Vol. 14, No. 1 April, 2019

ISSN 2091-1742

nec
A BRAND IN QUALITY ENGINEERING EDUCATION

SCITECH Nepal

"Journal of Scientific and Technical Studies"

Special Issue
Silver Jubilee Celebration of Nepal Engineering College

Nepal Engineering College
Changunarayan-4, Bhaktapur
Phone: 01-5221006, Fax: 01-5221001
(DAO Bhaktapur Regd No. 552-2060)

SCITECH Nepal

• April 2019 • Vol. 14, No. 1

Special Issue

Silver Jubilee Celebration of Nepal Engineering College

(DAO Bhaktapur Regd No. 552-2060)

Printed at:
Print and Art Service
Bagbazar, Kathmandu
Tel: 01 4244419, 4239154

From the Editor's Desk

Consistency and regularity are virtues. With the active support of the researchers-authors and conviction of the editorial members, Nepal Engineering College is proud to publish this issue of the SCITECH-Nepal Journal; this issue demonstrates regularity of the publication and consistency of the quality of the research papers. With on-line open access, international recognition and the increased visibility of this journal, we all are on the right tract.

The previous issue of the SCITECH-Nepal Journal was published about six months ago. The management aspect of the Nepal Engineering College has enhanced noticeably in the last six months, and the academic-research staffs of the college are now freer than ever to openly dedicate themselves in research activities. The sense of ownership of the college by the teaching and non-teaching staff has enhanced the research and academic environment in the last six months.

The open and official recognition of the Nepal Engineering College as a non-profit social institute by the Parliamentary Committee on Education and Health, the District Administration Office of Bhaktapur, Changunarayan Municipality, Nepal Engineering Council, Pokhara University, Ministry of Education, Science and Technology and finally by the Supreme Court of Nepal has further encouraged all the well-wishers of the College to dedicate themselves ever more in the development of the College as a sound academic and research institute. With this changed environment in the College, I am confident on the consistent rise in the quality of the published papers in the SCITECH Nepal Journal.

Interdisciplinary nature of the engineering and architecture education and research are well represented in the research papers in this issue. We vow to maintain and upgrade this tradition and also increase the accessibility of the journal and expand the network among the research paper authors, with an ultimate goal to have an impact of the published papers on the policy decisions and solution of everyday problem of an ordinary person.

Hari K Shrestha

Prof. Dr. Hari Krishna Shrestha
Chief Editor

Contents

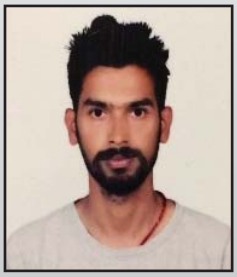
Optimal Allocation of Capacitor Bank for Loss Minimization and Voltage Improvement Using Analytical Method

Avinash Khatri KC[*], Tika Ram Regmi[+]

[*]Nepal Engineering College, Department of Electrical and Electronics Engineering, avinashkhatrikc@gmail.com

[+]Nepal Engineering College, Department of Electrical and Electronics Engineering, tikabikramregmi@gmail.com

Avinash Khatri KC *has received his Master Degree in Planning and Operation of Energy System from Kathmandu University. He is currently working at Nepal Engineering College as Jr. Prof. in Department of Electrical and Electronics Engineering. His research interests are hybrid energy, power system and artificial intelligence.*



Tikaram Regmi *has received his Master Degree in Electrical Power Engineering from Kathmandu University. He is currently working at Nepal Engineering College as Jr. Prof. in Department of Electrical and Electronics Engineering. His research interests are renewable energy, power system and control system.*

Abstract

An electric distribution system plays an important role in achieving satisfactory power supply. The quality of power is measured by voltage stability and profile of voltage. The voltage profile is affected by the losses in distribution system. As the load is mostly inductive on the distribution system and requires large reactive power, most of the power quality problems can be resolved with requisite control of reactive power. Capacitors are often installed in distribution system for reactive power compensation. This paper presents two stage procedures to identify the location and size of capacitor bank. In the first stage, the load flow is carried out to find the losses of the system using sweep algorithm. In the next stage, different size of capacitors are initialized and placed in each possible candidate bus and again load flow for the system is carried out. The objective function of the cost incorporating capacitor cost and loss cost is formulated constrained with voltage limits. The capacitor with the minimum cost is selected as the optimized solution. The proposed procedure is applied to different standard test systems as 12-bus radial distribution systems. In addition, the proposed procedure is applied on a real distribution system, a section of Sallaghari Feeder of Thimi substation. The voltage drops and power loss before and after installing the capacitor were compared for the system under test in this work. The result showed better voltage profiles and power losses of the distribution system can be improved by using the proposed method and it can be a benefit to the distribution networks.

Keywords : *quality; voltage stability; voltage profile; capacitor cost; voltage limits.*

I. INTRODUCTION

In context of Nepalese distribution system, the distribution lines are extended over long distances to feed loads scattered over large areas. Thus, the primary and secondary distributions lines in rural areas are largely radial laid usually extend over

long distances. This results in high line resistance and therefore high I^2R losses in the line. Not only significant loss due to the lengthy feeder, there will be the abrupt decrease in the voltage in the load end as the distance between source and load increases. Therefore load situated at the remote places suffers low voltage problems which arises different issues regarding power quality [1].

A major problem in a power distribution system is a voltage drop due to high-load, very long power distribution systems, and rapid load growth in a radial network. Voltage regulation is the primary constraint in the planning and sizing of rural distribution networks. The allowable voltage regulation limits and the assumptions adopted by distribution network planners and designers significantly affect the capital and lifetime costs of system. The problem of the losses and voltage drop in distribution feeders are dependent on each other and varies with the pattern of loading on the feeder. The distribution losses can be reduced by proper selection of distribution transformers, feeders, proper re-organization of distribution network, placing the shunt capacitor in appropriate places, and penetration of distributed generations at different point of distribution networks [2].

Capacitors are widely used in distribution systems for voltage stabilization and reactive power compensation as well as power/energy reduction, voltage regulation and system capacity release. The extent to which loss is minimized and voltage improves depends upon the proper selection of the size of capacitor bank and the position at which it is located [3].So, in order to maximize the benefits oriented from capacitor bank placement, the proper selection of the position and size is desirable.

II. LITERATURE REVIEW

A backward/forward sweep method is used to carry out the load flow of the radial distribution system. This sweep algorithm consists of two sets recursive equations. The first set of equations is used for the determination of the branch currents starting from the last branch and then proceeding to the root node in backward direction. The other equations to calculate the voltage magnitude and angle of each bus starting from central bus and moving towards the last bus in forward direction.

In backward sweep the branch current or the power flow is calculated starting from the branches in the last layer and moving towards the branches connected to the root node. The updated effective power flows in each branch are obtained in the backward propagation computation by considering the node voltages of previous iteration. It means the voltage values obtained in the forward path are held constant during the backward propagation and updated power flows in each branch are transmitted backward along the feeder using backward path. This indicates that the backward propagation starts at the extreme end node and proceeds towards source node [4].The forward sweep is basically for the voltage drop calculation. Bus voltages are updated in a forward sweep starting from branches from starting bus towards last bus. The central bus or source bus is set at its rated value. During the forward propagation the effective power in each branch is held constant to the value obtained in backward flow [5].

A number of optimization methods have been developed for solving different types of optimization problems. A two stage process to determine the optimum size and location of capacitor banks to achieve maximum energy savings [6]. Loss sensitivity factor (LSF) is calculated in the first stage to find the optimum location and in second stage of the method an analytical procedure is presented to determine the size of the capacitor. While considering the maximum energy savings, the installation and operation costs of the capacitor banks are also investigated. In order to reduce the loss of distribution radial feeder a capacitor bank allocation is done [7]. A load flow process is carried out before and after the placement of capacitor banks. The buses with the voltage level below the 0.95 pu are classified as weak buses. The different available size of capacitors is place at each weak buses and load flow is carried out. Then different cases are analyzed. Based on the lowest power loss, the optimal size and location for the capacitor bank is suggested. An analytical based solution model is proposed for capacitor placement to ensure power quality [6]. This method assumed that there are several identical capacitors are available at each and every bus and they can be placed one unit per bus at a time.

At every iteration run, a capacitor is added to the possible candidate bus and each time the value of objective function is evaluated. Operational and equipment constraints are also checked. The unit that reduces the overall objective function the most, without causing any limit violations, is chosen and placed permanently at the respective bus.

In recent years, the optimization process has been modified from traditional mathematical programming techniques called as modern or non-traditional optimization methods. Among them, artificial intelligent (AI) is widely used. AI such as artificial neural network (ANN), genetic algorithm (GA), simulated annealing algorithm (SA), ant colony Optimization (ACO) and particle swarm optimization (PSO) are widely used optimization methods in recent years. A novel emerging intelligence based on particle swamp optimization (PSO) for optimum selection of distribution capacitors is presented in [8]. The solution search algorithm provided a global optimal solution for the capacitor bank selections. To overcome the limitations of PSO, hybrid algorithms with genetic algorithm (GA) [8]. The basis behind such a hybrid approach was to have merits of PSO with those of GA. The purpose of applying mutation to PSO is to increase the diversity of the population and the ability to have the PSO to avoid the local maxima.

III. METHODOLOGY

A. Load Flow Analysis

Load flow process is vital to the power system. It gives the steady state solution of the network as well as helps in the continuous monitoring of the running state of system. Generally, the R/X ratio of the distribution system is very high resulting in the high voltage drop and power loss in the system. Due to this the conventional load flow algorithm such as Newton-Raphson and Fast Decoupled load flow methods may provide inaccurate results and also may fail to converge. In this work, in order to overcome convergence problem, a backward/forward sweep algorithm in [9] is used to carry out load flow of the distribution system.

B. Optimum Allocation of Capacitor by Analytical Method

The methodology based on analytical method discussed in the Fig. 1 determines the value of capacitor based on the cost associated with the active power losses and installation cost of capacitor bank in the distribution network. The process starts with normal load flow process. After carrying out load flow process the capacitor size ($\mathbf{Q_{VAR}}$) is randomly initialized between minimum and maximum value i.e. the minimum value is defined as zero and maximum value of capacitor is defined to be two times of the total reactive load of the system $\mathbf{Q_{load}}$.

$$\mathbf{0 < Q_{VAR} < 2Q_{load}}$$

The voltage of the bus is kept as constraint to be satisfied. Constraint defined for voltage regulation limit is considered from rule 40 [10] of Electricity Rules, 2050 (1993) and IEEE standard [11] i.e.

$$\mathbf{0.95 < V_{bus} < 1.05}$$

The objective function of this research work is to minimize the total cost base on the power losses and capacitor installation cost. The cost is divided into two parts.

i) Power loss Cost &
ii) Capacitor Cost

a. Cost of Losses

The power loss of the branch is calculated by line current I and branch resistance R as $\boldsymbol{P = I^2R}$. The total power loss is evaluated by adding all the line

$$\textbf{Total power loss} = \sum_{i=1}^{b} P_i$$

losses, i.e.

Where, b number of branches.

To evaluate the cost of the losses, the following data is considered. The power loss given by the load flow process is peak power loss. The average power loss is calculate by using loss of load factor (**LLF**), [13]

LLF = Average Loss / Peak Loss,

Therefore, the average loss is calculated as the peak power loss multiplied by LLF. The average power loss when multiplied by factor of 8760 gives the annual energy loss. The total cost of energy lost is calculated as given by the following equation.

Annual Total Loss Cost = LLF * Total Active Power Loss * Unit Energy Cost * 8760

b. Capacitor Cost

The cost of capacitor depends on the size of capacitor bank to be installed. Though cost of per unit size of capacitor varies from size to size i.e. per unit cost of capacitor of lower sized capacitor banks are more compared to per unit cost of large capacitor banks [6]. The cost for installation of capacitor bank in the distribution network is given as,

$$\text{Capacitor Cost} = k_c * Q_{VAR}$$

Where, k_c is the unit cost of capacitor \$/kVAR,

Q_{VAR} is size of capacitor to be installed in kVAR.

The life of capacitor bank installed is assumed to be for 10 years. So, the cost of capacitor installed per annum is given by annuity factor A, which is calculated as,

$$A = \frac{r}{1-(1+r)^{-n}}$$

Where, **r** is the rate of return on investment and **n** is number of years. Assuming the r = 10% and life of capacitor bank to be 10 years [13], thus annual Capacitor cost is given as,

Annual Capacitor Cost = A* Capacitor Cost

c. Objective function

The objective function is defined as,

Minimize, F= **Annual Total Loss Cost + Annual Capacitor Cost**

The value of objective function is calculated for each size of solution without violating the constraint limits. The solution with the minimum value of total value of objective function will be the optimized solution. The optimized solution gives the minimum system loss under voltage constraints with the size and bus number where the capacitor bank is to be placed.

C. Flowchart for optimum selection.

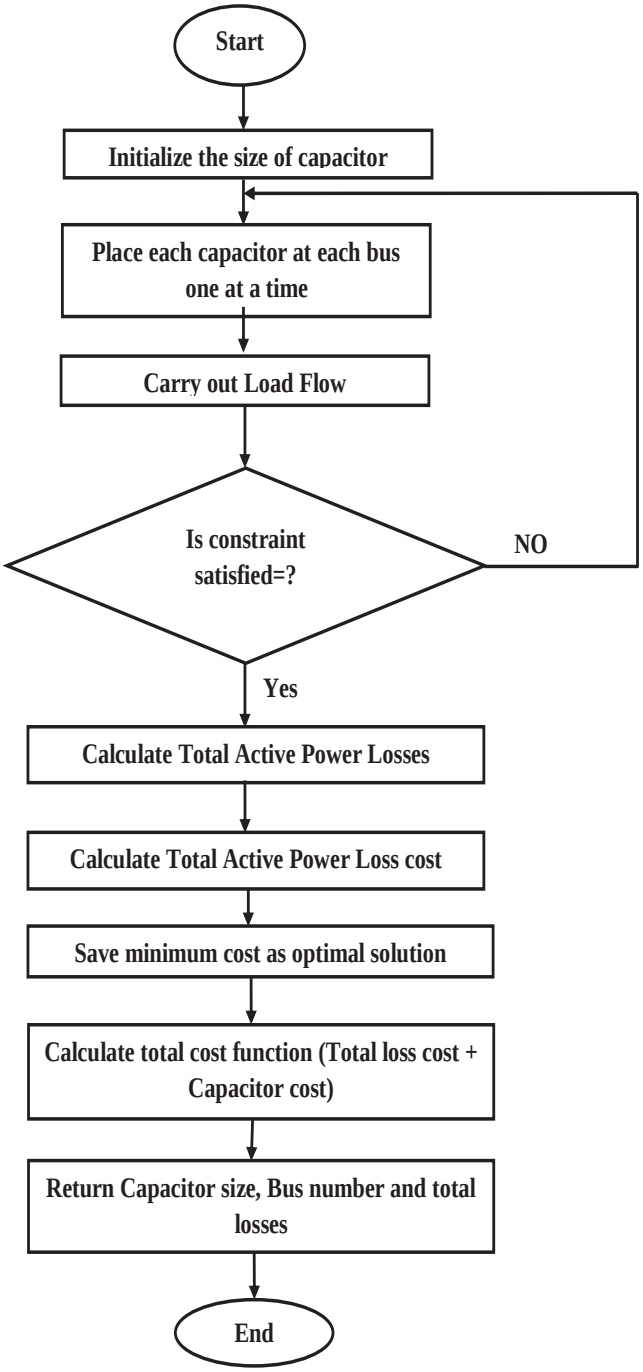


Fig.1 Flowchart for optimum size and location of capacitor

IV. RESULTS

The proposed method for optimal capacitor allocation for loss reduction was tested in the IEEE standard 12 bus radial distribution feeders and implemented on real radial distribution feeder of Nepalese system to calculate the total annual cost savings.

a. IEEE-12 bus radial distribution system

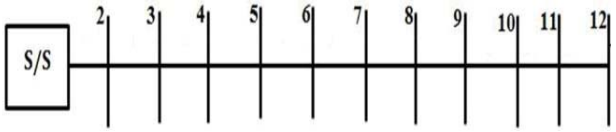


Fig.2 Single line diagram of IEEE-12 bus system

The proposed method for capacitor placement was implemented to a 12-bus radial distribution test system. The single-line diagram of the 11 kV, 12-bus system is illustrated in Figure 2. Using this method, the capacitors of rating 233 kVAR to be placed at the optimal location bus number 9. The initial power loss was 20.3 kW and it reduced to 11.86 kW after capacitor placement using the proposed method. The results of the proposed method are shown in Table 1.

Table 1 Voltage of different buses of IEEE 12 Bus with and without Capacitor Bank

Bus Number	Voltage (pu)	
	Without Capacitor Bank	With Capacitor Bank
1	1	1.0000
2	0.9944	0.9953
3	0.9891	0.9911
4	0.9808	0.9846
5	0.9702	0.9767
6	0.9670	0.9745
7	0.9643	0.9726
8	0.9537	0.9653
9	0.9505	0.9639
10	0.9477	0.9611
11	0.9469	0.9602
12	0.9467	0.9600

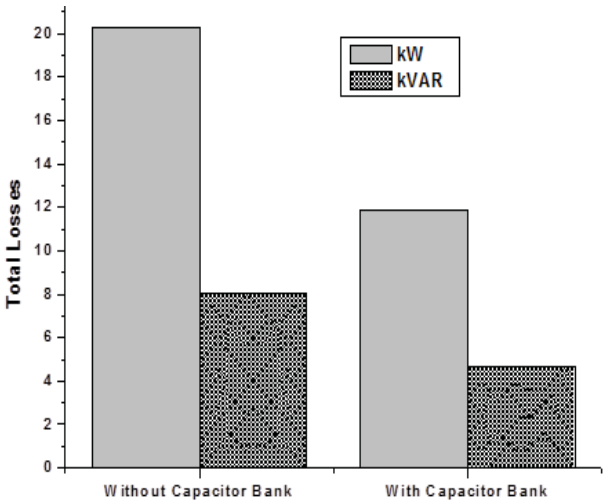


Fig.3 Total Losses of the system with and without Capacitor Bank

b. Nepalese Radial Distribution System

The figure 3 shows the single line diagram of section of Sallaghari feeder radial distribution feeder. The feeder consists of 11 buses. The system consists of two laterals from bus 2. The description of the buses is provided in the table 4.6. The feeder is supplied from Thimi through 11kV lines at 100 MVA base. The feeder section consists of 3 transformers of Nepal Electricity Authority (NEA) and seven private transformers. The feeder supplies to the residential load as well as for the industries. The power factor of the transformer supplying to the residential area was considered to be 0.8 whereas it was kept 0.6 for industrial areas. The loads on the feeders were lumped at 11 different places. The total load in the system were 1103kW and 1116kVAR.

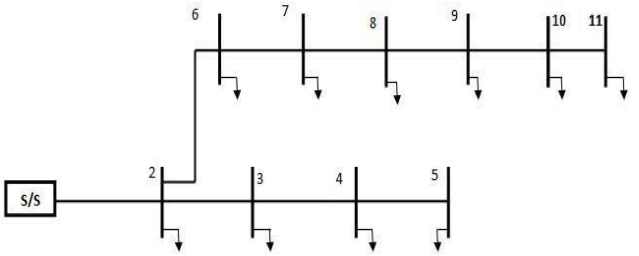


Fig. 4 Single line diagram for Sallaghari feeder

After running optimization program for the system, the required size of capacitor bank is found to be 907 kVAR and the optimum location, bus 7. The initial power loss was 55.23 kW and it was reduced to 31.32kW after capacitor placement using the proposed method. The results of the proposed method are shown in Table 2.

Table 2 Voltage of different buses of IEEE 12 Bus with and without Capacitor Bank

Bus Number	Voltage in pu	
	Before Capacitor	After Capacitor Placement
1	1.0000	1.0000
2	0.9686	0.9818
3	0.9627	0.9761
4	0.9587	0.9720
5	0.9563	0.9697
6	0.9606	0.9795
7	0.9524	0.9780
8	0.9490	0.9747
9	0.9463	0.9721
10	0.9441	0.9700
11	0.9433	0.9692

The load flow analysis gave the voltage profile of the system under study and different losses occurring at different segments of the system. The voltage was minimum at bus 11 which was equal to 0.9433pu and in a similar fashion the voltages at bus 8,9 and 10 were also below acceptable limits. The total loss of the system was equal to 55.23 kW and 51.68 kVAR.

The voltages at bus 8, 9, 10 and 11 which were below 0.95 pu previously improved above 0.95 pu after the introduction of capacitor bank at bus 7. The comparison of voltages at different buses before and after capacitor bank is represented by Fig.4.

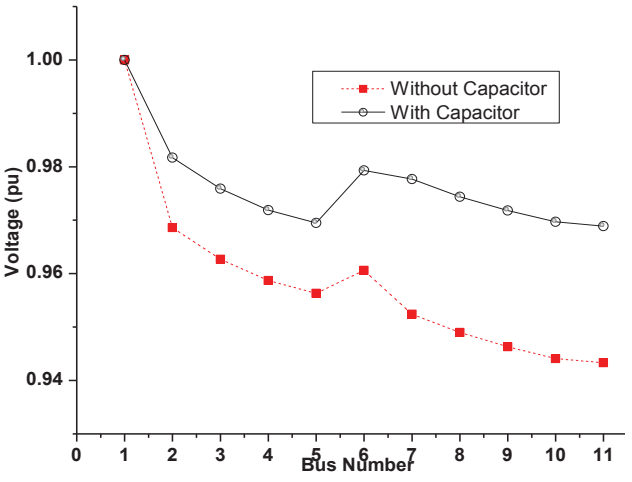


Fig.5 Voltage profile of feeder with and without Capacitor

Table 3 Summary of results before and after capacitor placement

		Base Case (Without Capacitor)	With Capacitor
Voltage (pu)	Min	0.9433	0.9686
	Max	0.9689	0.9817
Total Loss	Active(kW)	55.23	31.32
	Reactive (kVAR)	51.68	29.3
Capacitor	Size (kVAR)	-	907
	Location (Bus Number)	-	7
Loss Reduction	Active Power (kW)	-	23.91
Total Loss Cost($)		24190.74	13718.16
Capacitor Cost($)		-	72.56
Total Annual Cost($)		-	13790.72
Net Annual Savings ($)			10400.02

Table 3 summarizes the result for the load flow of section of Sallaghari feeder. The results show that the minimum voltage of the feeder was 0.9433 pu at bus 11 and maximum value of voltage was 0.9689 pu at bus 2. After the installation of the capacitor bank the minimum and maximum voltages were found to be 0.9686 and 0.9817 per unit at bus 11 and bus 2 respectively. Similarly, there is significant reduction in active as well as reactive power loss. The graph showing the comparison of loss with and without capacitor placement is shown in Fig.5 below. The installation of capacitor at the proposed location shows the annual profit of $10,400.

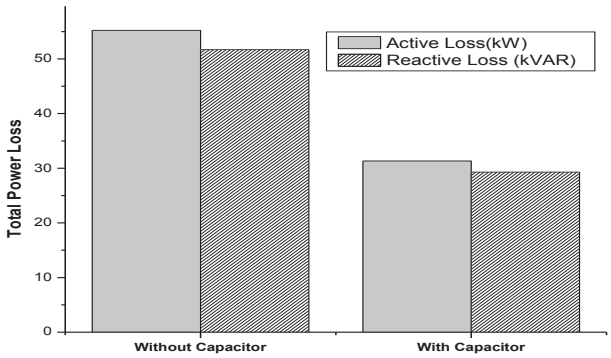


Fig.6 Total Losses with and without Capacitor Bank

V. CONCLUSIONS

Capacitors are often installed in distribution system for reactive power compensation to carry out power and energy loss reduction. Size and location are crucial factors in the application of capacitor for loss minimization. The appropriate placement of capacitors is also important to ensure that system power losses and total investment capacitor costs are minimal. The voltage drops and power loss before and after installing the capacitor were compared in this work. The result showed better voltage profiles and power losses of the distribution system could be improved by using the proposed method and it can be a benefit to the distribution networks.

VI. RECOMMENDATION

The approach described in this research work considers only the optimal size and location of capacitor bank; objective function can be formulated to determine the optimal number of locations for capacitor bank along with size and position. Future work will be dedicated to incorporate the different types of load models, load growth and multi-load level.

References

[1] K. V. S. R. Murthy and M. R. Raju, "Electrical Energy Loss in Rural Distribution Feeders- a Case Study," *Power*, vol. 12, no. 2, pp. 14–1, 2007.

