wjert, 2020, Vol. 6, Issue 2, 542-557. *Original Article* ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org SJIF Impact Factor: 5.924

ANALYSIS OF MALARIA DIAGNOSIS ON PATIENTS USING DATA MINING CLUSTERING TECHNIQUES

Mani Shanker Chaubey[1]*, Adelaja Adebayo Oluwaseun[2] and Nyaaku Oluwayemisi Esther[3]

[1*]MSc Computer Science and Information Technology, South Ural State University, Chelyabinsk, Russia.

[2]MSc Computer Science and Information Technology, South Ural State University, Chelyabinsk, Russia.

[3]MSc Computer Science, Lagos State University, Lagos, Nigeria.

Article Received on 15/02/2020 **Article Revised on 05/03/2020** **Article Accepted on 26/03/2020**

***Corresponding Author**
Mani Shanker Chaubey
MSc Computer Science and Information Technology, South Ural State University, Chelyabinsk, Russia.

ABSTRACT

The research was carried out on the malaria patients with some symptoms on high rate that shows positive +ve result while those with some symptoms on low rate that shows negative -ve result. KNIME data mining tool was used to build a comprehensive work flow model consisting of nodes with their respective functions. Fuzzy c-mean, k-mean and hierarchical clustering nodes were utilized to produce grouped subsets termed clusters from the malaria_result.csv file (training-set). A decision tree level classifier was designed from the patient's diagnosis of the malaria symptoms. Data Analysis Knowledge Discovery Process for the clustering was also built. The result obtained in this research shows statistical clustering means such as scatter plots, interactive histogram, clustered data table and interactive tables which will be helpful for future observations and predictions of malaria in health care.

KEYWORDS: KNIME, Fuzzy c-means node, k-mean node, hierarchical node, Knowledge Discovery.

INTRODUCTION

In most medical sectors across the world today especially in Nigeria, large amount of medical data related to patients suffering from malaria has been retrieved for medical history. The process of mining, analyzing these medical data records of malaria victims can effectively support medical practitioners and the health sector to predict likely occurrence of malaria thereby taking preventive measures to reduce the rate at which people contracted this ailment now and the future. The World Health Organization (WHO) statistical reports proved that malaria is the second leading cause of mortality from vulnerable disease in most African countries after HIV/AIDS.[1]

The *anopheles mosquitoes* bite on human transmits *genus of unicellular eukaryotes* (obligate parasites) called *plasmodium* is one of the main causes of malaria. There are different types of plasmodium parasites but only five types cause malaria in humans. *Plasmodium falciparum* which is the most common malaria parasite, mainly found in Africa also responsible for most malaria deaths worldwide while the *plasmodium malariae* is quite rare and found in Africa. The *plasmodium ovale* is a fairly uncommon parasite but usually found in countries located in the western part of Africa; this particular parasite can reside in the human's liver for several years without showing the victim any symptoms.[2] The *plasmodium knowlesi* is found in Southeast Asia. In some countries in Asia and the Southern part of America, the *plasmodium vivax* can be found, it causes milder symptoms than the *plasmodium falciparum* and it can stay in the liver for an approximately 3- 4 years which could result in relapses.[3]

The World Health Organization, World Bank, UNDP and UNICEF established a new initiative health sector-wide partnership known as the Roll Back Malaria RBM to support researches and discover products such as drugs to combat malaria at global, regional, country and local levels. The Roll Back Malaria movement saved an estimate of 3 million lives across the World; the global mortality rate drastically reduced by 43% while in Sub-Saharan Africa regions (such as Botswana, Gabon, Rwanda, Uganda, Senegal, Tanzania, Zimbabwe, Congo, Ghana and especially Nigeria) the rate decreased by 49%.[4] This paper focuses on using the data mining clustering technique to effectively analyze the large amount of data and medical examination history obtained from patients who were diagnosed of malaria to transform this data into useful knowledge. Data mining is a process used in extracting rules; predicting certain performances in numerous professions such as medicine and surgery, sciences, military, aviation, education, human resource and information technology.[5] Data mining can be described as the collection or gathering of pure driven data algorithms to achieve meaningful knowledge pattern from the raw data. Data mining also known as “data dredging” or “data fishing” or “knowledge mining in data” is a process that involves searching of large information of data or records to discover knowledgeable patterns and utilizing these patterns for predicting future occurrences especially in the medical area which is the main scope of this research.[6] As a result of technological advancement, it is possible to gather, store and easily retrieve large quantities of medical records containing vital information which could help in diagnosing and treating malaria. Data mining techniques such as clustering can extract intelligent knowledge through data analytics from medical data for convenient treatment and diagnosis. Clustering is a data mining task of assigning a set of objects into groups called *Clusters.* Clustering is a solution for data analytics which involves partitioning a set of data objects into subsets. In clustering techniques, objects in a subset are identical to one another yet dissimilar to objects in other subsets.[6] Clustering is an unsupervised learning performed on data sets by observations, this is before the semantic of the classes is not known beforehand. The main types of clustering methods include partitioning method, hierarchical clustering, fuzzy C-means clustering, density-based clustering, model-based clustering or grid-based clustering.[6] The main goal of clustering in data mining application particularly in the medical aspect is group both similar and different set of data objects derived for the medical results into the same clusters and different clusters respectively.[7] Clustering can also be defined as a process of grouping a given set of unlabeled patterns into a number of clusters such that identical patterns are attributed to one cluster.[8]

Decision Tree Level Classifer

Decision tree is a tree-like structure consisting of internal nodes, branches, and leaf node, in which each branch denotes an attribute value, each internal node denoted a test on an attribute which is used for and a leaf node represents the predicted classes or class distributions.[9] The decision tree have *nodes* that form *rooted tree*, this implies that decision tree is a *directed tree* with a node known as *root.*[10] Decision tree supports a predictive approach in data mining and machine learning.[11,12] The decision tree level classifier used in this research is an organized hierarchical structure which shows set of condition such as the list of malaria symptoms in the patient (high temperature/ fever, nausea, common cold, headache, body pain, diarrhea and vomiting), the decision and the result of the decision which states YES which depicts the test is positive ***+ve***; or NO which depicts the patient is being checked for other malaria symptom, else they are tested negative ***–ve***.

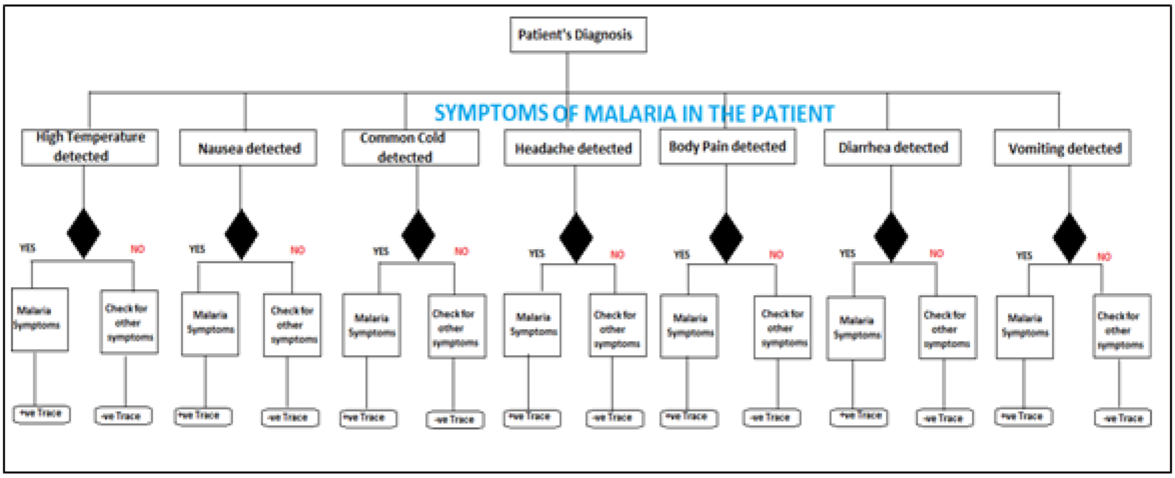


Figure 1: Hierarchical Structure Showing the Malaria Symptom and Results.

