

Using Data Mining to Predict and Control the Prevalence of Chronic Diarrhea

¹Amosa B.M.G., ²Onyeka Ndidi, ²Olatunji Funsho and ²Ugwu Jennifer

^{1*}Allbytes Research Group, Lagos, Nigeria ²Department of Computer Science, Federal Polytechnic Ede, Nigeria

Submitted: 10-01-2021

Revised: 23-01-2021

Accepted: 26-01-2021

ABSTRACT: Diarrhea is one of the most common health ailments. An estimated 2 billion cases of diarrhea disease is reported every year. Developing nations are mostly affected with around 1.9 million children below the age of 5 dying of diarrhea annually, which is the second foremost cause of death in this age group. The study was carried out with data collected from Ladoke Akintola University of Technology Teaching Hospital, Osogbo, Nigeria. A sample data of 300 patients with cases of diarrhea was used. The data mining techniques used are the K-nearest neighbor and C4.5 decision tree components of WEKA. The result shows that the Prediction and Control of Chronic Diarrhea can be effectively achieved with C4.5 decision tree algorithm and K-nearest neighbor (KNN).

KEYWORDS: Data Mining, Diarrhea, Predict, Algorithm

I. INTRODUCTION

Diarrhea is one of the most common health complaints. It ranges from a mild, temporary condition to an indisputable dangerous one. Internationally, an estimated two billion cases of diarrhea disease occur annually. Also, around 1.9 million children below the age of five (5) years, majorly in developing nations lose their lives from diarrhea every year [1]. These illnesses are particularly dangerous for young children, who are more liable to dehydration and diet losses during an outbreak of acute diarrhea. Around 90% of all diarrhea-related deaths occur in children below five years of age living in low- and middle-income countries [2]. Estimates of cases of diarrhea in both and adults children are staggering [3]. Internationally, some 2-4 billion cases of diarrhea occur annually [4] and this overwhelming level of morbidity has seen little decrease during the last four decades [3]. Currently, less than half of all young children with diarrhea receive appropriate

treatment [5], and in some low- and middle-income countries, the rates of appropriate case management are declining [6]. These scenarios are presented in a recent joint report by UNICEF and WHO, accompanied by a strategy to reduce the burden of diarrheal diseases [7]. Many of the risk factors for contracting diarrheal illnesses are associated with poor socioeconomic conditions, such as lacking access to safe water and sanitation, poor hygiene practices, and unsafe human waste disposal [8], [9], [10].

II. OVERVIEW OF DIARRHEA

Diarrhea is characterized by abnormally loose or watery stools. Most cases of diarrhea are due to bacteria, viruses, or parasites. Digestive system disorders can also cause chronic diarrhea. Many cases of diarrhea are due to an infection in the gastrointestinal tract. The microbes responsible for this infection include: bacteria, viruses and parasitic organisms. Some cases of chronic diarrhea are called "functional" because although all the digestive organs appear normal, they are not functioning as they ideally should. In the developed world, irritable bowel syndrome (IBS) is the most common cause of functional diarrhea. IBS causes many symptoms, including cramping, abdominal pain, and altered bowel habits, which can include diarrhea, constipation, or both. Inflammatory bowel disease (IBD) is another cause of chronic diarrhea. IBD describes either ulcerative colitis or Crohn's disease. Both conditions can also cause blood in the stool [11].

Some other major causes of chronic diarrhea include:

- a. Microscopic colitis: This is a persistent type of diarrhea that usually affects older adults. It develops due to inflammation and occurs often during the night.
- b. Malabsorptive and maldigestive diarrhea: The first is due to impaired nutrient absorption, and



the second is due to impaired digestive function. Celiac disease is one example.

- c. Chronic infections: A history of travel or antibiotic use can be clues in chronic diarrhea. Various bacteria and parasites can also be the cause.
- d. Drug-induced diarrhea: Laxatives and other drugs, including antibiotics, can trigger diarrhea.
- e. Endocrine-related causes: Sometimes, hormonal factors cause diarrhea. This is the case in Addison's disease and carcinoid tumors.
- f. Cancer-related causes: Neoplastic diarrhea is associated with a number of gut cancers.

