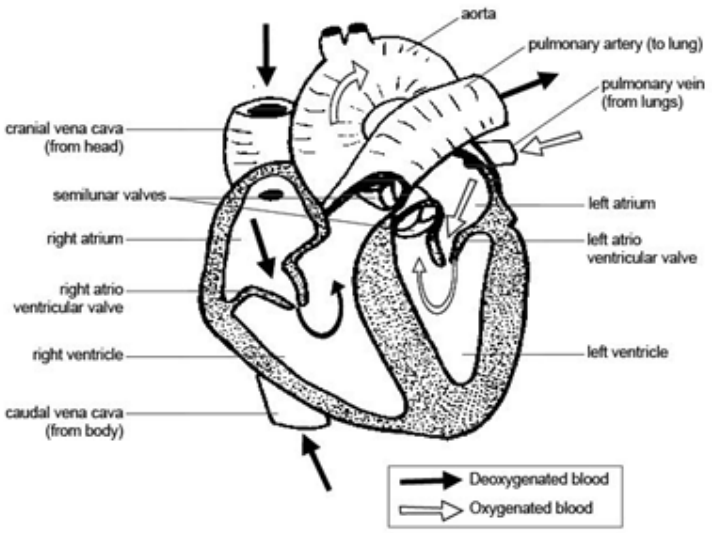


Figure1: Diagram of Human Heart

So, the main intension of this paper is to help in the decision making of a doctor for detecting the possibility or identifying the patient's suffering or going to suffer from heart problems. Apart from the above mentioned, this method should also help in diminishing the False Negative Rate of the prediction. It is the number of the actual positives which is negative through the prediction to the total negatives. In statistical hypothesis testing, this ratio is represented by the letter β. In the following sections we will discuss the different terminologies and factors related to this project and the methodology of HPPS, which can be a partner of the doctor in the decision making of whether the patient is going to suffer from any heart attack or not. In the next section we will discuss about the factors that we have taken for the survey and their correlations with the predictor output, followed by the proposed model and scenarios and lastly with the results for the selection of the algorithms.

2. DATASET DESCRIPTION AND ANALYSIS

The survey contains the data of 1094 patients from 5 different cities of India Delhi, Chennai, Bangalore, Kolkata and Hyderabad. The attributes that de ne as the features for the model are the different demographic details of the patients like Age and Sex with the different Risk Factors which we have defined previously. Here the predictor variable is Heart Problem or Not. Thus, there are many terminologies that de ne this. Some of them are:

1. Heart disease due to atherosclerosis [5]: In this case the walls of the arteries become stiff or hard due to the fatty deposits which in medical term known as plaques.

2. Cerebrovascular disease [6]: This is mainly due to the blockage in the blood flow through the blood vessels to the brain.

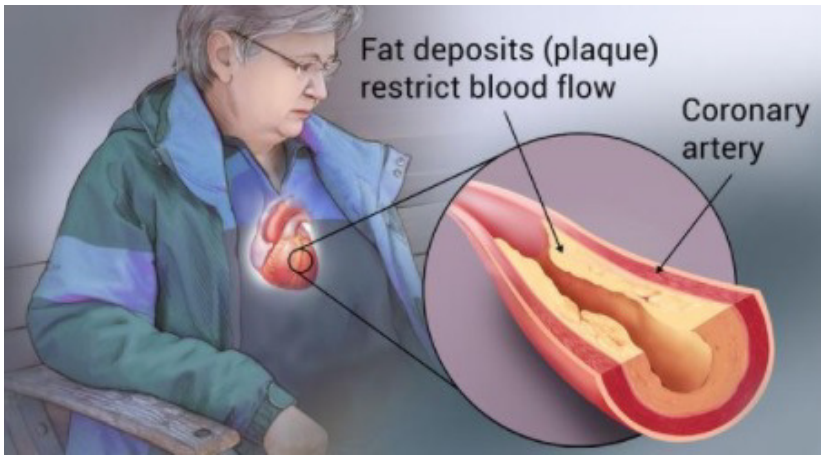


Figure 2: Heart Blockage

3. Ischemic heart disease [7]: This is mainly due to the deposit of the cholesterol on the walls of the arteries. Figure 2 shows how the deposit looks like in this similar case.

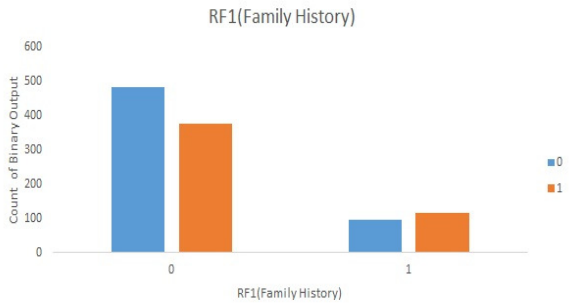
4. Hypertensive heart disease [8]: This happens mostly due to high blood pressure.

The above is some of the types of heart problems that we have discussed. There are many apart from the ones described before as the heart is one of the vital organs that help in the transportation of the oxygenated blood and nutrients and removal of wastes from the body. In the predicted value, we have given the value as 1 for the heart related problems and 0 as no problem in the heart.

Below is the analysis of the different risk-factors for the heart problem detection.

2.1. RISK FACTOR 1: FAMILY HISTORY

This is one of the important risk-factor as it depends on the hereditary behaviour of the heart [9]. Here, we have the values of 1080 patients and the rest are NA or No values. For those missing values we have assigned the value as 0 or the maximum of the value that appears in this risk factor. Which we will discuss in the results section.

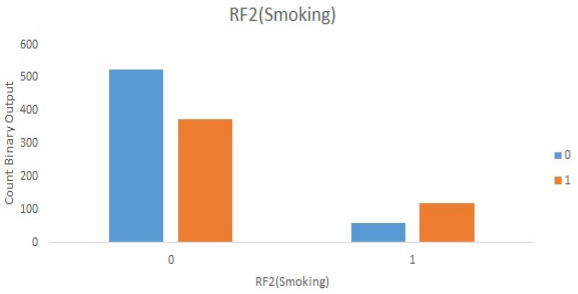


Family History	0	1	Grand Total
0	486	379	865
1	97	118	215

In the analysis we found that, when Family History is 1, then 118 out of 215 patient suffer from heart problem i.e 55%.

2.2. RISK FACTOR 2: SMOKING

It leads to the developing of the cardiovascular diseases, which includes heart attack and stroke. It leads to damaging the lining of the arteries which ultimately leads to atheroma. Below is the analysis of the data for the smoking that we have established.

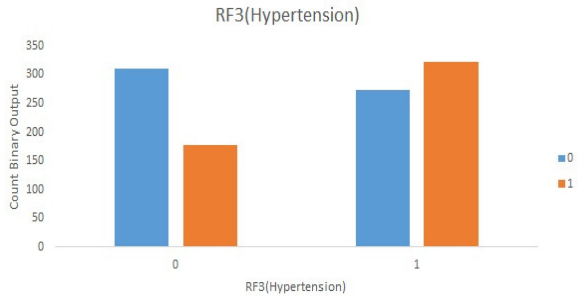


Smoking	0	1	Grand Total
0	524	375	899
1	59	121	180

The above curves show that if the patient has smoking as a characteristic, then 67.22% chances is, he/she will suffer from the heart related problems [10] [11].

2.3. RISK FACTOR 3: HYPERTENSION

This leads to the heart diseases that occur due to high blood pressure over a long period of time [12][13]. Due to blood pressure, the heart has to do pump more against this pressure, adding extra pressure to heart resulting into the thickening of the heart muscle.



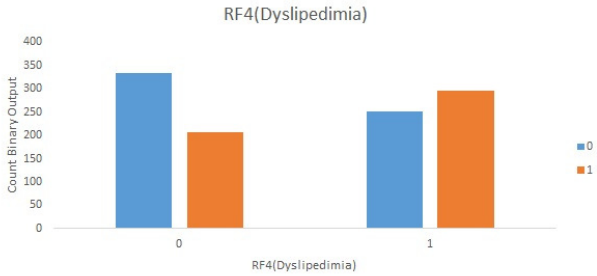
HyperTension	0	1	Grand Total
0	309	177	486
1	273	321	594

Figure 3: Hypertension

In the analysis done and represented in the figure 3, we can find that 54% chances is there for a hypertensive patient to suffer from any heart related problem.

2.4. RISK FACTOR 4: DYSLIPEDIMIA

This is a high level of lipids like cholesterol, triglycerides carried through the lipo-proteins present in the blood. The risk of *Atherosclerosis* increases due to the increase in the above-mentioned lipids in the blood leading to excessive pressure on the blood flow [14].

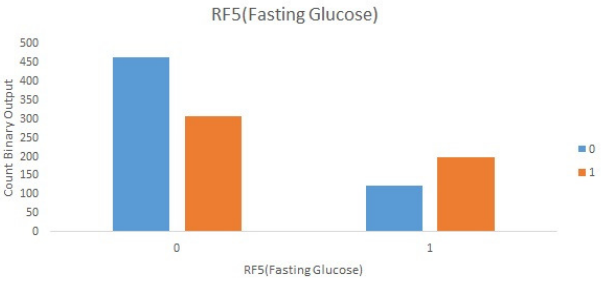


Dyslipedimia	0	1	Grand Total
0	334	208	542
1	252	296	548

In this analysis, we found that out of 1090 patients having the details of suffering from dyslipedimia which has been captured by the doctor, 548 suffer from the same. Out of 548, 296 patients suffered from heart related problems, which is a little over 54%.

