APPLICATION OF k- NEAREST NEIGHBOUR CLASSIFICATION IN MEDICAL DATA MINING IN THE CONTEXT OF KENYA

H. S. Khamis, K. W. Cheruiyot and S. Kimani
Institute of Computer Science and Information Technology, Jomo Kenyatta University of Technology, Nairobi, Kenya
E- mail: hassan@tum.ac.ke

Abstract
Medical data is an ever-growing source of information from hospitals in form of patient records. When mined, the information hidden in these records is a huge resource bank for medical research. This data contains hidden patterns and relationships, which can lead to better diagnosis. Unfortunately, discovery of these patterns and relationships often goes unexploited. Studies have been carried out in medical diagnosis to predict heart diseases, lungs diseases, and various tumors based on the past data collected from patients. However, they are mostly limited to domain-specific systems that predict diseases restricted to their area of operations. In retrospect, the performance of the k-nearest neighborhoods (k-NN) classifier is highly dependent on the distance metric used to identify the k nearest neighbors of the query points. The standard Euclidean distance is commonly used in practice. This study uses vast storage of information so that diagnosis based on historical data can be made. It focuses on computing the probability of occurrence of a particular ailment by using a unique algorithm. This k-NN algorithm increases the accuracy of such diagnosis. The algorithm can be used to enhance the automated diagnoses, which include diagnosis of multiple diseases showing similar symptoms. To validate the experimental results, a hypothesis was tested for the following variables: accidents, age, allergies, blood pressure, smoking habit, total cholesterol, diabetes and hypertension, family history of heart disease, obesity, and lack of physical activity. It was evident that there was a strong relationship between the above variables to the causes of common chronic diseases like: heart ailment, diabetes and cancer.

Key words: k-NN, classification, algorithm

1.0 Introduction

Medical professionals store significant amounts of patients' data that could be used to extract useful knowledge. Researchers have been investigating the use of statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of patients seeking emergency treatment. Statistical analysis has identified the risk factors associated with patients seeking emergency treatment to be accidents, age, allergies blood pressure, smoking habit, total cholesterol, diabetes, hypertension, heart disease, obesity and lack of physical activity. Knowledge of the risk factors associated with patients seeking emergency treatment can help medical professionals to identify patients at high risk of death.

Researchers have been applying different data mining techniques such as decision tree, naïve Bayes, neural network, bagging, kernel density, and support vector machines over different patients seeking emergency treatment datasets to help the medical professionals in the diagnosis. The results of the different data mining research cannot be compared because they have used different datasets. However, over time a benchmark data set has arisen in the literature: the Cleveland Heart Disease Dataset (CHDD). Results of trials on this dataset do allow for comparison.

Studies by the WHO (2009) indicate that over 75% of the patients that are taken to medical facilities who seek treatment succumb due to lack of information about their medical history or poor technique on how to mine the past medical data history. Accessing such information can sometimes prove to be futile since the patient maybe unconscious or may not be in a position to talk at such a time. In order to address this and reduce the mortality rate among patients seeking emergency treatment, there is need to develop a system that allows qualified medical doctors to update patient's records during any visit to a medical facility.

This project seeks to address the problem through a k-nearest neighbour classifier that eases data mining. It contains complete, up-to-date individual medical history records and close family members. This would be helpful especially when such patients need urgent medical attention yet they cannot talk. The individual as well as any

authorized physician registered with the Kenya Medical Practitioners and Dentists (KMPD) can view these records at any time and from any location that provides Internet access. The system aims at eliminating the monotonous and time consuming task of filling out numerous medical forms while visiting new physicians or new hospitals. Particularly, in the case of emergencies, where a patient might be unconscious, it is valuable to have his or her basic medical history available. An important feature of the system is that it captures all the medical history information filled in manually by a patient or physician in a normal doctor's office. It enables users to view their complete medical history, Emergency Medical Data, profile information, emergency contacts, and lists of allergies, current prescriptions, blood group, close relative blood groups and recent hospital visits. This medical expert system can also run as a standalone system.