[2] A. A. Abou El-Ela, M. T. Mouwafi, A.-M. Kinawy, and R. A. El-Sehiemy, "Optimal capacitor placement in distribution systems for power loss reduction and voltage profile improvement," *IET Gener. Transm. Distrib.*, vol. 10, no. 5, pp. 1209–1221, 2016.

[3] A. . S. Allw, M. S. Almurieb, and Z. A. Alamidy, "Reduction Power losses on distribution feeders in Electrical Distribution Systems."

[4] S. Sunisith and K. Meena, "Backward / Forward Sweep Based Distribution Load Flow Method," vol. 5, no. 9, pp. 1539–1544, 2014.

[5] J. Rupa and S. Ganesh, "Power Flow Analysis for Radial Distribution System Using Backward/ Forward Sweep Method," World Acad. Sci. Eng., vol. 8, no. 10, pp. 1621–1625, 2014.

[5] P. V. Babu and S. P. Singh, "Capacitor allocation in radial distribution system for maximal energy savings," *2016 Natl. Power Syst. Conf.*, no. February 2017, pp. 1–6, 2016.

[6] K. Tilakul, P. Buasri, A. Kaewrawang, and J. Triyangkulsri, "Capacitor Location and Size Determination to Reduce Power Losses of a Distribution Feeder in Lao PDR," vol. 4, no. 1, pp. 32–36, 2012.

[7] B. Gou and A. Abur, "Optimal capacitor placement for improving power quality," *Power Eng. Soc. Summer Meet. …*, pp. 488–492, 1999.

[8] A. M. Sharaf and A. A. A. El-Gammal, "Optimal selection of capacitors in distribution networks for voltage stabilization and loss reduction," *EPEC 2010 - IEEE Electr. Power Energy Conf. "Sustainable Energy an Intell. Grid,"* 2010.

[9] A. Wazir and N. Arbab, "Analysis and Optimization of IEEE 33 Bus Radial Distributed System Using Optimization Algorithm," *J. Emerg. Trands Appl. Eng.*, vol. 1, no. 2, 2016.

[10] Government of Nepal, "Electricity Rules," vol. 2050, no. 1993, pp. 1–68, 1993.

[11] "IEEE SA - 141-1993 - IEEE Recommended Practice for Electric Power Distribution for Industrial Plants." [Online]. Available: https://standards.ieee.org/findstds/standard/141-1993.html. [Accessed: 25-Aug-2017].

[12] K. Malmedal and P. K. Sen, "A Better Understanding of Load and Loss Factors," in *2008 IEEE Industry Applications Society Annual Meeting*, 2008, pp. 1–6.

[13] R. S. Rao and S. V. L. Narasimham, "Optimal capacitor placement in a radial distribution system using plant growth simulation algorithm," *Int. J. Electr. Comput. Energ. Electron. Commun. Eng.*, vol. 2, no. 9, pp. 1971–1978, 2008.

* * *

Implementing Neural Network and Multi resolution analysis in EEG signal for early detection of epilepsy

Sachin Shrestha*, Rupesh Dahi Shrestha+, Bhojraj Thapa**

Department of Electronics and Communication Engineering
Nepal Engineering College, Changunarayan, Bhaktapur, Nepal
**sachins@nec.edu.np, +rupeshd@nec.edu.np, ** anzanthapa@gmail.com*

Sachin Shrestha *has completed his Masters in Information and Communication Engineering from IOE, Pulchowk Campus. He has been involved in teaching profession for the last 11 years and is currently working as Assistant Professor at Nepal Engineering College, Changunarayan, Nepal. His research interests are ultra-low powered system design, biomedical signal processing, embedded system design and image processing.*

Rupesh Dahi Shrestha *received the BE degree in Electronics and Communication from Nepal Engineering College in 2005, and MSc degree in Information and Communication Engineering from Pulchowk Campus in 2013. He has experience of 13 years in teaching and is currently working as an Assistant Professor and the Head of the Department of Electronics and Communication Engineering, Nepal Engineering College. His research interests include signal processing, robotics and embedded system.*

Bhoj Raj Thapa *is an undergraduate student at Nepal Engineering College, Pokhara University. He has completed +2 from Liverpool International College under HSEB. He has done projects in his college life related to Android App Development, Signal Processing, and Embedded Systems.*

Abstract

Epilepsy is a neurological disorder of brain and the electroencephalogram (EEG) signals are commonly used to detect the epileptic seizures, the result of abnormal electrical activity in the brain. This paper focuses on the analysis of EEG signal to detect the presence of the epileptic seizure prior to its occurrence. The result could aid the individual in the initiation of delay sensitive diagnostic, therapeutic and alerting procedures. The methodology involves the multi-resolution analysis (MRA) of the EEG signals of epileptic patient. MRA is carried out using discrete wavelet transform with daubechies 8 as the mother wavelet. For EEG data, the database of MIT-BIH of seven patients with different cases of epileptic seizure was used. The result with one of the patients showed presence of a unique pattern during the spectral analysis of the signal over different bands. Hence, based on the first patient, similar region is selected with the other patients and the multi-resolution analysis along with the principal component analysis (PCA) for the dimension reduction is carried out. Finally, these are treated with neural network to perform the classification of the signal either the epilepsy is occurring or not. The final results showed 100% accuracy with the detection with the neural network however it uses a large amount of data for analysis. Thus, the same was tested with dimension reduction using PCA which reduced the average accuracy to 89.51%. All the results have been simulated within the Matlab environment.

Keywords *: Electroencephalography, Epileptic seizure, Discrete wavelet transform, Daubechies 8, PCA, Neural Network.*

1. INTRODUCTION

Electroencephalography (EEG) is the recording of electrical activity along the scalp. It measures voltage fluctuations resulting from ionic current flow within the neurons of the brain. In clinical context, it refers to the recording of the brain's spontaneous electrical activity over a short period of time, usually 20–40 minutes, as recorded from multiple electrodes placed on the scalp [1]. The diagnostic applications generally focus on the spectral content of EEG, that is, the type of neural oscillations that can be observed in EEG signals.

The EEG signal contains several spectral components. The amplitude of a human surface EEG signal ranges from 10 to 100 μV while the important frequencies from the physiological viewpoint lie in the range of 0.1 to 30 Hz. The standard EEG clinical bands are categorized as delta (0.1 to 3.5 Hz), theta (4 to 7.5 Hz), alpha (8 to 13 Hz), and beta (14 to 30 Hz) bands [2]. EEG signals with frequencies greater than 30 Hz are categorized as gamma waves.

Epilepsy is a chronic neuronal disorder or group of disorders, characterized by the recurrent (two or more) seizures that usually recur unpredictably in the absence of provoking factors. It is linked to an abnormal and excessive discharge from a set of neurons in a specific locus of the brain. This clinical manifestation consists of sudden and transitory abnormal phenomena which may include alterations in consciousness, motor, sensory, autonomic or psychic behaviours. It affects almost all groups of the society. The prevalence and incidence of epilepsy is varying among the countries with heterogeneous figures.

A. Motivation

About 1% of people worldwide (65 million) are suffering from epilepsy, [3] and nearly 80% of cases are in developing countries [4]. Epilepsy becomes more common with the growing age of people [5]. In the developed world, onset of new cases occurs most frequently in infants and the elderly [6] while in the developing world this is in older children and young adults [7], due to differences in the frequency of the underlying causes. It has been observed that about 5–10% of all people will have an unprovoked seizure by the age of 80 [8], and the chance of experiencing a second seizure is between 40 and 50% [9]. Epilepsy cannot be cured, but seizures are controllable with medication in about 70% of cases [10]. In those whose seizures do not respond to medication, surgery, neuro-stimulation or dietary changes may be considered. Not all cases of epilepsy are lifelong, and a substantial number of people improve to the point that medication is no longer needed.

Automated seizure onset detection will enable the engineering of novel therapeutic and alerting systems to ease the burden of such seizures. There might be therapeutic system capable of detecting and reacting to the onset of a seizure, a local electrical, thermal or neurochemical stimulus that can halt the progression of a seizure prior to the development of clinical symptoms. Moreover, just-in-time, local therapy could relieve patients of the toxic side effects that accompany systemic administration of multi anti-epileptic drugs. An alerting system equipped with seizure onset detection could warn the patient of seizure prior to the development of the debilitating symptoms, or could notify a family member so that the consequences of a seizure can be limited.

B. Onset detection of epileptic seizure

Various states of art designs have been implemented to analyse EEG signals. EEG signals have been used widely to analyse not just the disorders associated with the brain but also with the various states of the brain like the emotions associated with the nervous systems. Research into seizure detection methods began with the development of seizure event detectors [11], the difference between them being the increased accuracy performance with the later. The detectors developed were meant to detect the seizures of any individual with epilepsy i.e. they were patient non-specific. The variability within EEG severely limited the detection accuracy of such detectors. To improve the performance, investigators developed patient specific event detectors i.e. detectors that could be tailored to EEG of an individual [12]. Years later, the development of diagnostic and therapeutic applications that require initiation following seizure onset motivated the development of seizure onset detection algorithms.

One of the earliest patient non-specific seizure event detectors was the one developed by Gotman [11] in 1982. The Gotman algorithm searches

for the Hallmark sign of seizures i.e. rhythmic activity. The Reveal algorithm developed by Wilson [13] decomposes two second EEG epochs from each input channel into time-frequency atoms using the Matching Pursuit algorithm. Hand coded and neural network rules are then employed to determine whether features derived from the atoms of a channel are consistent with a seizure taking place on that channel. Saab developed a patient non-specific seizure onset detector [14] algorithm that uses features derived from a wavelet decomposition of each EEG channel to estimate the probability of a seizure event. Qu developed the first patient specific seizure onset detection algorithm [15][16][17][18]. Shoeb had proposed the machine learning based patient specific seizure onset detector [19]. The detector uses a support vector machine to classify a feature vector that automatically encodes the time evolution of spectral and spatial features within the scalp EEG. Sriram Ramgopal et al summarized the works done so far in the onset detection using different approaches [20]. Current trends are more inclined toward the supervised learning techniques however it is restricted to the epilepsy detection [21].

A lot of works has been done on the onset detection of the epilepsy however they are confined to the detection of the epilepsy rather than the effort towards the predictive model for the occurrence of the epilepsy. This paper is focussed more on the feature analysis for the earlier detection of the possible occurrence of epilepsy.

II. ALGORITHM DEVELOPMENT

The overall algorithm comprises of basically two major sections and is basically the continuation of the research work published in the September issue of SCITECH Nepal [27]. The first one is the extraction of the signal, its pre-processing and multi-resolution analysis of the signal and the second one is the classification based on extracted features using Neural Network and PCA. The research work is basically the continuation of the work being published in September 2018 issue of SCITECH Nepal [29].

A. EEG signal extraction and pre-processing

For this research paper, the database of MIT BIH has been used. The different cases presented in the database have been grouped into 23 cases, collected from 22 subjects (5 males, ages 3-22; and 17 females, ages 1.5-19) [19][22][23][24]. Instead of using data of all the patients, only those were selected who has effective number of seizure patterns and that they aren't close enough. Also, the seizures should occur within 1 hour. Based on this, only 7 patients were selected for the analysis. Here in this paper only one patient out of them has been chosen as specific case for the study purpose.

The technical details associated with the signal include the usage of sampling frequency of 256Hz, sampling period of 0.00390626 second, duration of 1 hour for most of the data and the 23 leads for the analysis of each case.

B. Discrete wavelet transform of the EEG signal

The Discrete Wavelet Transform analyses the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. DWT employs two sets of functions, called scaling functions and wavelet functions, which are associated with low pass and high pass filters, respectively. The decomposition of the signal into different frequency bands is simply obtained by successive high pass and low pass filtering of the time domain signal.

The original signal x[n] is first passed through a half band high pass filter g[n] and a low pass filter h[n]. After the filtering, half of the samples can be eliminated according to the Nyquist rule, since the signal now has a highest frequency of pi/2 radians instead of pi, where 2pi is the sampling frequency, fs in terms of radian. The signal can therefore be subsampled by 2, simply by discarding every other sample. This constitutes one level of decomposition and can mathematically be expressed as follows:

$$Y_{high} = g[k] = \sum_{n=0} x[n].g[2k-n] \quad \ldots\ldots (1)$$

$$Y_{low} = h[k] = \sum_{n=0} x[n].h[2k-n] \quad \ldots\ldots (2)$$

where, $Y_{high}[k]$ and $Y_{low}[k]$ are the outputs of the high pass and low pass filters, respectively, after subsampling by 2.

This decomposition halves the time resolution since only half the number of samples now characterizes

the entire signal. However, this operation doubles the frequency resolution, since the frequency band of the signal now spans only half the previous frequency band, effectively reducing the uncertainty in the frequency by half. The above procedure, which is also known as the sub-band coding, can be repeated for further decomposition. At every level, the filtering and subsampling will result in half the number of samples (and hence half the time resolution) and half the frequency band spanned (and hence double the frequency resolution).

With the subsampling operator ↓(down-sampling), shown in equation (3), equations (1) and (2) can be written more concisely, shown as equations (4) and (5).

$(Y{\downarrow}n)[k]=Y[kn]$ …… (3)

$Y_{low} = (x * g){\downarrow}2$ …… (4)

$Y_{high} = (x * h){\downarrow}2$ …… (5)

C. Calculation of energy of different bands of EEG

DWT decomposes original signal into two different frequency bands in its 1st level decomposition. In this research work, the decomposition has been done up to five levels resulting the six different frequency bands. The so obtained frequency bands are:

TABLE I
Frequency bands and associated range of frequency

Frequency bands	Range of frequency (Hz)
Band 1	0 to 4
Band 2	4 to 8
Band 3	8 to 16
Band 4	16 to 32
Band 5	32 to 64
Band 6	64 to 128

After obtaining the coefficients for these different frequency bands, the energy associated with each of them is calculated using the formulation based on Parseval's theorem as shown in equation (6).

$$EDi = \sum_{j=1}^{N} |D_{ij}|^2,\ i = 1, \ldots . l \ldots \ldots \ldots \quad (6)$$

$$EA_l = \sum_{j=1}^{N} |A_{ij}|^2 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \quad (7)$$

where i = l,.....,l is the wavelet decomposition level from level 1 to level l. N is the number of the coefficients of detail or approximate at each decomposition level. ED_i is the energy of the detail at decomposition level i and EA_l is the energy of the approximate at decomposition level l. The value of l is taken 5. Since the number of coefficients vary with the level, the values obtained have been scaled down by the number of coefficients.

In DWT, it is very important to choose a suitable wavelet alongside the number of levels of decomposition. Here, the Daubechies 8 has been chosen as a mother wavelet for DWT. The reason behind choosing this wavelet for EEG analysis is its smoothing feature which is suitable for detecting changes of the EEG signals. The scaling function and the wavelet function of this mother wavelet are shown in Fig. 1 and Fig. 2. These wavelets are asymmetric in nature and shows bi-orthogonal nature. This wavelet analysis can represent EEG sub-bands as a weighted sum of shifted and scaled version of an original wavelet without losing any information of energy. Since this transformation can retain the frequency as well time properties of the signal, this is considered superior to the other methods like Fourier transform for this research work.

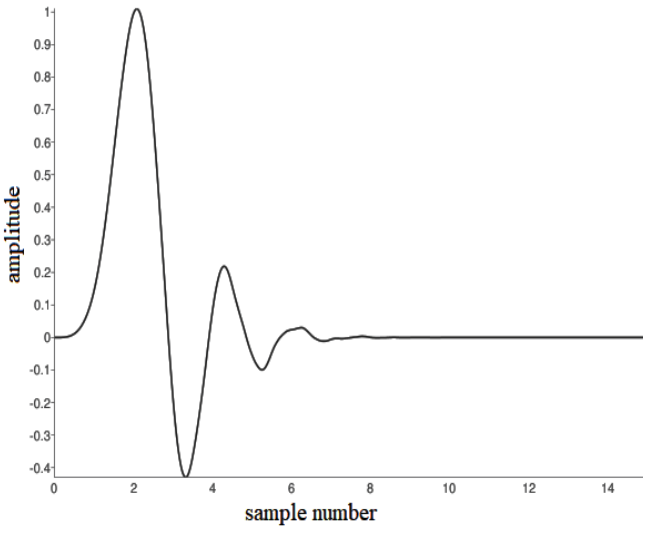


Fig. 1 Scaling function of Daubechies 8 wavelet

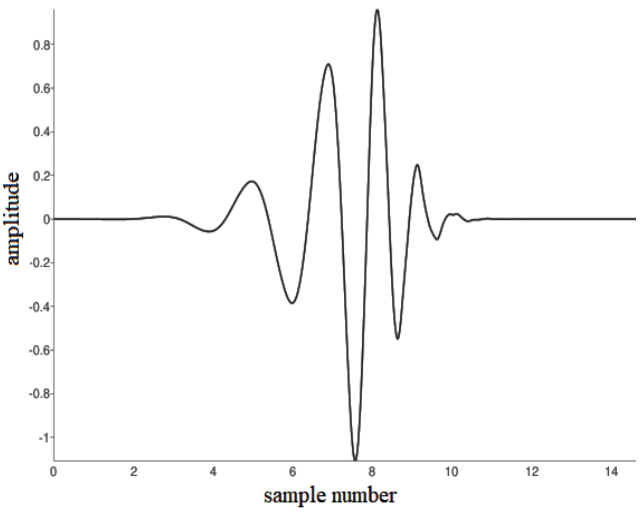


Fig. 2 Wavelet function of Daubechies 8 wavelet

D. Windowing and shifting for energy calculation

For the calculation of the energy of different bands, the calculation is done over a window of 256 samples which corresponds to the duration of 1 second. Next, the window is shifted by 128 samples that corresponds to the duration of 0.5 second. The window so taken is of rectangular nature and this window based calculation helps in obtaining better resolution for time. Next, the shift by 50% i.e. 128 samples with window size of 256 samples will help in eliminating the edge effect of the window. Thus the period of 1 hour i.e. 3600 seconds will give 6 x 7199 values corresponding to the each lead of EEG signal.

E. Neural Network and PCA

After the feature extraction process of the EEG signal extracted from CHB-MIT database, the most prior mechanism is to interpret those features and selection of the features. The extracted features are then analysed and classified on the basis of supervised learning classification technique. To extract the features using NN, energy values of the EEG signals was reshaped into a two dimensional matrix to create an input matrix. Then the machine was trained as per features to detect the unknown EEG signal if that is epileptic or non-epileptic. So, training and testing will be summarized process of the research work which will be done using Neural Network Toolbox of MATLAB.

The feedforward network was developed for the purpose that can be trained to classify inputs according to target classes. The network training function of the neural network updated weight and bias values according to the scaled conjugate gradient method. Backpropagation was used to calculate derivatives of performance with respect to the weight and bias variables [25]. The row vector of the hidden layer is sized as 50. Cross-entropy was used as performance calculation method. It calculates a network performance given targets and outputs, with optional performance weights and other parameters. The function returns a result that heavily penalizes outputs that are extremely inaccurate, with very little penalty for fairly correct classifications. Minimizing cross-entropy leads to good classifiers. The targets were divided into three sets using random indices as training, validation, and testing and the weightage of each was taken as 70%, 15% and 15% respectively.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The PCA has been chosen here to check if the reduced dimension, here it has been reduced to 24, can give the equally effective results. Here the implementation of PCA has centred the data and used singular value decomposition algorithm [26].

III. RESULT AND DISCUSSION

The result of MRA with one of the patients has already been published [27]. From the result of MRA of that patient, the conclusion was made that only 3 of the bands were more effective and hence only last three bands were used as features for the analysis. Next, the usage of only few of the leads instead of all could also generate equally effective results. Hence, only 8 different leads were used. The choice of the leads was made based on the fact that the data of all the patients didn't consist of data of all the leads in similar fashion. Hence, only those that could give us the effective similarity were used. This led to the choice of 8 leads. Instead of using all the patients, only those were selected who had effective number of seizure patterns and that they aren't close enough and were within 1 hour. Based on this, only 7 patients were selected for analysis.

TABLE II
Info of patients being used for analysis

Patient No	Gender	Age(Years)
1	Female	11
3	Female	14
5	Female	7
8	Male	3.5
14	Female	9
15	Male	16
18	Female	19

For the inputs to the neural network, and basically for the early onset detection of epilepsy, MRA results of EEG signals after windowing, of 1400th second prior to seizure till 400th second prior to seizure are taken. This forms the matrix input of [480000 x N], where N is the number of files present for each patient. Since, this could lead to need of larger input, this has been reduced to 24 dimension using PCA and the NN is further implemented.

A. Evaluation procedure

For the analysis of the result, positive predictive value (PPV), negative predictive value (NPV) and accuracy were calculated. The positive predictive value (PPV) is defined as the ratio of number of true positive (TP) by the summation of number of true positive and number of false positive (FP). A "true positive" is the event that the test makes a positive prediction, and the subject has a positive result under the gold standard, and a "false positive" is the event that the test makes a positive prediction, and the subject has a negative result under the gold standard. Likewise, the negative predictive value (NPV) is defined as the ratio of number of true negative (TN) by the summation of number of true negative and number of false negative (FN). A "true negative" is the event that the test makes a negative prediction, and the subject has a negative result under the gold standard, and a "false negative" is the event that the test makes a negative prediction, and the subject has a positive result under the gold standard. The accuracy is calculated as the ratio of summation of true positive and true negative to the summation of true positive, true negative, false positive and false negative [29]. All these could be presented well by the confusion matrix which is also known as error matrix and is widely popular to analyse problems in the field of machine learning and specifically the problem of statistical classification [28].

The confusion matrix for one of the patients after NN is shown in Fig. 3.

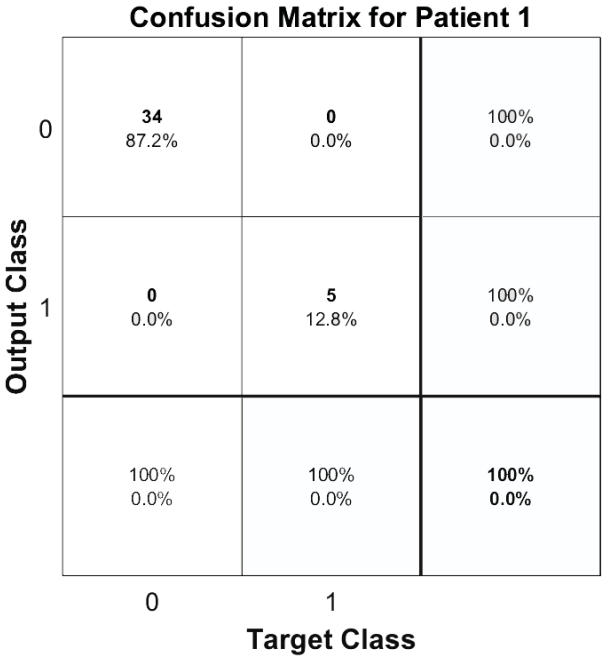


Fig. 3 Confusion matrix for Patient 1 without PCA

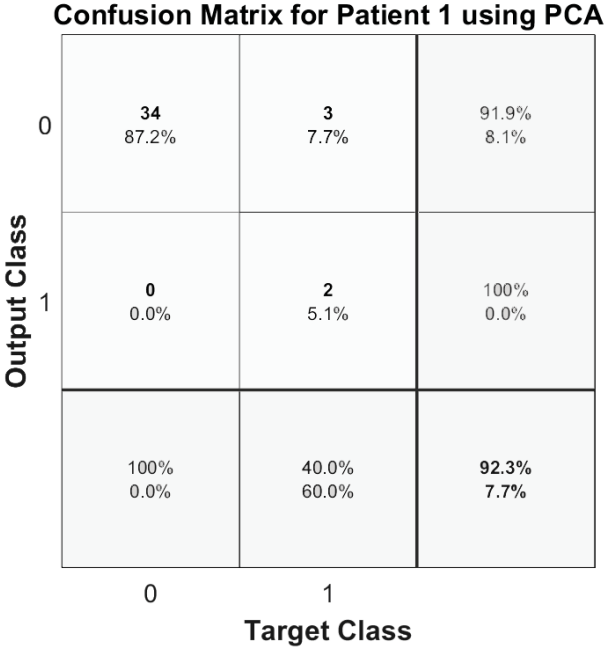


Fig. 4 Confusion matrix for Patient 1 with PCA

The summarized results for all the patients after using and not using PCA as depicted by the confusion matrix as shown in Fig 3 and Fig 4 is presented in the Table II and Table III.