2.5. Risk Factor 5: Fasting Glucose

Fasting Glucose greater than a certain value leads to type 2 diabetes and it is proved that type 2 diabetes increase marks the risk of Cardiovascular Disease(CVD) and ischemic heart disease(IHD) [15] [16] [17].

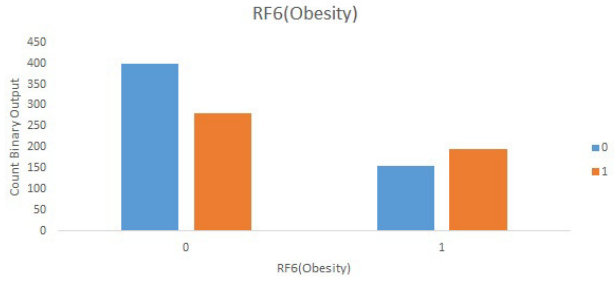


Fasting Glucose	0	1	Grand Total
0	461	306	767
1	122	197	319

According to our analysis, we found that 1066 data of the patients had this risk factor captured. Out of this, 319 had Fasting Glucose as marked 1. About 62% of those having 1 in this risk-factor suffered from the heart attack, which proved the analysis with that of the proven results.

2.6. Risk Factor 6: Obesity

The role of diet in the prevention of CVD is very crucial as it is a very key risk factor for CVD. Thus, obesity leads to the development of hypertension, diabetes, musculoskeletal disorder, thus putting in a high risk of CVD [18].

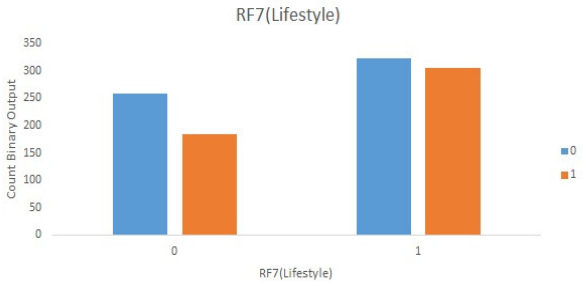


Obesity	0	1	Grand Total
0	399	280	679
1	156	194	350

According to the analysis, we found that 194 patients having Obesity suffered from heart related problems which accounts to 56%.

2.7. RISK FACTOR 7: LIFE STYLE

It is one of the most important factors in controlling the heart related problems. Some of the major lifestyle effects that can control in the prevention and keeping the heart in a good shape are Stop Smoking, Choosing Good Nutrition, High Blood Cholesterol, Lowering High Blood Pressure, Being Physically Active, Aiming for a healthy weight, Managing Diabetes, Reducing Stress and drinking alcohol etc. [19][20].



Life Style	0	1	Grand Total
0	259	185	444
1	323	306	629

In the analysis above, we find that 306 cases out of 629 marked as 1, suffered from heart related disease. Thus, this is around 49% of the cases. But if we see the two bar plots above we can find that the conversion of the heart problem is in a greater percentage in case of the bad life style. Thus, marking this risk factor to be one of the most important factors in determining the CVD.

2.8. RISK FACTOR 8: CABG

Coronary Artery Bypass Grafting is a kind of surgery done for those patients who have suffered from severe CHD.

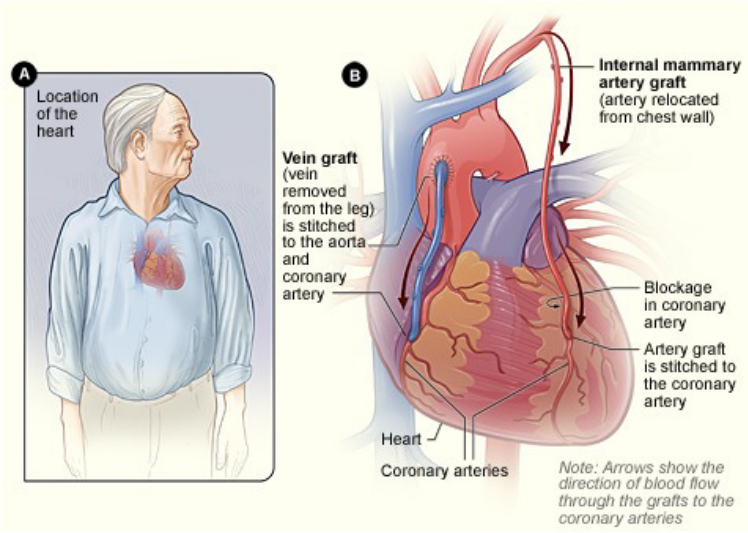
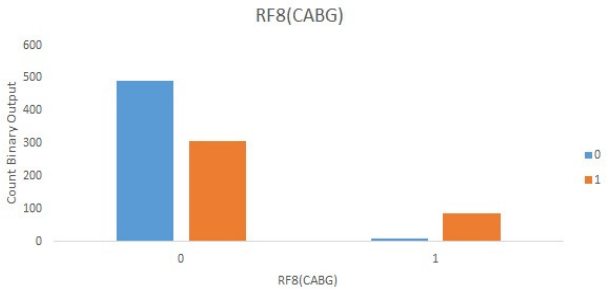


Figure 4: Bypass Grafts in heart

This is mostly whether a patient suffered from the serious heart attack and has a graft anywhere in the heart. Thus, this will be having a very high correlation for the heart being regularly checked up.

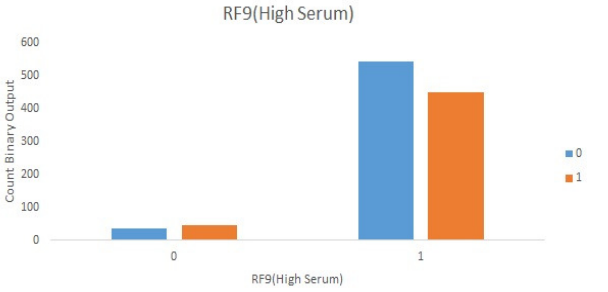


CABG	0	1	Grand Total
0	491	306	797
1	9	86	95

From our analysis done, we found that 95 of the cases of the patients had grafts present or this risk factor being high according to the doctor. Among the data, 86 do have severe heart problem and being asked for re-check-up.

2.9. RISK FACTOR 9: HIGH SERUM

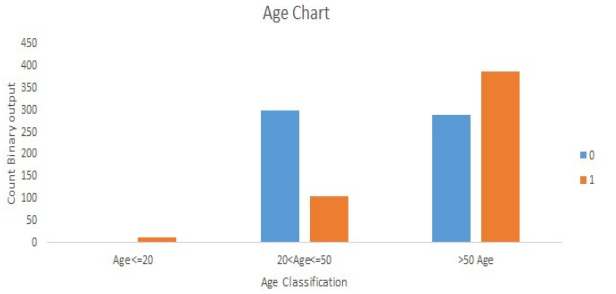
A Serum test is a measure of the amount of iron which is present in the left over liquid after the red blood cells and the clotting factors being removed from the blood. Hence having too much iron content in the blood can cause serious health problem. This has a direct correlation with the heart related problems [21].



High Serum	0	1	Grand Total
0	37	46	83
1	544	452	996

In this analysis, we found that 452 cases having suffered from heart problems out of 996 having High Serum.

Apart from the above Risk Factors we have different other attributes like Age, Sex, Location and Vascular Pattern. The analysis of the Age feature is shown below in binned form.



Age Range	0	1	Grand Total
Age <=20	3	12	15
20 < Age <= 50	299	106	405
> 50 Age	286	388	674

We have divided the age continuous values into three groups 'age < 20years', 'age between 20 years and 50 years' and 'age > than 50 years'. We can find from the analysis that most of the cases the age group more than 50 years have suffered from heart related problems which is not

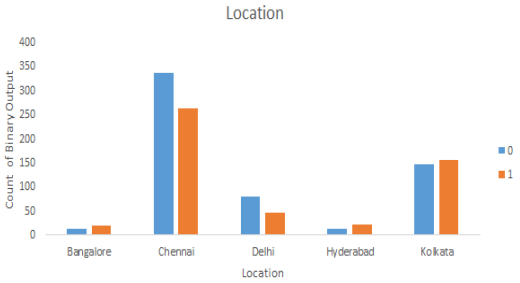
the case in case of the middle age group. Thus, the heart problem is skewed towards the more than 50 age group.

In the below graph we are showing the analysis for the heart problem with that of the gender or sex category of a patient.

Sex	0	1	Grand Total
Female	247	112	359
Male	341	394	735

Males are more prone to heart related problems than female as can be seen from the analysis.

In another analysis which represent the variation of the patients with the location of 5 different cities all over the India i.e Chennai, Delhi, Kolkata, Bangalore and Hyderabad, we find that the data is skewed towards Chennai as more data is available from that region and when the patient suffering from heart related problem is seen, Chennai, Delhi and Kolkata are having patients details more than 100 and out of them Kolkata region has more patients suffering from heart problem and is recorded as 51.66% which can be seen from the location graph below.