2.0 Objectives of the Study

The objectives of the study are:

i. To evaluate to what extent k-Nearest-Neighbour classifier enhance efficiency and accuracy amongst patients seeking emergency treatment in Kenya.
ii. To evaluate the factors affecting the implementation of k-Nearest-Neighbour mining technique in Kenyan hospitals.
iii. Design a repository with efficiently classified data for easy data mining.

2.1 k-Nearest Neighbour (kNN)

The *k*-nearest neighbours algorithm is one of the simplest machine learning algorithms. It is simply based on the idea that "objects that are 'near' each other will also have similar characteristics. Thus if you know the characteristic features of one of the objects, you can also predict it for its nearest neighbour." k-NN is an improvisation over the nearest neighbour technique. It is based on the idea that any new instance can be classified by the majority vote of its 'k' neighbours, - where *k* is a positive integer, usually a small number.

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance. A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. For example, in patients seeking heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. So the influence of the cholesterol measure will be higher than the age. To overcome this problem the continuous attributes are normalized so that they have the same influence on the distance measure between instances (Haines *et al*., 2010).

KNN usually deals with continuous attributes however it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances a2, b2 are different so the difference between them is equal to one otherwise it is equal to zero. A study Shouman *et al.,* (2012) shows the sensitivity, specificity, and accuracy results of KNN in the diagnosis of heart disease patients. The value of K ranged between one and thirteen. The accuracy achieved ranged between 94% and 97.4 % with different values of K. The value of K equal to 7 achieved the highest accuracy and specificity (97.4% and 99% respectively).

In another separate study by Shouman *et al.,* (2012) indicate that k-Nearest-Neighbour is one of the most widely used data mining techniques in classification problems. Its simplicity and relatively high convergence speed make it a popular choice. However a main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. When the sample is large, response time on a sequential computer is also large. Despite the memory requirement issue, it is showing good performance in classification problems of various datasets. Dividing the training data into smaller subsets and building a model for each subset then applying voting to classify testing data can enhance the classifier's performance.

2.2 Data Repositories

Medicine as a science, by definition, is a data gathering, experiment- or study-focused activity. Reported studies of the scholarly information gathering and usage behavior of scientists have provided information about how they access and use the professional and research literature (Davis, 2004). There has also been some work examining

their use of data repositories. A study of molecular biology graduate students (Bay, 2005) found that they used bioinformatics resources extensively, although they learned about these resources from laboratory colleagues rather than at the library. A 2003 bibliographic analysis study using SciFinder Scholar showed increased use of the GenBank data repository in the science journal literature at that time (Blaine, 2003).

A recent web site survey study of local institutional repositories at the top 11 biology graduate programs in the United States uncovered 93 biology-related repositories, suggesting the continuing importance of data archiving and sharing to the life sciences (Bay, 2005). Another web site survey (Marcial & Hemminger 2010) of 100 science data repositories (including several health, biological, and environmental science repositories) also discussed the growth in the number and importance of these resources. This study focused on the general structural characteristics of all types of data repositories, including data deposition and access, metadata, grant and contract support, sponsorship, and preservation policies, and used these observations to provide recommendations for the development and sustainability of science data repositories.

Patient registries have received significant attention in recent years and are being used for a variety of purposes. The concept of a repository of expired registries raises several questions related to the feasibility, value, and potential cost of such a repository, as well as issues related to research ethics, governance, data access and use, patient privacy, technical requirements, legal considerations, and incentives for donating data to the repository (Gliklich, 2012).

2.3 Data Mining

According to Brunassi et al. (2008) Most of the operations and activities of the public and private institutions are computationally registered and accumulate in large databases, the data mining technique - *Data Mining* (DM) - is one of the most effective alternatives to extract knowledge from the great volume of data, discovering hidden relationships, patterns and generating rules to predict and correlate data, that can help the institutions in faster decision-making or, even reach a bigger degree of confidence. Data mining means searching for certain patterns within large sets of data, which creates a lot of possibilities for business managers and decision makers.