Fuzzy C-Means Clustering

The fuzzy c-means clustering algorithm is a well-known unsupervised learning technique that is used to reveal the underlying data structure. Fuzzy clustering allows each data point to belong to several clusters, with a degree of membership to each one.[13] Fuzzy C-Means clustering is widely applied in medical diagnosis, shape analysis, image processing and analysis, target recognition, geology and engineering. It is a data clustering technique used for analyzing distance between various input data points.[14] Clusters are formed based on the distance between data points, likewise the centers of the cluster are formed for each cluster.[15]

Hierarchical Clustering

Hierarchical Clustering approach creates a decomposition of data sets (or objects) in multiple levels of hierarchies using some criterion. It is method of analyzing clusters with the aim of building the cluster in a hierarchical form.[7] There are two approaches to hierarchical clustering known as the *agglomerative/ bottom-top* and *divisive/ top-bottom.* The agglomerative hierarchical clustering starts with a single data point and merge two or more cluster in a recursive pattern.[7] The divisive hierarchical clustering starts with big cluster and splits this cluster into smaller clusters in a recursive pattern.[7]

K-Means Clustering

K-Means is one of typical partitioning clustering approach in which each cluster is represented by the centre of the cluster. It is a method of clustering observations into a specific number of disjoint clusters.[16] The aim of the algorithm is to minimize the measurement between the centroid of the clusters and a given observation by iteratively appending the observation to and clusters when the lowest distance is achieved. K-means performance is determined by initialization and appropriate distance measure.[17]

Data Analysis Framework

The result obtained from the malaria diagnosis conducted on the patient is termed as the RAW data; to derived meaningful information from this *RAW* data, we decided to perform data transformation and cleansing process before analyzing these data with three clustering methods (k-Means, Fuzzy C-means and Hierarchical clustering) respectively. The main target is to easily produce different set of clusters from the respective clustering method which we utilized. The final stage of this data analysis framework is the report patterns generated from the *interactive histograms, interactive tables and scatter plots* in the three clustering approaches employed in this work. From the outputs of the report patterns, we could discover knowledge due the *RAW* data which is being interpreted and effectively useful for most medical centre across the counties where malaria parasite is predominant for convenient diagnosis now and the future.

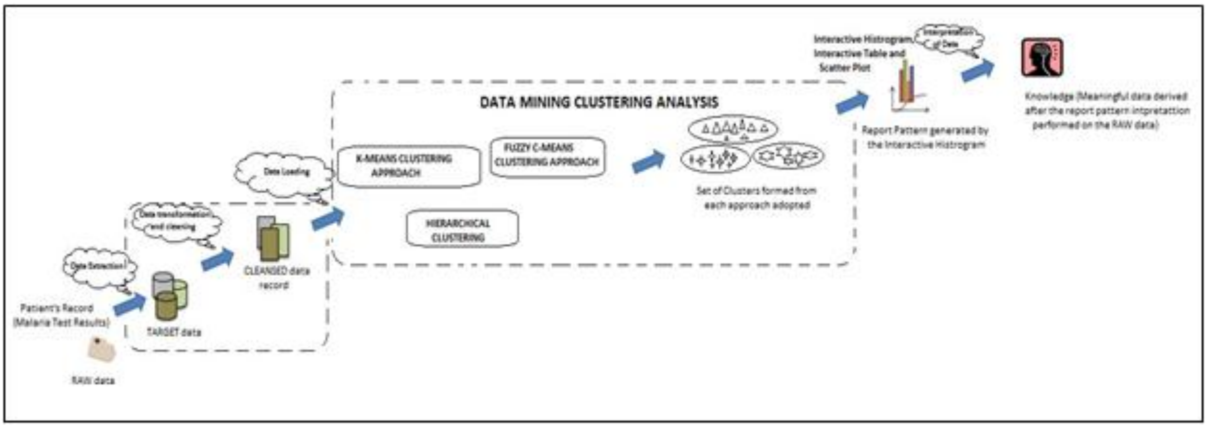


Figure 2: Data Analysis Knowledge Discovery Process on the Malaria Results.

RESULTS AND DISCUSSION

The goal of this research was to build a workflow model which reads the malaria patients data from **malaria_result.csv** file, perform three major clustering (fuzzy c-means, k-means and hierarchical) for the chosen attributes (Patient Medical File No, High Temperature, Nausea, Common Cold, Headache, Body Pain, Diarrhea, vomiting and the test result) hence displaying the results. The workflow model was build and the nodes were connected using the KNIME data mining software. Figure 3 shows the workflow model:

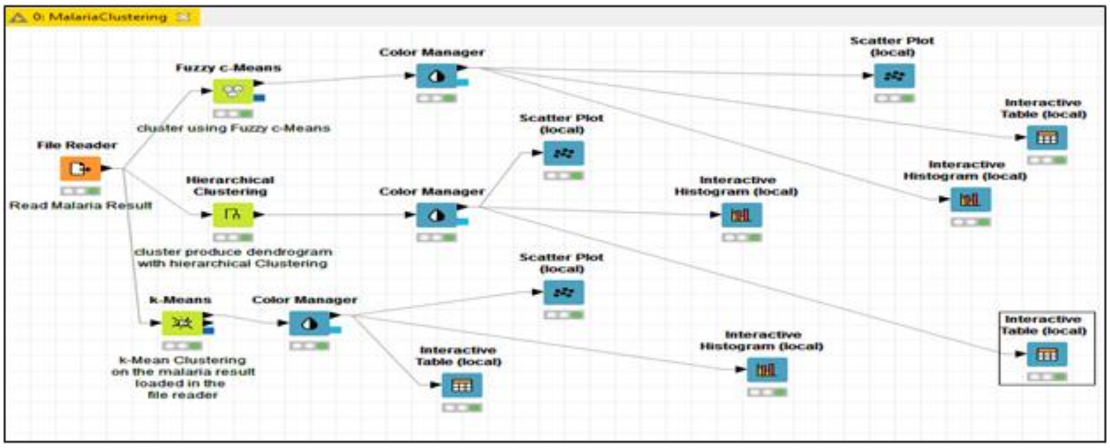


FIGURE 3: Workflow Model that read the malaria result and perform clustering.

3.0 Fuzzy C-Means Node Clustering on the Malaria Results

Fuzzy c-means clustering on the malaria data produced three clusters (cluster 0, cluster 1 and cluster 2) with their clusters centers on each attributes. Other results such as the within and between Cluster Variation, the value of partition coefficient, the value of the partition entropy, the Xie-Beni (XB) index and Fuzzy Hyper volumes for the three clusters were computed as shown in Figure 4:

FIGURE 4: Statistics View of the Fuzzy C-Means Clustering on the Malaria Results.

3.1 Hierarchical Clustering on the Malaria Results

Hierarchical clustering on the malaria data produced a dendrogram of the clusters on the row attributes, the distance between these clusters and the clustered data table produced by the dendrogram.