Location	0	1	Grand Total
Bangalore	13	20	33
Chennai	337	262	599
Delhi	79	46	125
Hyderabad	13	22	35
Kolkata	146	156	302

3. EXISTING PROCEDURE AND LITERATURE SURVEY

As talked earlier in this paper heart disease remains one of the main causes for deaths worldwide. About 7.4 million people died due to coronary heart disease, and 6.7 million were only due to stroke (WHO, 2015). In order to investigate the misfortune of heart attack, certain factors that are associated with different risks need to be addressed. Therefore, people with heart disease due to the presence of chest pain, resting blood pressure, cholesterol, fasting blood sugar resting electro cardiographic and maximum heart rate need early detection and prediction for better counseling and appropriate medicine. According to Anooj(2012) and Hedeshi and Abadeh(2014), the decision to make for the presence of any problem in heart sorely depends on the physicians intuition, experience and experience. This is a very challenging task and needs to take care of a number of factors. Mostly the work related to the prediction and figuring out the heart problem, many data driven techniques has been used in past and the work inclines towards the classification problem. This is a process used to tune a model and then predict the class for whether the patient is suffering from any heart related problem or not. To talk about the

intelligent methods in the medical sector, a vast number of related works has been performed (Muthukaruppan & Er, 2012; Sikchi et al., 2012; Kumar, 2013; Sikchi et al., 2013). The practitioners make use of these computerized intelligent methods for assist in the diagnosis to give suggestions with certain probability. In 2012 Opeyemi and Justice suggested one of the best and effcient technique to deal with the uncertainty by incorporating fuzzy logic and neural network. There are many diverse studies that tend to the ANFIS methodologies (Palaniappan & Awang, 2008; Patil & Kumaraswamy, 2009; Abdullah et al., 2011; Zhu et al., 2012; Kar & Ghosh, 2014; Mayilvaganan & Rajeswari, 2014; Yang et al., 2014). This research involves in the developing a framework that includes hybrid learning algorithms to find the least square estimates with gradient descent and Levenberg-Marquardt algorithms for training Statlog-Cleveland Heart Disease Dataset [24]. Some of the recent work on the heart problem prediction has been done using naive bayes [25] [26]. In [27], many classification algorithms like Naive Bayes, Decision Tree, K-NN and Neural Network is used for Prediction of Heart Disease and the result proves that Naive Bayes technique outperformed other used techniques. Similar to this [28] tree based algorithms J48, Bayes Net, Simple Cart, and REPTREE along with and Naive Bayes algorithm is used to classify and develop a model which diagnose heart attacks in the patient data. Three popular data mining algorithms (support vector machine, artificial neural network and decision tree) were employed to develop a prediction model using 502 cases for better prediction of heart problems [29]. SVM became the best prediction model followed by artificial neural network. In [30] a new concept of Weighted Associative Classifier was used where it was used to predict the probability of patients receiving heart attacks. In this Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. Coming forward a new approach different from above in [31] based on adaptive neuro-fuzzy models are presented was proposed. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work effciency in performing analysis for coronary heart disease occurrences [32].

4. PROPOSED SYSTEM

In the proposed model, we want to give a brief idea about how our system looks like and behaves. Below is the flow chart of our model:

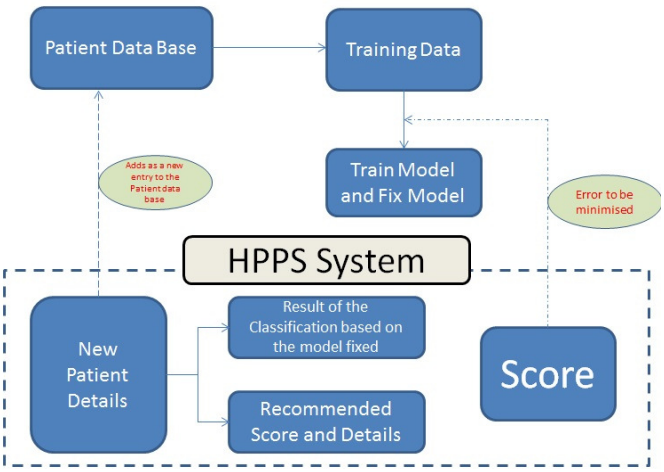


Figure 5: Flow Chart of HPPS

Dataset: There is a common database for the patient from where the data will be taken by the model to finalize the algorithm. The Database will be for a particular hospital where HPPS is being installed or it can be an on-line stored data where all the details of the patient for those

hospitals who use HPPS can access it. This will be helpful for both Classification Score and Recommendation [23].

Algorithms: We have used a wide range of algorithms and in the validation set, the algorithm that gives a better Selection Value i.e.

$$\text{Selection}_{\text{value}} = 0.6 * (1 - FNR) + 0.4 * Accuracy$$

Here, we have assigned 0.6 to the term having the FNR, as we wanted to diminish the False Negative Rate more than the Accuracy. Using the above metric, the algorithm which gives the maximum score in the validation set, is selected.

Recommender System: When a new patient detail is input to HPPS, using the risk factor combinations, all those similar patient details is made a clustered display using the cosine-similarity.

$$\text{score}_{\text{similarity}} = \frac{< patient_{new}, patient_{old} >}{|patient_{new}|_2^2 * |patient_{old}|_2^2}$$

Recommender Score: The Voted Output of the recommended patient details will be shown in the dash board [figure 6].

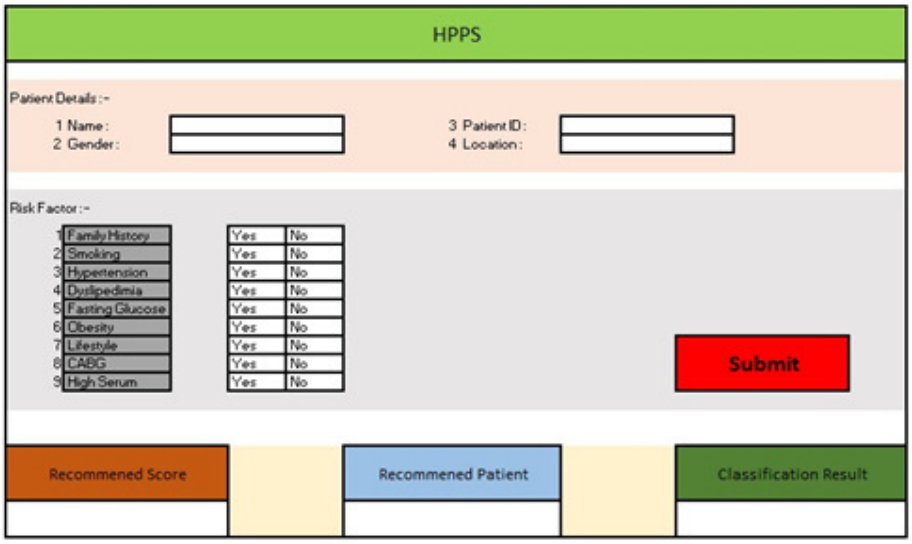


Figure 6: Dashboard of HPPS

Using the above information, the doctor will have multiple scenarios and also help him in aiding to his decision. This will also help to create a transparency among the doctor and the patient. So, here we want to showcase a system which can create a confidence in the patients mind for he/she is going to have any heart problem in the future or not, so as to take better care.

In the dashboard shown in figure 6, there are three sections. In the first section, the demographic details of the patient will be recorded and the Patient ID will be automatically filled. This will mostly depend upon the hospital id and the patient number. The second section is the risk-factors details section. Here the values that will be input will be mostly Boolean i.e. Yes or No. In the right side of this section one red button labelled Submit is present. Once the button is pressed, Section 3 will generate it's all relevant values.

5. RESULTS

In this section we have presented the results comparison for different algorithms that we used in the model. Here, we have analysed the details of 1094 patients having label as 1 or 0. Here 1 is represented for the patient's suffering from any kinds of heart disease and vice versa. Also, for small plaques, the label is given as 0. For training and validation to check how the algorithm is performing, we have used the holdout technique with 70:30 ratios. There are many others cross validation techniques but we have fixed our model to start the testing phase with the 70:30 percent holdout technique.

Matrix	Predicted NO	Predicted YES	Total
Actual NO	TN	FP	TN + FP
Actual YES	FN	TP	FN + TP
Total	TN + FN	FP + TP	TN + FP +FN + TP

Figure 7: Confusion Matrix

In figure 7, we present an example of a confusion matrix and the interpretation of it. The metric Accuracy is the ratio of the sum of TN and TP to the sum of TN, TP, FN and FP. Apart from the accuracy, we believe that we have to diminish the False Negative Rate which is the ratio between the FN and sum of TN and FN. Using these two metrics we define our own metric which we use to select the best algorithm i.e. $Selection_{value}$.

We want to penalize the model for predicting wrong for a patient having the chance for heart attack or heart problem but predictive No for that case. This we have taken into the consideration because the patients who have the chance of suffering from any heart problem cannot be predicted wrong. Using the above metric as Selection Value, we have found that particular algorithm in both the cases which gives that particular algorithm as a trade-off. Below are the results for the verification of different algorithms which are present in the model. All the accuracy that we present it here are the validation accuracy. It is how correctly the algorithm has predicted the validation set. 329 samples out of the total dataset is used for the validation set. The algorithm that we have used in our model are SVM-rbf, SVM-sigmoid, Logistic Regression, Decision Tree Classifier, Random Forest, Naive Bayes.

Algorithm	Validation Accuracy	Confusion Matrix	Predicted No	Predicted Yes
SVM-Sigmoid kernel	68.085	Actual No	130	44
		Actual Yes	61	94
SVM-RBF Kernel	74.16	Actual No	139	35
		Actual Yes	50	105
Logistic Regression l1	71.732	Actual No	135	39
		Actual Yes	54	101
Logistic Regression l2	72.34	Actual No	136	38
		Actual Yes	53	102
Decision Tree Classifier	64.74	Actual No	111	63
		Actual Yes	53	102
Random Forest	72.34	Actual No	132	42
		Actual Yes	49	106
Gaussian NB	71.12	Actual No	150	24
		Actual Yes	71	84
Multinomial NB	67.48	Actual No	139	35
		Actual Yes	72	83
KNN	70.212	Actual No	144	30
		Actual Yes	68	87
Bagging Classifier	72.34	Actual No	138	36
		Actual Yes	55	100
Ridge Classifier	66.56	Actual No	132	42
		Actual Yes	68	87
MLP Claasifier	71.12	Actual No	138	36
		Actual Yes	59	96

Figure 8: Results 1

In the figure 8, the results for the various algorithm is analyzed with the 0 as the imputation for the missing values. In this if we check the accuracy alone, SVM with rbf kernel gives a better result with 74.16 % accuracy followed by 72.34% with Random Forest, Bagging and Logistic Regression l2 norm. Apart from the above accuracy measure, we want to minimize the False Negative Rate i.e Actual is 1 but predicted is 0. The algorithm that best performed is SVMrbf with 29.118%.