A. Symptoms of Diarrhea

Diarrhea refers to watery stools, but it may be accompanied by other symptoms. These include: stomach pain, abdominal cramps, bloating, weight loss, fever, body aches and chills. Diarrhea is also a symptom of other conditions, some of which can be serious. Other possible symptoms are: blood or pus in the stool, persistent vomiting and dehydration.

B. Effects of Diarrhea

Enteric pathogens can cause dehydration and injury to the intestines by the processes described above. They can also result in damaging effects on other body systems. Numerous longlasting physical complications can result from infectious diarrhea. For example, bacteremia can result from infection by Salmonella or Campylobacter, and hemolytic uremic syndrome (HUS) can follow infection by Shigella or enterohemorrhagic E. coli (a subset of Shiga toxinproducing E. coli or STEC) [12]. Guillain-Barré syndrome, a debilitating auto-immune disorder, can occur subsequent to infection by Campylobacter jejuni [13], [14]. Reiter's syndrome, characterized by arthritis, urethritis, and conjunctivitis, may result from or accompany infection by Salmonella, Campylobacter, or Shigella [15]. Studies have documented long-term physical and cognitive deficits as a result of childhood diarrhea. Growth shortfalls have been found in children at the age of seven, who had suffered from diarrhea in early childhood [16]. Some particular enteric pathogens, such as EAEC and Cryptosporidium spp., can have negative effects on growth even when diarrhea is not evident [17], [18]. Diminished cognitive abilities have been observed in schoolchildren at 9 years of age who had suffered repeated episodes of diarrhea before the age of two [19], [20], [21], [22]. As illustrated by these reports, the possible effects of infectious diarrhea include long-lasting and

permanent detriment, which can have immense impacts on people's lives and consequently on society.

III. RELATED WORK

Data mining is the process of automating information (knowledge) discovery. Knowledge Discovery in Databases (KDD) is the process of getting high-level knowledge from low-level data. Data mining plays an important role in the KDD. Data mining is an interdisciplinary field. Its main aim is to uncover relationships in data and to predict outcomes. Researchers are trying to find satisfactory solutions in a reasonable time through search techniques as many problems are difficult to be solved in a feasible time by analytically [23]. Hence data mining got its importance. Data mining helps to extract patterns in the process of knowledge discovery in databases in which intelligent methods are applied. The emerging field of data mining promises to provide new techniques and intelligent tools which help the human to analyze and understand large bodies of data remains on difficult and unsolved problem. The common functions in current data mining practice include Classification, Regression, Clustering, Rule generation. Discovering association rules. modeling, summarization. dependency and sequence analysis. Classification is one of the important techniques of data mining. The input to the classification problem is a data-set called the training set having a number of attributes. The attributes are either continuous or categorical. Data mining techniques have been widely used in diagnostic and health care applications because of their predictive power. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion.

A. Educational Data Mining

TheEDMprocess converts raw data coming from educational systems into useful information that could potentially have a greater impact on educational research and practice" [24]. Traditionally, researchers applied DM methods like clustering, classification, association rule mining, and text mining to educational context [25]. A 2007, provided survey conducted in а comprehensive resource of papers published between 1995 and 2005 on EDM [26]. This survey covers the application of DM from traditional educational institutions to web-based learning management system and intelligently adaptive



educational hypermedia systems. In another prominent EDM survey [27], about 240 EDM sample works published between 2010 and 2013 were analyzed. One of the key findings of this survey was that most of the EDM research works focused on three kinds of educational systems. educational methods. namely. tasks, and algorithms. Application of DM techniques to study on-line courses was suggested [28]. They proposed a non-parametric clustering technique to mine offline web activity data of learners. Application of association rules and clustering to support collaborative filtering for the development of more sensitive and effective e-learning systems was studied [29].

B. Clustering Algorithms

Clustering simply means collecting and presenting similar data items. But what defines similarity? That is the key to understanding `clustering'. A cluster is therefore a group of items that are similar to each other within the group and dissimilar to objects belonging to other clusters. In statistical notation, ``clustering is the most important unsupervised learning algorithm" [30]. Being a pre-processing algorithm in the data mining process, clustering can significantly reduce the data size to meaningful clusters that can be used for further data analysis. However, one must be careful when reducing the data size because when representing data in the form of fewer clusters typically loses certain fine details similar to lossy data compression. The classification of clustering algorithms is imprecise because several of them overlap with each other. In traditional terms, clustering techniques have broadly been classified into two types, hierarchical and partitional.