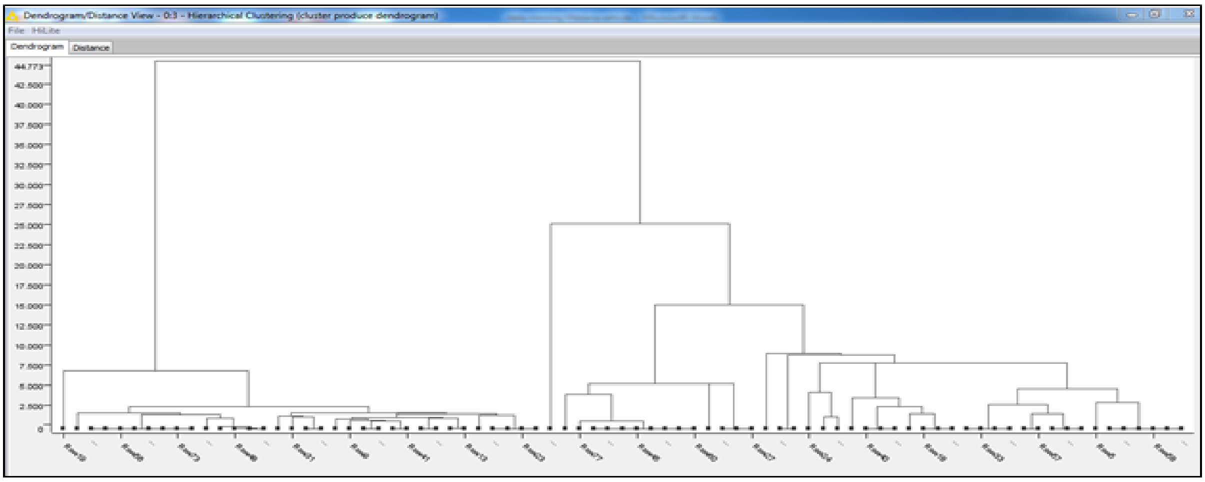


Figure 5: Hierarchical Clustering Dendrogram produced on the malaria result.

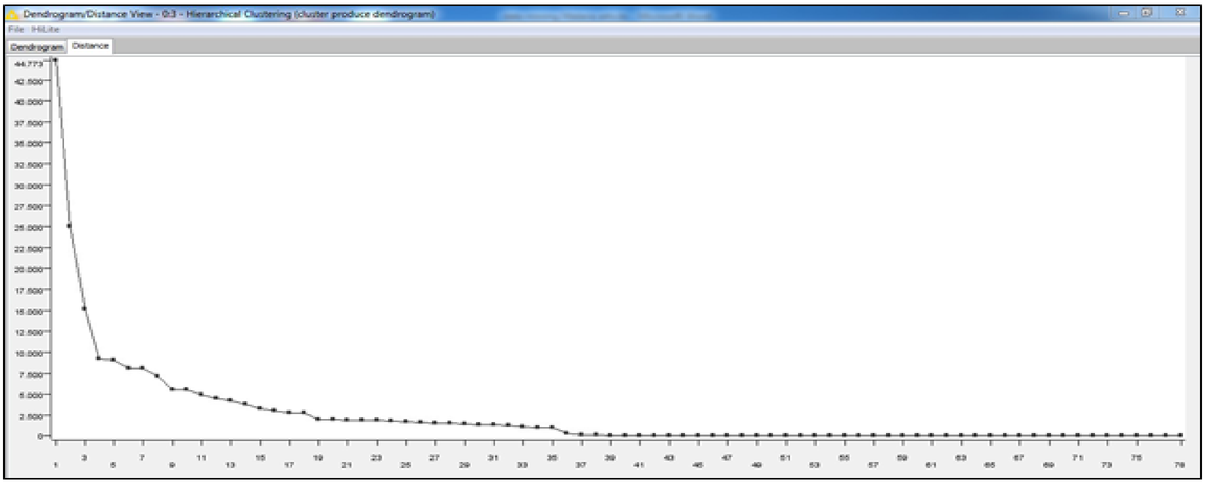


Figure 6: Hierarchical Clustering distance between the clusters produced from the data.

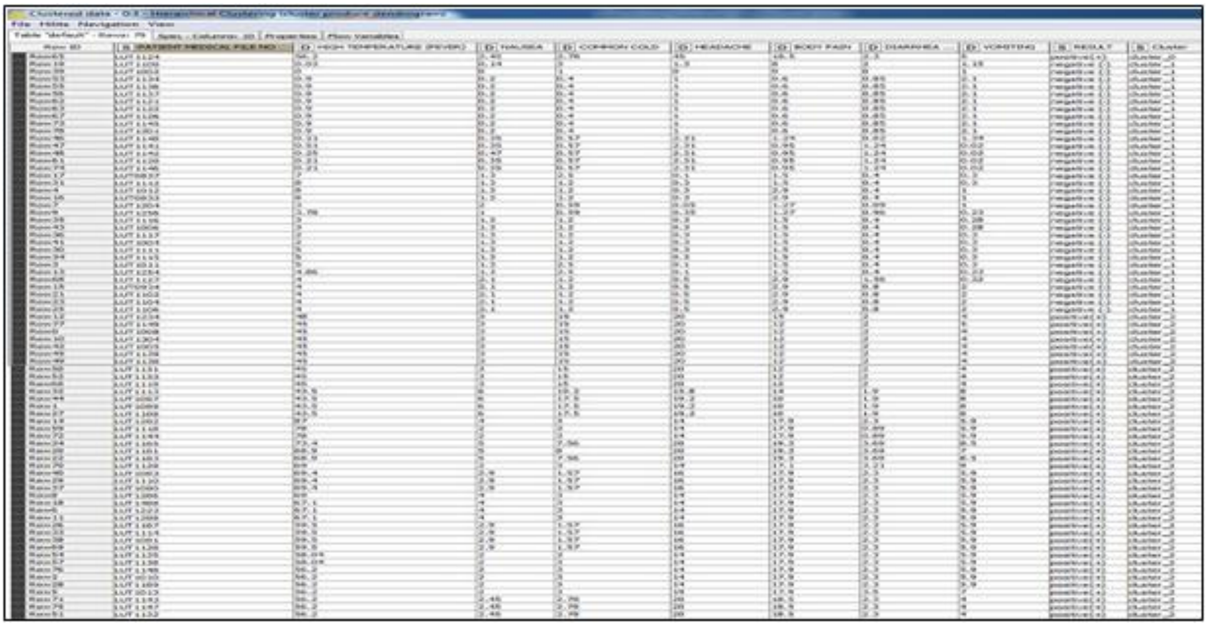
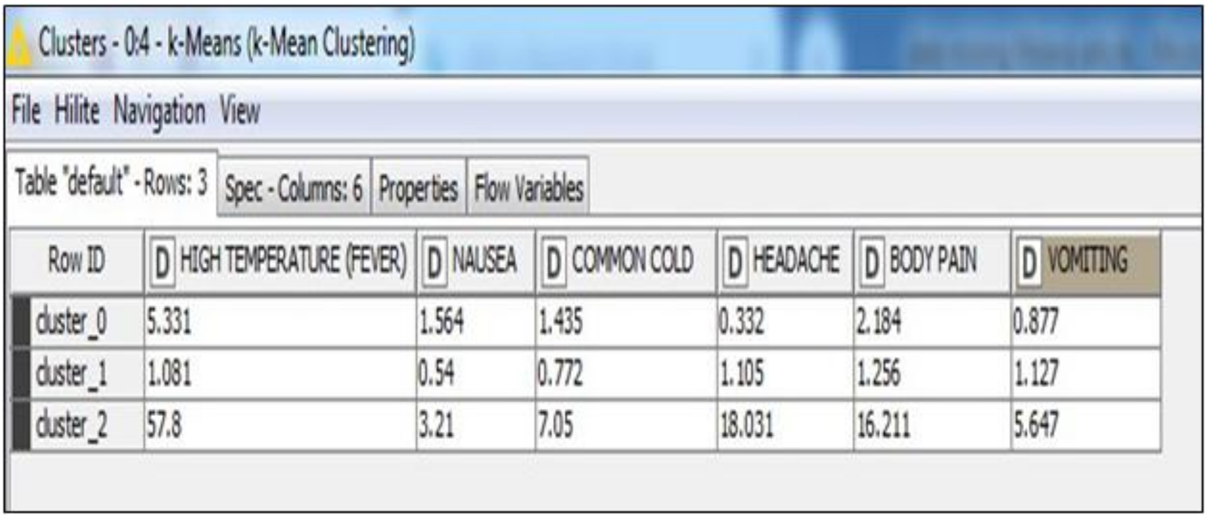


Figure 7: Clustered Data Table produced from dendrogram hierarchical clustering.

3.2 K-Means Clustering on the Malaria Results

Clusters produced on the malaria result from the k-Means clustering were cluster_0, cluster_1 and cluster_2. K-Means clustering on the training data load generated a cluster view and k-means cluster table.