TABLE III
Summarized results for all the patients without using pca

Patient	TN	FP	Sp	FN	TP	Se	FDR	NPV	Accuracy
1	34	0	100	0	5	100	100	100	100
3	30	0	100	0	4	100	100	100	100
5	34	0	100	0	3	100	100	100	100
8	14	0	100	0	5	100	100	100	100
14	17	0	100	0	6	100	100	100	100
15	26	0	100	0	9	100	100	100	100
18	28	0	100	0	4	100	100	100	100

TABLE IV
Summarized results for all the patients using pca

Patient	TN	FP	Sp	FN	TP	Se	FDR	NPV	Accuracy
1	33	5	86.8	1	0	0	97.1	0	84.6
3	30	0	100	0	4	100	100	100	100
5	34	2	94.4	0	1	100	100	33.3	94.6
8	15	2	87.5	0	3	100	100	60	89.5
14	16	2	88.9	1	4	80	94.1	66.7	87
15	24	6	80	2	3	60	92.3	33.3	77.1
18	26	0	100	2	4	66.7	92.9	100	93.8

From tables II and III, it can be concluded that the one without using Neural Network gives better result and in fact gave 100% accuracy however at the same time needs larger amount of data for analysis. With PCA implemented which could lead to lesser data for processing, the processing time could be highly reduced but it led to lower accuracy. The average accuracy is calculated as 89.51%.

Next, the delay associated with the detection of the onset is also crucial and with this methodology, since the data being used was 400 seconds prior to the seizure onset, with the high end faster processing unit, it could be done within the same seconds of delay.

IV. CONCLUSION

This paper discusses the multi-resolution analysis of the EEG signal under different bands for a specific patient prior to the occurrence of the epilepsy. The EEG signal is used as an important tool to inspect the cases of epileptic seizures. This research work intended to automate the process and perform the analysis of those signal frequency domain in order to extract the significant features out of it. Those features were used as important parameters deciding the occurrence of epilepsy.

The observation of 41 different cases of the patient has shown a unique pattern for the epileptic case and non-occurrence for the non-epileptic session with the PPV, NPV and accuracy as discussed in [27]. Similar unique pattern weren't observed in other patient. Hence based on the analysis of the first patient, similar region is used for other patients and PCA and NN were implemented. It had resulted in 100% accuracy for NN only however with large amount of data. With PCA the data input is reduced but has resulted in lower accuracy level. Since, the prediction of the occurrence of the epileptic seizure is more crucial to alert the patient or his/her near ones, this finding could be a boon for such and thus can be used as a very useful feature for the development of the prediction model for the detection of the onset of the epileptic seizure and alerting mechanism.

Acknowledgment

The authors would like to thank the University Grants Commission (UGC) of Nepal for funding this project [30]. Also, the authors are very grateful to the MIT BIH for availing the database of epileptic seizure for free for research purpose.

References

[1] Niedermeyer E. and da Silva F.L. (2004). Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Lippincot Williams & Wilkins. ISBN 0-7817-5126-8

[2] S. Tong and N.V.Thacor, Engineering in Medicine & Biology- Quantitative EEG Analysis Methods and Clinical Applications, Boston/London: Artech House, 2009.

[3] Thurman, DJ; Beghi, E; Begley, CE; Berg, AT; Buchhalter, JR; Ding, D; Hesdorffer, DC; Hauser, WA; Kazis, L; Kobau, R; Kroner, B; Labiner, D; Liow, K; Logroscino, G; Medina, MT; Newton, CR; Parko, K; Paschal, A; Preux, PM; Sander, JW; Selassie, A; Theodore, W; Tomson, T; Wiebe, S; ILAE Commission on, Epidemiology (September 2011). "Standards for epidemiologic studies and surveillance of epilepsy.".Epilepsia. 52 Suppl 7: 2–26.

[4] "Epilepsy". Fact Sheets. World Health Organization. October 2012. Retrieved January 24, 2013.

[5] Brodie, MJ; Elder, AT, Kwan, P (November 2009). "Epilepsy in later life". Lancet neurology 8 (11): 1019–30.

[6] Wyllie's treatment of epilepsy: principles and practice. (5th ed.). Philadelphia: Wolters Kluwer/ Lippincott Williams & Wilkins. 2010. ISBN 978-1-58255-937-7.

[7] Newton, CR (29 September 2012). "Epilepsy in poor regions of the world.". Lancet 380 (9848): 1193–201.

[8] Wilden, JA; Cohen-Gadol, AA (15 August 2012). "Evaluation of first nonfebrile seizures.". American family physician 86 (4): 334–40.

[9] Berg, AT (2008). "Risk of recurrence after a first unprovoked seizure". Epilepsia. 49 Suppl 1: 13–8.

[10] L Devlin, A; Odell, M; L Charlton, J. Koppel, S (December 2012). "Epilepsy and driving: current status of research" Epilepsy research 102 (3): 135–52.

[11] J. Gotman, Automatic recognition of epileptic seizure in EEG, *Electroencephalograpy and Clinical Neurophysiology*, 54:530-540, 1982.

[12] Hao QU and Jean Gotman, Improvement in seizure detection performance by automatic adaptation to the EEG of each patient, Electroencephalography and Clinical Neurophysiology, 86:79-87, 1993.

[13] Scott Wilson, Mark Scheuer, Ronald Emerson, and Andrew Gabor, Seizure detection: Evaluation of the reveal algorithm, *Clinical Neurophysiology*, 10:2280-2291, 2004.

[14] M.E. Saab and J. Gotman, A system to detect the onset of epileptic seizures in scalp EEG, *Clinical Neurophysiology*, 116: 427-442, 2005.

[15] Hao Qu, Self adapting algorithms for seizure detection during EEG monitoring, PhD dissertation, McGill University, 1995.

[16] Hao Qu and Jean Gotman, Improvement in seizure detection performance by automatic adaptation to the EEG of each patient, *Electroencephalography and clinical neurophysiology*, 86:79-87, 1993.

[17] Hao Qu and Jean Gotman, A seizure warning system for long term epilepsy monitoring, *Electroencephalography and clinical neurophysiology*, 45:2250-2254, 1995.

[18] Hao Qu and Jean Gotman, A patient specific algorithm for the detection of seizure onset in long term EEG monitoring, *IEEE transactions on biomedical engineering*, 44:115-122, 1997.

[19] Ali Shoeb, *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment,* PhD Thesis, Massachusetts Institute of Technology, September 2009.

[20] Sriram Ramgopal et al, Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy, *Epilepsy & Behavior*, 37 (2014) 291–307, 2010, Online: https://www.sciencedirect.com/science /article/pii/S1525505014002297, access date: May 2018.

[21] Nhan Duy Truong, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, Andrew Faulks, Omid Kavehei, Supervised learning in automatic channel selection for epileptic seizure detection, Expert systems with application, Volume 86, 15 November 2017, Pages 199-207, Online: https://www.sciencedirect.com/science/article/pii/S0957417417303779, accessed date: May 2018.

[22] CHB MIT Scalp EEG database, Online: http://physionet.org/physiobank/database/chbmit, access date: May 2018.

[23] Ali Shoeb, John Guttag, *27th International Conference on Machine Learning (ICML)*, June 21-24, 2010, Haifa, Israel.

[24] Ali Shoeb, Herman Edwards, Jack Connolly, Blaise Bourgeois, S. Ted Treves, John Guttag, Patient-Specific Seizure Onset Detection, *Epilepsy and Behavior.* August 2004, 5(4): 483-498. [doi:10.1016/j.yebeh.2004.05.005]

[25] Moller, *Neural Networks*, Vol. 6, 1993, pp. 525–533.

[26] Jolliffe, I. T. Principal Component Analysis. 2nd ed., Springer, 2002.

[27] Sachin Shrestha et. al, Analysis of EEG signal of specific epileptic patient prior to its occurrence, *SCITECH Nepal*, Vol 13, No 1, September 2018.

[28] Howard Hamilton, Confusion matrix, Online: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html, access date: December 2018.

[29] David Felson, Positive and Negative predictive values, Online: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Screening/ EP713_Screening5.html, access date: May 2018.

[30] University Grants Commission, Online: http://www.ugcnepal.edu.np, access date: May 2018.

* * *

Analysis of Consumer Data on Black Friday Sales using Apriori Algorithm

Menuka Maharjan
Department of Computer Engineering,
Nepal Engineering College
Changunarayan,Bhaktapur, Nepal
menukam@nec.edu.np

***Menuka Maharjan** is Assistant Professor at Department of Computer Science and Engineering, Nepal Engineering College. She holds M.E. in Computer Science from Nepal College of Information Technology, Pokhara University. She has been in the teaching field since the last 8 years. Her research interest includes data science and machine learning.*

Abstract

Ability to recognize and track patterns in data help businesses shift through the layers of seemingly unrelated data for meaningful relationships. Through this analysis it becomes easy for the online retailers to determine the dimensions that influence the uptake of online shopping and plan effective marketing strategies. This paper builds a roadmap for analyzing consumer's online buying behavior with the help of Apriori algorithm. The major factors that affect the consumer's online buying behavior are convenience, ease of use and perceived benefits. Security is also a major consideration when opting to conduct shopping activities online. This study will help in further analyzing the consumer online buying behavior towards Online shopping which will help the retailers to design appropriate marketing strategies for selling their products online which will further help in development of the country.

Keywords : *E-commerce, Data mining, Consumer Behavior, Personal perceived values, Website quality*

I. Introduction

The increase of data available and its dynamic growth will bring new challenges to the load profiling and consumer characterization. The new tools must be able to treat large amounts of data with all the problems common to real databases, like noise, missing values and outliers. These tools must be flexible, robust and able to provide an easy actualization of the knowledge as new data is available. This paper presents a framework developed to support the retail and distribution companies on the extraction of knowledge. from electricity consumption data. This is based on the application of data mining techniques to the determination and characterization of a set of load profiles, representing the different consumption patterns of a sample of consumers [1]. The data has been taken from the Black Friday dated on November 23, 2018, the day after Thanksgiving. It's traditionally the busiest shopping day of the year because it kicks off the holiday season. This season is crucial for the economy because around 30 percent of annual retail sales occur during the holiday season. For some retailers, such as jewelers, it is even higher, at almost 40 percent [2].

II. Literature Review

Black Friday Sales and Deals

The best Black Friday deals are, surprisingly, not on Black Friday. Many retailers, including Amazon, offer deals earlier and earlier, upstaging Black Friday itself. The competition this year is so fierce, stores are innovating new ways to get your dollar. Research reveals that the most deals for electronics are offered at the beginning of November. The best day for Christmas decor is November 22. Discounts are 23 percent on average. The best day to buy toys is the day before Thanks giving. The number of deals hit their peak the week before Thanksgiving.

The average in-store discount is 20 percent for the entire week. That discount increases to 37 percent on Thanksgiving, Black Friday, and Saturday. Online sales are the best on Thanksgiving Day, not Black Friday. The average discount is 24 percent. It's the best day to find online discounts for sporting goods, computers, apparel, and video games. But Black Friday itself is the best day for online deals on TVs, tablets, appliances, and jewelry. The days from Thanksgiving through Cyber Monday capture 20 percent of all holiday online shopping. Use these statistics to get the best Black Friday deals. In addition, don't forget deals offered on Green Monday. It's the last day to shop online and be sure that your package arrives before Christmas [1].

Statistics Canada Internet survey also stated that over one-half of the connected households used more than one type of device to go online in 2010. Over one-third of all Internet users in Canada went online using wireless devices (such as laptop, PDA, tablet). In recent years, smart phone becomes much more affordable and popular, and it becomes another main entry point for Internet world. The survey also shows that the countries that have relatively lower percentage of internet users also experienced significant increase in terms of internet penetration rate [10].

III. Methodology

The framework is able to treat different data sets in an easy and efficient way and provides results like consumer classification rules according to their ages, gender, and product category or purchase amount. These results can be updated as new data is collected.

The paper is organized as follows: In Section I the consumers' Attributes are described and preprocessed and cleaned the data to apply it on the algorithm. In Section II the the data is implementing in Apriori algorithm with the minimum support and confidence. In section III the evaluation and testing of the rule that is generated by the Apriori algorithm.

Section I

There are total of 537577 instances and 12 attributes. They are listed below. The types of data are ordinal, nominal and numeric but to apply it in Apriori algorithm all the attributes are converted into nominal. Before applying the apriori rule the data is being converted in arff file to apply in weka.

User_ID: Numeric
Produc_ID: Nominal
Gender: Nominal
Age Nominal
Occupation: Numeric
City Category: Nominal
Stay_In_Curretnt_City_Years: String
Marital Status: Numeric
Product_Category_1: Numeric
Product_Category_2: Numeric
Product_Category_3: Numeric
Purchase: Numeric

Section II

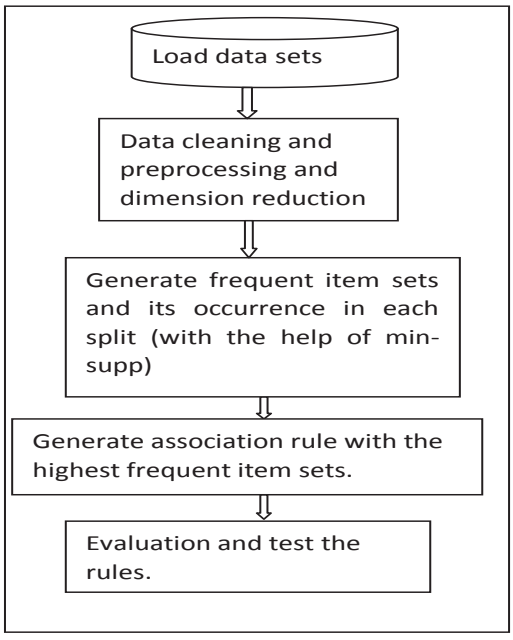


Figure 1: Architecture for consumer behavior analysis

Apriori algorithm

General Process Association rule generation is usually split up into two separate steps: 1. minimum support is applied to find all frequent item sets in a database. 2. these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straight forward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over I and has size $2n − 1$ (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, efficient algorithms can find all frequent item sets.

Apriori Algorithm Pseudo code

Procedure Apriori (T, minSupport) { //T is the database and minSupport is the minimum support

L1= {frequent items};

for (k= 2; Lk-1 !=∅; k++) {

Ck= candidates generated from Lk-1

//that is Cartesian product Lk-1 x Lk-1 and eliminating any k-1 size item set that is not

//frequent

for each transaction t in database do{

#increment the count of all candidates in Ck that are contained in t

Lk = candidates in Ck with minSupport

}//end for each

}//end for

return U;

}

As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [7].

Section III

The best-known constraints are minimum thresholds on support and confidence.

A. Support

The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set

Supp(X) = no. of transactions which contain the item set X / total no. of transactions

In the example database, the item set {Gender=M, Product_Category_2=0, Product_Category_3=0} has a support of 4 /15 = 0.26 since it occurs in 26% of all transactions. To be even more explicit we can point out that 4 is the number of transactions from the database which contain the item set {Gender=M, Product_Category_2=0, Product_Category_3=0} while 15 represents the sample of total number of transactions

B. Confidence

The confidence of a rule is defined:

For the rule {Gender=M, Product_Category_2=0} ==> {Product_Category_3=0} we have the following confidence:

Supp ({Gender=M, Product_Category_2=0, Product_Category_3=0}) / supp({Gender=M ,Product_Category_2=0}) = 0.26 / 0.4 = 0.65

This means that for 65% of the transactions containing milk and bread the rule is correct. Confidence can be interpreted as an estimate of the probability P(Y | X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

C. Lift

The lift of a rule is defined as:

The rule {Gender=M, Product_Category_2=0}=>{ Product_Category_3=0} has the following lift:

Supp ({Gender=M, Product_Category_2=0, Product_Category_3=0}) / supp ({Product_ Category_3=0}) x supp({Gender=M, Product_ Category_2=0})= 0.26/0.46 x 0.4= 1.4

D. Conviction

The conviction of a rule is defined as:

The rule {Gender=M, Product_Category_2=0} => {Product_Category_3=0} has the following conviction:

1 – supp ({Product_Category_3=0})/ 1- conf ({Gender=M, Product_Category_2=0}=>{ Product_Category_3=0}) = 1-0.46/1-0.65 = 1.54 The conviction of the rule X=>Y can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions.

In this example, the conviction value of 1.54 shows that the rule {Gender=M, Product_ Category_2=0}=>{ Product_Category_3=0} would be incorrect 54% more often (1.54 times as often) if the association between X and Y was purely random chance.

IV. Implementation

Out of 537577 instances only 1550 instances is picked out randomly in weka and missing data is replaced by zero and for the algorithm to implement in Apriori, all the data must be in nominal form so it is converted using the supervised filter in weka.

Apriori output

Minimum support: 0.1 (155 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large item sets:

Size of set of large item sets L(1): 20

Size of set of large item sets L(2): 59

Size of set of large item sets L(3): 47

Size of set of large item sets L(4): 11

Best rules found:

1. Product_Category_2=0 500 ==> Product_ Category_3=0 500 <conf :(1)> lift :(1.46) lev :(0.1) conv :(156.77)
2. Gender=M Product_Category_2=0 374 ==> Product_Category_3=0 374 <conf :(1)> lift :(1.46) lev :(0.08) conv :(117.27)
3. Marital_Status=0 Product_Category_2=0 285 ==> Product_Category_3=0 285 <conf:(1)> lift:(1.46) lev:(0.06) conv:(89.36)
4. City_Category=B Product_Category_2=0 237 ==> Product_Category_3=0 237 <conf :(1)> lift:(1.46) lev:(0.05) conv:(74.31)
5. Gender=M Marital_Status=0 Product_ Category_2=0 218 ==> Product_Category_3=0 218 <conf:(1)> lift:(1.46) lev:(0.04) conv:(68.35)
6. Marital_Status=1 Product_Category_2=0 215 ==> Product_Category_3=0 215 <conf:(1)> lift:(1.46) lev:(0.04) conv:(67.41)
7. Product_Category_1=8 Product_Category_2=0 203 ==> Product_Category_3=0 203 <conf:(1)> lift:(1.46) lev:(0.04) conv:(63.65)
8. Gender=M City_Category=B Product_ Category_2=0 180 ==> Product_Category_3=0 180 <conf:(1)> lift:(1.46) lev:(0.04) conv:(56.44)
9. Product_Category_1=5 Product_Category_2=0 179 ==> Product_Category_3=0 179 <conf:(1)> lift:(1.46) lev:(0.04) conv:(56.13)
10. Age=26-35 Product_Category_2=0 178 ==> Product_Category_3=0 178 <conf:(1)> lift:(1.46) lev:(0.04) conv:(55.81)

V. Evaluation of the result

We can see the rules are presented in antecedent => consequent format. The number associated with the antecedent is the absolute coverage in the dataset (in this case a number out of a possible total of 1550). The generated total 4 item sets 20,50,47,11 out of these item sets it generated the best rule with its confidence, lift, level and conviction. The number next to the consequent is the absolute number of instances that

match the antecedent and the consequent. The number in brackets on the end is the support for the rule (number of antecedent divided by the number of matching consequents). We can see that a cutoff of 91% was used in selecting rules, mentioned in the "Associator output" window and indicated in that no rule has coverage less than 0.91.I don't want to go through all 10 rules, it would be too onerous. Here are few observations:

- We can see that all presented rules have a consequent of "Product Category_3".
- All presented rules indicate a high total transaction amount.
- We can analyze the bought product categories according to the city, gender and age from the rules formed above with the minimum support 0.1 and confidence 0.9.

We have to be very careful about interpreting association rules. They are associations (think correlations), not necessary causally related. Also, short antecedent are likely more. If we are interested in total for example, we might want to convince people that buy product category 1and product category _3 are the most bought out of 500 instances with convince values 156.77 (Rule #1). This may sound plausible, but is flawed reasoning. The product combination does not cause a high total, it is only associated with a high total. What might be interesting to test is to model the path through the store required to collect associated items and seeing if changes to that path (shorter, longer, displayed offers, etc) have an effect on transaction size or basket size.

VI. Conclusion

The power of automatically learning association rules from the large datasets can provide much more efficient approach by using the Apriori algorithm rather than deducing rules by hand. The method discovered the interesting trends in customer's buying patterns from the datasets implementing the association rules for consumer's online buying behavior which will help the retailers to design appropriate marketing strategies for selling their products online. Utilizing this approach, ecommerce house can determine key indicators to enhance sales on special occasions like Black Friday and festivals.

References

[1] [Online]. Available: https://www.thebalance.com/what-is-black-friday-3305710.

[2] A. S. AI-Malaise, "Implementaion of Apriori Algorithm to Analze organization Data: Building Decision Support System," International Journal of Computer Application, vol. 66, no. 9, p. 27, 2013.

[3] M. dagdoug, "www.kaggle.com," 2018. [Online]. Available: https://www.kaggle.com/mehdidag/black-friday.

[4] "thebalance.com," [Online]. Available: http://www.thebalance.com/what-is-black-friday-3305710.

[5] [Online]. Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf.

[6] A. R. Kharwar and N. P. P. Viral Kapadia, "Implementing Apriori algorithm on web server log," in National conference on recent trends in Engineering and Technology, Gujarat,India, 2016.

[7] S. K. Manpreet Kaur, "Market Basket Analysis: Identify the changing trends of market data using association rule mining," in International COnference on Computational Modeling and Security, Sangrur,India, 2016.

[8] K.-C. C. Nan-chen Hsieh, "Enhancing Consumer Behaviour Analysis by Data Mining Techniques," International Journal of Information and Management Sciences, no. 20, p. 15, 2009.

[9] S. Anita, "Performane predication of association rule mining algorithm using machine learning tool weka," Mathematical Sciences International Research Journal, vol. 6, no. 2, p. 5.

[10] Z. Y. Ling Liu, "Improving online shopping experience using data mining and statistical techniques," Journal of convergence information technology, vol. 8, no. 6, p. 9, 2013.

* * *

Hand Gesture Vocalizer for Dumb and Deaf People

Sanish Manandhar[1], Sushana Bajracharya[2], Sanjeev Karki[3], Ashish Kumar Jha[4]

[1,2,3]Department of Computer Science, Nepal Engineering College, Changunarayan, Bhaktapur, Nepal

[1]sanish.manandhar20@gmail.com, [2]sushanabajra96@gmail.com, [3]sanjeevkarki77@gmail.com

[4]Assistant Professor, ashishkj@nec.edu.np

***Sanish Manandhar** is a student of BE computer Engineering, Batch 2014 at Nepal Engineering College and recently he has received his bachelor degree from Pokhara University. He is currently working as web developer in a prominent software company.*

***Sushana Bajracharya** is a student of BE computer Engineering, Batch 2014 at Nepal Engineering College and recently she has received her bachelor degree from Pokhara University. She is currently working as web developer in a prominent software company.*

***Sanjeev Karki** is a student of BE computer Engineering, Batch 2014 at Nepal Engineering College and recently he has received his bachelor degree from Pokhara University. He is currently working as web developer in a prominent software company.*

***Ashish Kumar Jha** received his Master's degree in Computer Science specialization in Networking from Sharda University in 2017. He has involved in software development and teaching profession since 2013, currently working as assistant professor in Nepal Engineering College. His research interest includes Internet of things, Image Processing and Pattern Recognition.*

Abstract:

The main purpose of this paper is to confer the system that converts a given sign used by disabled person into its appropriate textual, audio, and pictorial form using components such as Arduino Mega, Flex sensors, Accelerometer, which could be understand by a common person. A wearable glove controller is design with flex sensors attached on each finger, which allows the system to sense the finger movements, and a Gy-61 accelerometer, which are uses to sense the hand movement of the disabled person. The wearable input glove controller sends the collected input signal to the system for processing. The system uses Random forest algorithm to predict the correct output to an accuracy of 85% on current training model.

***Keywords:** Random Forest, Speech synthesis, Flex sensors, Data glove, Smart Glove.*

I. Introduction

Humans possess the voice capability for interaction and communication among each other. Unfortunately, not everybody has the capability of speaking and hearing. Sign language used among the community of people who cannot speak or hear as the means of communication. Sign language is a gesture representation that involves simultaneously combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions to express fluently with a speaker's thoughts. The people who cannot speak makes use of the sign languages to communicate with other fellow vocally impaired person and even with other normal people who knows the meanings of sign languages or an interpreter is needed to translate the meanings of sign languages to other people who can speak and do not know the meanings of sign languages. However, it is not always possible for an individual to be around all the time to interpret the sign languages

and not everybody can learn the sign languages. Thus, another alternative is that we can use a computer or a smart phone as a mediator. The computer or a smart phone could take an input from the vocally impaired person and give its textual as well as and audio form of output.