C. Spatio-temporal data mining (STDM)

Spatio-temporal data mining (STDM) is becoming growingly important in the big data era with the increasing availability and importance of large spatio-temporal datasets such as maps, virtual globes, remote-sensing images, the decennial census and GPS trajectories. STDM has broad various domains applications in including environment and climate (e.g. wind prediction and precipitation forecasting), public safety (e.g. crime prediction), intelligent transportation (e.g. traffic flow prediction), human mobility (e.g. human trajectory pattern mining), etc. Classical data mining techniques that are used to deal with transaction data or graph data often perform poorly when applied to spatio-temporal datasets because of many reasons. [31] reviewed the state-of-the-art in STDM research and applications, with emphasis

placed on the data mining tasks of prediction, clustering and visualization for spatio-temporal data. [32] reviewed STDM from the computational perspective, and emphasized the statistical foundations of STDM. [33] reviewed the methods and applications for trajectory data mining, which is an important type of ST data [34]

D. PyClustering

PyClustering is an open source data mining library written in Python and C++ that provides a wide range of clustering algorithms and methods, including bio-inspired oscillatory networks. PyClustering is mostly focused on cluster analysis to make it more accessible and understandable for users [35].

e. PPDM

Privacy-preserving data mining (PPDM) offers the possibility of using data mining methods without disclosing private information. PPDM approaches include data perturbation (data modification) [36], [37] and encryption [38]. Cryptographic methods are renowned for securing data. Literature provides many examples where PPDM effectively utilize cryptographic methods [39, [40].

f. Other Data mining progams

Several open-source general purpose data mining libraries or programs have been developed such as Knime [41], Mahout [42], and Weka [43]. Although these software programs provide algorithms for many data mining tasks, they provide very few algorithms for mining frequent patterns in databases. Prediction and Control of Stroke by Data Mining is presented in [44], while [45] examined the analysis of Heart diseases Dataset using Neural Network Approach. Also the Diagnosis of Diabetes using Classification Mining Techniques was analysed in [46]. Cancer diagnosis using Data Mining Technology was discussed in [47].

None of the above literature addresses the issue of the application of data mining for predicting and controlling diarrhea disease. The aim of this research is to develop a data mining system for predicting and controlling the prevalence of chronic diarrhea using the K-nearest neighbor and C4.5 decision tree components of WEKA 3.6 software.

IV. METHODOLOGY

Pattern Classification is a two-step process based on a set of attributes namely assigned as class label and sample data. A clustering algorithm

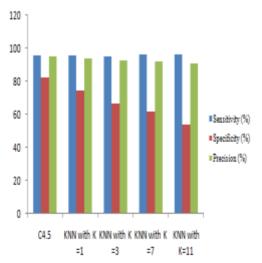


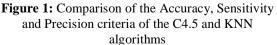
which makes a model from the analysis of training data set of class labels and predefines concepts is presented in the Learning phase, while the classification of accuracy of the model measured by the use of test data is presented in the Test phase. To actualize the of this research, data of diarrhea patients was collected from Ladoke Akintola University Teaching Hospital Osogbo. Literature on Diarrhea and Data mining was also studied. Because some of the collected records have unspecified values, the following techniques were used for the validation and analysis: Mean value of feature was used to fill theunspecified values; it is a method that is commonly used for continuous data like height and weight. Mean values in line was used to make sure that their class was equal to lines with unspecified values. It is a method commonly used for continuous features. WEKA 3.6 data mining software was used for data analysis, while C4.5 and the K-nearest neighbor algorithms were used on the data respectively.

	Table 1: Classification Algorithm of Diarrhea with data sets							
Algorithm	TN	FP	FN	TP	Sensitivity (%)	Specificity (%)	Precision (%)	Classification Accuracy (%)
C4.5	97	14	22	644	95.33	82.5	95.27	93.88
KNN with K =1	86	35	20	646	95.53	74.4	93.72	92.67
KNN with K =3	77	42	21	643	95.32	66.2	92.48	90.35
KNN with K =7	69	49	11	654	96.44	61.7	91.84	91.46
KNN with K=11	64	57	14	6 55	96.40	53.8	90.56	90.81

V.RESULTS AND ANALYSIS

The different values of the criterions used for the algorithms are presented in Table 1. The comparisons of the criteria in the two methods are as in Figure 1, while the Specificity criterion is in Figure 2.