Row ID	D HIGH TEMPERATURE (FEVER)	D NAUSEA	D COMMON COLD	D HEADACHE	D BODY PAIN	D VOMITING
cluster_0	5.331	1.564	1.435	0.332	2.184	0.877
cluster_1	1.081	0.54	0.772	1.105	1.256	1.127
cluster_2	57.8	3.21	7.05	18.031	16.211	5.647

Figure 8: K-Means Clustered Data Table on the attributes.

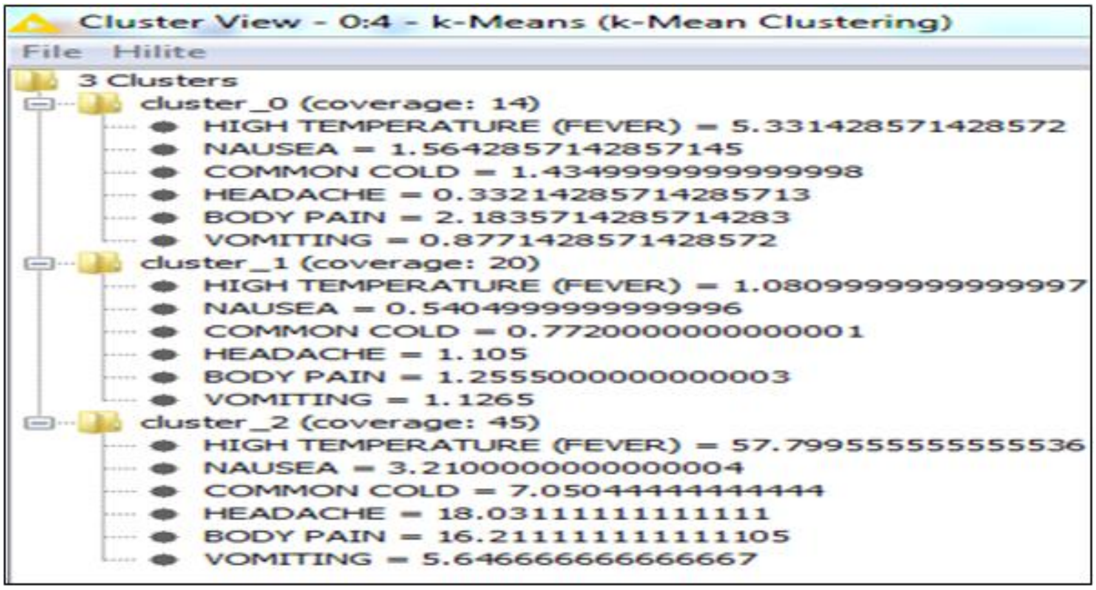


Figure 9: K-Means Cluster View on the attributes.

3.3 Scatter Plot (Fuzzy C-Means, K-Means, And Hierarchical) Clustering.

Fuzzy C-means Clustering Scatter Plot on the attributes.

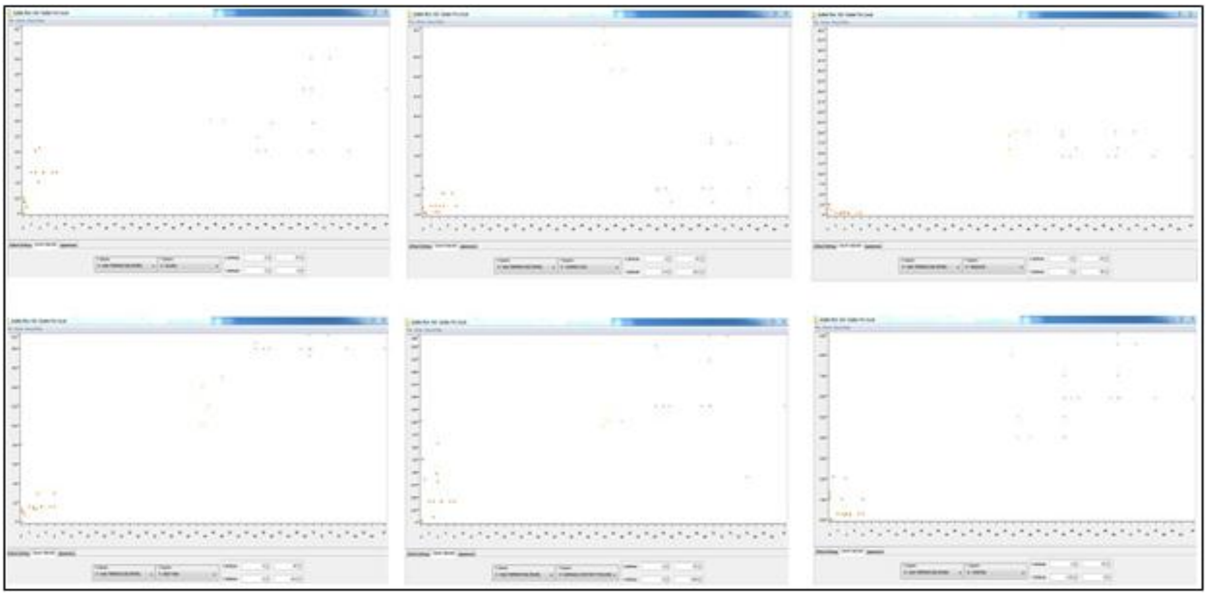


Figure 10: Scatter Plot generated by Fuzzy C-Means on the attributes.

Hierarchical Clustering Scatter Plot on the attributes.

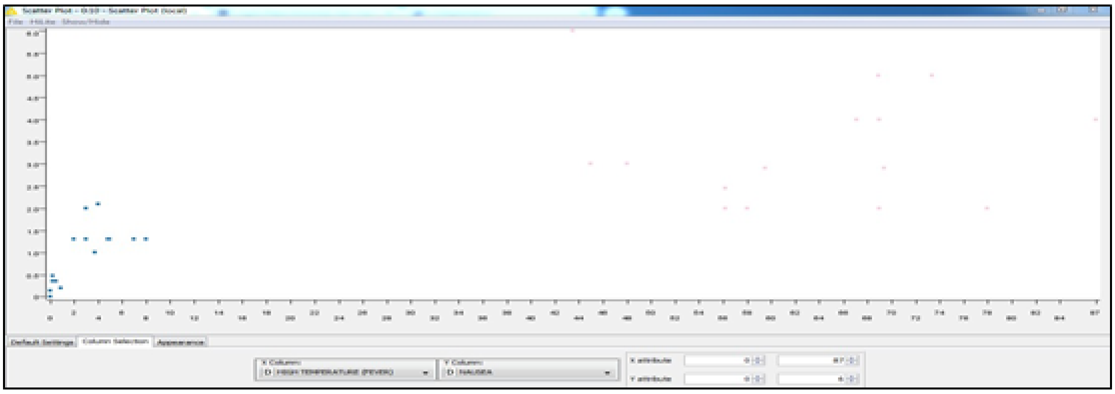


Figure 11: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Nausea).

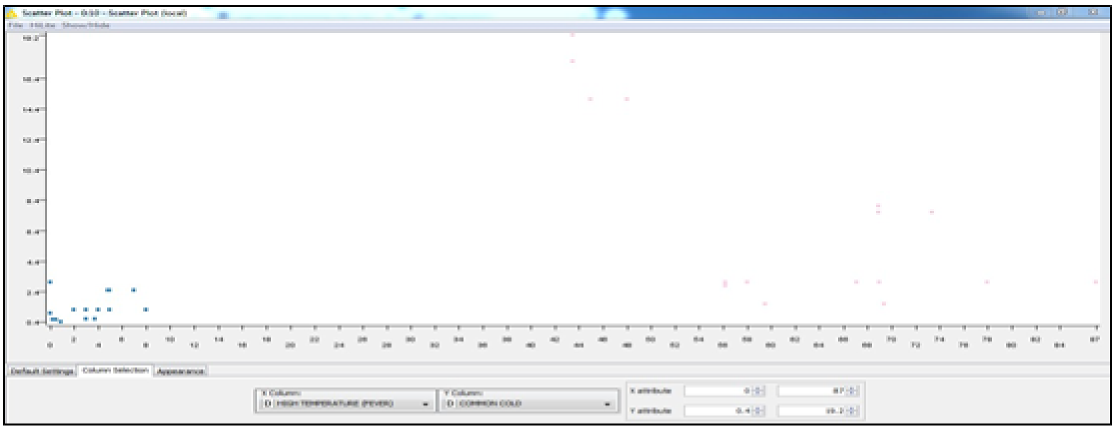


Figure 12: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Common Cold).