By analyzing those patterns, better business decisions can be made in order to enable businesses to achieve greater financial and entrepreneurial success (Goldschmidt and Passos, 2005). Nowadays, information and knowledge are legal, strategic and indispensable prerogatives in search for greater autonomy in the actions of the health companies, social control and decision-making with time getting shorter and shorter. Because of this, several national and international companies of production, consumption, financial market, teaching institutions and libraries have already adopted in their routines, data mining to monitor funding, client consumption, prevent fraud and foreseeing market risks, among others. In the health sector, mainly the public one, the application is being accepted as a way of accelerating the search for knowledge. Besides, the use of data mining in the big hospital databases or even in the information systems of public health contributes to discover relationships so that they can make a prevision of future tendencies based on the past, best characterizes the patient that seeks for assistance, identifies successful medical therapies for different diseases and shows patterns of new injuries (Cardoso and Machado, 2008).

However, several managers and health professionals are concerned with the understanding of the data and in using the information and knowledge of the health databases to promote the information management and the quality of care. This probably occurs due to the fast rhythm of data generation, which produces a natural incapacity in the human being to explore, extract and interpret these data to obtain knowledge of these bases (Brunassi *et al.,* 2008).

In this sense, the informatics and the technologies directed to the collection, storage and data availability has been developing and making available techniques, methods and automatic computational tools, capable of helping in the extraction of useful information inside this great volume of complex data (Goldschmidt and Passos, 2005).

However, to attend this new context, the health informatics has been using these methodologies of computing science to accomplish its studies. Among them, the methodology *Knowledge Discovery in Databases* (KDD), that is,

discovery of the databases knowledge, and the data mining, which is one of the most important stages of KDD (Cardoso and Machado, 2008).

As the theme is "pulverized" in the most diverse areas of knowledge, this article, aimed at presenting a literature review of the main indexed databases and some books published on the subject, thus presenting the use of the technique of data mining, concepts, tasks and methods(Brunassi *et al*., 2008).

2.4 Data Modeling for Emergency Records

In 1900s medical professionals used patient records to study and learn from their patients and from the medical records of their patients, in order to improve their knowledge of diseases. In the 2000s, as in the 1900s, physicians continue to initiate this learning process by taking a history of the patient's medical problems, performing a physical examination of the patient, and then recording the history and physical examination findings in the patient's medical record. To confirm a preliminary diagnosis and to rule-out other possible diagnoses, physicians refer the patients for selected tests and procedures that usually involve the clinical laboratory, radiology, and other clinical-support services. After reviewing the information received from these services, physicians usually arrive at a more certain diagnosis, and then prescribe appropriate treatment. For an unusual or a complex medical problem, physicians may refer the patient to appropriate medical specialists, and may also review evidence-based reports of appropriate therapies by consulting relevant medical literature and bibliographic databases.

2.5 Origin of Medical Database

Lindberg (1979) described the degrees of difficulty in the development of medical innovations in the grades of their complexity: (1) the easiest was the automation of a simple function such as providing a patient's billing for services; (2) more difficult was the automation of a more complex function such as collecting and storing a patient's medical history; (3) very difficult was constructing a very complex function such as a medical database; and (4) the most difficult was developing the highly complex medical information and database-management system for a hospital, as Starr (1982) had aptly ranked the hospital to be the most complex organizational structure created by man.

Databases were defined by Frawley *et al.* (1992) as logically integrated collections of data in one or more computer files, and organized to facilitate the efficient storage, change, query, and retrieval of contained relevant information to meet the needs of its users. Frawley estimated that the amount of information generated in the world doubled every 20 months, and that the size and number of computer databases increased even faster. Medical repositories was the term proposed by Johnson (1996) as more accurately representing a shared resource of patient data that was collected for the purpose of supporting medical care. Johnson advised that a large scale, medical repository required a data model to define its functional requirements and to produce a formal description, a conceptual schema of all the data generated in the enterprise and how it was all related, and a database structural design to define its technical requirements. Since a medical database usually operated within a medical database-management system, the database needed to be compatible with the information system of the enterprise of which it was a part; and it also needed to be operationally and structurally independent of all subsystems and applications programs. The evolution, design, implementation, and management of computer-stored databases were described in some detail by Connolly and Begg (1999), Collen (1986, 1990, 1994, 1995); and also by Coltri (2006) who considered computer-stored databases to be one of the most important developments in software engineering.