II. Related Work

A Lexicon Statics Survey of Signed Language in Nepal This thesis proposed by Hope M. Hurl but after his field study about the variety of sign languages used in Nepal and its similarity or dissimilarity with the American Sign Language (ASL) and British Sign Language (BSL) through the collection of world list and interviews. The results of the world list comparisons showed that there are at least three signed languages in the country [1]. In more recent years NpSL has been introduced to many Hearing Impaired Associations have undertaken to open classes for hearing impaired children in many places where the government does not yet have a school for the hearing impaired. In this course, a research conducted in the following cities of Nepal – Kathmandu, Surkhet, Jumla, Pokhara, Ghandruk, Dharan, and Rajbiraj where it was discovered that there was one national sign language used by all hearing impaired named the NpSL. The survey further concluded that, there were 3 village sign language that were used in the Indian Sign Language (ISL), ASL and BSL and the most widely used technique was the sign writing which is somewhat similar to imitate the object used in spoken language.

In this area, there has been a lot work done with implementation different sign language and technologies few work are discuss below:

A. Kinetic sign language translator [2]

This project was a result of collaboration, facilitated by Microsoft Research, between the Chinese Academy of Sciences, Beijing Union University, and Microsoft Research Asia, in which all the organization has made crucial contributions. The system understands the gestures of sign language and converts them to spoken and written language and vice versa. This system captures a conversation from both sides, displays the signer and renders a written and spoken translation of the sign language in real-time. It also takes the spoken words of non-signer and turns them into accurate, understandable sign language.

B. Sign Language Recognition System [3] [4]

Yang Quan, a Chinese student, defined a Basic Sign Language Recognition system that is able to translate a sequence of signs into the commonly used speech language and vice versa. The sign language/speech bidirectional translation (from signs to speech and speech to signs) focused on the Chinese Manual Alphabet where every single sign belongs to a single letter from the alphabet. The system was composed of a camera, a video display terminal (VDT), a speaker, a microphone, and a keyboard. Two kinds of data were used which were; vector of hand gestures and vector of lip actions. In order to characterize these vectors, they used the Normalized Moment of Inertia (NMI) algorithm and Hu moments. The former attempts to solve translation invariability, size invariability, anti-gradation distortion, and so on and the latter is a set of algebraic invariants that combines regular moments and is useful to describe the information of the image. They do not vary under change of size, translation, and rotation and have been widely used in Pattern Recognition and successfully proven in sign language letter recognition. As said before, they combine the hand gestures recognition with the lips movement reader in order to make the system more accurate. By using a multi-features SVMs classifier trained with a linear kernel, the 30 letters from the Chinese manual alphabet recognized with average accuracy of 95.55%.

C. Microcontroller and Sensors Based Gesture Vocalizer [5]

Sensors Based Gesture Vocalizer describes the design and working of a system that is useful for dumb, deaf and blind people to communicate with each another and with the normal people as well. Gesture Vocalizer is a large-scale multi-microcontroller based system that designed to facilitate the communication among the dumb, deaf and blind communities with the normal people. In the project a data glove is used which can detect almost all the movements of a hand and microcontroller based system converts some specified movements into human recognizable voice. The data glove is equipped with two types of sensors, which are; the bend sensors and accelerometers as tilt sensors. This system is beneficial for dumb people and their hands will

speak having worn the gesture, vocalizer data glove. This project is similar to our proposed project but it does not make use of any smart phone, instead it uses external speakers and a LCD display of outputs.

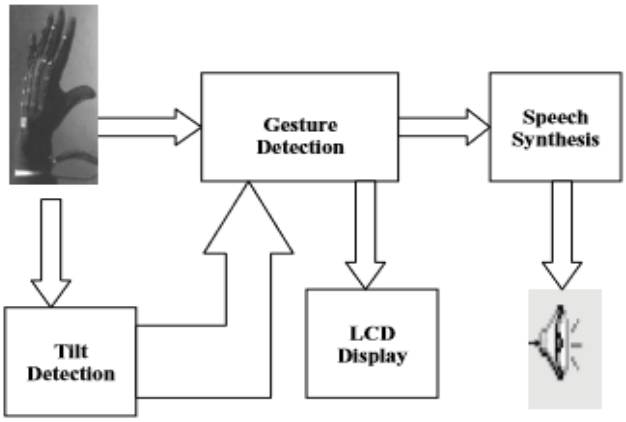


Figure 1: Block diagram of Microcontroller and Sensors Based Gesture Vocaliser

D. Hand Gesture Recognition Using KINECT [6]

This is an automatic Gesture Recognition System, which uses Microsoft Kinect for capturing the raw image of the gesture. Originally developed for gaming, the Kinect's sensors read a user's body position and movements and, with the help of a computer, Gesture Recognition System translates them into words. This system thus has tremendous potential for understanding the complex gestures and for translating the signs into spoken or written words. It check textual character in train data and display as output if present. If no match found, then the system allows the user to store the new gesture and its corresponding textual form. The system allows the user to auto-create a dictionary. The main motive of this project is to make interaction possible between an unimpaired and hearing-impaired individual.

III. Hardware Details: Flex Sensors

The flex sensors also referred as the bend sensor is a type of sensor that measures the amount of bending or deflection or flexing as the name itself has it. As the sensors are bent the sensors provides us with an electrical resistance value, the sensor gives so the more the sensors are bent, more resistance value. When measured with a multi meter, the flex sensor at a flat and steady position gives a resistance value near to 25k ohm whereas when the sensors are fully bent they give a resistance value near to 72k ohm. They are usually in the form of a thin strip, are very comfortable to use, as they are very light in weight, and are easily bent [8].

Figure 2: Flex sensor

One side of the sensor is printed with a polymer ink that has conductive particles embedded on it, and note that the flex sensors gives a proper change in resistance value only when it is bent away from the ink like and provides very minimum change in resistance value when it is bent in the reverse direction.

The flex sensor comes in various sizes and made by various companies, and in our research, we have used 10 flex sensors of 2.2 inches each placed on five fingers. Each flex sensor consists of two strips, one is a thin strip connected to the high voltage (that can be 5v or 3.3v) and another is a fat strip consisting of the polymer ink connected to ground through a resistor and provides the resistance output as well.

By connecting the flex sensor with a static resistor to create a voltage divider, it produces a variable voltage that can read by a microcontroller's analog-to-digital converter [9].

The accelerometer can use to measure the static acceleration of gravity in tilt- sensing applications and can used to measure dynamic acceleration resulting from motion, shock, or vibrations. By measuring the amount of static acceleration due to gravity, we can find out the angle the device tilted at with respect to the earth and by sensing the amount of dynamic acceleration, you can analyze the way the device is moving. There are many different ways to make an accelerometer.

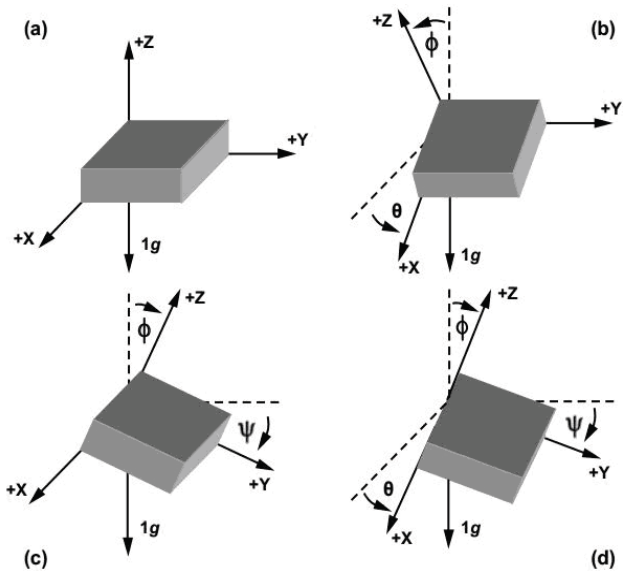


Figure 3: Tilt and inclination of the accelerometer

Acceleration along the axes deflects the movable mass. This displacement of moving plate (mass) unbalances the differential capacitor, which results in sensor output. Output amplitude is proportional to acceleration. 16-bit ADC used to get digitized output. The full-scale range of acceleration is +/- 2g, +/- 4g, +/- 8g, +/- 16g. It measures in g (gravity force) unit. When device placed on flat surface it will measure 0g on X and Y-axis and +1g on Z-axis [11].

IV. Methodologies

The classification of this article is based on Random Forest Algorithm.

A. Random forest algorithm

Random forests or random decision forests are an ensemble learning method or supervised classification algorithm for classification, regression and other tasks, that operate by constructing a large number of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This algorithm creates a forest with a number of trees and the more trees in the forest the more robust the forest looks like.

Random decision forests correct the decision trees' habit of over fitting to their training set. Ensemble algorithms are those, which combine more than one algorithms of same or different kind for classifying objects [13]. For example, running a prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object.

Similarly, in the random forest classifier, higher number of trees in the forest results in higher accuracy of the outputs. Random forest classifier creates a set of decision trees from randomly selected subset of training set and then it aggregates the votes from different decision trees to decide the final class of the test object.

Suppose training set is given as: [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4], random forest may create three decision trees taking input of subset for example [14], [X1, X2, X3], [X1, X2, X4]& [X2, X3, X4]

So finally, it makes predictions based on the majority of votes made from each of the decision trees. This algorithm works well because a single decision tree may be prone to a noise, but aggregation of many decision trees reduce the effect of noise gave more accurate result for a different sets of input. The random forest can apply weight concept for considering the impact of result from any decision tree. Tree with high error rate give low weight value and vice versa. There are two stages in this algorithm, one is random forest creation, and the other is to make a prediction from the random forest classifier created in the first stage.

The whole process is explained below.

Random Forest creation pseudo code:

1. Randomly select “k” features from total “m” features.
   a. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

Random forest prediction pseudo-code:

- Takes the test features, use the rules of each randomly created decision tree to predict the outcome, and stores the predicted outcome (target).
- Calculate the votes for each predicted target.
- Consider the high voted predicted target as the final prediction from the random forest algorithm.

The over fitting problem will never come when we use the random forest algorithm in any classification problem.

- The same random forest algorithm can be used for both classification and regression task.
- The random forest algorithm could be used for feature engineering.
- It means identifying the most important features out of the available features from the training dataset.

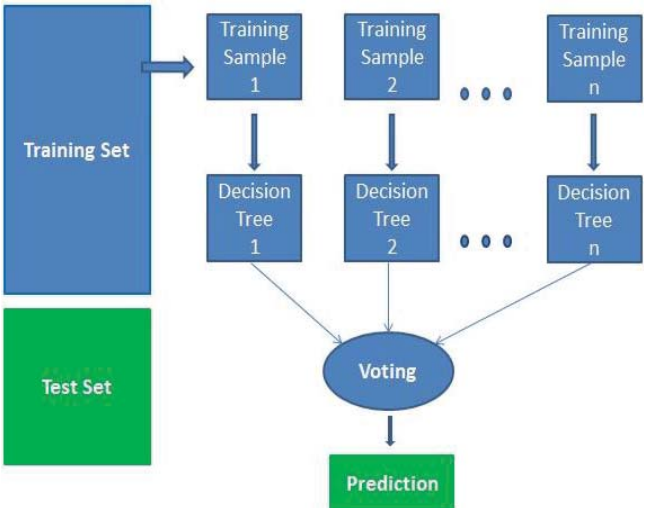


Figure 5: Working of Random Forest algorithm [15]

Random forest considered as a highly accurate and robust method. It does not suffer from the over fitting problem, this algorithm can be used in both classification and regression problems.

Random forest can also handle missing values, with two ways: using median values to replace continuous variables, and computing the proximity- weighted average of missing values.

Random forest is slow in generating predictions because it has multiple decision trees. This algorithm is time – consuming. The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

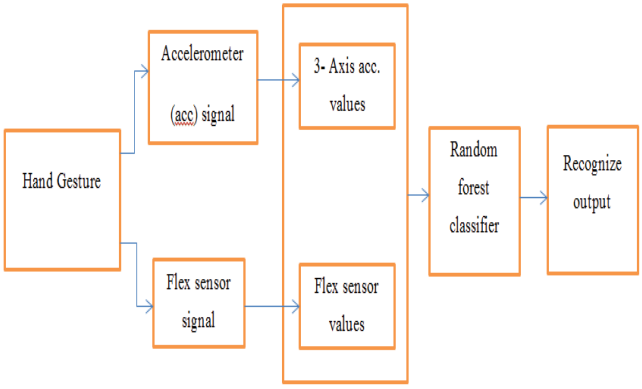


Figure 6: Feature Extraction Diagram

When a user makes a hand gesture, eight user inputs from the glove controller has given to the system in which there is 3- axes accelerometer signal and five flex sensor signals. Three axes values and five flex sensor values that has taken from the user acts as the features for recognizing a particular gesture. These features for the gesture has passed through random forest classifier, then the gestures has classified according to the features, and the output is recognized.

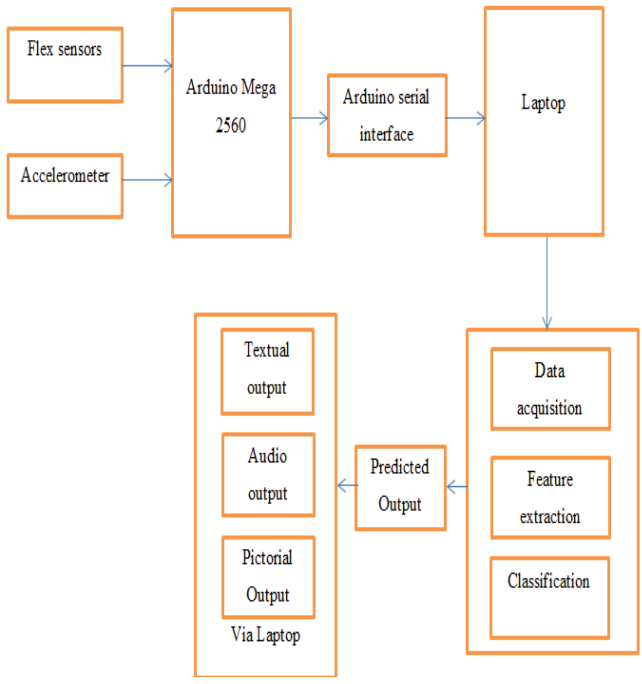


Figure 7: System Block Diagram

With the machine, has trained with those dataset, for the real time application of the trained machine, for a particular gesture the flex sensor and accelerometer data were collected from the glove movement and processed by the Arduino converting those raw data to meaningful data. These values help in feature extraction process. Those data has sent from the Arduino to the serial port, has collected by the python through the serial interface and saved as temporary dataset (CSV format) form. Finally, the temporary dataset passed through the machine such that the machine extracts the features of those dataset and predicts the appropriate output for that gesture with reference to the Random Forest Classifier model. As the output has recognized by the model, laptop screen and speaker has used to display the output to the user.

V. Algorithms

A. Algorithm for dataset preparation at Arduino Mega 2560

Step 1: Start

Step 2: Input: For a particular gesture, 8 sensor data inputs (5 flex sensors data and 3 accelerometer data)

Step 3: Convert raw sensor data to meaningful data

Step 4: Set the baud rate

Step 5: Send data to the serial port

Step 6: Is Arduino buffer flushed?

If yes, go to step 7

Else, wait for buffer to flush then continue

Step 7: Add delay of 10 milliseconds.

Step 8: Go to step 2.

B. Algorithm for real time application

Step 1: Start

Step 2: Train the model using the training dataset.

Step 3: Set the baud rate similar to Arduino serial baud rate and open the serial connection.

Step 4: Is there data at serial port?

If no, wait till data available then continue.

Else, go to step 5.

Step 5: Input: Collect data from the serial port at real time.

Step 6: Open new file in write mode.

Step 7: Write the collected data the new file in CSV format.

Step 8: Flush the Arduino buffer.

Step 9: Close the Serial port connection.

Step 9: Pass the temporary dataset through the model.

Step 10: Trained machine predicts the output.

Step 11: Display the prediction on the screen and play the audio of the predicted word.

Step 12: Add 1-second delay

Step 13: End

VI. Results

Finally, we have obtained a system that can read the values for a particular gesture done by the user, predict the output for the gesture, display it on the laptop screen, and provide an audio output via a desktop GUI. The dataset for all the alphabets and frequently used words has created. Using those datasets, the machine has trained and the model created. Then dataset for all the alphabets combined and shuffled in order to train the machine with reduced variance and make sure that model remains general and over fit less. When the input has taken for sign language of the alphabet or word, the input passed through the trained machine and the closest value predicted display and the audio played as output. Then the correlation plot of each alphabet was compare with other alphabet to check the level of closeness of alphabets with each other and predict the accuracy of the model. The accuracy for our model founded to be 96.8 %.

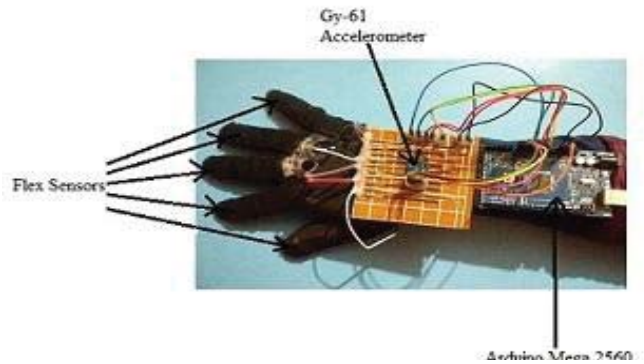
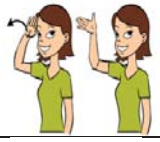


Figure 6: Glove Controller

List of Signs our system is train to predict

Sign	ASL	Sign	ASL
	A		AND
	B		NO
	C		EAT
	D		I NEED TO GO TO WASHROOM
	E		I
	OKAY		STOP
	HELLO		THANK YOU
	CAN		BYE

Some Dataset Values for Some Gestures of our dataset

Index	SVTHUMB	SVINDEX	SVMIDDLE	SVRING	SVPINKEY	X-axis	Y-axis	Z-axis	Cl
0	277	68	151	131	168	390	359	354	A
1	281	71	139	132	174	383	322	365	A
2	272	80	142	141	185	392	383	348	A
3	286	78	145	142	179	399	323	338	A
4	270	100	166	158	203	398	356	352	A
5	288	66	148	135	150	394	389	383	A
6	271	75	155	142	159	374	299	386	A
7	246	70	144	135	151	388	339	350	A
8	272	69	139	128	151	388	353	370	A
9	287	84	155	144	172	374	299	386	A
10	251	74	148	134	161	388	339	350	A

Figure: Dataset Values for alphabet A

Index	SVTHUMB	SVINDEX	SVMIDDLE	SVRING	SVPINKEY	X-axis	Y-axis	Z-axis	Class
208	240	165	184	182	214	388	351	367	C
209	224	158	179	175	208	392	351	349	C
210	234	163	182	179	226	393	352	361	C
211	218	143	173	165	210	392	352	362	C
212	226	162	180	176	217	393	352	344	C
213	223	148	176	170	204	392	351	362	C
214	228	156	180	175	210	393	352	367	C
215	196	125	151	136	189	392	356	363	C
216	213	143	163	157	218	391	358	365	C
217	181	135	160	142	199	395	353	357	C

Figure: Dataset Values for alphabet C

Index	SVTHUMB	SVINDEX	SVMIDDLE	SVRING	SVPINKEY	X-axis	Y-axis	Z-axis	Class
816	269	219	242	248	264	390	366	351	Hello
817	260	219	240	240	259	398	342	348	Hello
818	266	221	244	250	284	394	310	345	Hello
819	264	223	242	252	275	376	382	350	Hello
820	262	207	239	228	252	399	322	340	Hello
821	268	217	238	239	249	388	380	343	Hello
822	259	215	239	238	264	397	349	348	Hello
823	253	219	244	242	267	392	304	355	Hello
824	275	219	240	243	257	376	378	352	Hello
825	273	211	241	238	265	393	352	350	Hello

Figure: Dataset Values for Word Hello

Index	SVTHUMB	SVINDEX	SVMIDDLE	SVRING	SVPINKEY	X-axis	Y-axis	Z-axis	Class
506	226	119	255	277	320	395	544	351	Okay
507	229	124	254	269	313	391	347	372	Okay
508	234	132	255	272	310	396	349	355	Okay
509	232	125	257	274	311	391	350	369	Okay
510	230	131	257	265	294	398	345	352	Okay
511	226	132	230	236	282	396	342	360	Okay
512	229	131	249	269	299	390	349	370	Okay
513	224	127	263	278	319	387	346	375	Okay
514	224	103	259	275	311	391	343	371	Okay
515	187	121	233	242	281	390	361	357	Okay

Figure: Dataset Values for Word Okay

VII. Conclusion

Smart gloves which take input from the users' gestures and gives a textual output. It helps the mute people to communicate with one another and with the normal people. The mute people use their standard sign language to communicate which is not easily understandable by common people without the knowledge of the sign language. Thus, this system converts the sign language into audio, pictorial and into text form, improving their life in a significant way Image processing can be applied to take in new gestures from the user themselves and store it in the dataset. Make proper, well finished and a wireless wearable technology to bridge the communication gap with accurate and precise result.

Refrences

[1] Hope M Hurlbut, "A Lexicostatistic Servey of the Signed Languages in Nepal," SIL Electronic Survey Reports, p. 23, June 2012.

[2] X Chai et al., "Sign Language Recognition And Translation with Kinect," Institute of Computing Technology, CAS, Microsoft Research Asia, Beijing, China, October 2013.

[3] Y Quan and P Jinye, "Application of improved sign languages recognition and synthesis technology," Industrial Electronics and Application, 2008, ICIEA 2008, 3rd IEEE Conference, pp. 1629-1634, June 2008.

[4] D Jun, "Gesture Recognition," Shan Hai Maritime University, China, May 2004.

[5] Ata Ur Rehman, Salman Afghani, Muhammad Akmal, and Raheel Yousaf, "Microcontroller and Sensors Based Gesture Vocalizer," Rachna College of Engg. and tech., Air University Islamabad, Islamabad, Research Paper 2008.

[6] Prerana K.C, Nishu Mahato, and Bishruti Satyal, "Hand Gesture Recognition using KINECT," Nepal Engineering College, Baktapur, Bachelor Thesis P3-I.0-14-4, 2015.

[7] M. S Kasar, Anvita Deshmukh, Akshada Gavande, and Priyanka Ghadage, "Smart Speaking Glove-Virtual tongue for Deaf and Dumb," Internaltional Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 5, no. 3, p. 7, Mar. 2016.

[8] Sparkfun. (2015) learn.sparkfun.com. [Online]. https://learn.sparkfun.com/tutorials/flex-sensor-hookup-guide/all.

[9] Robotpark. robotpark.com. [Online]. http://www.robotpark.com/GY-61-DXL335-3-Axis-Accelerometer-Module.

[10] Arthur Samuel, "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229, 1959.

[11] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, pp. 832-844, 1998.

[12] Savan Patel. (2017, May) medium.com. [Online]. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1

[13] Avinash Navalani. (2018, May) datacamp.com. [Online]. https://www.datacamp.com/community/tutorials/random-forests-classifier-python

* * *

Socioeconomic Impacts of FMIS after Rehabilitation: Case Study of Bakultar Irrigation Project, Mahadevsthan, Dhading

Dharma Raj Bagale[1], Lal Krishna BC[2]

[1]Department of Civil & Rural Engineering, Nepal Engineering College (nec), Changunarayan, Bhaktapur, Nepal, dharma.bagale@gmail.com

[2] Central Campus of Humanities and Social Sciences, Mid-Western University, Birendranagar, Surkhet, Nepal, lk_bc8@yahoo.com

Dharma Raj Bagale *has completed M.A. in Rural Development from Mid-Western University, Surkhet and B.E. in Agriculture Engineering from Institute of Engineering, Tribhuvan University. He is currently working as Junior Professor in the Department of Civil & Rural Engineering at Nepal Engineering College since March 2016. His field of research interest is Integrated Water Resource Management (IWRM). Apart from academic activities, he is also involved in the study of various irrigation and renewable energy projects.*

Lal Krishna BC *is an Assistant Professor at Central Campus of Humanities and Social Sciences, Mid-western University, Surkhet. He has received his Master's degree of Arts in Rural Development from Tribhuvan University.*

Abstract

The main purpose of the study is to focus on the socioeconomic impacts of rehabilitation of Bakultar Irrigation Project in the Mahadevsthan, Dhading district, Nepal. Primary information was collected through household survey, Key Informant Interview, and FGD. Both qualitative and quantitative techniques were used for data analysis. The rehabilitation intervention of the irrigation project has changed the cropping pattern and increased the crop productivity. The cropping intensity has increased to 244.6 % from 174 %. The study shows that rehabilitation of irrigation project has positive impact over poverty reduction, food sufficiency, gender equity, migration control and reduction of women's drudgery in household works.