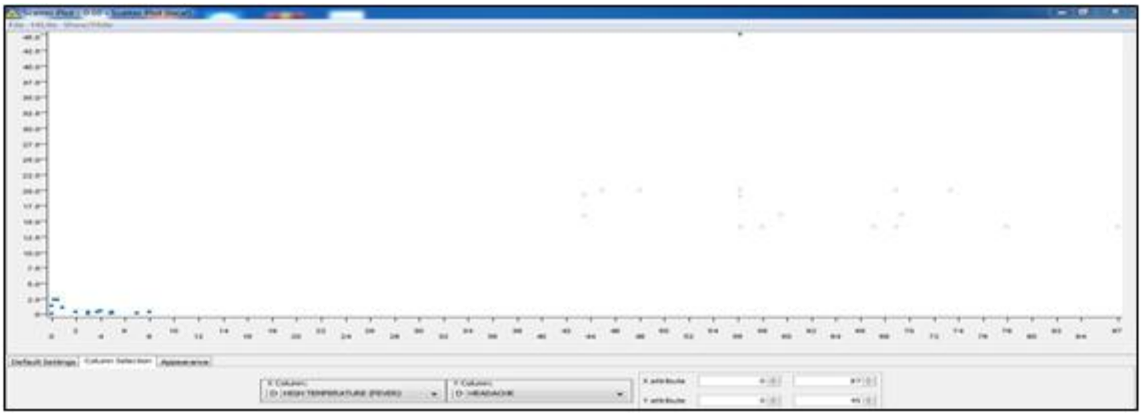


Figure 13: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Headache).

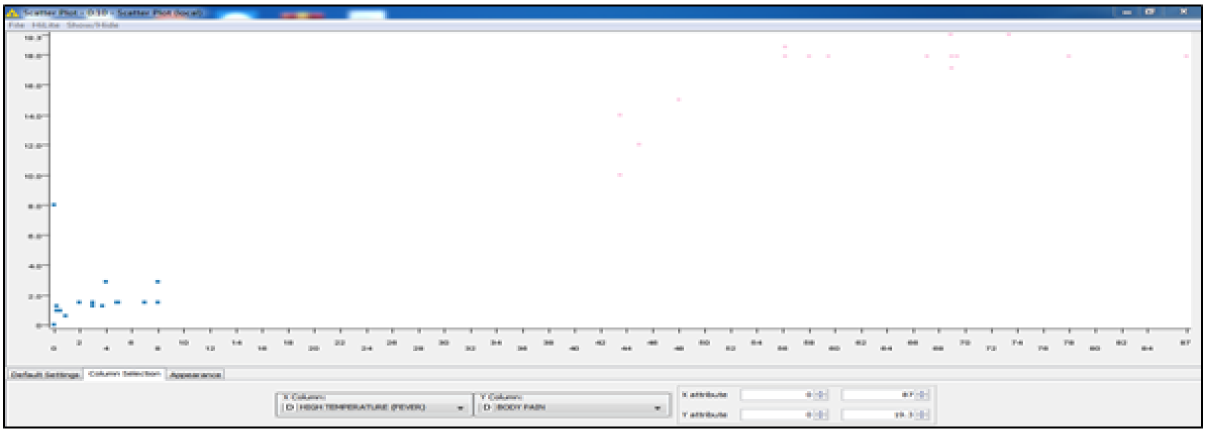


Figure 14: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Body Pain).

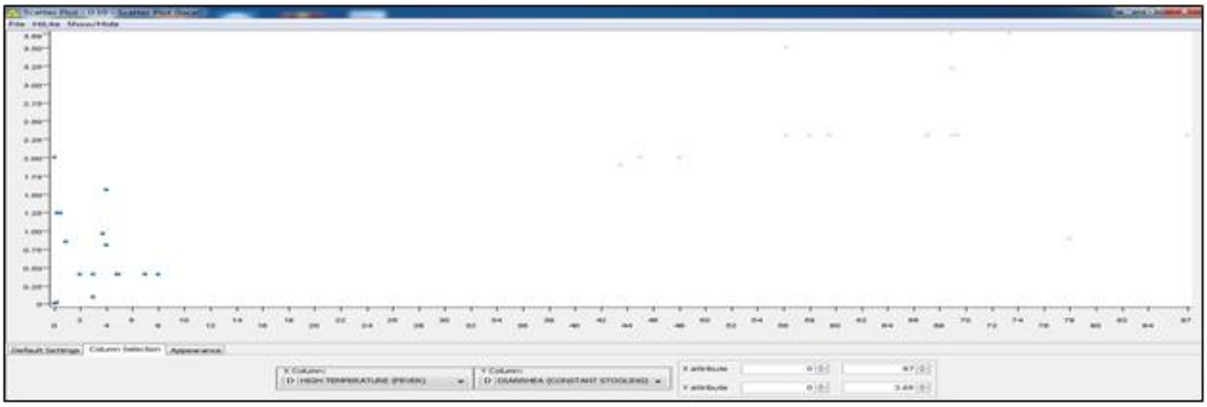


Figure 15: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Diarrhea).

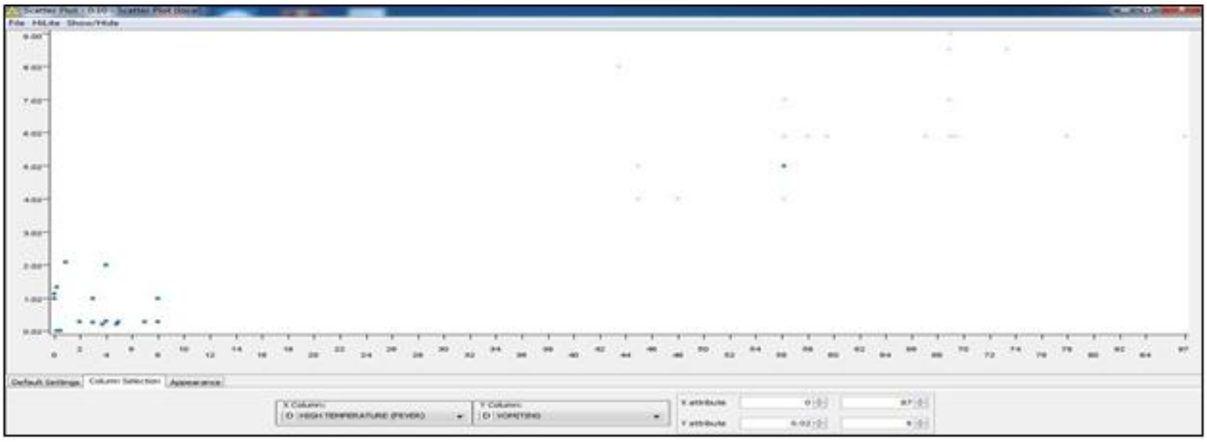


Figure 16: Scatter Plot generated by Hierarchical clustering on attributes (High temperature and Vomiting).

K-Means Clustering Scatter Plot on the attributes.