Database-management systems soon replaced the earlier file-based systems that often stored the same data in multiple files, and where it could be more difficult to retrieve and coordinate a patient's data. A database-management system was defined by Blum (1986) as software consisting of a collection of procedures and programs with the requirements for: entering, storing, retrieving, organizing, updating, and manipulating all of the data within its database; managing the utilization and maintenance of the database; including a meta database to define application- specific views of the database; entering data only once, even though the same data might be stored in other subsystems; retrieving, transferring, and communicating needed data in a usable format, and having the ability to create inverted files indexed by key terms; maintaining the integrity, security, and required level of confidentiality of its patients' data; and fulfilling all management, legal, accounting, and economic requirements.

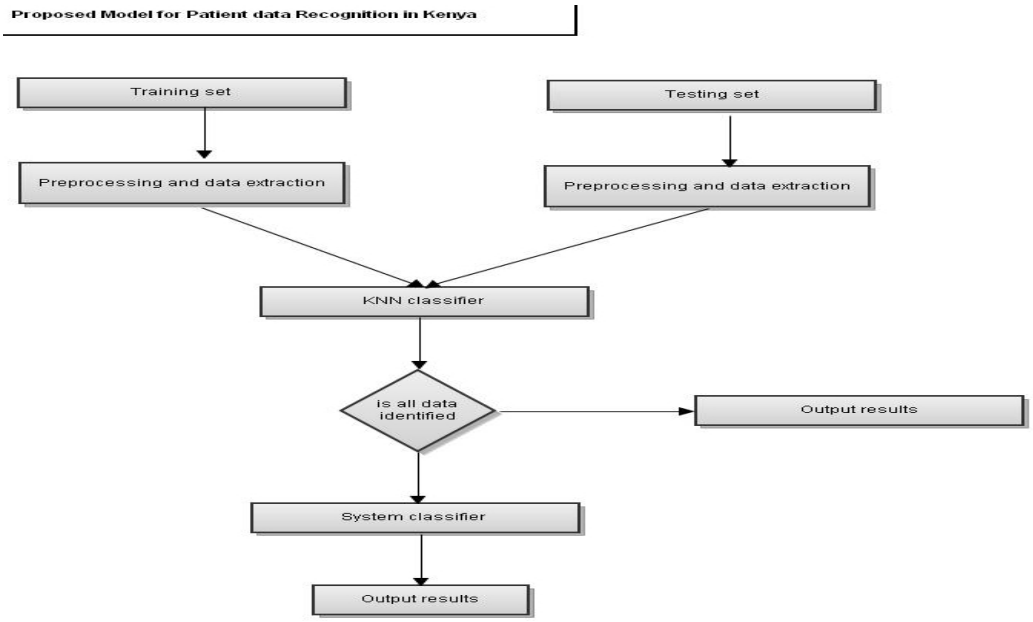
2.6 Emergency Data Modeling Architectures

Data-modeling designs to provide the conceptual schema that represented the information in clinical repositories were advocated by Johnson (1996) to be as important for large medical databases as were their structural designs. He defined the conceptual schema for patient care as a representation of all of the data types required to manage the health-care process, whether using a hierarchical, a relational, or an object-oriented structural database design, or a combination of database structural designs. He advised that the structural design of a database needed to be able to provide rapid retrieval of data for individual patients and to have the capability to adapt to changing information needs of growth and new technology; yet he emphasized that the primary purpose of the database structural design was to implement the conceptual schema. To properly build a database, Johnson (1996) proposed that it was necessary to first develop a model of the database that defined its functional requirements, its technical requirements, and its structural design.