In the figure 9, we have imputed the missing values if present in the data, with the maximum frequency present and we see increased values or accuracies for all the algorithms and SVM-rbf performed better with 75.68% accuracy. But if we check the False-Negative Rate, Random Forest performed better in this category. Even in the previous scenario, Random forest had this actual number lesser but the rate was higher. When checked with the Selection Value, Random Forest is the better algorithm with selection probability of 0.741 in comparison to 0.738 of SNM-rbf. These results will pop up in the section 3 of the Dash board and will take a decision making in case of the prediction.

Algorithm	Validation Accuracy	Confusion Matrix	Predicted No	Predicted Yes
SVM-Sigmoid kernel	69.3	Actual No	129	45
		Actual Yes	56	99
SVM-RBF Kernel	75.68	Actual No	151	23
		Actual Yes	57	98
Logistic Regression l1	74.16	Actual No	139	35
		Actual Yes	50	105
Logistic Regression l2	74.16	Actual No	139	35
		Actual Yes	50	105
Decision Tree Classifier	65.05	Actual No	121	53
		Actual Yes	62	93
Random Forest	73.25	Actual No	130	44
		Actual Yes	44	111
Gaussian NB	70.51	Actual No	147	27
		Actual Yes	70	85
Multinomial NB	64.74	Actual No	123	51
		Actual Yes	65	90
KNN	70.82	Actual No	145	29
		Actual Yes	67	88
Bagging Classifier	73.86	Actual No	138	36
		Actual Yes	50	105
Ridge Classifier	68.08	Actual No	131	43
		Actual Yes	62	93
MLP Claasifier	73.86	Actual No	138	36
		Actual Yes	50	105
Voting	75	Actual No	144	30
		Actual Yes	53	102

Figure 9: Results 2

Thus, with the view of the above results, we have used the type-2 case for the data processing and as from the validation score from the Selection Value, Random Forest as the brain behind the model. The algorithm can vary whenever a new patient details is fed into the system.

6. CONCLUSION

In the above procedure, we not only want to maximize the accuracy of the algorithm that we select to help the doctor take a decision rather, we want to decrease and penalize the model for having a bad prediction for the cases where the patient has a high probability for the heart attack but the model predicting for no heart problem. We hence stated one new metric called Selection Value which takes care of these scenarios and selects that algorithm which gives maximum S.V. We do not want to bias the doctor with the results of the classification rather as discussed in the proposed scenario section; we try to give the doctor with the better option with the history similar data results. Using these data, the doctor can have a transparency with the patient and the patient won't feel cheated at the end. With the more amounts of data being fed into the data base, the system will be very intelligent.

REFERENCES

[1] Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms, Sanjay Kumar Sen

[2] Peylan-Ramu, Nili, et al. "Abnormal CT scans of the brain in asymptomatic children with acute lymphocytic leukemia after prophylactic treatment of the central nervous system with radiation and intrathecal chemotherapy." New England Journal of Medicine 298.15 (1978): 815-818.

[3] Decramer, Isabel, et al. "Effects of sublingual nitroglycerin on coronary lumen diameter and number of visualized septal branches on 64-MDCT angiography." American Journal of Roentgenology 190.1 (2008): 219-225.

[4] Alkhorayef M, Babikir E, Alrushoud A, Al-Mohammed H, Sulieman A. Patient radiation biological risk in computed tomography angiography procedure. Saudi Journal of Biological Sciences. 2017;24(2):235-240. doi:10.1016/j.sjbs.2016.01.011.

[5] Diaz, Marco N., et al. "Antioxidants and atherosclerotic heart disease." New England Journal of Medicine 337.6 (1997): 408-416.

[6] Rodgers, Anthony, et al. "Blood pressure and risk of stroke in patients with cerebrovascular disease." Bmj 313.7050 (1996): 147.

[7] Gertler, Menard M., et al. "Ischemic heart disease." Circulation46.1 (1972): 103-111.

[8] Diamond, Joseph A., and Robert A. Phillips. "Hypertensive heart disease." Hypertension research 28.3 (2005): 191-202.

[9] Leander, Karin, et al. "Family history of coronary heart disease, a strong risk factor for myocardial infarction interacting with other cardiovascular risk factors: results from the Stockholm Heart Epidemiology Program (SHEEP)." Epidemiology 12.2 (2001): 215-221.

[10] US Department of Health and Human Services. "The health consequences of smoking: a report of the Surgeon General." (2004): 62.

[11] Hjermann, I., et al. "Effect of diet and smoking intervention on the incidence of coronary heart disease: report from the Oslo Study Group of a randomised trial in healthy men." The Lancet318.8259 (1981): 1303-1310.

[12] Collins, Rory, et al. "Blood pressure, stroke, and coronary heart disease: part 2, short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context." The Lancet 335.8693 (1990): 827-838.

[13] Wolf, Philip A., Robert D. Abbott, and William B. Kannel. "Atrial fibrillation as an independent risk factor for stroke: the Framingham Study." Stroke 22.8 (1991): 983-988.

[14] Miller, M. "Dyslipidemia and cardiovascular risk: the importance of early prevention." QJM: An International Journal of Medicine 102.9 (2009): 657-667.

[15] Haffner, Steven M., et al. "Reduced coronary events in simvastatin-treated patients with coronary heart disease and diabetes or impaired fasting glucose levels: subgroup analyses in the Scandinavian Simvastatin Survival Study." Archives of Internal Medicine 159.22 (1999): 2661-2667.

[16] Emerging Risk Factors Collaboration. "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies." The Lancet 375.9733 (2010): 2215-2222.

[17] Jee, Sun Ha, et al. "A coronary heart disease prediction model: the Korean Heart Study." BMJ open 4.5 (2014): e005025.

[18] Poirier, Paul, et al. ”Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss.” Circulation 113.6 (2006): 898-918.

[19] Ornish, Dean, et al. ”Can lifestyle changes reverse coronary heart disease?: The Lifestyle Heart Trial.” The Lancet336.8708 (1990): 129-133.

[20] Villareal, Dennis T., et al. ”Effect of lifestyle intervention on metabolic coronary heart disease risk factors in obese older adults.” The American journal of clinical nutrition 84.6 (2006): 1317-1323.

[21] Killip, Thomas, and Mary Ann Payne. ”High serum transaminase activity in heart disease.” Circulation 21.5 (1960): 646-660.

[22] Sowjanya, K., Ayush Singhal, and Chaitali Choudhary. ”MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices.” Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015.

[23] Pazzani, Michael J., and Daniel Billsus. ”Content-based recommendation systems.” The adaptive web. Springer, Berlin, Heidelberg, 2007. 325-341.

[24] Sagir, Abdu Masanawa, and Saratha Sathasivam. ”A Novel Adaptive Neuro Fuzzy Inference System Based Classification Model for Heart Disease Prediction.” Pertanika Journal of Science & Technology 25.1 (2017).

[25] Pattekari, Shadab Adam, and Asma Parveen. ”Prediction system for heart disease using Nave Bayes.” International Journal of Advanced Computer and Mathematical Sciences3.3 (2012): 290-294.

[26] Medhekar, Dhanashree S., Mayur P. Bote, and Shruti D. Deshmukh. ”Heart disease prediction system using naive Bayes.” Int. J. Enhanced Res. Sci. Technol. Eng 2.3 (2013).

[27] Peter, T. John, and K. Somasundaram. ”An empirical study on prediction of heart disease using classification data mining techniques.” Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on. IEEE, 2012.

[28] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. ”Prediction of heart disease using classification algorithms.” Proceedings of the world Congress on Engineering and computer Science. Vol. 2. 2014.

[29] Xing, Yanwei, Jie Wang, and Zhihong Zhao. ”Combination data mining methods with new medical data to predicting outcome of coronary heart disease.” Convergence Information Technology, 2007. International Conference on. IEEE, 2007.

[30] Ratnaparkhi, Devendra, Tushar Mahajan, and Vishal Jadhav. ”Heart Disease Prediction System Using Data Mining Technique.” International Research Journal of Engineering and Technology (IRJET) 2.08 (2015): 2395-0056.

[31] Sagir, Abdu Masanawa, and Saratha Sathasivam. ”A Novel Adaptive Neuro Fuzzy Inference System Based Classification Model for Heart Disease Prediction.” Pertanika Journal of Science & Technology 25.1 (2017).

[32] Sen, Ashish Kumar, Shamsher Bahadur Patel, and D. P. Shukla. ”A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level.” International Journal of Engineering and Computer Science 2.9 (2013): 1663-1671

AUTHORS

Nimai Chand Das Adhikari received his Master's in Machine Learning and Computing from Indian Institute of Space Science and Technology, Thiruvananthapuram in the year 2016 and did his Bachelor's in Electrical Engineering from College of Engineering and Technology in the year 2011. He is currently working as a Data Scientist for Philips Lighting (SS Supply Chain Solutions Pvt. Ltd.). He is a vivid researcher and his research interest areas include computer vision, health care and deep learning.