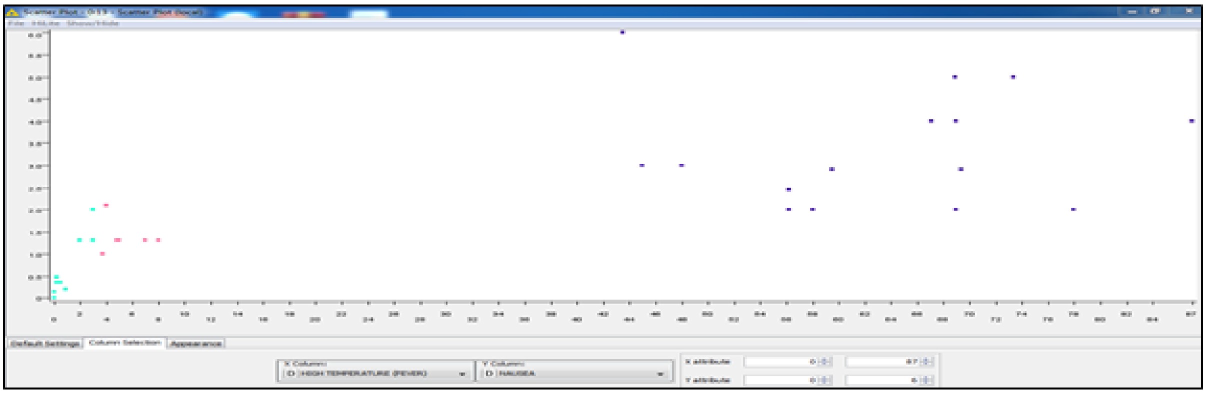


Figure 17: Scatter Plot generated by k-Means clustering on attributes (High temperature and Nausea).

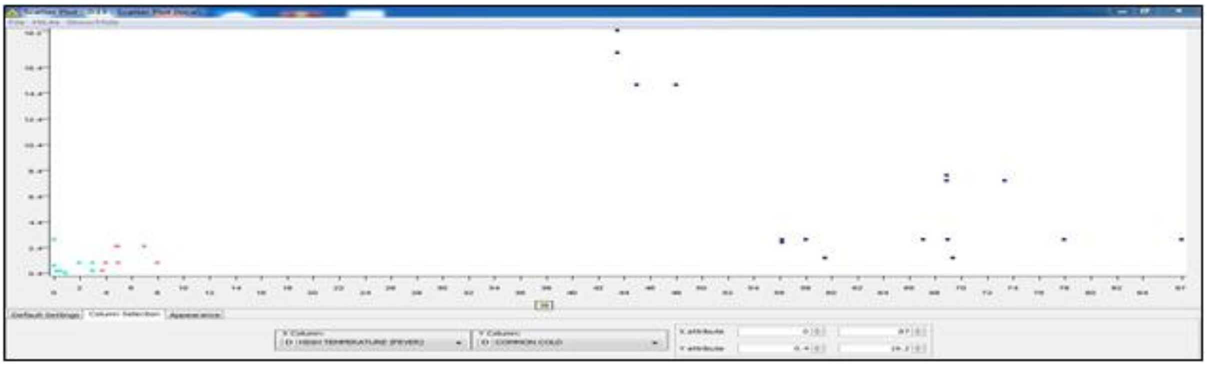


Figure 18: Scatter Plot generated by k-Means clustering on attributes (High temperature and Cold).

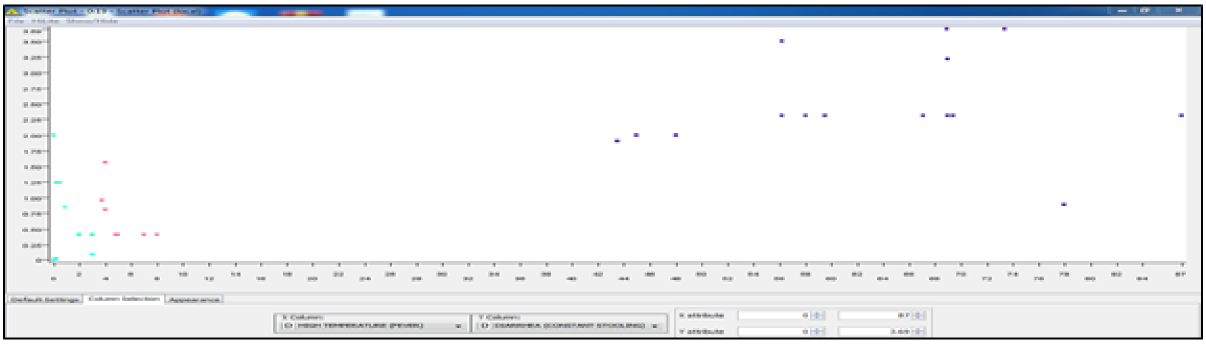


Figure 19: Scatter Plot generated by k-Means clustering on attributes (High temperature and Diarrhea).

3.4 Interactive Table And Histogram Fuzzy C-Means Clustering.

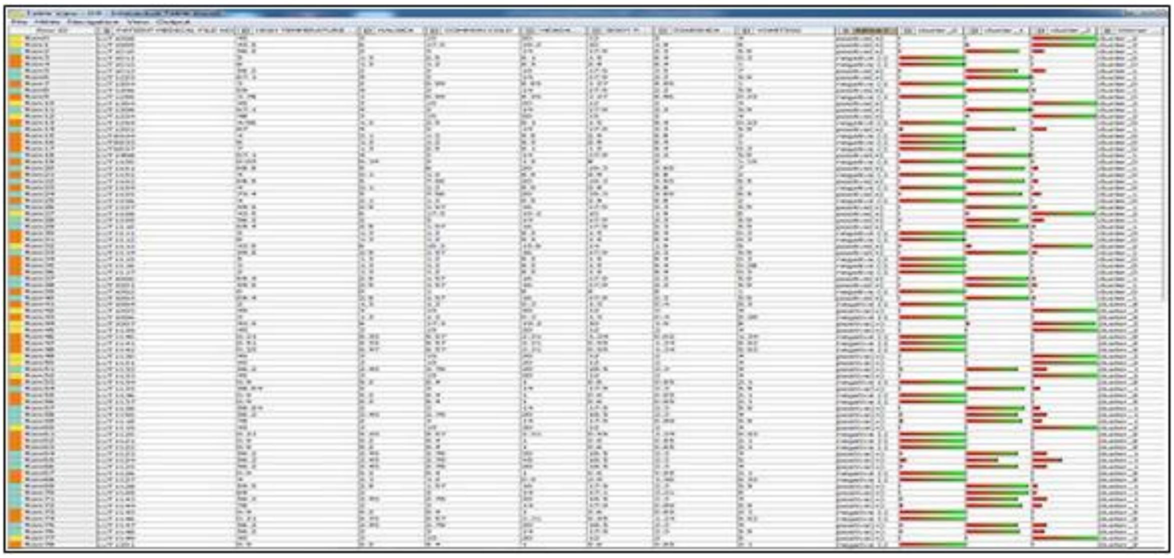


Figure 20: Interactive Table generated by Fuzzy C-Means clustering on all attributes.

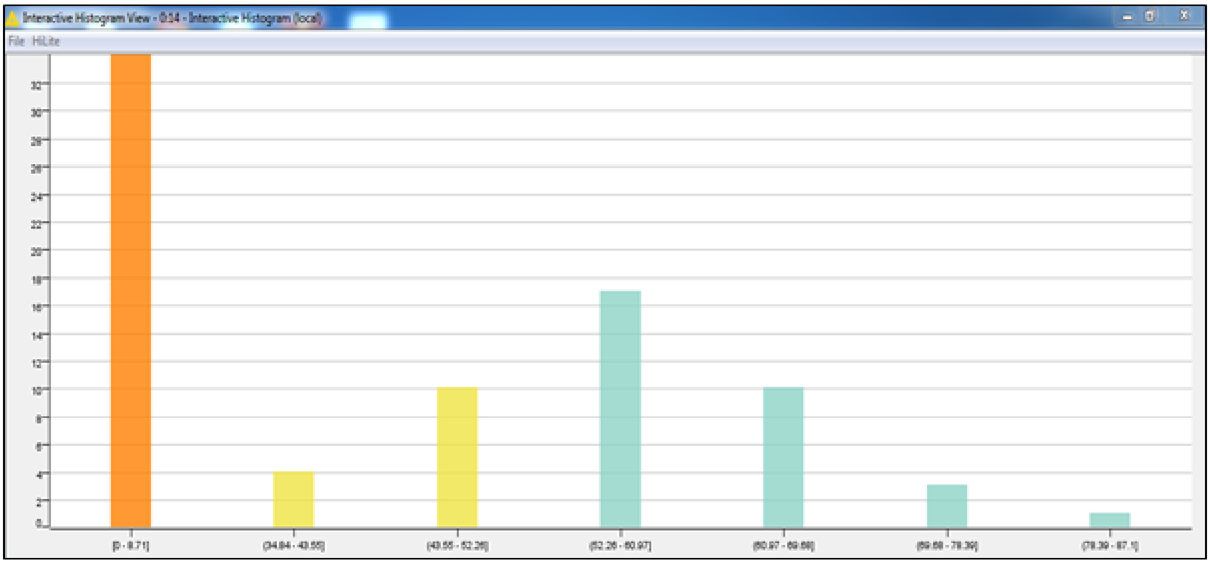


Figure 21: Interactive Histogram View of Fuzzy C-Means clustering.