The database model needed to produce a formal description, a conceptual schema of all the data generated in the enterprise, and how all of the data were related. Thus the users of a medical database needed to define its functional requirements as to exactly what they wanted the database and its database-management system to do. Since a medical database usually operated within a larger medical-information system, the functional requirements of the medical database needed to be compatible with those of the medical enterprise of which it was a part. Whether a medical database served as the primary electronic medical record (EMR), or served as a secondary medical database, such as a clinical research database with its data derived from the EMR, both had some similar basic functional requirements. Davis and Terdiman (1974) recommended that as a minimum, the major goals of a medical database should be to maintain readily accessible all of the relevant data for each patient served, and to provide a resource for the systematic retrieval of all relevant data from all patients' records for any desired primary purpose, or for a secondary administrative or a research purpose.

2.7 Conceptual Framework for Data Recognition

The proposed k-Nearest Neighbour model is given in Figure 5. In the proposed method, the given data is pre-processed to extract all the metadata. KNN is used to find the closest neighbors of the given data with all the available training data. If a label is found then the algorithm quits, otherwise the system classifier is applied. The proposed algorithm was used to recognize the object. The results are compared to those obtained with single system classifier and KNN.



Source: UML designed

Figure I: KNN data recognition model for Kenyan patients

2.0 Methodology

Coast province has one provincial referral hospital (Coast general hospital) and seven district hospitals: Kilifi district hospital, Kwale district hospital, Lamu district hospital, Malindi district hospital, Mombasa district hospital, Taita Taveta district hospital and Tana River district hospital.

Part A of the questionnaire will be concerned with collecting respondent's demographic data. Part B of the questionnaire will be concerned with evaluating the extent to which k-Nearest-Neighbour classifier enhance service delivery to patients seeking emergency treatment in Kenya. Whereas part C of the questionnaire will be concern with evaluating the factors affecting the implementation of k-nearest neighbor mining technique in Kenyan hospitals and any other services that me be available.

Quantitative data collected from part A of the questionnaire such as the demographic characteristics of the respondents was analyzed using descriptive measures that is mean, standard deviation and variance. The data will be then presented using frequency distribution tables. Data collected from part B and C was analyzed using the same method.

3.0 Results

3.1 Gender of the Respondents

From the data presented in Figure II, it is evident that majority of the respondents were male who represented 68% of the respondents.

Various studies have shown that female are more averse to adopting the use of technology compared to male and this could have effects on the finals result as it could indicate biasness.

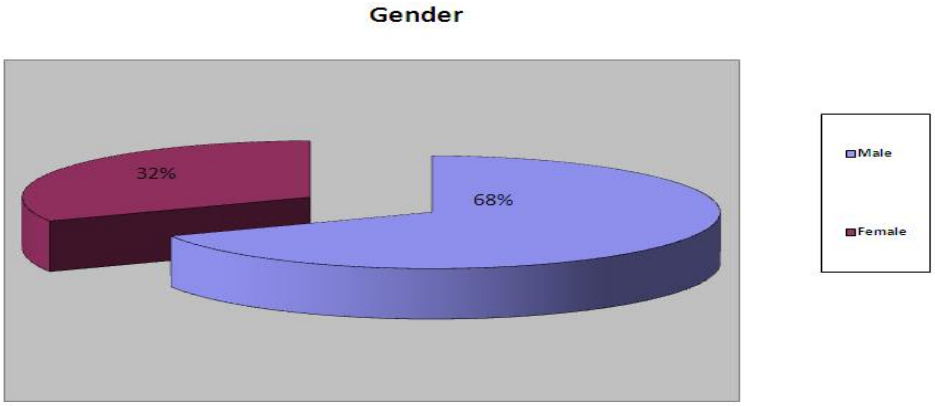


Figure 2: Gender

3.2 Working Experience

Figure 3 shows that majority of respondents are young and inexperienced. According to WHO (1997) report has adopted a model to how rate medical doctor experience – 25 years. Most of the doctors who are now independent owe to be under some experience medical doctor tutelage but due to inadequacy of personnel they are trusted to do work that owe to be done by more experienced practitioners.