Keywords *: Agriculture, Irrigation, FMIS, Rehabilitation, Socioeconomic impacts*

I. INTRODUCTION

Agriculture forms the backbone of Nepali economy, contributing about 34.35% in the GDP and providing direct or indirect employment for two-third of the population [4]. Irrigation plays vital role in agricultural productivity enhancement. Of the country's total cultivable area 2,641,000 hectares, two-third is potentially irrigable but irrigation facility is available to only a total area of 1,362,000 hectares and merely 36% of irrigated area has year round irrigation [3]. The crop productivity under rain fed condition is fairly low as compared to irrigated. Therefore, it is necessary to increase areas under irrigation through new projects or improving the efficiency of existing projects. The development of irrigation projects in Nepal are mainly done by Government Organizations, Non-government Organizations and water users. The government's ultimate aim is to provide year round irrigation facilities and increase - irrigated areas of the country. But the impact of irrigation in any area needs to be assessed for further investments.

This study focused on the socioeconomic impacts of "Bakultar Irrigation Project" located at Mahadevsthan, Dhading after its rehabilitation. It has a net command area of 29 ha, extracting 75 lps water from perennial source of the Malekhu River.

II. METHODOLOGY

A. Research Design

The study is based on descriptive as well as exploratory research design. It describes the situation of farmers and assesses the socioeconomic impacts after the rehabilitation of the irrigation project.

B. Study Area

The study area lies in the Mahadevsthan of Dhading district, Nepal. It is accessible through a gravel road of about 10km distance from Malekhu in the Prithivi Highway.

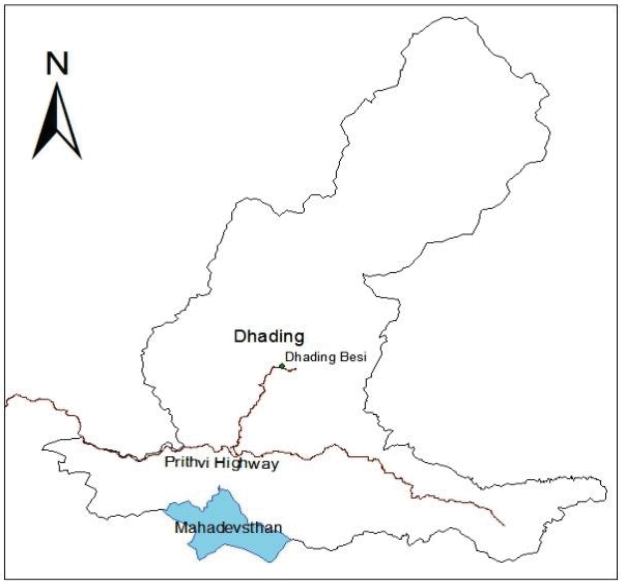


Fig. 1: Location Map of the Study Area

C. Universe and Sample

The project area covered a total of 58 households (HHs) only, therefore all the HHs were selected for the study adopting census method. The perception of the people at head, middle and tail of the irrigation system was given equal attention as the availability of water may usually low at tail than head reach.

D. Methods of Data Collection

The study is based on both primary and secondary data collected during the period of May to July 2016.

1. Primary Data:

The primary information was collected via household survey, Focus Group Discussion with water user association (WUA), Key Informants Interview with project officials.

2. Secondary Data:

Secondary data regarding the rehabilitation of the irrigation project was obtained from Department of Irrigation (DOI), Irrigation Development Division (IDD) and District Agriculture Development Office (DADO), Dhading.

E. Data Processing and Analysis

After collecting data, the calculations related to analysis were performed with the help of MS Excel. Both qualitative and quantitative techniques were used for data analysis.

III. RESULTS AND DISCUSSION

A. Overview of the Irrigation System

Bakultar Irrigation Project is a Farmer Managed Irrigation System (FMIS). According to local people, this is a century old system but not functioning since some decades. The farmers started functioning of the canal again in the year 2062 B.S. and also constructed the main canal by mobilizing local resources themselves. But most part of the main canal was unstable due to seepage and landslide. The rehabilitation work was started on 15thAshwin 2068 B.S. and was completed on 15thChaitra 2069 B.S. partly by the WUA and partly by the contractor [2].

B. Landholding and Land Tenure System

Most of the households in the project area were land owners but there were tenants too, who work as labor force. The land distribution in the command area was not uniform. Figure 2 shows land holding size of farmers.

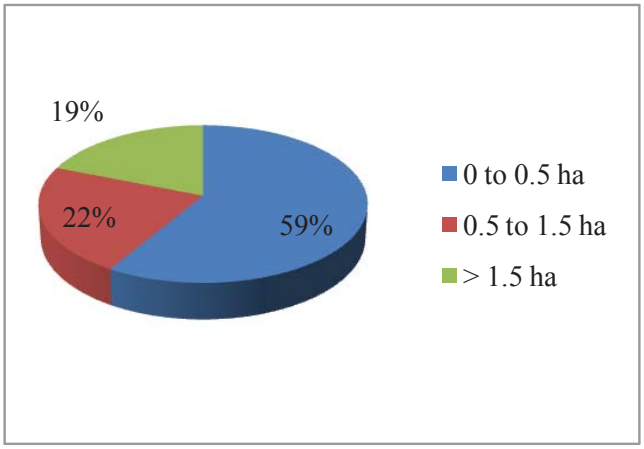


Fig.2: Land Holding Size of Farmers

C. Occupation, Income and Expenditure

Regarding the source of income, it was found that agriculture was the main source of income. It is followed by services, foreign employment, business, wages and others. Figure 3 presents the occupation of farmers.

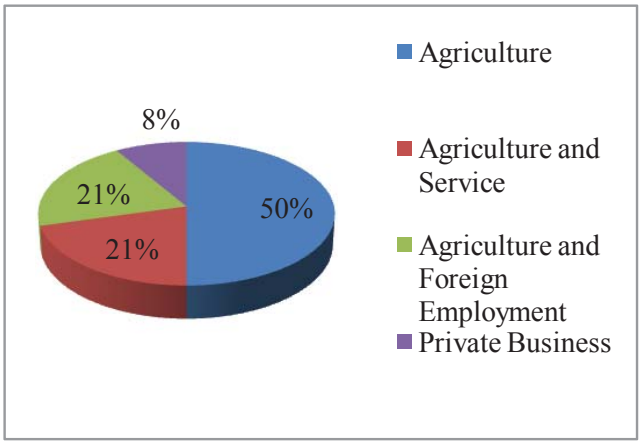


Fig. 3: Source of Income of Farmers

The average annual household income was found to be NRs 274,700 while the average annual expenditure was NRs 196,225.

D. Change in Cropping Pattern

It was found that the cropping pattern has been changed in the area due to the rehabilitation of irrigation project. Paddy–Vegetable, Paddy-Fallow-Maize, and Paddy-Vegetable-Maize were the general cropping pattern in the project area. Before irrigation, hardly two crops were grown in a year. The winter crops like mustard, wheat and vegetables; and paddy, also in spring, are possible only after rehabilitation of irrigation system. Paddy-Mustard-Paddy, Paddy-Mustard-Maize, Paddy–Wheat-Maize and Paddy-Mustard-Vegetables are the cropping pattern after the rehabilitation of the project. The vegetables grown in the project area mainly include cabbage, cauliflower, sponge gourd, bitter gourd, okra, potatoes, etc.

E. Change in Cropping Intensity

Cropping intensity is the ratio of total area of harvested crop to total available land, and is reported in percent. The change in cropping intensity is only possible due to irrigation facility. So, the direct effect of irrigation can be visualized with increase in cropping intensities. The figure 4 demonstrates the comparison of cropping intensity during different season before and after rehabilitation of the project.

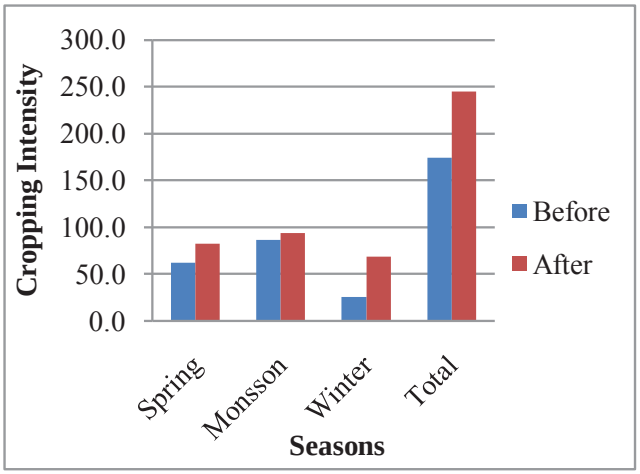


Fig. 4: Cropping Intensity Before and After Project Rehabilitation

The cropping intensity changed to 244.6 % from 174.1 % after project rehabilitation. The increase in cropping intensity is due to availability of water during winter and spring season.

F. Agriculture Production and Productivity

Paddy and Maize are the major crops grown during spring season. The average production of paddy increased to 3.9 from 2.8 mt/ha and that of maize to 1.9 from 1.1mt/ha. During monsoon season, paddy is the major crop grown. The average production of paddy increased to 3.8 from 1.3mt/ha. Mustard, wheat, vegetable are major crops grown during winter season. The average

productivity of mustard, and wheat are 0.4 and 1.5 mt/ha, respectively. The average productivity of vegetables increased to 9.2 from mt/ha after the rehabilitation of the project. The increase in crop productivity is mainly due to water availability, use of hybrid varieties of paddy and maize replacing the local varieties, and the use of fertilizers.

G. Socioeconomic Impacts

The long term impacts due to rehabilitation of irrigation project are poverty reduction in the project area, sufficiency of food items in farmers' household, shift of occupation from farm to off farm, reduction of women's drudgery in household works, gender issue and migration control. The socioeconomic impacts of Bakultar Irrigation Project after rehabilitation are described as follows.

1. Income from Agriculture Production:

Most of the farmers usually depend on agriculture for their main income. After rehabilitation of the irrigation project, agricultural production increased, which ultimately raised the income of individual farmer. The increase in income is dependent on size of landholdings. The farmers with large land holdings have given their land in lease and got involved in animal rearing and horticulture. Similarly, the farmers with small landholding have taken others' land in lease to meet their food demand. However, the income is not substantial as the farmers are lagging in the diversification of cropping and problem to access direct market. Each individual farmer have to take his/her agricultural production in the Malekhu Bazar, some 10 km away from the project area. This is due to unavailability of collective method of marketing. This deprived farmers of getting fair price for their production. The comparison of household with income from agriculture before and after project rehabilitation is illustrated in Table I.

TABLE I
INCOME FROM AGRICULTURE BEFORE AND AFTER THE PROJECT REHABILITATION

Income from Agriculture (NRs '000)	0 – 20	20-50	50-100	above 100
Nos. of HHs Before	28	16	11	3
Nos. of HHs After	8	12	22	16

2. Poverty Reduction

The reduction in poverty is directly related with income from agriculture production, employment opportunity created by the project, sufficiency of food items, level of expenditure on education, sanitation, health services, purchasing of durable goods and construction and maintenance of houses. From the data presented above, reduction in poverty in the study area is as shown in figure 5.

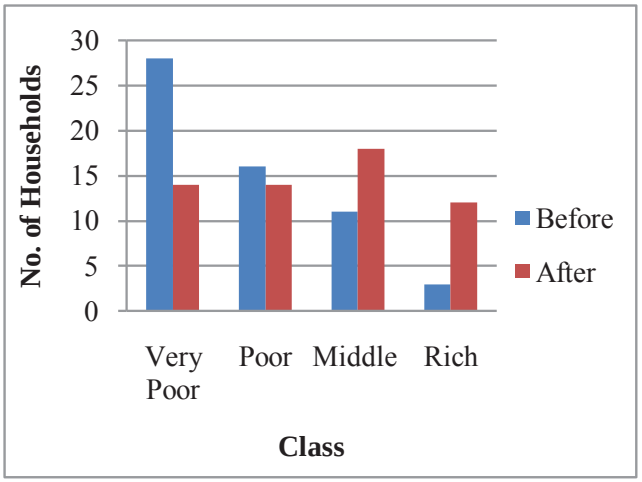


Fig. 5: Comparison of Poverty Level before and after Rehabilitation

3. Employment:

The rehabilitation of the irrigation project created both short term employment during construction period and long term employment due to productive use of new irrigated land and intensification of the crop production. Almost 20-30% of the total project cost was covered by labour (both skilled and unskilled).The rehabilitation project also generated employment opportunities for inactive work force, semi-skilled and mostly unskilled, striving for employment in project area and neighboring villages during the construction period. The land areas that remained fallow during winter season were cultivated with mustard, wheat and vegetables. Most of the farmers in project area, who used to work as labour force in neighboring village, are now engaging in their farm land throughout the year.

4. *Food sufficiency:*

The comparative situation of food sufficiency due to the agricultural production before and after rehabilitation of irrigation system is as shown in Figure 6.

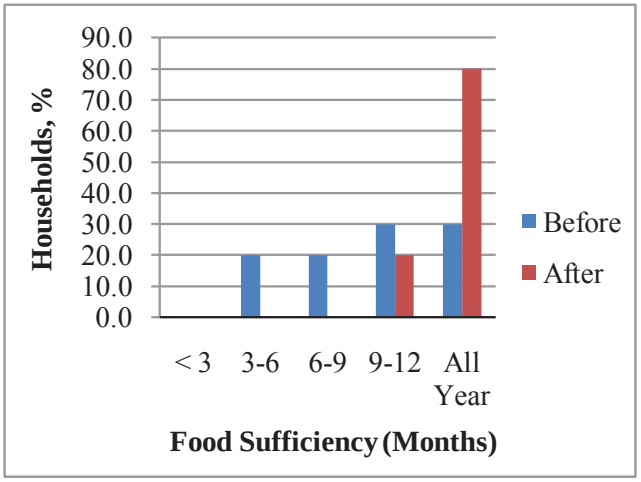


Fig. 6: Comparison of Food Sufficiency

The food sufficiency in Bakultar area indicates good progress and is increasing satisfactorily. The above figure clearly shows that the satisfactory change in food sufficiency due to rehabilitation of the project.

5. *Expenditure Level of Farmers:*

The basic education for children and health services are provided by government. But, for higher education and complex health problem, people in the area have to visit either district headquarter or Kathmandu, which is an economical burden for them and have to depend on banks, cooperative and relatives for loan. However, with the increase in the income level, the farmers have ability to disburse for basic as well as other requirements for their family members. The earning from the sales of agriculture product is mainly expended on the purchase of daily consumable food items insufficient from their farm production.

6. *Women's Drudgery*

Women in project area are main workers in agricultural farms. Due to the rehabilitation of the irrigation project, the women's drudgery has been reduced radically. Workload of most of the women is reduced due to lining of canal and ultimately lesser regular repair & maintenance of canal. Irrigation helped timely planting of crop; reduce weed infestations especially in paddy thereby reducing the labor requirement. The time required for fetching water to their livestock is reduced.

7. *Gender Issue*

Reference [1] emphasizes on the women's involvement in water user association. But, women's involvement in decision-making does not seems enough since the beginning of the project rehabilitation. Only two women were included in the first 11 member executive committee holding one major post, i.e. vice-chairperson. Similarly, only three women were included in second executive committee of 9 members with one major post, i.e. treasurer.

However, women played active role during construction work. The majority of physical works was done by women. They participated both as labour contributions in WUA works, and as paid labour in WUA payable and contractor's works. Women are also contributing in regular repair and maintenance of the system.

Only few women are involved in meetings, trainings, and exposure visits. So, they still need support in capacity building and income generation activities.

8. *Migration*

The migration is a very challenging for developing countries like Nepal. The economically active age group is attracted for better opportunity in and outside of county, which ultimately created the manpower problem in any area.

The Bakultar Irrigation Project is somehow able to control such migration, mainly due to active population engagement in food production and employment opportunity created in the area. The seasonal migration, primarily in winter season in the Kathmandu and India is also largely reduced as the farmers are busy in their own field works.

9. *Shift of Occupation*

Due to the sufficiency of food from the farm after the rehabilitation of project, many farmers with large landholding are giving their land in lease to

marginal farmers and they are involved in private business like vehicle services in the village, shop, animal rearing, horticulture, etc. The number of such households increased to 11 from 3 households.

IV. LESSONS LEARNED

There are many FMISs in hilly regions of Nepal which are not working efficiently and farmers are facing huge problems to irrigate their fields. Rehabilitation of these systems can be done successfully with the participation of the water users. Increase in water availability and reliability increases the income from agriculture and ultimately can have positive impact in socioeconomic of the area. However, the WUAs are not capable for smooth operation and maintenance of the irrigation system as the Irrigation Service Fee (ISF) collection is in very poor condition. Mainly, the "water guard" appointed for operation of the system is unable to get payment on time.

The farmers are not able to earn sufficient income from their farm although the productivity has increased. It is due to the problems in market and pricing system. The middle-man gets more profits than the producers. Therefore, appropriate mechanism is needed so that farmers can get fair price for their production, through which sufficient ISF can be collected for operation and maintenance of the system.

V. CONCLUSION & RECOMMENDATIONS

A. Conclusion

The rehabilitation of Bakultar Irrigation Project has been able to increase socioeconomic condition of the area. Poverty reduction, sufficiency of food items, shift of occupation from farm to off farm, reduction of women's drudgery in household works, gender inclusion and reduction in migration are such impacts.

B. Recommendations

Improvement in agricultural inputs, such as use of hybrid seed, proper dose and timing of fertilizer etc. are highly recommended.

The farmers should focus the activities to develop the project area as organic farming center.

Agriculture Cooperative is needed to be formed in the project area to organize the farmers and solve the market problem, so that the gap between the farmer and the market is reduced and farmers can enjoy the fair price of their produce.

REFERENCES

[1] GoN, *Irrigation Policy*. Kathmandu: Government of Nepal,2060.

[2] IDD Dhading,. *Bakultar Irrigation Project*:*Detail Project Report*. Dhading: Department of Irrigation,2011.

[3] NPC, *Three Years Plan (2014-2016)*. Kathmandu: National Planning Commission, 2014.

[4] World Bank, *Nepal Report*. Washington DC: The World Bank, 2014.

* * *

An Insight on Social Media Stream Mining

Rojina Deuja[1], Krishna Bikram Shah[2]
Department of Computer Science and Engineering, Nepal Engineering College
Changunarayan-5, Bhaktapur, Nepal
[1]rojinadeuja33g@gmail.com, [2]krishnabs@nec.edu.np

***Rojina Deuja** has received her Bachelor of Engineering in Computer from Nepal Engineering College, Pokhara University, Nepal. She has been working on Data Warehousing and Business Intelligence for over a year. Her interests lie in Data mining, Machine Learning and Artificial Intelligence.*

***Krishna Bikram Shah** is Assistant Professor at Department of Computer Science and Engineering, Nepal Engineering College. He holds M. Tech. in Computer Science and Engineering from Sikkim Manipal University, India. He has served at several computing departments in India. His research interest include Artificial Intelligence, Bioinformatics and Machine Learning.*

Abstract

Data stream mining is one of the realms gaining upper hand over traditional data mining methods. Transfinite volumes of data termed as Data Streams are often generated by Internet traffic, Communication networks, On-line bank or ATM transactions etc. The streams are dynamic and ever-shifting and need to be analysed online as they are obtained. Social media is one of the notable sources of such data streams. While social media streaming has received a lot of attention over the past decade, the ever-expanding streams of data presents huge challenges for learning and maintaining control. Dealing with billions of user's data measured in petabytes is a demanding task in itself. It is indeed a challenge to mine such dynamic data from social networks in an uninterrupted and competent way. This paper is purposed to introduce social data streams and the mining techniques involved in processing them. We analyse the most recent trends in social media data stream mining to translate to the detailed study of the matter. We also review innovative implementations of social media stream mining that are currently prevalent.

I. Introduction

Data Streams are high-speed, continuous flow of data. These streams are generated from an array of different sources like online searches, social media, business transactions, phone conversations, network traffic etc. These data are mined by programmers to maintain records, which is eventually used to acquire useful information. This process of extracting knowledge structures from data streams is termed as Data Stream Mining. Data stream mining operates under the common canopy of data mining, machine learning and knowledge discovery.

The process of deriving actionable trends and patterns by analysing raw social media data streams is called Social media mining. It may require either human data analysts or automated software programs to sift through huge amounts of social media data (social media usage, opinion mining, content sharing, online customer behaviour, etc.) in order to discern patterns and trends. These patterns and trends are usually of interest to businesses, companies, governments and

even non-profit organizations, as these systems can use the data patterns and trends to design their strategies or introduce new programs [1].

A major challenge is that social media data streams possess multiple unparalleled properties like transfinite length, concept-drift, concept-evolution, feature-evolution and limited labelled data. Moreover, the volume and speed of the stream itself stimulates added complexities that are not present in traditional data mining. Such streams may have different types of features in addition to the large volume of data. On top of that, the concepts and features tend to evolve throughout the stream. A concept-drift takes place when the underlying concept of data changes over a certain period. Feature-evolution is said to occur when a feature set varies over time. Similarly, whenever new classes evolve in the streams, concept-evolution takes place. There is also a lack of labelled data since manual labelling is infeasible at large scale. Due to these reasons, traditional data mining techniques are indeed incapable of addressing social data streams.

In this paper, we discuss different techniques used for Social Network Analysis (SNA). We also examine the implications for mining social media streams, and enlist four popular online services working on the discovery and analysis of useful information from social media.

II. Tracking data flow

Unlimited units of data are produced in social media sites every second. The essential information that propagates between the users in a social network can be in the form of posts, messages, comments, links, tags and multimedia. There are billions of discrete users who have their own connections, behaviours, preferences and trends.

In order to mine such data, it is necessary to track units of information that correspond to pieces of information (posts, events, articles, etc.). These data vary within a very short time and need to be handled at a very large scale. Below are a few ways in which this can be done:

A. Cascading links to article

A hyperlink is a reference to a document or a resource typically activated by clicking or tapping. Hyperlinks are reliable to trace since they travel relatively unchanged. They are widely used in blogs, microblogs, social networking sites and internet forums. Users write posts that refer (link) to other posts and in this way the information propagates. The transfer from such hyperlinks can be identified and cascade graphs can be constructed by a time ordered propagation of information as shown in Fig. 1.

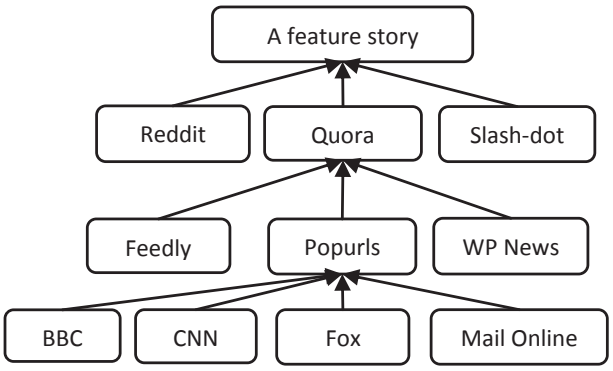


Figure 1: A cascade of articles tracing a feature story in the web

Previously, hyperlinks were limited to designing the structure of networks, where posts were modelled as nodes and hyperlinks as edges in a directed graph. Jackson [2] suggested that social network analysis (SNA) can be carried out using hyperlinks in websites. He contented hyperlink-based network analysis as a strong approach for studying representations of online networks. Adamic and Glance [3] analysed the connectedness of political blogs by studying the linking behaviour in the blog posts. Miller et al, [4] used a combination of graph theory and sentiment analysis fields to track sentiment flow in blog post networks traced through hyperlinks. Park [5] presented hyperlink network analysis (HNA) for the study of online social (or communication) networks. He also explained various data-gathering techniques used for HNA.

B. Keywords and Hashtags

Keywords are words or phrases which are used as a metadata tag in websites. At one time, keywords were one of the most important elements used in web searching and text mining. Even today, search strings are matched with keywords.