3.5 Interactive Table and Histogram K-Means Clustering.

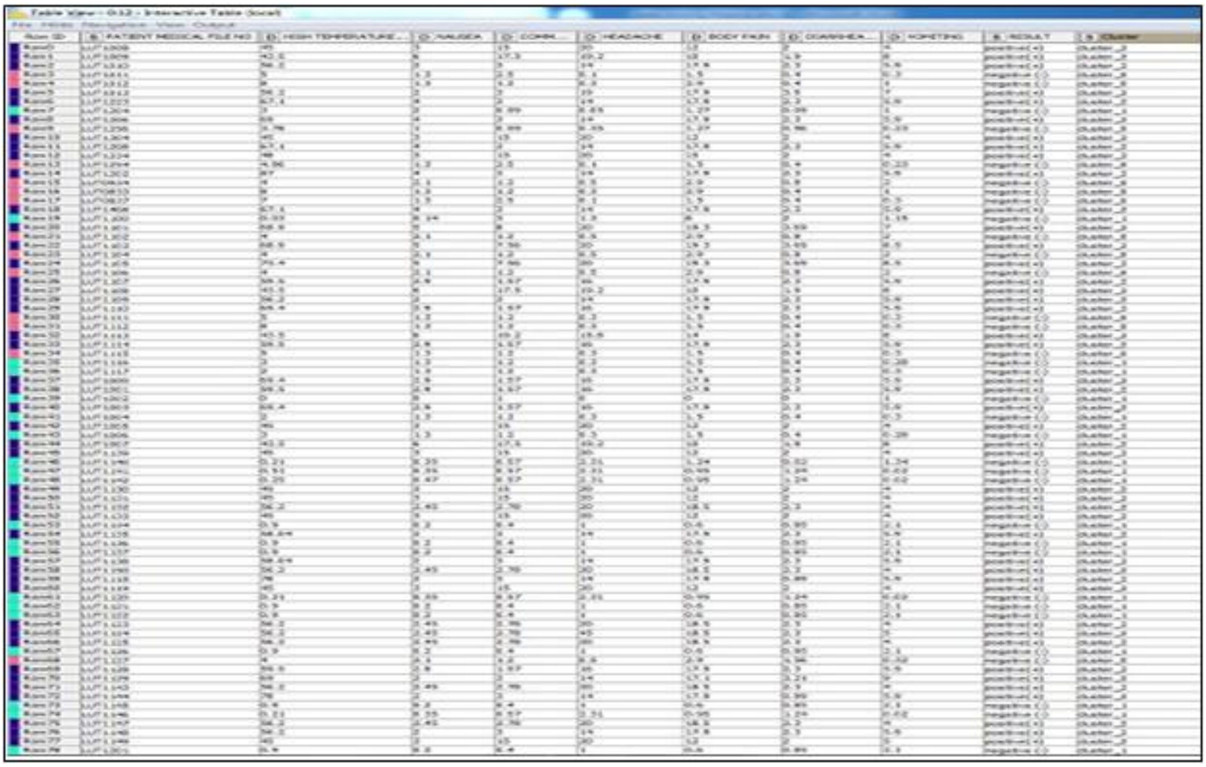


Figure 22: Interactive Table generated by K-Means clustering on all attributes.

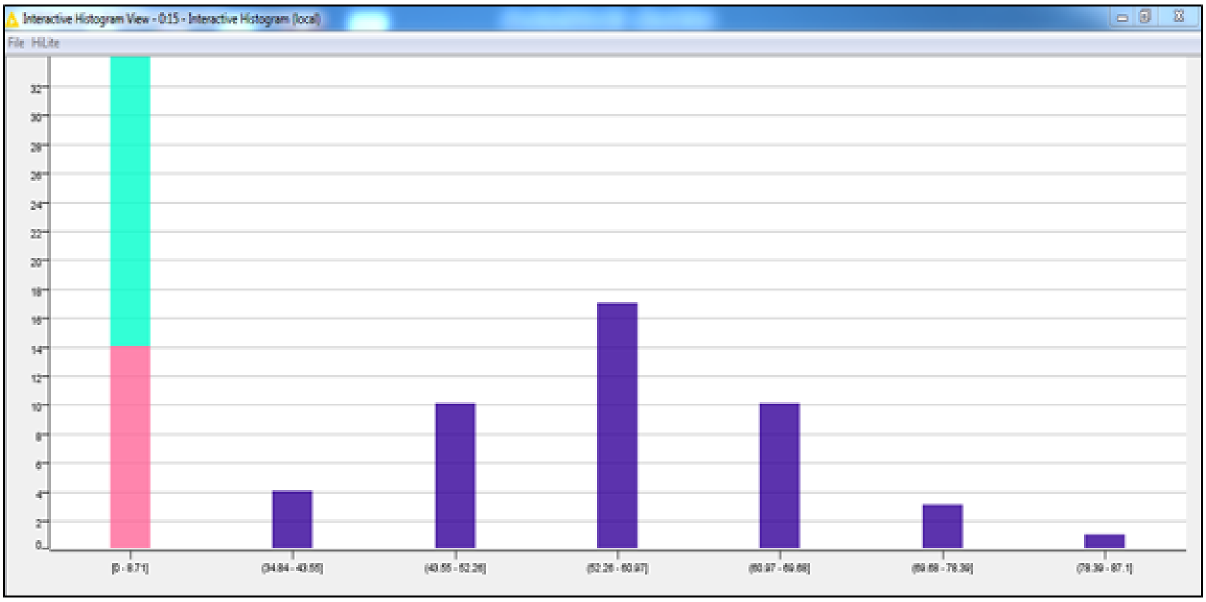


Figure 23: Interactive Histogram View of K-Means clustering.