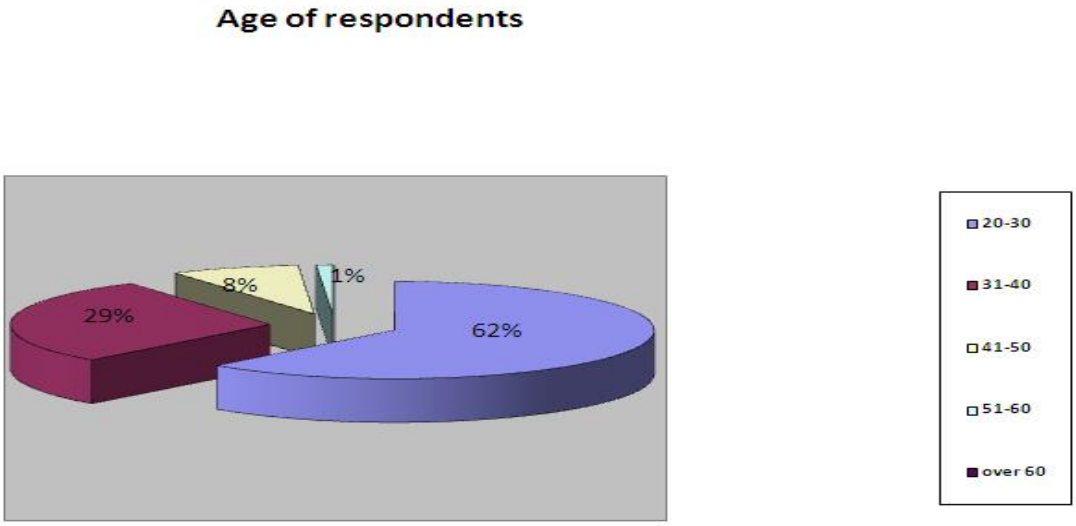


Figure 3: Age of respondents

3.3 Level of Education

Data in Figure IV show that majority of the respondents have the minimum qualifications of being a medical doctor – 89% with a few who have specialized. Most studies indicate that medical doctor performance is always in tandem with education qualification.

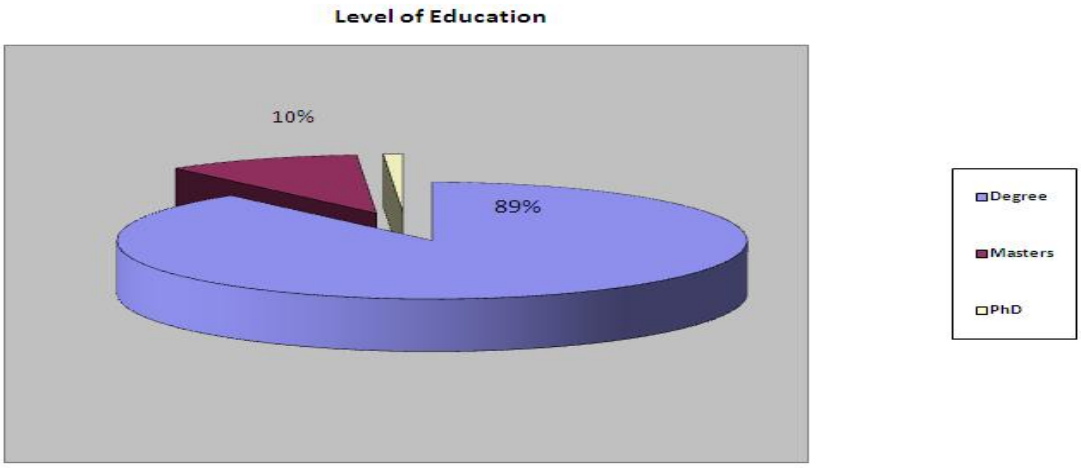


Figure 4: Level of education

3.4 Age of the Hospital

Data in Table 1 shows that majority of the medical facilities under study are old enough to have established a system that could help them mine data. In addition given the age of the hospitals, it is expected that the medical facilities have adequate patient data to mine from.