Arpana Alka received her Master's in Machine Learning and Computing from Indian Institute of Space Science and Technology, Thiruvananthapuram in the year 2017 and did her Bachelor's in Computer Science Engineering from National Institute of Technology, Surat in the year 2014. She is currently working as a Data Science Engineer for Busigence Technologies. Her interest areas include deep learning, video analytics, medical application and NLP.

Rajat Garg received his Bachelor's in Biotechnology Engineering from National Institute of Technology, Jalandhar and is currently working as Data a Scientist in Philips Lighting (SS Supply Chain Solutions Pvt. Ltd.). His interest areas include Machine Learning, Computer Vision and Data Analysis.

INTENTIONAL BLANK

SOFTWARE QUALITY IMPROVEMENT THROUGH STATISTICAL ANALYSIS ON PROCESS METRICS

Karuna Prasad, Divya MG, Sarat Chandrababu and Mangala N

C-DAC, Bangalore, Karnataka, India

ABSTRACT

Software Quality can be considered as totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs. The Quality of any software can be achieved by following by well-defined software process. These software process results into various metrics like Project metrics, Process metrics and Product metrics. Process metrics are very useful from management point of view. Process metrics can be used for improving the software development and maintenance process for defect removal and also for reducing the response time.

This paper describes on importance of capturing the Process metrics during the quality audit process and also attempts to categorize them based on the nature. To reduce such defect, corrective actions are recommended.

KEYWORDS

Software Metrics; ISO; Software Quality Audit; Process Metrics

1. INTRODUCTION

The quality of software is of utmost importance in the field of software engineering. Software quality also depends on the process which is carried out to design and develop the software. Even after the process is followed with minute care, the errors and defects may still exist. The quality of a software product is mainly determined by the quality of the process used to build it. Measurement and analysis will help in determining the status of the software process in terms of whether the process is followed and the functioning is as intended. Verification is the similar type of control from the management perspective. To meet such goals, quality audit for software process are conducted time to time. By measuring the errors and defects, we can take steps to improve the process.

The improvement of process will depend on metrics captured in the lifecycle of software. Software metrics can be classified into Project metrics, Process metrics and Product metrics [1]. Process metrics are management metrics which are used for improving the software development and maintenance process for defect removal and reducing response time of the process. Process metrics are invaluable tool for an organization who are wanting to improve their process. Usually

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 39– 48, 2017. © CS & IT-CSCP 2017 DOI: 10.5121/csit.2017.71804

these process metrics are not used mostly because of uncertainty about which metrics to use, how to perform measurements and how to overcome such defects.

For software process improvement, there are many models which are available for example Capability Maturity Model (CMM), Bootstrap, Personal Software Process (PSP), IT Infrastructure Library (ITIL), IEEE, Six Sigma and ISO 9000 quality management system. These models evaluate the software product, quality and their drawback. Moreover locally designed actions can be initiated in areas where improvement is needed. The software process must be defined and documented. In addition to the processes, standards for the different work products to be defined, e.g. coding and document standards.

The rest of this paper is organized as follows. In section II we have presented our approach and objectives. In section III we have presented the literature review which is basis of our work. In section IV quality process is explained, in next section categorization of errors and defects are presented. In section VI we have presented corrective actions. In section VII data collection methodology is explained. In section VIII results and the analysis are discussed. Finally we have provided conclusion in section IX.

2. APPROACH AND OBJECTIVE

In this paper we have applied statistical quality assurance to the errors and defects reported during the quality audit for the year 2015 and 2016 in our organization. This has been done in view to improve the quality of software development process and hence the software products. We are presenting that by measuring the errors and defects we can take actions to improve them. We are also presenting how each and every errors and defects are grouped. There after each of them is categorized with severity like minor, moderate or serious. The data collected over a period of two years has been analysed and presented. The analysis also describes recommended actions for the corrective action. The idea has been inspired from the software engineering practitioners Roger S Pressman and Bruce R Maxim [2].

Broadly we are trying to address 3 objectives namely quality improvement, categorizing of errors and recommendation of corrective actions.

3. RELATED WORK

In [3] the authors have presented the mechanism of how software engineering capabilities relate to the business performance. They have proposed a structural model including the Software Engineering Excellence indicator which consisted of deliverables, project management, quality assurance, process improvement, research and development, human resource development and customer contact.

In [4] the author has shared how NASA's Johnson Space Center developed a 'statistical method' to determine sample size for the number of process tasks to be audited by SQA. The goal of this work is to produce a high quality product which is cost effective.

In [5] authors have said that technological choices are fundamental for project planning, resource allocation, and quality of the final software product. For analysis they have taken open source web applications available in SourceForge. Authors aim to provide tools to support project

managers. They have said that there is need to set thumb rule to guide technological choices to increase the quality of software artifacts.

This paper [6] is related to software product quality modelling and measurement. The outcome of the research is grouped as system-level software quality models, source code element-level software quality models and applications of the proposed quality models.

Our work focuses on applying statistical quality assurance to improve the quality of software products.

4. QUALITY PRACTICES BEING FOLLOWED AT OUR ORGANIZATION

International Organization for Standards (ISO) is an independent body that provides requirements, specifications, guidelines, characteristics etc that can be used consistently to ensure that materials, products, processes and services are fit for their function. ISO International Standards ensure that products and services are reliable and of good quality. The technical committees are made up of experts from the relevant industry, but also from consumer associations, academia, NGOs and government [7].

ISO 9001:2008 standards set out the criteria for a quality management system and are the only standard in the family that can be certified to. It can be used by any organization, large or small, regardless of its field of activity. In fact is implemented by over one million companies and organizations in over 170 countries. This standard is based on a number of quality management principles including a strong customer focus, the motivation and implication of top management, the process approach and continual improvement [7].

Our organization is ISO 9001:2008 certified. For the development of software, ISO 9001 process is been followed. The ISO related activities are mainly carried out by the quality assurance team. The main role of quality assurance team is ensuring Quality Management System conformance, promoting customer focus, and reporting on Quality Management System performance. A quality manager is traditional employee who has been given this responsibility. Monitoring the quality objectives that have been established and reporting this to top management is another traditional role of the quality manager. Having one person focus on the management of this important activity is a good idea to provide focus and direction.

Quality manager is also responsible for internal audit planning & management. Internal audit is the disciplined approach to evaluate and improve the effectiveness of software quality processes. The scope of internal audit is mainly risk management, control and governance of software processes.

5. CATEGORIZATION OF ERRORS AND DEFECTS

Software metrics is a standard of measure of a degree to which a software system or process possesses some property. It can be classified into three categories: Project metrics, Process metrics and Product metrics [1]. Project metrics are those that describe the project characteristics and execution example resource requirement, hardware requirement etc. Process metrics are statistical software quality assurance (SQA) data or management metrics which are used for improving the software development and maintenance process. The Process metrics is usually

captured in the software quality audit process such as deviation form process, effectiveness of defect removal during development, propagation of error from phase to phase. Product metrics focus on the quality metrics of deliverables and are used to measure the properties of software like lines of code (LOC), defects/KLOC, defect density, customer satisfaction etc.

Process metrics is collected through the SQA audits. The error and defects so found are categorized in 12 types [2]. Most of the categories are self-explanatory however we have listed here them for the purpose of more clarity. All the errors and defects reported in "Auditor Note Sheet" are categorized as IID, IES etc depending upon the nature of error and defect.

1. Incomplete or erroneous specifications (IES) - Any specification incompletion is captured in this category. Any deviations from the process manual or specification like approval missing, partial implementation etc are included. If any missing metrics in the specification/template is also considered as IES.

2. Misinterpretation of customer communication (MCC) - Any deviation from customer requirement, feedback, suggestion etc not captured are categorized in this category.

3. Intentional deviation from specification (IDS) - IDS relates to deviation from process manual, software requirement specification etc due to lack of suitable reasons.

4. Violation of programming standards (VPS) - Any deviation from standards or introduction or modification can be counted in this category.

5. Error in data representation (EDR) - Any deviation from data formats as declared in specification.

6. Inconsistent competent interface (ICI) - Any deviation from recommended interface related errors.

7. Error in design logic (EDL) - Any deviation from committed logic eg DFDs, UML or ER diagram.

8. Incomplete or erroneous testing (IET) - Any errors and defects reported in testing by stakeholder/ customer/ third-party user etc. after completion of testing.

9. Inaccurate or incomplete documentation (IID) - Any missing sub sections of process manual or incomplete documentation.

10. Error in programming language translation of design (PLT) - Any design feature not captured while implementation which can cause defects in products.

11. Ambiguous or inconsistent human/computer interface (HCI) - Any error or defects in graphical user interface.

12. Miscellaneous - Any other errors and defect not captured in above mentioned categories.

All of the above categories are further classified based on the severity of the error/defects. They are labelled as minor, moderate and serious. It is classified as minor if the error/ defect not critical

to impact the process. Similarly, the defect is classified as moderate if the process is observed to be followed but cannot be evidenced. If the error or defect is observed to have major deviation from process then it is categorized as serious.

6. CORRECTIVE ACTIONS

For each of the error and defect categorized above, a corrective action is recommended as discussed below;

1. Incomplete or erroneous specifications (IES) - Effective Peer Review to be conducted.

2. Misinterpretation of customer communication (MCC) - Effective implementation of requirement gathering techniques to be adhered to improve the quality of customer communication and specification.

3. Intentional deviation from specification (IDS) - Reasons to be captured for intentional deviation and same to be reviewed.