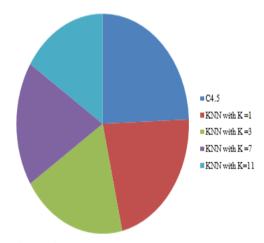


Figure 2; Comparison of specificity criteria of C4.5 and KNN algorithms

As presented in the figures of K-nearest algorithm, Specificity and Precision decreased after K increment, and then Accuracy decreased a bit. Moreso, after K increment, Sensitivity increased a bit this indicates that the number of patients diagnosed with diarrhea increased but at the same time the false-positive patients also increased. This led to Precision reduction as K increased. By implication the C4.5 algorithm is better than the K-nearest neighbor algorithm in the Accuracy, Precision and Specificity criteria with a slight value. This slight difference in classification is



important in medical diagnosis because accurate prediction of illness very is germane.

VI. CONCLUSION

In this research, C4.5 decision tree algorithm and K-nearest neighbour have been used to predict diarrhea among high risk community. The decision-tree algorithm was used because of its higher accuracy. This is depicted in the efficiency of the Classification of the Sensitivity and Accuracy of the algorithms. Besides the experience and knowledge of medical practitioners, the use of Data mining techniques can also be incorporated in the diagnosis and management of diarrhea. This is a useful tool in predicting and controlling the prevalence of diarrhea including the chronic types. Future research in the area of decision support systems incorporating data mining techniques is solicited.

REFERENCES

- Black, R., Fontaine, O., Lamberti, L., Bhan, [1]. M., Huicho, L., El Arifeen, S., Masanja, H., Walker, C. F., Mengestu, T. K., Pearson, L., Young, M., Orobaton, N., Chu, Y., Jackson, B., Bateman, M., Walker, N., & Merson, M., 2019, "Drivers of the reduction in childhood mortality 1980-2015 diarrhea and interventions to eliminate preventable diarrhea deaths by 2030". Journal of global health, 9(2), 020801. https://doi.org/10.7189/jogh.09.020801
- [2]. Kallander K., 2007, "Infektionstermometern - en rapport ominfektionssjukdomarnas inverkan på hälsoläget i världen. Boken om smittsamma sjukdomar". Stockholm: Världsinfektionsfonden.
- [3]. Kosek M, Bern C, & Guerrant R. L., 2003, "The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000". Bull World Health Organ 2003; 81(3): 197-204.
- [4]. Farthing M.J., 2007, "Diarrhoea- a significant worldwide problem". Int J Antimicrob Agents; 14(1): 65-9.
- [5]. Forsberg B.C, Petzold M.G, Tomson G, & Allebeck P., 2007, "Diarrhoea case management in low- and middle-income countries-an unfinished agenda". Bull World Health Organ; 85(1): 42-8.
- [6]. Ram P.K, Choi M, Blum L.S, Wamae A.W, Mintz ED, & Bartlett AV., 2008, "Declines in case management of diarrhoea among children less than five years old". Bull World Health Organ; 86(3): E-F.