Table 1: Level of education

Age of the Hospital		Below 10	11-20	21-30	Over 30
	Freq.	18	16	3	1
	%	47.3	42.1	7.8	2.6

3.5 Computer and Internet Usage

Graph V indicates that 7.1% of the respondents use computers and Internet to a small extent, 21.4% to a medium extent and 60.7% to a large extent. It is evident that almost 90% of the respondents use computers and Internet - knowledgeable about them. Since most application of k-mean is both ICT intensive, this shows that most of the employees may not need a lot of training when the organization adopts the k-mean hence saved time and costs.

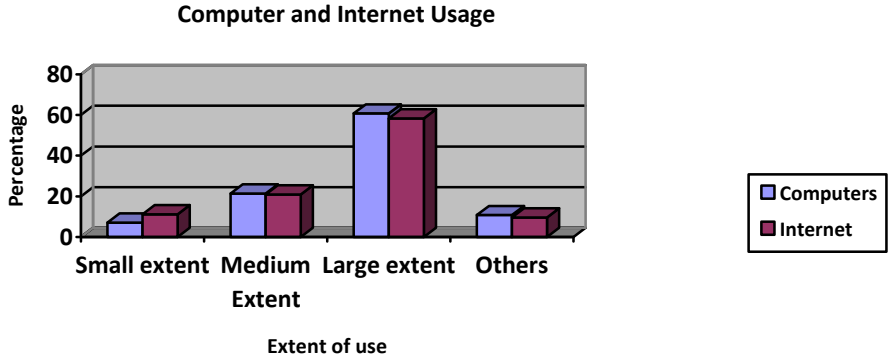


Figure 5: Computer and internet usage

3.6 Adoption of k-mean Services

This section addresses the first objective on the extent to which data mining has been adopted in medical facilities in Kenya. This was achieved through analyzing the extent to which some services offered by k-mean have been

adopted by doctors in Kenya. The results are in Table 2 below.

Table 2: The extent of k-mean services usage

No.	Services offered by mining system	Mean	StDev	Variance
7	Classification Accuracy	1.63	0.44	0.194
8	Single data classifier (kNN)	2.25	1.11	1.232
9	Efficiency in data accessing	1.75	0.65	0.423
	Overall mean	**1.88**	**0.73**	**0.616**

The extent of adoption was classified as: no extent– 5, very small extent - 4, small extent – 3, large extent - 2 and very large extent - 1. With a mean of 1.88, it means that the level of adoption of k-NN services within the medical facilities in Kenya is still very low.

Factors affecting the use of k-mean systems

This section is aimed at addressing the second objective of the research, that is, to establish the factors that determine the extent of k-mean services usage. The respondents were asked to indicate the extent to which they agree to various variables as affecting the usage of services offered by k-mean. The responses were classified as: strongly disagree – 5, disagree - 4, undecided – 3, agree - 2 and strongly agree – 1.

3.7 Memory Space and Availability of Data to be mined

Table III indicates that availability of data to mine had a mean of 2.05 which indicate that most respondent agree that it was a major reason for mining data. Given that mean and the standard deviation, it's obvious that there is a strong correlation between availability of data and adoption of mining data in medical facilities. On the other hand, memory space seems not to have a lot of effect in the determining adoption of a mining algorithm.

Table 3: Memory space and availability of data to be mined

Memory space and availability of data to be mined	Mean	StDev	Variance
k-nearest neighbor classifier is associated with large memory space	1.73	1.07	1.145
The hospital decision to adopt the use of k-*nearest neighbor classifier* was informed by availability of Data to be mined	2.34	0.75	0.563
Average	**2.05**	**0.854**	**0.91**

3.8 Administrative Costs

From the data presented in Table IV, the decision for the to adopt the use of data mining being determined by cot of its implementation, maintenance and training of staff had a mean of 2.97 and shows a strong correlation between the use of a data mining technique and administrative costs. It was also evident that most respondents believed that implementing a k-mean system for data mining could reduce time wastage in dealing with patients, reduce cost of employees whilst improving customer service.