4. Violation of programming standards (VPS) - Reason to be captured for intentional violation and same to be reviewed.

5. Error in data representation (EDR) - Recommend to use tools for data modelling also perform more stringent data design reviews.

6. Inconsistent competent interface (ICI) - Recommend more appropriate technical reviews and trainings.

7. Error in design logic (EDL) - Recommend more appropriate technical reviews and trainings.

8. Incomplete or erroneous testing (IET) - Recommend to adopt more appropriate testing methodologies with proper test plans.

9. Inaccurate or incomplete documentation (IID) - Recommend to use tools for documentation and reviews.

10. Error in programming language translation of design (PLT) - Cross reference with design requirements and appropriate tools usage to be recommended.

11. Ambiguous or inconsistent human/computer interface (HCI) - Graphical user requirement techniques and technology to be recommended.

7. DATA COLLECTION : A USE CASE

At C-DAC [8] quality audit is conducted every quarter. Audit is conducted for every project which is in design & development phase or maintenance phase. Quality assurance team rolls out the schedule with project name, auditee, auditor, date, time, and venue. With this auditee will keep ready all document and details required for audit. After the audit auditor will submit “Auditor Note Sheet” to quality assurance team. Auditor note sheet contains audit errors and

defects, if any. Quality assurance team publishes the entire "Auditor Note Sheet" in ISO related intranet web site where all C-DAC members have access to these Note Sheets.

Table 1. Error categorization for year 2015.

Error Type	Serious Errors	Moderate Errors	Minor Errors
MCC	0	0	0
IES	1	2	6
VPS	0	0	0
EDR	0	0	0
ICI	0	0	0
EDL	0	0	0
IET	0	0	0
IDS	3	1	0
IID	0	0	0
PLT	0	0	0
HCI	0	0	0
MIS	0	0	0
	4	3	6

Table 2. Error categorization for year 2016.

Error Type	Serious Errors	Moderate Errors	Minor Errors
MCC	0	0	0
IES	1	2	11
VPS	0	0	0
EDR	0	0	0
ICI	0	0	0
EDL	0	0	0
IET	0	0	0
IDS	2	1	0
IID	0	0	0
PLT	0	0	0
HCI	0	0	0
MIS	0	0	0
	3	3	11

For our experiment we have taken 2 years data namely Year 2015 and Year 2016. Based on our quality assurance guidelines of our organization these errors and defects are grouped as serious, moderate and minor which is described in section V. Also based on its nature every error or defect is categorized as IID, IES etc, same is recorded in the Table 1 and Table 2. Figure1 and Figure2 capture the severity of the errors thus categorized.

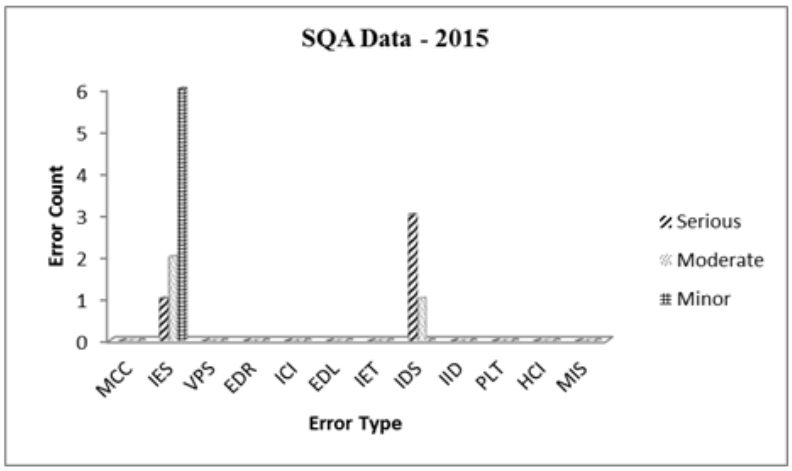


Figure 1- Severity of errors captured for year 2015

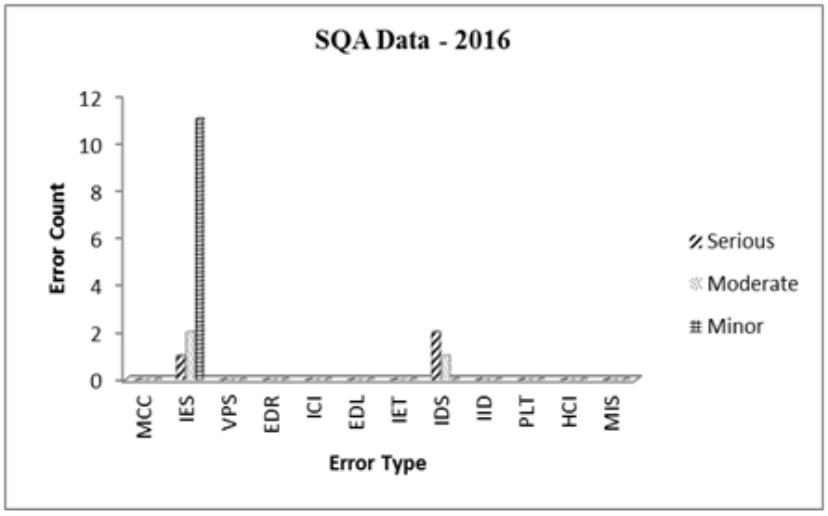


Figure 2- Severity of errors captured for year 2016

8. ANALYSIS AND RESULT

Every year three internal audits and one external audit's are conducted. Internal audit is conducted by Software Quality Assurance team of C-DAC, external audit is conducted by third party. During the audit, auditors will recode their observation, errors and deviations. This is termed as "Non Conformity- (NC)" in "Auditor Note Sheet" statement. We have collected all the NC's reported, same is categorized as per section V and grouped as serious, moderate and minor. From the analysis recorded at Table 1 & Table 2 the total errors and defects are presented in Table 3. The total serious, moderate and minor errors of both the years are represented in Table 4. Figure3 and Figure4 projects the cumulative errors for two years.

Table 3 – Cumulative errors for 2 years

Sl No	Year	Total errors
1	2015	13
2	2016	17

Table 4 – Severity of Cumulative Errors

Type of errors	Year 2015	Year 2016
Serious	4	3
Moderate	3	3
Minor	6	11

Figure 3 – Projection of errors

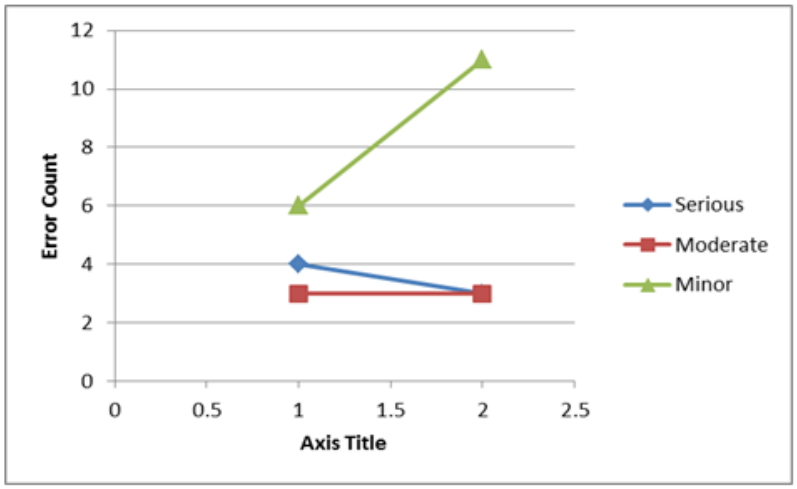


Figure 4 – Cumulative Projection of Severity errors

All the errors and defects are categorized and grouped mainly to know the statistics of software quality of projects. The data represented in Table 3 for the year 2015 is collected from 9 projects. The projects are either in design, implementation or maintenance states. The projects belong to various domains such as distributed computing, cryptography, high performance computing, Internet of things, mobile applications etc. These projects are implemented in programming languages java, c, python and other scripting languages. Some of these are using databases.

In Table 3, it is recorded that in year 2015 total error reported was 13. Out of which 4 are serious, 3 are moderate and 6 are minor type. The one serious error was due to Incomplete or erroneous specifications- effective 'peer review process' was recommended. Remaining 3 serious errors was due to Intentional deviation from specification – reason was WBS not updated, approval was not taken in time etc. All the causes of error was analyzed and training provided on quality process.

Also, there were 2 moderate and 6 minor errors due to 'Incomplete or erroneous specifications' and one more was due to 'Intentional deviation from specification'. In both the case effective peer review process and training on quality process was recommended. Similar analysis was carried for the year 2016.

The objective of the paper is to measure the errors and defects (non conformity) of all the projects, review it and recommend the appropriate corrective action. So that the project development cost will not over shoot, it can be delivered in time hence the quality of the project will improve. Hence software quality of products delivered by organization improves.

9. CONCLUSION

To improve the software quality, we collected software Process metrics. Our focus was mainly towards collecting metrics obtained through the quality control process. The errors and defects found through the software quality audits was our prime focus. These defects were subsequently categorized into 11 types. An analysis of such defects was conducted and recommendation for improving such defect and process are suggested. It was found that after implementing the recommendation the defects captured for the next subsequent year was reduced.

ACKNOWLEDGEMENT

We thank Centre for Development of Advanced Computing (C-DAC), the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY) for supporting us to carry this work. We thank the Ms Veena KS from Software Quality Assurance team,C-DAC, Bangalore for sharing data.