3.6 Interactive Table and Histogram Hierarchical Clustering.

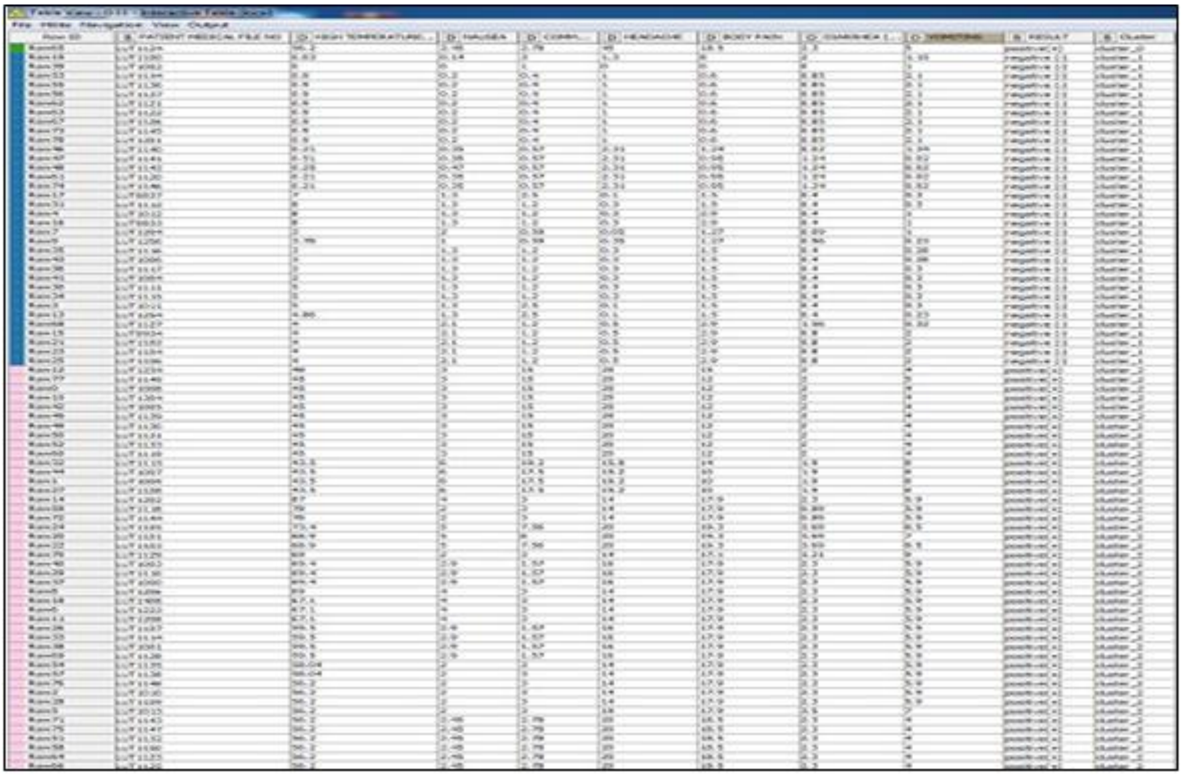


Figure 24: Interactive Table generated by Hierarchical clustering on all attributes.

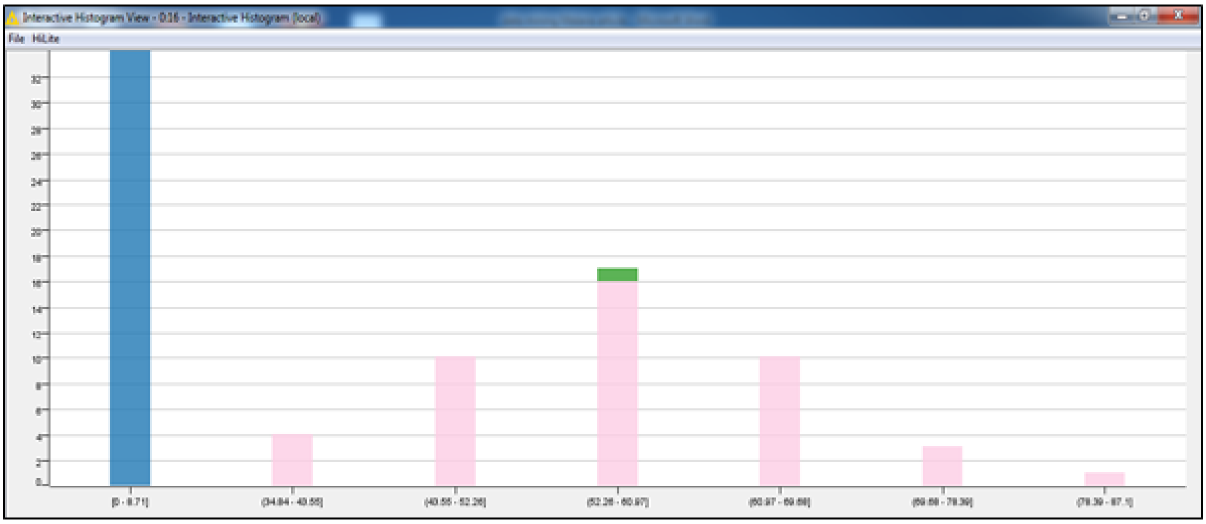


Figure 25: Interactive Histogram generated by Hierarchical clustering.

CONCLUSION

Fuzzy c-means, hierarchical and k-Means clustering approach were implemented to analyze the malaria data (training set). The clustering analysis produced different results respectively. We deducted that the Xie-Beni (XB) index value produced by the fuzzy c-means clustering is 0.27220555023892895. This denotes that the separation between the cluster's distances to the cluster's centre is a minimal range and compactness between the different clusters is low. The hierarchical clustering on malaria results produced a dendrogram (divisive/top-down tree-like structure), and computed the distances between each clusters. The k-means clustering computed the coverage for cluster_0, cluster_1 and cluster_2 as values of 14, 20 and 45 as shown in figure We could conclude that the cluster_2 with the highest value (45) shows that patients within the attributes ranges tends to have a positive trace of malaria symptoms while cluster_0 and cluster_1 with values of 14 and 20 tends to read negative trace of malaria symptoms. The results can help the medical sectors to predict future occurrences of malaria in most countries.

REFERENCES

1. Dr Margaret Chan, "*WHO Global Malaria Programme, World Malaria Report",* WHO Press, Geneva, Switzerland, *2013;* 1-286.
2. David Tordrup, Jakod Virenfeldt et al, *"Variant Plasmodium ovale isolated from patient infected in Ghana" Malaria Journal BioMed Central*, January, 2011; 1-5.
3. National Health Service (NHS), "*Causes of Malaria",* Department of Health, Website: https://www.nhs.uk/conditions/malaria/causes/.
4. The World Health Report, "*Rolling Back Malaria (The Challenge Of Malaria)"*, 1999; 49-63.
5. Al-Radaideh Q.A & Al Nagi E, *"Using Data Mining Techniques to build a classification model for predicting employee's performance."* International Journal of Advanced Computer Science and Applications, 2012; 3(2): 144-151.
6. Alkadhwi Ali Hussein Oleiwi and Adelaja Oluwaseun Adebayo, *"Data Mining Application Using Clustering Techniques (K-Means Algorithm) In The Analysis Of Student's Result"* ISSN: 2458-925X, May-2019; 5(5): 2587-2593.
7. Alkadhwi Ali Hussein Oleiwi, Adelaja Oluwaseun Adebayo and Ali Alkattan Hussein, *"Data Mining using Hierarchical Clustering Techniques on the position of employees in an information technology firm.".* Global Scientific Journals, June-2019; 7(6): 429-437.
8. Ming Chaun Hung and Don Lin Yang, *"An Efficient Fuzzy C-means Clustering Algorithm",* IEEE, 2001; 225-232.
9. Kiruthika Devi. S, Krishnapriya.S and Dristipona Kalita, *"Prediction of Heart Disease using Data Mining Techniques"*, Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016; 1-5.
10. DavinderKaur, Rajeev Bedi and S.K Gupta, *"Review of decision tree data mining*

algorithms: ID3 and C4.5", July 2015; 5-8.

11. Lior Rokach and Oded Maimon, *"Decision Trees",* Data mining and Knowledge discovery handbook, 166-192.
12. Adelaja Oluwaseun Adebayo and Mani Shanker Chaubey, *"Data Mining Classification Techniques on the Analysis of Student's Performance",* Global Scientific Journals, Vol 7, Issue 4, ISSN 2320-9186, April-2019; 79-95.
13. Open for Innovation KNIME, *"Fuzzy c-Means Clustering",* KNIME AG, Technoparkstrasse 18005 Zurich, Switzerland.
14. Yong. Y, Chongxun.Z, and Pan.L, *"A novel Fuzzy C-Means Clustering Algorithm for image thresholding",* Measurement Science Review, 2004; 4(1).
15. Rui. A and Sousa. J. M. C, *"Comparison of fuzzy clustering algorithms for Classification"*, International Symposium on Evolving Fuzzy Systems, 2006; 112-117.
16. K-Means clustering algorithm, Website: http://kom.aau.dk/group/04gr742/pdf/ kmeans_worksh eet.pdf.
17. Chen Ke, *"Machine Learning K-means clustering"*, University of Manchester, 1-22.