Hashtags are similar to keywords, but they are preceded by a '#' (hash) symbol. Hashtags are mostly associated with driving engagements on a topic. For example, the hashtag #MeToo was used to spark up a movement against sexual violence. Hashtags have been incorporated in social media like Twitter, Facebook, Instagram, Snapchat etc.

The information network that is built with the help of keywords and hashtags can be very crucial in studying online information flow. Techniques like Double Ranking (DR) technique [6], Search set (S) retrieval [7], context-sensitive topical PageRank (CTPR) [8] etc. have been proposed to extract topical keywords in Twitter. Meanwhile, [9] analysed the motivation, goals and usage patterns of social media users through the hashtags that they used. A comparison of users' search behaviours between Web search and Twitter search was done in [10] while the difference between traditional search engines and social media search associated with health related information was explored in [11].

C. Memes

Meme can be regarded as a unit of information that spreads between various users in any social media network through their social ties. Memes can be used to derive valuable data and also to study user or community behaviour in any social network. When a certain meme gets viral, it can be used to attract attention for marketing and advertisement.

Memes can be, thus analysed to follow iterations of any story as they evolve. Leskovec[12] introduced a meme-tracking approach to identify news cycles- which are daily rhythms in the online new outlets. The Memetracker clusters together short, distinctive phrases that characterize a specific topic. It then creates groups of news on topics that can be tracked over time to find diffusion patterns in the news cycle. One difficulty in clustering memes is that the actual granularity of a meme is very difficult to capture with the help of general entities such as tags and keywords. Thus, the messages are grouped on the basis of existing atomic entities called Protomemes. Protomemes are aggregated to place them into cohesive groups that represent a consistent topic of interest. Protomeme detection is a very common pre-clustering procedure used for isolating memes before processing them and has been utilized in works including [13] and [14].

III. Maintaining data quality

After tracking data flow, the data from different sources are gathered using varying methods. Such data do not necessarily come in a pre-processed form. A huge amount of the collected data may be incomplete, incorrect, misleading or inadequate for the required application. For social media streams, the quality of data that is being tapped is much more significant than the scale of the data being accumulated. In order to assure the quality of data, the following discrepancies need to be removed:

A. Noise

Noise is the distortion of the data that is present in the social networks. The signal-to-noise ratio is often very high in social media streams. There are four main sources of noise in social media data:

- Fake/Unused accounts: A lot of social media accounts are not owned or managed by a person of its true identity. Similarly, many accounts in different sites are created by the users but never used again.
- Spam: Spams are mainly characterized by malicious links, malware injection attempts, unsolicited and poorly targeted marketing messages, phishing etc. Spams are undesirable and their appearance impacts user's sentiment analytics and keyword/brand trends.
- Deceptive content: Deceptive content includes information that is produced with a motive to deceive the users and passed on as valid information. For example, in 2016, the news of the death of renowned celebrity Johnny Depp took over Facebook. Many fans, news channels alike were fooled by the news which was later found to be a hoax.
- Duplicate data: Duplicate data are formed when there are multiple instances with the exact same characteristic values. Multiple accounts of the same user, Retweets, Re-posts, Shares, Forwards, Auto blogs, Splogs, Tweetbots, Cross-posting etc., can lead to duplicate copy of the same data. Duplicate data hampers the magnitude and validity of statistics in a sample stream.

Clearly, noise in data streams is undesirable and needs to be filtered out. Even so, there are a few filtering algorithms that are effective in combating the adverse effects of noise. In [15], the noise contributed by users was detected with the help of data collected from Twitter and forums. Several

classical strategies along with graph algorithm were used to filter out noises from the data. In another work, user interests were evaluated in order to address dynamic issues in noise web data reduction [16]. The authors contend that noise in online data should be isolated by taking user interest levels on web data into account. Here, a tool called the Noise Web Data Learning (NWDL) - capable of learning noisy web data in user profiles was presented. Similarly, in [17] a Noise Removal and Structured Data Detection (NR-and-SDD) algorithm was proposed, which reduces processing cost of noisy data and improves structured data detection in relevance to personality features.

B. Missing Data

Missing data are those feature values that are missing in instances. The missing information may either be that of an actor (user) or the link with other users. For example, individuals may avoid reporting profile information on social media, such as their age, location, or hobbies. Missing values limit the capacity to describe the network context of the actors and their neighbours.

There are two distinct forms of missing data: node level and tie level. Node level is when the data of a particular person or object is missing, as if they were not part of a study at all. Node level missing data is characterized by unit non-response. Unit non-response is when the actor does not participate. Tie level missing data occurs when any unit connecting the social contacts of that unit to other units in the network is unavailable [18].

If not excessive, the instances containing the missing values can be removed. Another strategy is to entirely ignore the missing node or tie, as if it never existed. But ignoring missing data can have a large effect on descriptive analysis of social networks. An additional approach is to impute the data that is missing. Imputing involves assuming or guessing a missing node and its properties. Several methods have been suggested for this purpose: One is to use an ERG model to model the dataset, then filling in the missing data through maximum likelihood estimates which are based on the ERGM parameters. K-tree model [19] was used to accurately estimate properties of complete cascade with up to 90% data missing. Similarly, Ward and Hoff [20] used Latent space models which are based on logistic regression to replace the missing data.

IV. Analysing data streams

The data that retrieved from social streams presents new opportunities to study human behaviour, while still challenging our capability to analyse them. It is necessary to analyse the underlying mechanisms within the data streams to keep with the real-time spread of information through the on-line networks. Predictive models of social interactions can be constructed if we understand how messages flow through social networks and how we can identify networks over which important messages spread. Below is a list of techniques that have been widely implemented in analysing dynamic social media streams.

A. Graph theory

Graph theory relates to the study of graphs using mathematical or pictorial representation of a network that models the relationship between various objects. A graph constitutes of a set of vertices or nodes, with certain pairs of these nodes connected by edges (undirected) or arcs (directed).

Graph theory is credibly one of the widely-used methods for social network analysis since the early times of social network. In such networks, a vertex denotes a user and an edge denotes any form of association between the corresponding vertices (e.g. Fig. 2). One major advantage of modelling relationships in a graph is that the edges and vertices can be easily added or removed.

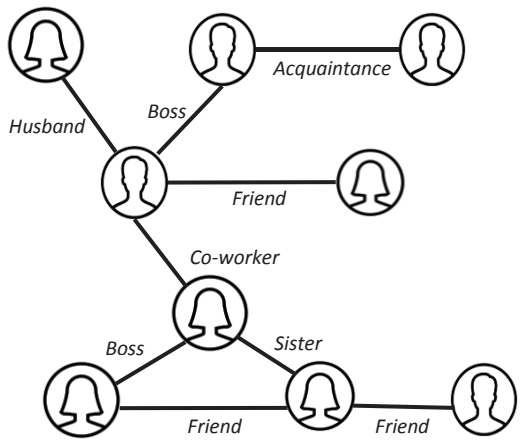


Figure 2: A graph showing different types of associations between users

Network models like Random graphs, Erdos–Rényi model, Barabási – Albert model, Agent based, Fitness model and Boolean Network are commonly used for the purpose of social analysis. Räisänen [21] employed random graphs as models of Peer-to-Peer social networks. Three random graph models along with their degree distributions were studied by performing numerical simulations for each model. Ioannides [22] wrote about ongoing research on networks by illustrating three main strands of literature on social networks. He outlined different graph theories, old and new, and then discussed Markov random graphs of social interaction. Hamill and Gilbert [23] compared the characteristics of four standard models (regular lattice, random, small-world and preferential attachment) to conclude that they do not reproduce social networks quite well. They proposed an agent-based model based on the ideas of social space and social distance termed as a social map to properly replicate the basic natures of large social networks.

B. Community Extraction

Community can be categorized as a group of users sharing common interests in any social group. Communities showcase a distinct character of a user group in social networking sites. For mining the data from a particular social service, it is essential that the communities of users in that particular social network are identified. However, it is no easy take to detect and model the communities formed among the dynamic group of users. The majority of such methods are based on partitioning an entire network into communities, with the expectation of multiple ties within communities and a few ties between [24]. An example of such structure is shown in Fig. 3.

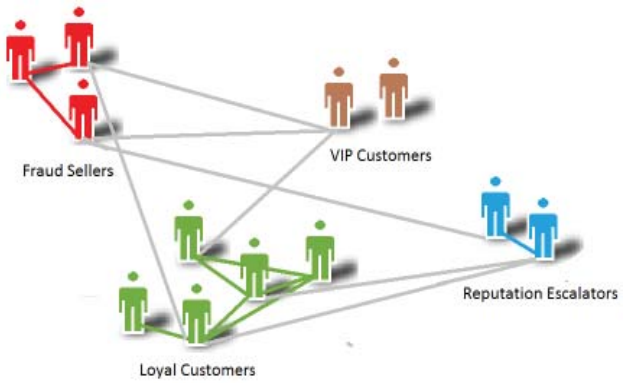


Figure 3: Example of community extraction in a retail business using clustering

Detecting the communities and extracting their behaviour is necessary to study their domain. Several popular link analysis techniques can be used for this purpose which includes HITS algorithm and several Graph Clustering techniques. Many clustering techniques have been applied in works like [25], [26] and [27] for the extraction of community from social network, with hierarchical clustering being mostly used. The hierarchical clustering approach is a combination of techniques where network nodes are grouped to uncover their individual strengths, which is then used to extract communities from the overall network [28].

C. Link Prediction

The data produced in social media sites are complex and ever-changing. The association between specific nodes (user or objects) is comparatively easier to study. While studying social networks, the need of foreseeing the unobserved social links between different users may arise. This can be further described using a set of interesting questions such as: Can associations develop between any two nodes? What factors cause such associations? How do any patterns in the association between nodes evolve over time? Does the association between two nodes affect other nodes?

The challenges that these questions present is to predict the likelihood of a future association between the nodes that is different from the current association between them. This problem is often termed as the Link Prediction problem [29]. For example, in any given social network at time t, it may be needed to predict the social link between actors at time t+1. According to the classical approach for link prediction, the social network is fit into a model and then the model is used for link prediction. Some of the notable models for link prediction are p* model [30], Latent Space model [31] and Dynamic Latent Space model [32]. Besides these, other models have also been proposed that include [33]; [34].

D. Semantic Web of Social Network

The Semantic Web (SW) can be regarded as an extension of the current World Wide Web. The SW environment allows data in different web pages to be structured and tagged in a way that the data can be shared and reused across different applications.

SW has applications in areas including Search Engine Optimization (SEO), Data Integration (DI) systems, Decision Support Systems (DSSs), Seamless applications and Intelligent Agents.

SW technologies are implemented with the help of ontologies to model user data. An ontology is simply a formal conceptualization of a domain. It specifies classes of objects, their instances and relationships between them. Noy [35] demonstrated the evolution of applications of ontology for user modelling and adaptation in relation to the Semantic Web. Plumbaum et al. [36] implemented 17 distinct social applications to identify requirements and attributes for developing a common user model called the Social Web User Model (SWUM), that aggregates information from distributed users. Mika [37] used a community-based ontology extraction that was implemented as part of the Flink system.

Today, one of the most widely used ontologies is FOAF (Friend of a Friend). The FOAF ontology specifies a vocabulary that can be used to define and maintain information describing users, their attributes and their connections with other users. The FOAF is further explained in works including [38], [39] and [40].

V. Examples of social media mining

Trend analysis on social media provides a strong foundation for strategic recommendations. Many applications for processing massive volumes of user data generated by microblogging and social networking services are prevalent. We list a few examples of popular online mining services that use social media streams:

A. Google Hot Trends

Google Trends is a Google Inc. product that is publicly available to everyone. It can be used to find out about how often a particular search is carried out of the total searches across different parts of the world. The popularity of a particular search-item is also categorized in terms of location and language. Then Google launched Google Hot Trends as an add-on to Google Trends.

Google Hot Trends can be installed as an iGoogle gadget. It is can also be utilized as an hourly Atom web feed [41]. Google Hot Trends is slightly more versatile than Google Trends. Google Hot Trends provides insight with a 24-hour search-volume graph on online search results, news, articles, blogs etc. Google Hot Trends collects online data and models them to present them to its users. Hence, it is an excellent example of the implementation of social media data mining.

B. Cloud4Trends

Cloud4Trends is a system that identifies trend and relevant topics in social network through clustering in real-time. It was built as a framework for collecting and analysing user generated content (UGC) in microblogging and blogging spaces. This trend detection application has a 3-tier design: (1) data collection in streaming fashion; (2) application of an online clustering technique on the data to detect recent trends; and (3) refinement and ranking of clusters for data visualization [42].

C. IceRocket

IceRocket is another popular Internet search engine that caters its users with real-time search facilities. IceRocket was launched in 2004, originally with features purported to make online searching easier on a Personal Digital Assistant (PDA) [43]. Today, IceRocket specializes in blog searching. However, IceRocket is also widely employed in mining data streams from social networking websites like Twitter and Facebook. It also mines data from news and articles on the internet. The Big Buzz feature of Ice-Rocket provides all of the previously mentioned services under a single page. Ice-Rocket is a free public web facility that is ad supported. IceRocket owns an API that it licenses to firms and companies monitoring social media streams.

D. SPOONS

SPOONS (Swift Perceptions of Online Negative Situations) are a Netflix service outage detection system via real-time stream analysis [44]. The SPOONS system was developed as an early warning mechanism for outages. It tracks Twitter for those situations when the users are complaining about the same thing at the same time. Apparently, the main motive of the SPOONS system is to use Twitter posts to determine when Netflix users are

reporting a problem with any of the Netflix services. SPOONS work with a combination of term-frequency measurement and sentiment analysis.

V. Conclusion

Social media has become an inseparable part of the lives of people around the world. The data streams from social sites possess great value for companies, scientists and individuals the same. If we are able to accurately analyze the social media streams and acquire useful information from them, then data stream mining will be a hero in true terms. To fully realize the potential of social media streams, new data mining systems need to be developed. Incremental improvements will, of course, continue to help. This will assure that we get huge amount of information from social media users and reap benefits from it.

References

[1] A. Weiler and M. Grossniklaus R. Popovici, Online Clustering for Real-Time Topic Detection in Social Media Streaming Data in 23rd International World Wide Web Conference, Seoul, Korea, 2014.

[2] M.H. Jackson, "Assessing the Structure of Communication on the World Wide Web," Journal of Computer-Mediated Communication (3:1), 1997.

[3] L.A. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election in Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05, pp. 36–43, 2005.

[4] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, J.and C. Potts, Sentiment FlowThrough Hyperlink Networks in ICWSM, 2011.

[5] H.W, Park, "Hyperlink Network analysis: A New Method for the Study of Social Structure on the Web", 2003.

[6] W. Shuai, C. Zhiyuan Chen, L. Bing and S. Sherr, Identifying Search Keywords for Finding Relevant in AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence pp. 3052-3058, 2016.

[7] G. King, P. Lam and M. Roberts, "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text," American Journal of Political Science, 61, 4, pp. 971-988, 2017.

[8] W.X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.P. Lim and X. Li, "Topical keyphrase extraction from twitter in ACL," pp. 379–388, 2011.

[9] J. Sang, C. Xu, and R. Jain. Social multimedia mining: From special to general in Multimedia (ISM), 2016 IEEE International Symposium on , pp 481–485. IEEE, 2016.

[10] J. Teevan, D. Ramage, and M M. Ringel, Twittersearch: a comparison of microblog search and web search in Proceedings of the fourth ACM international conference on Web search and data mining, pp. 35–44. ACM, 2011.

[11] M. D. Choudhury, M. R. Morris, and R. W. White, Seeking and sharing health information online: Comparing search engines and social media in Proceedings of the 32nd annual ACM conference on Human factor s in computing systems, pp. 1365–1376, ACM, 2014.

[12] J. Leskovec, L. Backstrom, and J. Kleinberg, Meme-tracking and the dynamics of the news cycle in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 497–506. ACM, 2009.

[13] E. Ferrara , M. J. Asbagh , O. Varol , V. Qazvinian , F. Menczer , A. Flammini, Clustering Memes in Social Media in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.

[14] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams in Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, 2009.

[15] A. Stoica, Filtering Noisy Web Data by Identifying and Leveraging Users' Contributions in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[16] J. Onyancha, V. Plekhanova , Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests in International Journal of Computer and Information Engineering Vol:12, No:1, 2018.

[17] M. F. Uddin, Noise Removal and Structured Data Detection to improve search for personality features in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, USA, 2016.

[18] M. Huisman, "Treatment of Missing Data in Longitudinal Network Studies," in SIENA workshop, 2009.

[19] E. Sadikov, M. Medina, J. Leskovec, and H. G. Molina, Correcting for missing data in information cascades, WSDM'11, Hong Kong, China, 2011.

[20] P. D. Hoff and M. D. Ward, "Modeling Dependencies in International Relations Networks," Cambridge University Press, Vol. 12, No. 2 (Spring 2004), pp. 160-175

[21] J. Räisänen, Random graphs as models of Peer-to-Peer social networks in U.U.D.M. Project Report, 2012

[22] Y.M. Ioannides, "RANDOM GRAPHS AND SOCIAL NETWORKS: An Economics Perspective," Department of Economics, Tufts University yannis, 2015.

[23] L. Hamill and N. Gilbert, "Simulating Large Social Networks in Agent-Based Models: A Social Circle Model," Centre for Research in Social Simulation, University of Surrey, UK, 2010.

[24] Y. Zhao, E. Levina and J. Zhu, Community extraction for social networks in Proc Natl Acad Sci USA. 2011 May 3; 108(18): 7321–7326. Published online 2011.

[25] M. Girvan, M. E. Newman, Community structure in social and biological networks in Proceedings of the National Academy of Sciences, 2002.

[26] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, "Community detection in social media," Data Mining and Knowledge Discovery, pp. 515-554, 2012.

[27] M. Newman, "Networks: An introduction," Oxford University Press, 2010.

[28] M.A. Olowe, M.M. Gaber, F. Stahl, "A Survey of Data Mining Techniques for Social Network Analysis," pp. 4-7.

[29] D.L. Nowell and J. Kleinberg, The Link-Prediction Problem for Social Networks in Journal of the American society for information science and technology, 2007, pp. 1019-1031.

[30] S. Wasserman, P. Pattison, "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*," Psychometrika, 1996, Volume 61, Number 3, p 401.

[31] P. D. Hoff, A. E. Raftery and M. S. Handcock, Latent Space Approaches to Social Network Analysis in Journal of the American Statistical Association Volume 97- Issue 460, pp. 1090-1098, 2002.

[32] P. Sarkar and Andrew W. Moore Dynamic social network analysis using latent space models in ACM SIGKDD Explorations Newsletter Volume 7 Issue 2, pp. 31-40, 2005.

[33] M. Giatsoglou and A. Vakali, Social networking trends and dynamics detection via a cloud-based framework design in Mining Social Network Dynamics Workshop (MSDN 2012), WWW 2012 Conference, 2012.

[34] T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measures in IEEE/WIC/ACM international conference on web intelligence. IEEE, pp. 85-88, 2007.

[35] N. F. Noy Semantic integration: a survey of ontology-based approach in ACM Sigmod Record 33(4), pp. 65–70, 2004.

[36] T. Plumbaum, S. Wu, E.W.De Luca, S. Albayrak, "User modeling for the social semantic web," SPIM, pp. 78–89, 2011.

[37] Mika, Peter. "Ontologies are us: A unified model of social networks and semantics." J. Web Sem, pp. 5-15, 2007.

[38] L. Ding, T. Finin, and A. Joshi, "Analyzing Social Networks on the Semantic Web," IEEE Intelligent Systems, 2005.

[39] J. Golbeck and M. Rothstein, Linking Social Networks on the Web with FOAF: A Semantic Web Case Study in Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, USA, 2008.

[40] D. Kavitha, SURVEY OF DATA MINING TECHNIQUES FOR SOCIAL NETWORKING WEBSITES in International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April 2017.

[41] (2019) Google Trends [Online]. Available: https://en.wikipedia.org/wi-ki/Google_Trends

[42] S. Volkova, Data Stream Mining: A Review of Learning Methods and Frameworks, 2012.

[43] (2017) IceRocket [Online]. Available: https://en.wikipedia.org/wiki/-IceRocket

[44] E. Augustine, E., C. Cushing, A. Dekhtyar, K. McEntee, K. Paterson and M. Tognetti, Outage detection via real-time social stream analysis: leveraging the power of online complaints in Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion, pp. 13–22, New York, NY, USA, 2012.

* * *

Spreadsheet Implementation for Optimization of Power Flow Network Problem: A Case Study of Kathmandu Valley

Sujan Acharya[1], Anand Tewari[2]
Department of Electrical and Electronics, Nepal Engineering College
Changunaryan Bhaktapur, Nepal
[1]sujanbgl123@gmail.com, [2]mail2anandtewari@gmail.com

Sujan Acharya *has been pursuing Master's degree in Energy System Planning & Management and completed his bachelor's degree in Electrical Engineering from IOE, Pulchowk Campus. He is working as a Junior professor at Nepal Engineering College. His research interest includes demand side power management and fuel cell engineering.*

Anand Tewari *has received his M.Tech degree in Power System Engineering from Lovely Professional University, India and B.Tech degree in Electrical Engineering from Uttar Pradesh Technical University, India. He is currently working as an Assistant Professor in Nepal Engineering College. His research interest includes FACTS technology and Artificial Intelligence.*

Abstract

This paper is about the optimization of power loss in the power transmission network of the Kathmandu valley using spreadsheet modeling as an optimal power flow software. It is helpful for identifying the best path for the optimum power flow in dry as well as wet season. The available generating stations and substations taken as supply and demand nodes near Kathmandu valley are considered and the network is optimized as a transshipment flow problem. The main goal in this network flow model is to determine how much power should be allowed to flow across each transmission lines with minimal losses. In addition, optimization is performed considering the power import from India as well. Considering active power flow with a different scenario, and depending upon the load demand, the power system network is modeled in order to find out the best possible routes for efficient power flow with minimum loss.

Keywords : **Optimization, electricity, optimal power flow, generators, transmission, network, spreadsheet modeling, energy imports.**

I. Introduction

Nepal has a huge hydro-electrical potential expected to be at around 83000 MW[1]. Currently, the power generation is about 1100 MW[2]. The Kathmandu valley needs 315 MW peak power during the dry season and 275 MW in wet season[3]. Kathmandu valley is currently served by a total of 14 grid substation and 13 generators have been utilized for the energy supply system. The accurate and reliable power flow has to be maintained for the stable and secure power system. The optimum power flow problems have a long history in its development. A generalized formulation of the economic dispatch problem including voltage and other operating constraints was introduced and later named the optimal power flow problem (OPF). The load flow problem in an electric power system is concerned with

solving a set of static nonlinear equations describing the electric network performances [4].

Spreadsheet Modeling is one of the strong tools for optimization incorporated with a linear programming module [5]. Excel possesses a plug-in named Solver which is a powerful tool for constrained optimization and solving of non-linear equations. The solver can be used in power network calculations for both practical and educational purposes[6]it is often necessary to solve a large system of non-linear transcendent equations. Therefore, the program rou-tines are complex to develop, while specialized commercial soft-ware packages are expensive and not easily available. On the other hand, Microsoft Excel as a part of Microsoft Office software package is widely-spread in the world and rela-tively easily available. Excel possesses a plug-in named Solver which is a powerful tool for constrained optimization and solving of non-linear equations. Solver can be used in power network calculations for both practical and educational purposes. Solver is also applicable to the other problems concerning electrical theory and practice (circuit theory, electromagnetics, etc.. Although there are numerous OPF algorithms to find the optimal solution, there is one common feature that permeates all these OPF tools, incorporation of a model of the power network.

Power flow studies are one of the most important aspects of power system planning and operation in terms of security and reliability[7]. The load flow gives us the sinusoidal steady state of the entire system-voltages, real and reactive power generated and absorbed and line losses. Power flow analysis is the backbone of power system analysis and design. They are necessary for planning, operation, economic scheduling and exchange of power between utilities.

Panda power is one of the tools based on Python aimed for optimization of balanced power systems[8]. The unified power flow controller (UPFC) is a FACTS device for the control of active and reactive power flow and can also be used for improving the quality of supply[9].

II. Motivation

Power flow analysis holds a key for planning, operation, economic scheduling and exchange of power between the utilities. Several optimization tools are available for the load flow. Some of the tools require high memory and take more time for solving each iteration as like Newton Raphson and Gauss-Seidel method. But, the spreadsheet simulation requires less computation time and less memory allocation. The power system needs dynamic control systems for reliable and secure operation. Therefore, this modeling can be one of the best optimization tools available for load flow.