REFERENCES

[1] Ashwin Tomar and V. M. Thakare, "The Survey of Metrices on Software Quality Assurance and Reuse", National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013)

[2] Roger S. Pressman, Bruce R. Maxim, "Software Engineering: A Practitioner S Approach, Eighth Edition", 2015, McGraw-Hill Education.

[3] Yasuo Kadono, Hiroe Tsubaki, Seishiro Tsuruho. 2008. "A Study on Management of Software Engineering in Japanese Enterprise IT Industry", 978-1-4244-3397-1/08/$25.00 ©2008 IEEE

[4] Neera Bansal Talbert, Paramax Space System, Texas. “Representative Sampling within Software Quality Assurance”, 1063-677393 $3.00 0 1993 IEEE

[5] Valentino Sartori, Birhanu Mekuria Eshete, Adolfo Villafiorita. 2011. “Measuring the Impact of Different Metrics on Software Quality: a Case Study in the Open Source Domain”, ICDS 2011 : The Fifth International Conference on Digital Society, ISBN: 978-1-61208-116-8

[6] Peter Hegedus, University of Szeged, Hungary. 2015. “Advances in Software Product Quality Measurement and Its Applications in Software Evolution”, ICSME 2015, Bremen, German.

[7] www.iso.org

[8] www.cdac.in

[9] Chandramouli Subramanian, Chandramouli Seetharaman, B. G Geetha Saikat Dutt, “Software Engineering”, 2015, Pearson India Education Service

[10] Rajib Mall, “Fundamentals of Software Engineering”, 4th Edition, PHI Learning Private Limited

RUNWAY DETECTION USING K-MEANS CLUSTERING METHOD USING UAVSAR DATA

Ramakalavathi Marapareddy and Sowmya Wilson Saripalli

School of Computing, University of Southern Mississippi Hattiesburg, MS 39406-0001,USA

ABSTRACT

Remote sensing data gives the essential and critical information for detecting or identifying an object, a place, image fusion, change detection, and land cover classification of selected area of interest. The runway detection is an important topic because of its applications in military and civil aviation fields. This paper presents an approach for runway detection using Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVAR) data by implementing K-means clustering method. The obtained results reveal that we can obtain better detection, for the 9 and 11 classes, with iterations set to 10. In this work, the effectiveness of algorithm was demonstrated using quad polarimetric L-band Polarimetric Synthetic Aperture Radar(polSAR) imagery from NASA Jet Propulsion Laboratory's (JPL's) Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR). The study area is Louis Armstrong New Orleans International Airport, LA, USA.

KEYWORDS

Remote sensing, Runway detection, K-means clustering, polSAR

1. INTRODUCTION

Remote sensing is the acquisition of information about an object without making physical contact with the object and thus in contrast to on-site observation. Remote sensing is used in numerous fields, which includes geography, land surveying, military, intelligence, economic, planning, humanitarian applications, and so on. Remote sensing images contain large amount of geographical environmental information, giving new prospects in the field of the automatic detection of geospatial objects for multiple purposes [1]. Among these objects, runways have been the focus of consideration because of their significance in civil and military applications.

There are some literatures about remote-sensing imagery usage for detection and identification of airport runways in complex airport scenes, aerial optical imagery, and in synthetic aperture radar images. In main features of the runway, the most obvious feature is a straight line, so the runway target detection problem turns into, how to detect straight lines in the image. Generally, Hough transform was used to detect airport runway. The main advantage of the Hough transform is not sensitive to noise, better able to handle partial occlusion in the image and covering other issues. However, because it is a type of exhaustive search, so its computational complexity and space complexity is very high, which cannot meet the requirements of real-time systems [2]. One way to solve this would be to use cluster analysis. Cluster analysis is an unsupervised process of grouping observations (i.e., pixels) into classes or clusters, so that observations in the same class

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 49– 54, 2017. © CS & IT-CSCP 2017 DOI: 10.5121/csit.2017.71805

share more common features than to those in other classes. Various algorithms perform this process, one of them is K-means clustering algorithm, which is the most popular method. The main advantage of K-means is: If variables are huge, then K-means may be computationally faster than hierarchical clustering (if K is small).

In this paper, runway detection is done using polarimetric synthetic aperture radar (polSAR) imagery. The detection of polSAR image data is performed to compute and analyze runways using K-means unsupervised clustering algorithm.

2. METHOD

2.1 OVERALL PROCEDURE

The proposed methodology for the detection of Runway is shown in Figure 1, is based on image data collected from polSAR format. Subset images or select region of interest (ROI) to detect the structural data of a runway. Basic filter application is performed to extract better image visualization and analyzation. And then K-means clustering is performed by taking classes and iterations as parameters, and this process is used for the detection of the runway.

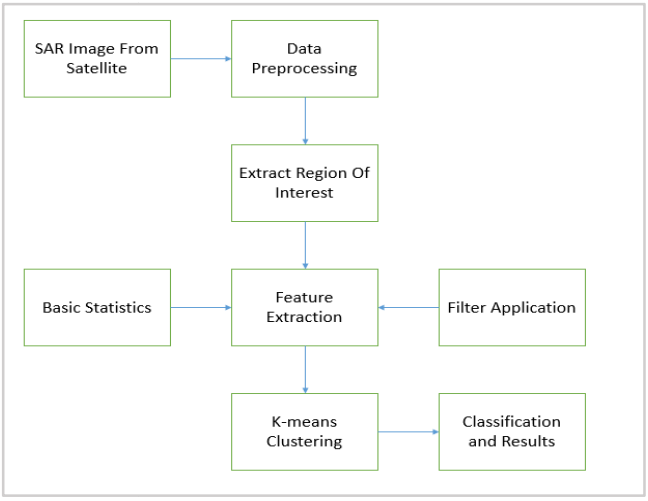


Figure 1. Overall procedure for the detection of Runways.

2.2 REGION OF INTEREST (ROI)

UAVSAR is a sensor that captures polSAR data in different polarizations. UAVSAR data is available in cross-polarized (HHHV, HHVV, HVVV) and co-polarized (HHHH, HVHV, VVVV) we have selected a fully cross-polarized image HVVV to perform our research work [3]. SAR polarimetry using quad-polarization data is the HV-polarization base in which an antenna transmits and receives horizontally and vertically polarized and different polarizations of the backscatter signal are detected as: VV (vertical transmit and vertical receive), HV (horizontal transmit and vertical receive), and HH (horizontal transmit and horizontal receive).

The Keyhole Markup Language (KML) file is used to find the test area from a satellite image, Keyhole Markup Language Zipped (KMZ) files represent the ground projected data, is used to display on Google earth map, as shown in Figure 2.

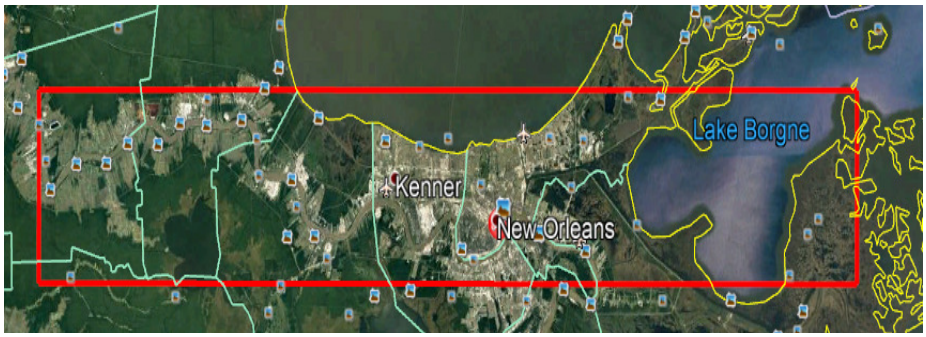


Figure 2. KML file of UAVSAR data on google earth

ROI are selected samples of a raster, such as areas of water that are identified for a particular purpose [4]. After creating ROI in polygon shape, we sub set the data and mask the pixels outside the ROI with value 0 by using an image detection software. Equalization filter and Contrast up to 70% are applied to have a better image detection. ROI masking pixels outside region is a polygon which covers the runway, as shown in the figure 3.

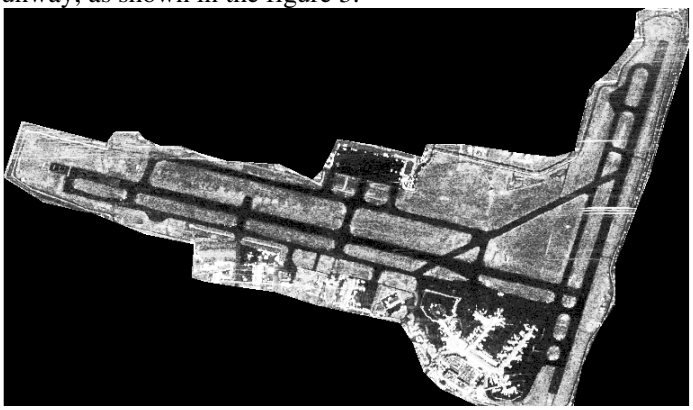


Figure 3: Region of Interest

2.3 K-MEANS CLUSTERING

K-means is a clustering method that aims to find the positions of the clusters that minimize the distance from the data points to the cluster [5]. Let X = {x_i}, I = 1, . . , n be the set of n d-dimensional points to be clustered into a set of K clusters, C = {c_k, k = 1, . . . , K}. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of cluster c_k. The squared error between μ_k and the points in cluster c_k is defined as

$$J(c_k) = \sum_{x_i \epsilon c_k} \|x_i - \mu_k\|^2$$

K-means will minimize the sum of the squared error over all the K clusters resulting,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon c_k} \|x_i - \mu_k\|^2 \quad [6]$$

The reason for choosing K-means algorithm is due to its popularity for the following reasons:

1. Its time complexity is O(nkl), where n is the number of patterns, k is the number of clusters, l is the number of iterations taken by the algorithm to converge [7].