- [7]. Wardlaw T, Salama P, Brocklehurst C, Chopra M, & Mason E., 2010, "Diarrhoea: why children are still dying and what can be done". Lancet; 375(9718): 870-2
- [8]. Daniels D.L, Cousens S.N, Makoae L.N, & Feachem R.G., 1990, "A case control study of the impact of improved sanitation on diarrhea morbidity in Lesotho". Bull World Health Organ; 68(4): 455-63.
- [9]. Manun'ebo M.N, Haggerty P.A, Kalengaie M, & Ashworth A, Kirkwood B.R., 1994, "Influence of demographic, socioeconomic and environmental variables on childhood diarrhoea in a rural area of Zaire". J Trop Med Hyg; 97(1): 31-8.
- [10]. Graf J, Meierhofer R, Wegelin M, & Mosler H.J., 2008, "Water disinfection and hygiene behaviour in an urban slum in Kenya: impact on childhood diarrhoea and influence of beliefs". Int J Environ Health Res; 18(5): 335-55.
- [11]. Medical News Today, 2019, "What you should know about diarrhea". Retrieved on 10th December, 2019. https://www.medicalnewstoday.com/articles/ 158634
- [12]. Rudolph J.A, & Cohen M.B., 1999, "New causes and treatments for infectious diarrhea in children". Curr Gastroenterol Rep; 1(3): 238-44.
- [13]. Hughes R.A, Cornblath D.R. & Guillain-Barre, 2005, Lancet; 366(9497): 1653-66.
- [14]. Allos B.M., 1997, "Association between Campylobacter infection and Guillain-Barre syndrome". J Infect Dis; 176(Suppl 2): S125-8.
- [15]. Camilleri M, & Murray J.A., 2008, "Diarrhea and Constipation". In: Fauci A.S, Kasper D.L, Hauser S.L, Longo D.L, Jameson J.L, Loscalzo J, Eds. Harrison's principles of internal medicine. 17th ed. New York: McGraw-Hill Medical
- [16]. Moore S.R, Lima A.A, Conaway M.R, Schorling J.B, Soares A.M, & Guerrant R.L., 2001, "Early childhood diarrhoea and helminthiases associate with long-term linear growth faltering". Int J Epidemiol; 30(6): 1457-64
- [17]. Checkley W, Gilman R.H, Black R.E, et al., 2004, "Effect of water and sanitation on childhood health in a poor Peruvian periurban community". Lancet; 363(9403): 112-8.
- [18]. Checkley W, Gilman R.H, Epstein L.D, et al., 1997, "Asymptomatic and symptomatic cryptosporidiosis: their acute effect on



weight gain in Peruvian children". Am J Epidemiol ; 145(2): 156-63.

- [19]. Checkley W, Epstein L.D, Gilman R.H, Black R.E, Cabrera L, & Sterling C.R., 1998, "Effects of Cryptosporidium parvum infection in Peruvian children: growth faltering and subsequent catch-upgrowth". Am J Epidemiol; 148(5): 497-506.
- [20]. Guerrant D.I, Moore S.R, Lima A.A, Patrick P.D, Schorling J.B, & Guerrant R.L., 1999, "Association of early childhood diarrhea and cryptosporidiosis with impaired physical fitness and cognitive function four-seven years later in a poor urban community in northeast Brazil". Am J Trop Med Hyg; 61(5): 707-13.
- [21]. Lorntz B, Soares A.M, Moore S.R, et al., 2006, "Early childhood diarrhea predicts impaired school performance". Pediatr Infect Dis J; 25(6): 513-20.
- [22]. Niehaus M.D, Moore S.R, Patrick P.D, et al., 2002, "Early childhood diarrhea is associated with diminished cognitive function 4 to 7 years later in children in a northeast Brazilian shantytown". Am J Trop Med Hyg; 66(5): 590-3.
- [23]. Usama Fayyad, Gregory Piatetsky-Shapiro, & Padhraic Smyth, 1996, "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial, 17(6); 37 -54.
- [24]. Mohsin, M.F.M., Norwawi, N.M., Hibadullah, C.F. & Wahab, M.H.A., 2010, "Mining the student programming performance using rough set," presented at the Int. Conf. Intell. Syst. Knowl. Eng. (ISKE), 478-483.
- [25]. Romero C. & Ventura S., 2007, "Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, 135-146.
- [26]. Peña-Ayala, A., 2014, "Educational data mining: A survey and a data mining based analysis of recent works," Expert Syst. Appl., 41(4): 1432-1462
- [27]. Zaiane, O.R., & Luo, J., 2001, "Web usage mining for a better Webbased learning environment," in Proc. Conf. Adv. Technol. Edu., pp. 60-64.
- [28]. Zaiane, O.R., 2002, "Building a recommender agent for e-learning systems," in Proc. Int. Conf. Comput. Edu., pp. 55-59.
- [29]. Madhulatha, T.S., 2012, "An overview on clustering methods." [Online]. Available: https://arxiv.org/abs/1205.1117