III. Literature Review

Joram Schito it – Information Technology 2017; 59: 31–39 [10]: In planning transmission lines with the use of Geographic Information Systems, the use of the Least Cost Path (LCP) algorithm has been established while relevant criteria are modeled using Multi-Criteria Decision Analysis (MCDA). First results are promising since a stable algorithm has been developed that computes a cost surface, a Least Cost Corridor (LCC), an LCP, and the transmission towers' positions by simple additive weighting based on user's weights. The findings are integrated into a 3D Decision Support System which aims at facilitating the work of TL planners by realistic modeling and by reducing the approval process for new TL, that optimize a given quantity, possibly subjected to some restrictions on the allowed parameter ranges.

Zhigang Li, Member, IEEE, Jinyu Yu, and Q. H. Wu, Fellow, IEEE(2017)[11]: It proposes an approximate LPF method that considers active and reactive power and transmission losses simultaneously. A general LPF model considering tap changers and phase shifters are derived using the LTVM. An approximation for branch power loss is developed, and approximate complex branch flows considering power loss are formulated. The test results demonstrate that the proposed cold-start LPF-NIC and warm start LPF-IC models approximate the PF solution more accurately than several state-of-the-art linear models in terms of the voltage magnitudes, branch load flows, and transmission power losses without additional computational burden.

Junchao Ma, Student Member, IEEE, Liqiang Yuan, Member, IEEE, Zhengming Zhao, Senior Member, and Fanbo He, Member, IEEE(2016)[12]: It is an OPF strategy for DC micro-grids based on hierarchical control is proposed to minimize the transmission loss in a generic DC micro-grid. Both converter loss and Ohmic line loss are taken into consideration. Simulation results show that the proposed method can realize both DC bus voltage regulation and OPF condition at the

same time in both steady and dynamic situations.

B. Hayes, I. Hernando-Gil, A. Collin, G. Harrison vol. 29, no. 4, 2014[13]: It applies the OPF technique to evaluate which demand-responsive loads in a given distribution network provide the maximum "value" to the system, in terms of their ability to relieve upstream network constraints and provide ancillary services, such as operating reserve. Detailed, aggregate load models of each of the main load sectors are applied, and a time-sequential OPF is used to determine the effective MVA reserve which could be provided at the GSP interface by each load type. The OPF analysis is carried out using a standard industry tool (PSS/E), and therefore can be readily be applied by network operators to determine which DSM loads, or combination of DSM loads, can provide the maximum benefit to the system in a range of operational scenarios.

IV. Methodology

A number of practical decision problems in business fall into a category known as network flow problems. These problems share a common characteristic—they can be described or displayed in a graphical form known as a network. This modelling focuses the network flow problems as a transshipment problem. The following procedure has been followed in order to find out the best path for the optimal power flow considering supply and demand nodes.

A. Data Acquisition

The basic data about generated power, power flow, the load demand and supply of various load centers, area of conductor, conductor size chart, temperature value, resistance, length of line etc. for setting up the objective functions and constraints, were acquired from the system operation department and grid operation department of Nepal Electricity Authority (NEA)[2].

B. Modeling

The transmission system contains a total of 14 substations that have been connected to the various supply nodes and connected within themselves. The voltage level considered for all of the connections is 132 kV and 66 kV. The different incomers for each substation have been considered along with the maximum current at each incomer. Also, the demand during dry and wet season has been calculated. The average demand has been taken for the peak load period in the evening. A simple set of rules, known as the Balance-of-Flow Rules, applies to formulate the constraints for minimum cost network flow problems. For minimum cost network, apply the balance of flow rule as follows:

Flow Problems Where:	At Each Node:
Total Supply > Total Demand	Inflow −Outflow ≥ Supply or Demand
Total Supply < Total Demand	Inflow −Outflow ≤ Supply or Demand
Total Supply = Total Demand	Inflow −Outflow = Supply or Demand

Table 1: Balance of flow rule

The network diagram shows that there are altogether 13 generating stations and 14 grid substations of Kathmandu valley, which are interconnected through different transmission networks.

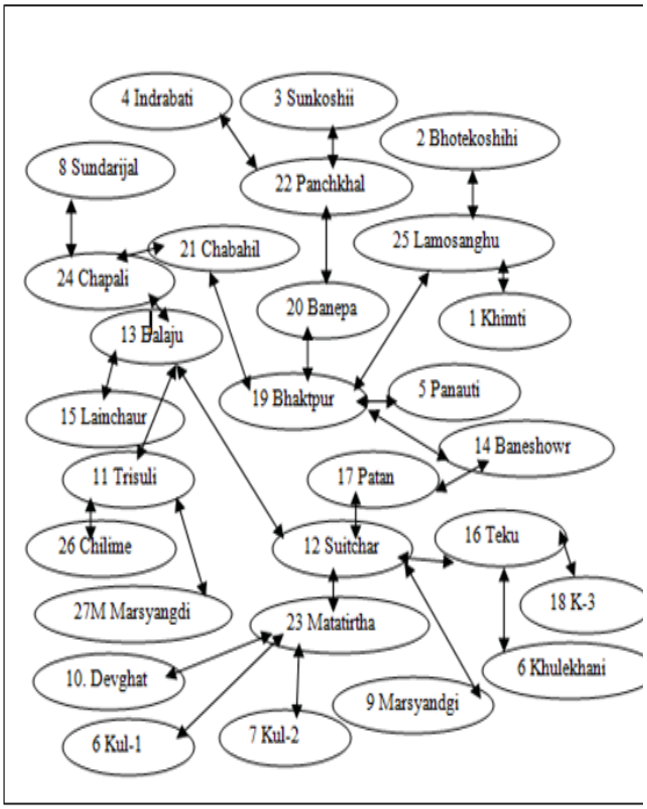


Fig1: Network diagram Demad and supply nodes

1. Conversion to Transshipment Problem

Generating stations (13-hydropower) are the supply nodes or sending nodes, and 14 substations are the demand nodes as well as transshipment

nodes or receiving nodes. Substations nodes have the capability to consume power as well as transfer power to other substations. The positive number in the nodes signifies demand and negative number signifies supply nodes. The objective function is to minimize the line loss which is analogous to transportation cost. In such a way, it is possible to minimize line loss optimally.

2. Setting up Decision Variables

The goal in a network flow model is to determine how much power flows across each of the arcs or networks. For each arc in a network flow model, we define one decision variable as:

X_{ij} = amount of power flowing from node i to node j of power that in a network flow problem usually incurs some loss of power.

3. Setting up an Objective Function

The objective is to minimize the total loss that might be in the form of cost, distance or penalty factor. Loss minimization in each branch or lines is proportional to power flow and the line resistance. This is also called a minimum cost of the network flow problem.

X_{ij} = amount of power flowing from node i to node j (MW)

$$Minimize: Power\ loss(L_{ij}) = \sum [\frac{X_{ij} * 1000}{1.732 * V_{ij} * p.f}]^2 * R_{ij}$$

L_{ij} = loss of power flowing from node i to node j

V_{ij} = voltage level of node i to node j (KV)

R_{ij} = Resistance of node i to node j

p.f. = power factor

4. Setting up Constraints

The network considered here is a dynamic network which is dependent upon numerous constraints such as thermal limit, resistance value, power flow, current; voltage, minimum demand availability of each load center etc., which are applied in the network. The number of nodes determines the number of constraints. Here, one constraint is applied for each node. A simple set of rules, known as the Balance-of-Flow Rules (Table2) is applied to formulate the constraints for the following minimum cost network flow.

Net flow $\leq$ Supply/Demand

Substation Demand $\leq$ Load Constraints

Total Demand with Load = Total Supply

Current across nodes $\leq$ Current limit of nodes

5. Extracting the solutions

The distance between the nodes and the corresponding transmission line resistances are assigned with specifying demand and supply capacities. The network flow problem is formulated according to the data available, constraints and objective function and solved in a spreadsheet. The designed flow problem is solved in order to find the optimal path in the following different scenarios.

i) Wet season (Nepal Power Only)

ii) Dry season (Nepal Power Only)

iii) Wet season (With India and Power)

iv) Dry season (With India Power)

V. Spreadsheets Results

A. Wet Season (Nepal Power Only)

The peak load of Kathmandu is at 275.20 MW during the wet season[2].

The following table summarizes the result.

Total Line Loss In Mw	6.6588
Total Supply	281.86
Loss %	2.36%
Total Demand	275.20
Demand Met	275.20
Total Demand with Loss	281.86
% of Demand Met	100.00%
Power Delivered by Kulekhani (MW)	38.16

Table 2: Result after solving the model at the wet season

B. Dry Season (Nepal Power Only)

In the dry season, the peak load of Kathmandu is at 315.60 MW. The total supply during the dry season is about 214.84MW[2].

The following table summarizes the result.

Total Line Loss In Mw	3.3907
Total Supply	214.84
Loss %	1.58%
Total Demand	315.60
Demand Met	211.45
Total Demand With Loss	214.84
% of Demand Met	67.00%
Evening Load Shedding (MW)	104.15

***Table 3**: Result obtained from spreadsheet modeling at dry season*

C. Wet Season (With power import from India)

Due to the acute power shortage and a huge gap between supply and demand, the country severely suffered from load shedding. To solve the problem of load shedding, Nepal has been importing power from India. Nepal has been importing up to 450 MW from India from various transmission line networks.

The following table summarizes the result.

Total Line Loss (MW)	6.1518
Total Supply	281.35
Loss %	2.19%
Total Demand	275.20
Demand Met	275.20
Total Demand with Loss	281.35
Total Import From India	12.35
% of Load Shared from India	4.49%
Power Delivered by Kulekhani (MW)	25.30

***Table 4:** Result obtained after solving the model at the wet season*

D. Dry Season (With power import from India)

The power shortage during the dry season is severe. The power import from India during dry season holds the key to meet the demand. Kathmandu has been free from load shedding for over a year, partly with the help of imported power from India. The runoff river power plants are operated in full capacity whereas the power imported from India and the power generated from Kulekhani have been taken to minimize the total loss.

The following table summarizes the result.

Total Line Loss (MW)	5.7030
Total Supply	321.30
Loss %	1.77%
Total Demand	315.60
Demand Met	315.60
Total Demand with Loss	321.30
Total Import From India	114.61
% of Load Shared from India	36.31%
Power Delivered by Kulekhani (MW)	63.85

***Table 5:** results obtained after solving the model using the spreadsheet*

During the wet season, the total demand for the Kathmandu at peak load condition cannot be met with the power generation from Run Off River power plants only. The only storage plant of Nepal, Kulekhani (38.16 MW) has to be used in the wet season to meet the peak demand. During the dry season, the total demand of Kathmandu at peak load condition cannot be met with the power generation from Nepal along with the use of Kulekhani in full capacity. The evening load shedding at the peak time would be 104.15 MW. The power imported from India is also used to meet the unmet demand during the wet season. For optimum performance and to minimize the loss, 25.3 MW of power has to be generated by Kulekhani whereas 12.35 MW of imported power is to be used. The power imported from India is used to meet the unmet demand during the dry season. For optimum performance to minimize the loss, 63.85 MW of power has to be generated by Kulekhani whereas 114.61 MW of imported power is to be used. The contribution of power import from India equals to 36.31% of total demand.

VII. Conclusion and Discussion

This research has been done for the optimization of power flow with minimum losses that corresponds to minimum cost considering the various transmission line routes in Kathmandu valley. Power flow can be optimized for the dry as well as a wet season with power generation from Nepal and import from India. The spreadsheet modelling is taken as a transshipment flow network problem in order to identify the optimal path for optimal Power flow. During the dry season, the optimum new path is the flow as shown in network

diagram bypassed power flow between Syuchatar to Matatirtha, Baneshwor to Patan, Baneshwor to Bhaktapur, Chabahil to Bhaktapur, Chabahil to Chpali. In this case, there is a minimum loss of 1.58%(about 3.39 MW) with a supply of 214.84MW. During the wet season minimum loss is 2.36% (about 6.5 MW) off supply 281.86MW. In the optimal operation, the following path is bypassed: Bhotekoshi to Lamosangu, Panauti to Bhaktapur, Syuchatar to Matatirtha, Balaju to syuchatar, Banepa to Pachkhal, and Chapali to Chabahil. The power imported from India is connected to Khimti and hetauda substation. Considering power from India to be 12.35MW during wet season, the minimal loss is about 2.19%. During the dry season, 114.6MW power is imported from India accounts for 1.77% power loss. This analysis holds a key for planning, operation, economic scheduling and exchange of power between utilities. The selection of proper conductor and route helps to minimize power loss which could play a vital role in our country. The sole authority responsible for the design and operation of the transmission line, Nepal Electricity Authority (NEA), can use such type of optimization technique for the fast and reliable power system operation and control.

Acknowledge

The author greatly appreciates to Prof. Dr. Amrit Man Nakarmi and friends Manisha shahi, Sumesh Raut and Ramesh Adhikari for the kind co-operation during paper completion.

[1] H. M. Shrestha, "Cadastre of Hydropower Resources," *Hydro Nepal*, no. 15, pp. 87–89, 1991.

[2] Nepal Electricity Authority, "Annual Report Fiscal year 2017/18," pp. 1–196, 2017.

[3] W. and E. C. Secretariat, "Government of Nepal Water and Energy Commission Secretariat Electricity Demand Forecast Report," no. January, 2017.

[4] J. J. Grainger and J. WUliam D. Stevenson, *POWER SYSTEM ANALYSIS*. 1994.

[5] C. T. Ragsdale, *Spreadsheet Modeling and Decision Analysis, Fifth Edition*. Virginia Polytech Institute and State University.

[6] Č. Zeljković and M. Gaćanović, "an Example of Using Microsoft Excel Solver for Power Network Calculations," pp. 2–6.

[7] R. Billinton and R. N. Allan, *Reliability Evaluation of Power Systems*, vol. 2nd Editio. 1984.

[8] L. Thurner *et al.*, "Pandapower - An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6510–6521, 2018.

[9] A. Alidemaj, V. Komoni, G. Kabashi, and I. Krasniqi, "Control Active and Reactive Power Flow with UPFC connected in Transmission Line," *8th Mediterr. Conf. Power Gener. Transm. Distrib. Energy Convers. (MEDPOWER 2012)*, pp. 31–31, 2012.

[10] J. Schito, *Modeling and optimizing transmission lines with GIS and Multi-Criteria Decision Analysis*, vol. 59. 2017.

[11] Z. Li, J. Yu, and Q. H. Wu, "Approximate Linear Power Flow Using Logarithmic Transform of Voltage Magnitudes with Reactive Power and Power Loss Consideration," *IEEE Trans. Power Syst.*, vol. 8950, no. c, pp. 1–1, 2017.

[12] J. Ma, L. Yuan, Z. Zhao, and F. He, "Transmission Loss Optimization-Based Optimal Power Flow Strategy by Hierarchical Control for DC Microgrids," *IEEE Trans. Power Electron.*, vol. 32, no. 3, pp. 1952–1963, 2017.

[13] B. Hayes, I. Hernando-Gil, A. Collin, G. Harrison, and S. Djokić, "Optimal power flow for maximizing network benefits from demand-side management," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1739–1747, 2014.

* * *

Enabling and Limiting factors in eXtreme Programming (XP) with Evaluation Framework

Sundar Kunwar
sundark@nec.edu.np
Department of Computer Science and Engineering, Nepal Engineering College, Pokhara University
Changunarayan, Bhaktapur, Nepal

***Sundar Kunwar** is working as an Assistant Professor in the Department of Computer Science and Engineering at Nepal Engineering College, where he has been since 2007. He has received BE in computer engineering from Tribhuvan University in 2006 and Master in Software Development from University of Tampere, Finland in 2013. His research interest is in improving the agile software development methodologies through agile modeling. In addition he is also interested and has made numerous contributions to robotics projects. He had successfully coordinated robotics competition organized by IOE held on 2067 on theme "Working together for better Nepal" and had stood first position among all the engineering colleges of Nepal.*

Abstract

As agile software development methodologies are used in many domains and come with different shapes and sizes, it is one of the complex human endeavours. Extreme Programming (XP) is one of the well-known agile software development methodologies and is driven by a set of values including simplicity, communication, feedback and courage, but lacks the mechanism to measure these values demanding the evaluation framework to make it measurable and attainable. The main aim of this study is to build the software process improvement model that can be used for evaluating XP values and practices. The proposed XP evaluation framework in this study is XP focused and evaluates the XP project, product and practices. The XP evaluation framework is a collection of some new and validated metrics used for evaluating XP projects, XP practices, XP products and some additional factors concerned with XP. The evaluation framework for extreme programming is basically based on the assessment and evaluation of various project characteristics, extreme programming characteristics, product characteristics and other additional characteristics. The metrics used for assessments and evaluations of XP are designed to be simple, precise, understandable, economical, timely, consistent, accountable, unambiguous, suitable and reliable.

*[**Keywords** : Agile, eXtreme Programming (XP), evaluation framework, metrics, lightweight requirement, onsite customer, pair programming]*

I. INTRODUCTION

One of the major challenges of agile software development methodologies is to develop a mechanism to measure the various aspects of software development process [1]. Therefore, there is always a need of such measurement mechanism that could quantify the various aspects of software development methodology and final product of the development process. To evaluate the XP, a framework that contains various metrics to capture information about development team, development process, development tools and the final product is proposed in this study. This is useful to those organizations which have adapted or willing to adapt XP methodology. Measurement is important in software projects because it keeps us involved in it, informs about the current status

and provides the guidelines to process further. There are many evaluation frameworks available to evaluate different practices of XP. Usually measurement encompasses of qualitative evaluation and measures in term of numerical values to show the assessment results [2]. A quantitative evaluation framework was proposed for agile methodologies and was based on the four postulates of Agile Manifesto [1]. The quantitative evaluation framework based on four postulates of Agile Manifesto cannot evaluate the practices of methods on which it is used. It can only tell about the agility of the agile methods evaluated. The evaluation framework initiated by William [4] is more general agile evaluation framework with no XP focused features. The proposed XP evaluation framework in this study is XP focused and evaluates the XP project, product and practices.

According to Fenton and Pfleeger [4], "measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules". An entity can be anything like time, event, commodity, thing, place or person. Measurement is extensively used in most of the production and manufacturing area to estimate costs, calibrate equipment, assess quality and monitor inventories [5]. Science and engineering disciplines are incomplete without measurement tools and techniques. Why measurements are used? The most general four reasons for measurements are: to characterize, evaluate, predict and improve the existing or proposed system. As shown in Figure 1, attributes of the entity are taken into consideration for the propose of measurement and are assigned with numbers or symbols.

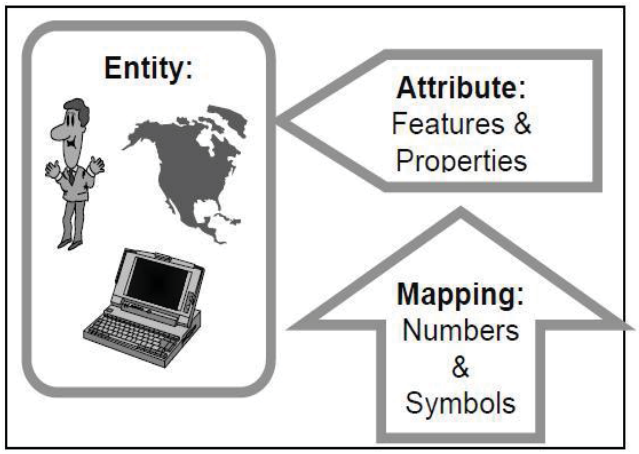


Figure 1: Measurement of entity [5].

This measurement does not give any meaning unless we express with the mapping system like height is 5.9 feet and weight is 65 kg. Software metrics are the integral part of the state of the practice of software engineering. Many customers specify software and quality metrics as a part of their contractual requirements. As all the attributes of software are difficult to measure, software measurements do not seem to have fully penetrated into industry practices A metrics is a quantifiable measurement of software products, process, or project that is directly observed, calculated, or predicted. As shown in Figure 2, software metrics are the measurement based techniques applied to software process, products and services to supply or to improve the engineering and management information.

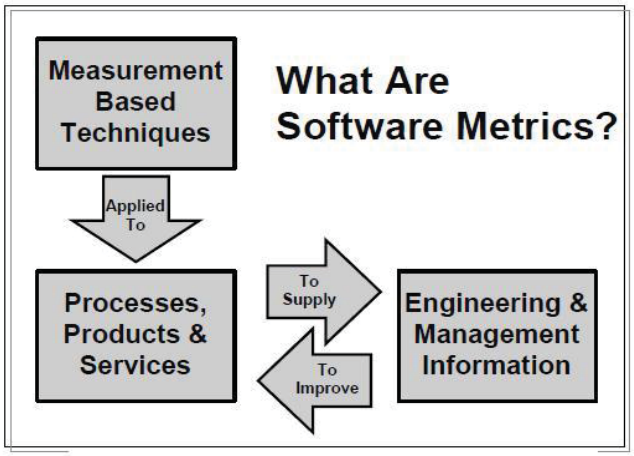


Figure 2: Software Metrics [5]

They are useful in predicting outcomes as well as decisions when required. Metrics need to be defined clearly before using it. Following are the elements that should be clearly defined before using metrics. [6]

- **Metrics Name**: Appropriate name that has something to do with its functionalities should be given.
- **Metrics Description**: Description of what is being measured.
- **Measurement Process**: How metrics is used for measurement?
- **Measurement Frequency**: How often measurement is used?
- **Threshold Estimation**: How are thresholds calculated?
- **Current Thresholds**: Current range of values considered normal for metrics.

Several studies have shown that there are enabling as well as limiting factors in extreme practices of XP. A detail study about the rules and practices of XP was carried out through interpretive approach and some enabling and limiting factors were discovered and the most criticized factors such as lightweight requirements, onsite customer and Pair Programming are taken into account to make XP practices more realistic and practical.

II. RESEARCH METHODOLOGY

The work is more concerned with the development of evaluation framework with enabling and limiting factors in XP. Most three criticized extreme practices of XP- user stores, pair programming and online customers are mainly taken into account as an initial research framework for discovering enabling and limiting factors and evaluating various aspects of it. An interpretive approach was followed to conduct a literature review. A research can be interpretive if it builds on the assumptions that humans learn about the reality from the meaning they assign to social phenomena such as language, consciousness, shared experiences, publications, tools, and other artefacts [7]. The most fundamental principle of the interactive research approach is a hermeneutic cycle derived from documents and literary analysis. The different components of the hermeneutic cycle are illustrated in Figure 3. The first component of the hermeneutic cycle is concerned with the pre-understanding of researchers on the subject matter and the second component is concerned with the absorption of more knowledge from different sources to widen knowledge to expand the researcher's interpretation potential. The third component is concerned with theory building on the basis of an interpretation of knowledge, explanation attempts and missing knowledge. The last component is concerned with documenting the new theories and knowledge acquired through interpretive research approach.

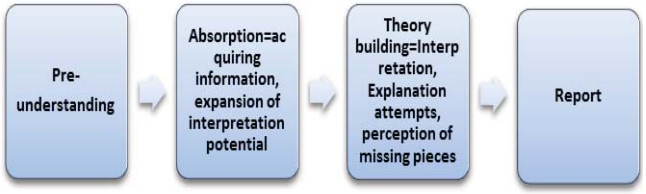


Figure 3. Hermeneutic Cycle [3]

III. PROPOSED EVALUATION FRAMEWORK FOR XP

The measurements in physical systems are rigidly defined and do not require more effort to quantify them. However, the measurements in software engineering are not so rigidly defined as in physical systems and take a lot of effort to quantify them. Software engineers make very difficult and critical decisions based on the result of such measurements. The evaluation framework for extreme programming is basically based on the assessment and evaluation of various project characteristics, extreme programming characteristics, product characteristics and other additional characteristics. The metrics used for assessments and evaluations of XP are designed to be simple, precise, understandable, economical, timely, consistent, accountable, unambiguous, suitable and reliable. The proposed extreme programming evaluation framework consists of four sections with numbers of subsections. The general block diagram of the proposed XP evaluation framework is shown in Figure 4:

XP Project Records	XP Practice metrics	XP Product metrics	Additional XP metrics
Project Detail Member Detail Client Detail	Various validated and proposed matrices for XP practices	Product detail Product Quality Product productivity	Additional metrics

Figure 4: Proposed XP evaluation framework

Proposed XP evaluation framework design is more specific to extreme programming. It is a collection of some validated and proposed metrics. As illustrated in the figure, proposed XP evaluation framework consists of four sections with some subsections. Subsections of each section are more concerned with both validated and proposed metrics. The first section is Project evaluation which is used for recording and measuring the project and project members' details. The second section is XP practice metrics which contains validated as well as proposed metrics for assessment and evaluation of XP practices used for software development process. The third section is XP product metrics which contains validated as well as proposed metrics for final product assessment and evaluation. The fourth section is Additional XP metrics which contains some validated

as well as some proposed metrics for assessment and evaluation of additional information on XP that are not covered in other sessions of proposed XP evaluation framework.