2. If variables are huge, K-means is faster computationally, then hierarchical clustering, keeping k small [8].

3. K-means produces tighter clusters than hierarchical clustering [8].

3. RESULTS AND DISCUSSION

The goal of this research is to detect the runway of the airport using polarimetric SAR data by applying K-means clustering with different classes and iterations. Detecting the runways from satellite and aerial images is a complicated task, but this data can be analyzed by clustering. The runways are uniform, they have a gray level and this valid feature is used to distinguish runways from other landforms.

In this research, we will define number of classes, number of iterations, and set the threshold value to 5, the change threshold is used to end the iterative process when the number of pixels in each class changes by less than the threshold. K-means unsupervised classification calculates initial class means evenly distributed in the data space, then iteratively clusters the pixels into the nearest class using a minimum distance technique. Each iteration recalculates class means and reclassifies pixels with respect to the new means. All pixels are classified to the nearest class unless a standard deviation or distance threshold is specified, as we set threshold to 5, this process continues until the number of pixels in each class changes by less than the selected pixel change threshold that is 5 or the number of iterations is reached [4]. The following figures show how we used K-means in detection of Runway, using different classes and iterations: Figure 4 shows K-means classification with 5 class and with (a) Iterations 1 (b) Iterations 10 (c) Iterations 100 (d) Iterations 1000. Figure 5 shows K-means classification with 7 class and with (a) Iterations 1 (b) Iterations 10 (c) Iterations 100 (d) Iterations 1000. Figure 6 shows K-means classification with 9 class (a) Iterations 1 (b) Iterations 10 (c) Iterations 100 (d) Iterations 1000. The image of 9 class with 10 iterations shows clear runway, and gives the good detection compared to 5 class and 7 class. Figure 7 shows K-means classification with 11 class (a) Iterations 1 (b) Iterations 10 (c) Iterations 100 (d) Iterations 1000. 11 class with iterations 10 shows good runway detection along with 9 class with iterations 10.

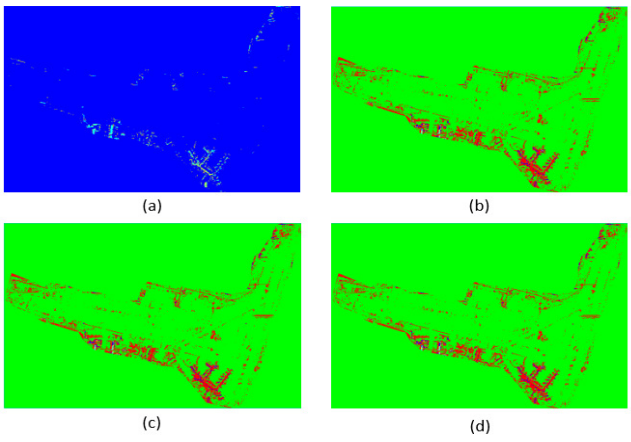


Figure:4 K-means classification with 5 classes, with Iterations: (a) 1 (b) 7 10 (c) 7 100 (d) 1000

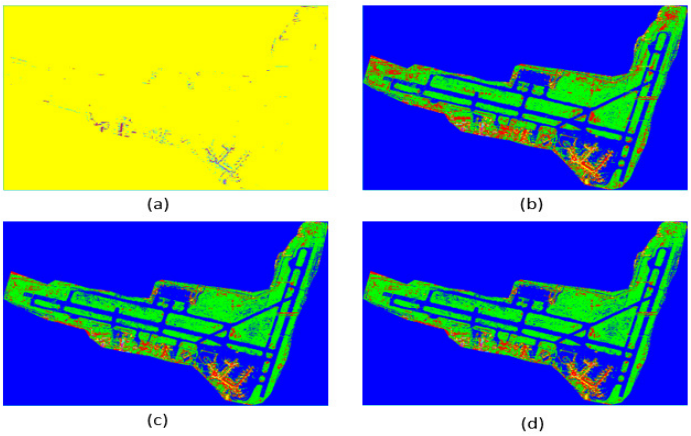


Figure:5 K-means classification with 7 classes, with iterations: (a) 1 (b) 7 10 (c) 7 100 (d) 1000

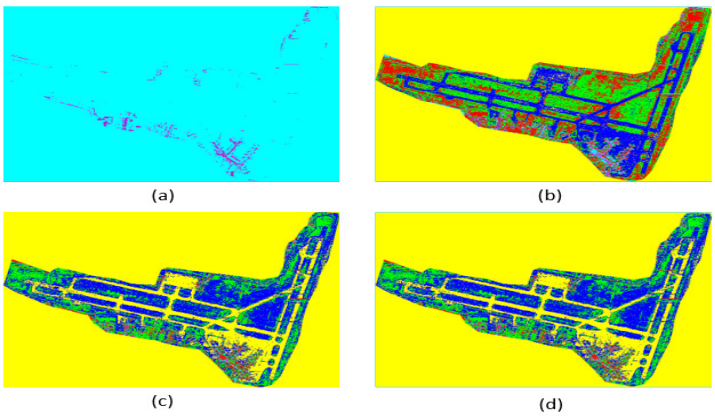


Figure:6 K-means classification with 9 classes, with iterations: (a) 1 (b) 7 10 (c) 7 100 (d) 1000

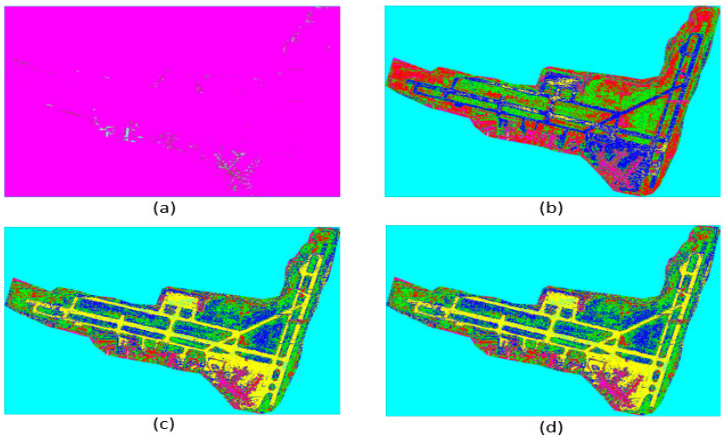


Figure:7 K-means classification with 11 classes, with iterations: (a) 1 (b) 7 10 (c) 7 100 (d) 1000

We have experimented with classes ranging from 5 to 11 and with iterations 1, 10, 100 and 1000. For each class varying the iterations, we observed that the image is better analyzed (has good resolution) when the iterations is equal to 10, for iterations below and above 10, clustering is not

that great. And especially for 9 & 11 classes, with iterations 10, gives the best image showing the runway of the airport.

4. CONCLUSION

This paper explains an approach for runway detection using remote sensing images by implementing K-means clustering classification. The K-means algorithm has been implemented on quad polarimetric L-band polSAR image from NAS JPL's UAVSAR. The study area is Louis Armstrong New Orleans International Airport, Louisiana, USA. We worked with classes 5, 7, 9 and 11, and with iterations from 1, 10, 100 and 1000.The obtained results show that, we have better detection of runways when we take 9 & 11 classes and iterations as 10. For iterations equals to 1, we observed that classification ends after 1 iteration, irrespective of the threshold value taken. And for iterations 100 and 1000, we observed the increased number of pixels reaching the threshold value and only the pixels that do not reach the threshold value continues until number of iterations is reached or until the number of iterations is completed.

ACKNOWLEDGEMENT

We thank our team and colleagues for their suggestions and advices. We also thank Alaska Satellite Facility (ASF) and NASA Jet Propulsion Laboratory for imagery.

REFERENCES

[1] Bala, P., Tom, S., and Shinde, R., "GIS and Remote Sensing in Disaster Management," Imperial Journal of Interdisciplinary Research(IJIR), vol.3, no.5, 2017.

[2] ZhuZhong Yang., JiLiu Zhou., and FangNian Lang., "Detection Algorithm of Airport Runway in Remote Sensing Images," TELKOMNIKA Indonesian Journal of Electrical Engineering, vol.12, no.4, pp.2776-2783.

[3] Dataset: UAVSAR, NASA 2011. Retrieved from ASF DAAC 7 /, (accessed on October 2017).

[4] http://www.harrisgeospatial.com/docs/, (accessed on September 2017).

[5] http://www.onmyphd.com/?p=k-means.clustering, /, (accessed on October 2017).

[6] Jain, K., A., "Data clustering: 50 years beyond K-means," In Pattern Recognition Letters, vol. 31, no. 8, 2010, pp. 651-666, October 2017.

[7] Sarada, W., and Kumar, P., V., "A Review on Clustering Techniques and Their Comparison," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol.2, no.11, November 2013.

[8] http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm, /, (accessed on October 2017).

AUTHORS

Ramakalavathi Marapareddy (Kala) received Ph.D. from Mississippi State University (MSU), in 2015, in Electrical and Computer Engineering (ECE). BS & MS degrees from Jawaharlal Nehru Technological University Hyderabad, in 2000 & 2003, respectively, both in ECE. At present, she is working as an Assistant professor, at School of Computing, The University of Southern Mississippi.

Sowmya Wilson Saripalli is masters student in computer secinece at School of Computing (SoC), The University of Southern Mississippi (USM).

AUTHOR INDEX