- [30]. Ashish Dutt, Maizatul Akmar Ismail, & Tutut Herawan, 2017, "A Systematic Review On Educational Data Mining". IEEE Access vol 5. pp. 15991-16005
- [31]. Shekhar, S., Jiang, Z., Ali, R.Y., Eftelioglu, E., Tang, X., Gunturi, V.M.V. & Zhou, X., 2015, "Spatiotemporal data mining: A computational perspective". ISPRS International Journal of Geo-Information, vol. 4, pp. 2306–2338.
- [32]. Mazimpaka, J.D & Timpf, S., 2016, "Trajectory data mining: A review of methods and applications". Journal of Spatial Information Science, vol.13, pp. 61– 99.
- [33]. Li, X., Zhao, K., Cong, G., Jensen, C.S., & Wei, W., 2018, "Deep representation learning for trajectory similarity $34^{\acute{th}}$ computation". In 2018 IEEE International Conference on Data Engineering (ICDE), pp. 617-628.
- [34]. Andrei V. Novikov, 2019, "PyClustering: Data Mining Library". Journal of Open Source Software, 4(36), 1230. https://doi.org/10.21105/joss.01230
- [35]. Kerschbaum, F., & Harterich, M., 2017, "Searchable encryption to reduce encryption degradation in adjustably encrypted databases". In IFIP Annual Conference on Data and Applications Security and Privacy (pp.325{336). Springer doi:https://doi.org/10.1007/978-3-319-61176-1_18.
- [36]. Kieseberg, P., & Weippl, E., 2018, "Security challenges in cyber-physical production systems". In International Conference on Software Quality (pp. 3{16). Springer. doi:https://doi.org/10. 1007/978-3-319-71440-0_1.
- [37]. Li, P., Li, J., Huang, Z., Gao, C.-Z., Chen, W.-B., & Chen, K., 2017, "Privacypreserving outsourced classification in cloud computing". Cluster Computing, (pp. 1{10.). doi:https://doi.org/10.1007/s10586-017-0849-9.
- [38]. Chen, K., & Liu, L., 2005, "A random rotation perturbation approach to privacy preserving data classification". The Ohio Center of Excellence in Knowledge-Enabled Computing,
- [39]. URL:https://corescholar.libraries.wright.edu/ knoesis/916/.
- [40]. Chen, K., & Liu, L., 2011, "Geometric data perturbation for privacy preserving outsourced data mining". Knowledge and Information Systems, 29, pp. 657{695.



doi:https://doi.org/10.1007/s10115-010-0362-4.

- [41]. Berthold, M. R. et al., 2009, "KNIME the Konstanz information miner: version 2.0 and beyond". SIGKDD Explorations, 11(1): 26
- [42]. Mahout Software available at: http://mahout.apache.org/
- [43]. Witten, I. H., & Frank, E., 2005, "Data mining: practical machine learning tools and techniques". Morgan Kaufmann.
- [44]. Leila Amini, Reza Azarpazhouh, Mohammad Taghi Farzadfar, Sayed Ali Mousavi, Farahnaz Jazaieri, Fariborz Khorvash, Rasul Norouzi, & Nafiseh Toghianfar, 2013, "Prediction and Control of Stroke by Data Mining". International Journal of Preventive Medicine, 8th Iranian Neurology Congress, vol 4, pp. S245-S249
- [45]. Usha Rani, 2011, "Analysis of Heart diseases Dataset using Neural Network Approach". International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, No.5, pp 1-8. DOI : 10.5121/ijdkp.2011.1501 1
- [46]. Aiswarya Iyer, S. Jeyalatha & Ronak Sumbaly, 2015, "Diagnosis of Diabetes using Classification Mining Techniques". International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, pp 1-14. DOI : 10.5121/ijdkp.2015.5101 1
- [47]. Muhammad Shahbaz, Shoaib Faruq, Muhammad Shaheen, & Syed Ather Masood, 2012, "Cancer diagnosis using Data Mining Technology". Life Science Journal, 2012;9(1), pp. 308-313

International Journal of Advances in Engineering and Management ISSN: 2395-5252

IJAEM

Volume: 03

Issue: 01

DOI: 10.35629/5252

www.ijaem.net

Email id: ijaem.paper@gmail.com