A. Project Records

Project records are designed in order to evaluate the project and member details. Personnel and team makeup are documented as top risk factors in software development.

B. XP Practices Metrics

XP has its roots spread in information technology system development where it makes the development process more responsive to changing business requirements [8]. The fourteen principles of XP are: Humanity, Economics, Mutual Benefit, SelfSimilarity, Improvement, Diversity, Reflection, Flow, Opportunity, Redundancy, Failure, Quality, Baby Steps, and Accepted Responsibility [9]. However, there are no any measuring means to assess all these practices and principles. Therefore, the proposed XP practice metrics play a vital role to assess the effectiveness of these practices and they are discussed below:

1. Sit Together Attendee

Sit together is one of the simplest but most difficult XP practices. XP advocates the entire team members must be present but it is not always possible. Therefore, sit together attendee records the name and of the absentee team member in the meeting.

2. Number of Requirements (User Stories)

The size of the project mainly depends upon the number of user stories which serve as a lightweight requirement to software development process. Simply, it counts the number of user stories in the project.

3. Requirement Complexity

Requirement complexity qualifies how complex is each user story to implement. It can be qualified as low, medium and high.

4. XP Stakeholders

It is used for recording all the concerned stakeholders and their roles in the XP project.

5. Project Velocity

Project velocity is the measure of the time taken (in days) and the number of stories completed in a single iteration. It measures the length of the iteration in days and the tasks completed.

6. Automated Unit Tests per User Story

It quantifies the total number of automated unit tests carried out per user story. The main objective of this metrics is to know how many unit tests are created for each user story before they are implemented.

7. Frequency of Automated Unit Test

It shows how often the automated unit tests are carried out. It can be calculated as FAUT= (total number of unit tests/total number of classes) per user story*100%.

8. Acceptance Tests

It keeps all the necessary information about acceptance tests.

9. Number of iterations per user story

Implementation of a user story may or may not be fully implemented in iteration. Therefore, it measures the numbers of iterations taken by user story to get fully implemented

10. Onsite Customer Availability

Onsite is very simple but difficult practice of XP. It is the measure of how often the customer is available on onsite of development. It can be qualified as Full time, Part time and Never.

11. Pairing Frequency

In Pair Programming, one programmer is driver who writes code while the other is observer or navigator who reviews the code as it is typed in. The two programmers switch roles frequently. Pairing frequency measures how often the role of driver and navigator changes in Pair Programming.

C. XP Product Metrics

XP product metrics are concerned with measuring the product related measurements.

1. Number of Component, Methods and Lines of Codes

Number of components, methods and lines of codes determine the size of the project.

2. Productivity Metrics

Halstead proposed the coding productivity metrics and the idea was to determine the productivity from the numbers and types of words used in the program. It is also referred as a token count measure. It can be calculated using the following formula. [9]

Volume = length*log2 (vocabulary)

Where length = N1 + N2

Vocabulary = n1 + n2

n1 = the number of unique operators

n2 = the number of unique operands

N1 = the total number of operators

N2 = the total number of operands

3. Difficulty and Effort Metrics

IBM researchers developed difficult metrics which measure the effort required to understand code and maintain a piece of software. It is calculated as follows. [10]

Difficulty = n1/2*N2/n2

Effort=difficulty*volume

Where,

n1 = the number of unique operators

n2 = the number of unique operands

N2 = the total number of operands

Volume = length*log2 (vocabulary)

4. Defect Removal Effectiveness

Defect Removal Effectiveness (DRE) is defined as the ratio of defects removed during the development phase to defects latent in the product and it is usually expressed in percentage [11].

5. Constraint

Constraints are the limitations or restrictions present in the project. It lists all the known present in the system.

D. XP Additional Metrics

There are many metrics that can be put under additional metrics which can be used for evaluating and measuring various aspects of XP. Some of them are discussed below:

1. Customer Problem Metrics

The customer problem metrics is generally expressed in terms of problems per user month (PUM).

PUM = Total problems that customers reported (true defects and non-defect-oriented problems) for a time period /Total number of licenses-months of the software during the period.

2. Customer Satisfaction Metrics

Customer satisfaction is measured in term of results obtained from customer surveys. The result is analysed in term of following five levels: Very satisfied, Satisfied, Neutral, and Dissatisfied and Very dissatisfied.

3. Estimation of Number of Defects

It was first proposed by Jones [12] for the estimation of the number of defects based on the numbers of functional points of the system. It is calculated as:

Potential Number of Defects=$FP^{1.25}$

Where FP is the functional points of the system

4. Halstead Metrics for Effort

It was Halstead [9] who proposed an effort metrics to determine the effort spent. It is calculated as:

E=V/L

Where,

E = effort

L=NLog2n

V=Program Volume

N=Program Length

n=Program Vocabulary

IV. Enabling and limiting factors in xp

Several studies have shown that there are enabling as well as limiting factors in extreme practices of XP. A detail study about the rules and practices of XP was carried out through interpretive approach and some enabling and limiting factors were discovered and the most criticized factors such as lightweight requirements, onsite customer and Pair Programming are taken into account to make XP practices more realistic and practical. The lightweight requirement is one of the most criticized extreme practices of XP. This study proposes the scenario based requirements engineering practices for XP with stakeholder analysis to overcome the defects in the requirement practices of XP. It is known fact that the unclear and deficient requirements create more problem than they solve. As very lightweight requirement engineering practices are followed in drafting requirement in XP, there is always danger of drafting unclear and defective requirements. The unclear and defective requirements result the propagation of error throughout the software development cycle. This may result final product with undiscovered errors which is one of the risk factors for customers and software developers. The most common enabling and limiting factor of the requirement process in XP is listed below:

Enabling factors of requirement in XP

- Lightweight process.
- Divide and conquer approach.
- Less effort and time.
- Emphasis on oral communication over written documentation.

Limiting factors of requirement in XP

- It is very difficult to find the real representative of customer business.
- Single person (onsite customer) is responsible for making decisions about the business.
- High chances of unclear and defective requirement collected from a single person.
- Bypassing the requirements engineering practices.

The limiting factors seem to affect more than an enabling factor of the requirement process in XP. Therefore, to eliminate all the limiting factors, new approach for collecting requirements in XP is proposed in this study and the approach is called scenario based requirement engineering process where all the related use cases are collected from the real world working environment. The realistic scenarios are generalized for requirement analysis to get the requirements from it. There are some scenario based tools that make the process more organized and simple. As automated tools are present to facilitate the scenario based requirements, it can be successfully implemented into XP without making it heavyweight methodology. For example CREW SAVRE version 2.1 built on Window NT platform supports scenario based requirement engineering such as incremental specification of use cases and high level requirements, automatic scenario generation from use cases, description of use cases and scenario of historical data, user walk-through and validation support among others [13]. With the scenario based approach stakeholder identification and analysis becomes easier and simpler. In most of the cases, it is possible to identify and analyse the stakeholders and their roles from real world scenarios. This makes the requirements stronger and realistic. Stakeholder analysis is performed to understand the system with stakeholders staked to it, their relationships, interests and expectation. It helps to avoid the expectation gap between developers and customers with different interests. As the requirement is obtained through intensive communication process in XP, it will definitely help to improve the requirement process in XP. And then the detail user story is drafted in electronic form that is made available through web pages which will act as written requirement specification in future.

Onsite customer practice is also one of the most criticized extreme practices of XP. Onsite customer is responsible for drafting a user story, sitting together with the whole team. User story acts as requirement specification in XP. He/she is also responsible for user story prioritization that defines the priority of user story to be implemented

and development of acceptance tests with developers. It is also believed that onsite customer is courageous enough to make a business decision.

Many studies show that onsite customer practice is effective but unrealistic and impractical. The most common enabling and limiting factors of onsite customer are listed below:

Enabling factors of onsite customer

- Team oriented practices.
- Provides business values.
- Timely decision.
- Bearing responsibilities for failure or success of project.

Limiting factors of onsite customer

- Full time availability.
- Inadequate domain knowledge.
- Decision making authority on single people.

There were not so many studies performed relating onsite customer extreme practices of XP. Out of several alternative solutions to onsite customer, two conceptual models were taken into consideration. First is multiple customer representative models where single customer is replaced by a multiple concerned customers who can provide all the necessary information that the developer is looking for. Second is segregating customer model where the domain experts act as customer in case real customer are inaccessible. Especially, it can be practiced in outsourcing projects.

Pair Programming (PP) is another the most criticized extreme practice of XP. It has been claimed that PP improves software development process in many ways. However, some studies and researches show that two developers working together cannot be productive, economical and chances of delay if developers have strong disagreements on some issues.Two alternative solutions to Pair Programming: Distributed Pair Programming Model and Collaborative Adversarial Pair (CAP) Programming model are proposed in this study.

Enabling factors of Pair Programming

- Collaborative and supportive effort.
- Feel of code ownership.
- Reluctant to interruption-single person can be easily interrupted than a pair.
- Pairs are less likely to go down Gopher Holes and Blind Alleys.
- Two minds are always better than single.

Limiting factors of Pair Programming

- Differences in programming and communication skills.
- Antisocial or anti personalities.
- Perception of cost and time.
- Common schedule and agreement.
- Discourage in pairing.

The personal traits development training is proposed to inexperienced and resistant programmers to help in cultivation of two personalities making them right pair. It helps to improve communication skills, to make more comfortable, confident and comprising which are suitable personal traits for Pair Programming. Two models for improving Pair Programming were proposed. First is Distributed Pair Programming (DPP) when programmers are located geographically apart and the second is a Collaborative Adversarial Pair (CAP) to take the merits and downplay the demerits of PP. There are some studies that examine the enabling or/ and limiting factors of XP. Some of the analytical studies present the alternative solution to limiting factors of XP to improve the XP software process. Table 1 shows the analysed enabling and limiting factors of User Story of XP. Similarly, Table 2 shows the analysed enabling and limiting factors of Pair Programming and Table 3 shows the analysed enabling and limiting factors of onsite customer.

XP Practices	Enabling Factors	Limiting Factors	Remedy/Remedies	Ref.
User Story	**Clear vision:** The customer has a clear vision of business processes, product requirements and product background.	**Deficient Requirement:** Customers are not able to give complete requirements to developers. **Flood Requirement:** Customer has high expectations exaggerating the capacity of computer. **Frequent Changes:** Frequent changes in requirement will lead stagnation, modify and even abandon the finish work. **Negative Influence** The contradiction between customers and developers has a negative influence on the demand of high quality.	**i. Kano Model Analysis** for measuring customer feeling and measuring effects of the product or software quality. **ii. High Quality Requirement Analysis** to measure the customer wish and developer need. **iii. XP Demand Module** It is established with Kano Model thinking and High Quality Requirement Analysis to explore the high quality requirements with customer awareness and reduce the misunderstanding in software development process and hidden threats.	**[14]**
User Story	**Not stated**	**Single Customer** The assumption that, in the planning game, the business could be represented by just one customer. **Non-functional requirements** The lack of consideration of non-functional requirements from the standpoint of the business.	i. A **process** and a **representation** are proposed for writing the stories and tasks cards. ii. Also include **non functional** requirements as user stories. iii. The word should be **underlined** to show that it has an **explicit link** with other underlined word.	**[15]**
		Linkage The lack of explicit links between stories and tasks cards to the code **Process** The lack of a process for producing stories and tasks.	iv. The process is described using **SADT** diagram to **verification** and **validation.**	
User story	**Rapid** Rapid response to changing requirements.	**Defects** Less predictable, less stable, less reliable and less quality assurance requirements. **Informal requirements definition** User stories drafted by customer are prioritised, but no formal documentation.	**Mapping extreme practices to ISO Process Model**	**16]**
User story	Unambiguous, Correct, and Understandable Modifiable, Verifiable and Annotated by Relative Importance Complete and Concise Requirements	**Not Stated**	**Not Necessary**	**[17]**

Table 1: *Enabling and Limiting factors of user story found in different studies.*

XP Practice	Enabling Factors	Limiting Factors	Remedy/Remedies	Ref.
Pair Programming	**Counter Balance** The detrimental effects of paired programming are counterbalanced by other XP best practices such as common metaphor, simple design, unit tests, coding standard and the reverse is true.	**Productivity** Two developers working together cannot equal the productivity of the same two developers working in parallel. **Cost** It has been statistically shown that paired programming costs approximately 15% more time than traditional programming **Personal Characteristics** Effective paired programming is difficult to achieve and requires a careful cultivation of personalities within the development team. **Dynamic interchange** The dynamic interchange of roles is one major problem in PP.	**Personalities Traits** It was noticed that certain personality traits are beneficial for paired programming. **Improvement in interview technique** It can be used for ensuring the traits of pair programmers during their interviews.	**[18]**
Pair Programming	**Defects** The end defect content is statistically lower. **Faster** The pair solves the problem fast. **Code Review** Mistakes can be found during coding. **Learning** People learn more about the system and software development. **Communication** It provides an opportunity to improve the communication skills. **Understanding** Project end with many people understanding the software product.	**Cost** The development cost for Pair Programming enabling factors is only 15%. **Wrong Perception** Managers view programmers as a scarce resource, and are reluctant to "waste" such by doubling the number of people needed to develop a piece of code. **Tradition** Programming has traditionally been taught and practiced as a solitary activity. **Reluctant** Many experienced programmers are very reluctant to program with another person.	It is only the study of cost and benefits of Pair Programming. **No remedy** is provided to address its costs.	**[19]**
Pair Programming	**Better code** Its premise—that of two people, one computer—is that two people working together on the same task will likely produce better code than one person working individually **Benefits** Faster software development, higher quality code, reduced overall software development cost, increased productivity, better knowledge transfer, and increased job satisfaction are some benefits of PP.	**Time schedule and agreement** It requires that the two developers be agreed for the same place at the same time. **Management prospective** It requires an enlightened management that believes that letting two people work on the same task will result in better software than if they worked separately. **Cost** The cost of Pair Programming is higher than that of sole programming. **Paring Up** Novice-expert and expert-expert pairs have not been demonstrated to be effective.	**Collaborative Adversarial pair (CAP) programming** The main objective is to take the merits of Pair Programming while at the same time downplay with its demerits**.** The main idea is to design together, construct test and code independently and then test together.	**[20]**

Table 2: *Enabling and Limiting factors of Pair Programming found in different studies*

XP Practice	Enabling Factors	Limiting Factors	Remedy/Remedies	Ref.
Pair Programming	**Counter Balance** The detrimental effects of paired programming are counterbalanced by other XP best practices such as common metaphor, simple design, unit tests, coding standard and the reverse is true.	**Productivity** Two developers working together cannot equal the productivity of the same two developers working in parallel. **Cost** It has been statistically shown that paired programming costs approximately 15% more time than traditional programming **Personal Characteristics** Effective paired programming is difficult to achieve and requires a careful cultivation of personalities within the development team. **Dynamic interchange** The dynamic interchange of roles is one major problem in PP.	**Personalities Traits** It was noticed that certain personality traits are beneficial for paired programming. **Improvement in interview technique** It can be used for ensuring the traits of pair programmers during their interviews.	**[21]**
Pair Programming	**Defects** The end defect content is statistically lower. **Faster** The pair solves the problem fast. **Code Review** Mistakes can be found during coding. **Learning** People learn more about the system and software development. **Communication** It provides an opportunity to improve the communication skills. **Understanding** Project end with many people understanding the software product.	**Cost** The development cost for Pair Programming enabling factors is only 15%. **Wrong Perception** Managers view programmers as a scarce resource, and are reluctant to "waste" such by doubling the number of people needed to develop a piece of code. **Tradition** Programming has traditionally been taught and practiced as a solitary activity. **Reluctant** Many experienced programmers are very reluctant to program with another person.	It is only the study of cost and benefits of Pair Programming. **No remedy** is provided to address its costs.	**[22]**

XP Practice	Enabling Factors	Limiting Factors	Remedy/Remedies	Ref.
Pair Programming	**Better code** Its premise-that of two people, one computer-is that two people working together on the same task will likely produce better code than one person working individually **Benefits** Faster software development, higher quality code, reduced overall software development cost, increased productivity, better knowledge transfer, and increased job satisfaction are some benefits of PP.	**Time schedule and agreement** It requires that the two developers be agreed for the same place at the same time. **Management prospective** It requires an enlightened management that believes that letting two people work on the same task will result in better software than if they worked separately. **Cost** The cost of Pair Programming is higher than that of sole programming. **Paring Up** Novice-expert and expert-expert pairs have not been demonstrated to be effective.	**Collaborative Adversarial pair (CAP) programming** The main objective is to take the merits of Pair Programming while at the same time downplay with its demerits**.** The main idea is to design together, construct test and code independently and then test together.	**[23]**

Table 3: *Enabling and Limiting factors of onsite customer found in different studies.*

During this study, following are the most remarkable enabling and limiting factors noticed and the alternative solutions are proposed to limiting factors to improve the XP software process. It is shown in Table 4.

Extreme Practice	Enabling factors	Limiting factors	Remedy	Remarks
Lightweight Requirements (User story)	**Lightweight process** **Divide and conquer approach** **Less effort and time** Emphasis on **oral communication** over written documentation**.**	High chances **of unclear and defective requirement** collected from a single person. Bypassing the **Requirement Engineering Practices.**	**Requirement Specifications** are collected from **Scenario Based Requirement Engineering (SBRE)** Practices.	**SBRE** is not so heavyweight method. Processes are simple and easy to practice. However, it is not as simple as user story. Further improvements and modifications are necessary to make the process lightweight.
Onsite customer	**Team oriented** practices. Provides **business values** **Timely decision** **Bearing responsibilities** for failure or success of project	**Full time availability.** **Inadequate domain knowledge.** **Decision making authority** on single people	**Multiple Customers Representative Model** **Surrogate Customer Model**	**Multiple customers** having adequate domain knowledge are dealt based on their **priority**. Customers are surrogated by **domain experts** according to need and necessity.

Pair Programming	**Collaborative and supportive** effort Feel of **code ownership** **Reluctant to interruption**-single person can be easily interrupted than a pair Pairs are less likely to go down **Gopher Holes and Blind Alleys**. **Two minds** are always better than single.	Differences in **programming and communication** skills **Antisocial or anti personalities** Wrong **perception of cost and time** **Common schedule and agreement** **Discourage** in pairing	**Personality traits development trainings to pair resistant.** **Distributed Pair Programming (DPP)** **Model.** **Collaborative Adversarial Pair Programming (CAPP) Model**	Training is only provided to those who are found to be **pair resistant**. **DPP** is practices when the developers are geographically apart. **CAPP** is validated model to take the merits and downplay the demerits of Pair Programming.

Table 4: *Remarkable Enabling and Limiting factors observed with alternative solutions.*

V. Conclusion

The study proposes evaluation framework for evaluating XP project with different existing and proposed metrics in order to evaluate it. The evaluation framework consists of enough room to include the desired metrics on specific field of XP project. It is more concerned with the XP project which cannot be applied for other methodologies. Software metrics were chosen or proposed to evaluate the XP practices. However, the agility of agile software development methodologies can be somehow affected by the XP evaluation framework. The proposed XP evaluation framework is a comprehensive tool for agile software development to evaluate XP practices without imposing excessive burden. With the improvement in XP practices and process, the metrics can also be further modified or added. An active continuation of research is needed for refining and validating the XP evaluation framework to make it possible to implement practically in real projects. This can be done through the international collaboration with software industries to refine and validate the study. After the refinement and validation, it can be used as standard XP evaluation framework in real projects. There are many numbers of enabling as well as limiting factors in XP. This study is concerned only with some extreme practices of XP although there are many other extreme practices to be studied. The study concentrates on only three the most criticized practices-lightweight requirement, onsite customer and Pair Programming of XP. In future, further study about other extreme practice can be carried out to refine the practices and make them simple, practicable as well as effective.

REFERENCES

[1] E. Karla, C. Pablo, & F. Estevez, "A Quantitative Framework for the Evaluation of Agile Methodologies", Journal of Computer Science and Technology, vol. 10, no.2, pp. 68–73, 2010.

[2] N. Ahmad, "Software Measurement and Metrics: Role in Effective", International Journal of Engineering Science and Technology (IJEST), vol 3, no. 1, pp. 671–680, 2011.

[3] L. Williams, W. Krebs, L. Layman, A. I. Antón, and P. Abrahamsson, "Toward a Framework for Evaluating Extreme Programming," Proceedings of the 8th International Conference on Evaluation and Assessment in Software Engineering (EASE '04), pp. 11–20, 2004

[4] N. Fenton and J. Bieman, Software metrics. Boca Raton, Fla. [u.a.]: CRC Press, 2015.

[5] L. Westfall, The Certified Software Quality Engineer Handbook, ASQ Quality Press, 2009.

[6] N. Ahmad, "Software Measurement and Metrics: Role in Effective", International Journal of Engineering Science and Technology (IJEST), vol. 3, no. 1, pp. 671–680, 2011.

[7] E David, Research Methods for Political Science: Quantitative and Qualitative Approaches, 2nd ed. Reading, ME Sharpe, 2004, [E-book] Available: http://books.google.com/books?id=8PJYznDXQIcC&pgis=1

[8] G. Meszaros, J. Andrea & S.Smith, "Framework XP – Building Frameworks using XP", The Pennsylvania State University, Pennsylvania, 2002

[9] Maurice H. Halstead, "Elements of Software Science (Operating and Programming Systems Series", Elsevier Science Inc., New York, NY, USA, 1978.

[10] T. Andersson, "A Survey on Software Quality Metrics", Åbo Akademi University, Department of Computer Science,Finland, 1990.

[11] E. Karla, C. Pablo & F. Estevez, "A Quantitative Framework for the Evaluation of Agile Methodologies", Journal of Computer Science and Technology,vol. 10, no. 2, pp. 68–73, 2010

[12] C. Jones, "Software Estimation Rules of Thumb", Proceedings of the 1998 IFPUG conference, vol. 10, no. 3, pp. 1–11, 1998.

[13] N.A. Maiden , S. Minocha, K. Manning & M. Ryan, "CREWS-SAVRE: Systematic Scenario Generation and Use", Proceeding in the 1998 third international conference on Requirements Engineering, vol. 12 no. 5, pp. 148-155, 1998.

[14] Z. Li-li, H. Lian-feng & S. Qin-ying, "Research on Requirement for High-quality Model of Extreme Programming", Proceedings of the 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 4 no. 3, pp. 518–522, 2011.

[15] R. De. Janeiro, "Extreme Requirements (XR)", Proceedings of the 2001 Requirements Engineering Conference Applied Science, pp. 1–13, 2001.

[16] E. Erharuyi, "Combining Extreme Programming with ISO 9000 : 2000 to Improve Nigerian Software Development Processes", Blekinge Institute of Technology, Sweden, 2007.

[17] R. Duncan, "The Quality of Requirements in Extreme Programming", Software Defence Engineering, pp. 19–22, 2001.

[18] A. J. Dick, & B. Zarnett, "Paired Programming & Personality Traits", Proceedings of the 2002 Workshops on Database Theory, 2002.

[19] A. Cockburn, & L. Williams, "The Costs and Benefits of Pair Programming", In Extreme Programming and Flexible Processes in Software Engineering XP, pp. 1–11, 2000.

[20] R. Swamidurai & D. Umphress, "Collaborative-Adversarial Pair Programming", ISRN Software Engineering, pp. 1–11, 2012.

[21] A. J. Dick, & B. Zarnett, "Paired Programming & Personality Traits", Proceedings of the 2002 Workshops on Database Theory, 2002.

[22] A. Cockburn, & L.Williams, "The Costs and Benefits of Pair Programming", In Extreme Programming and Flexible Processes in Software Engineering XP, pp. 1–11, 2002.

[23] R. Swamidurai & D. Umphress, "Collaborative-Adversarial Pair Programming", ISRN Software Engineering, vol. 5, no. 8, pp. 1–11, 2012.

* * *

Nepal Engineering College

Changunarayan-4, Bhaktapur
Phone: 01-5221006, Fax: 01-5221001
www.nec.edu.np