Table 4: Administrative cost

Administrative Cost	Mean	StDev	Variance
It is expensive to develop, maintain and train staff	2.97	0.5	0.25
Average	**2.97**	**0.5**	**0.25**

Classification Efficiency

Table V shows that classification efficiency factor in determining data mining decision with a mean of 2.69. Speed of accessing information has a mean of 2.41 which shows that most respondents agree.

Table 5: Classification efficiency

Classification efficiency	Mean	StDev	Variance
K-*nearest neighbor classifier* ensures quick access to when mining a patient's data.	2.41	1.22	1.49
K-*nearest neighbor classifier* ensures fast access to related data when mining a patients data	2.97	0.92	0.85
Average	**2.69**	**1.07**	**1.17**

4.0 Conclusions

In light with the findings of the first objective, To evaluate to what extent k-Nearest-Neighbour classifier enhance efficiency and accuracy amongst patients seeking emergency treatment in Kenya, it was evident that application k-NN algorithm can greatly help reduce errors in diagnosis, reduce time spent on diagnosing whilst improving efficiency and effectiveness in treatment.

In light with the findings of the second objective: To evaluate the factors affecting the implementation of k-Nearest-Neighbour mining technique in Kenyan hospitals. It was evident that k-NN mining technique was mostly affected by administrative costs and classification efficiency.

5.0 Recommendations

Given the dynamic trends in k-NN classifying systems, it could be important to carry-out regular studies to ascertain the current position thereby increasing the potential acceptance. Although this study was more concerned with doctors, it would also be interesting to try and evaluate the systems efficacy in other contexts like different industries or countries. It would be interesting also to carry out similar research but using other data mining techniques like: K-means algorithm and Bayesian algorithm.

5.1 Limitations of the Study

The research suffers from the quality of the composition of the sample. By virtue of selecting a sample of 8 hospitals, this research was not sufficiently heterogeneous. The limited heterogeneity in respondents' characteristics could have affected both the nature and the extent of the predictor variables. The large scope of this research project and the complex multidisciplinary nature of the study was also a major challenge. In addition, this researcher did not consider all parties that could be the most common consumers of this system like administrators and patients.

Reference

Bay, S. D. (2005). Nearest neighbour classi_cation from multiple data representations. Master's thesis, University of Waterloo, Department of Systems Design Engineering.

Blaine GI. (2003). Networks and distributed systems. A primer. Proc MEDINFO. 1983:1118–21.

BRUNASSI, L.A. (2008). Development of a fuzzy system forestrous detection. 2008. 101f. Dissertation (Master in Agricultural Engineering) – Campinas University, Campinas, SP.

Coltri, A. (2006). Databases in health care, chap 11. **In**: Lehman HP, Abbott PA, Roderer NK, et al., editors. Aspects of electronic health record systems. 2nd ed. New York: Springer; 2006. p. 225.

Davis, L. S. (1975). Prototype for future computer medical records. Comput Biomed Res. **3**, pp 539–54.

Davis, L. S. (1973). A system approach to medical information. *Methods Inf Med.,* **12**: pp 1–6.

Davis, L. S. (1970). Prototype for future computer medical records. *Comput Biomed Res*., **3**, pp 539–54.

Gliklich, R. (2012). President, Quintiles Outcome, Harvard Medical School, MD.

Goldschmidt, R. and Passos, E. (2005). Data Mining – Um Guia Prático. Rio de Janeiro, editora Campus.

Johnson, S. B. (1996). Generic data modeling for clinical repositories. *J Am Med Inform Assoc*., **3**, pp 328–39,

Lindberg, D. A. B. (2005). The growth of medical information systems in the United States. Lexington: Lexington Books.

Marcial, L. H. and Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, **61**(10), pp 2029–2048.