UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A DECISION SUPPORT SYSTEM FOR THE DIAGNOSIS OF BREAST CANCER USING FUZZY LOGIC AND CASE BASED REASONING

BY

WANYONYI PETER SIMEON

REGISTRATION No. P54/66259/2013

SUPERVISOR: DR. CHRISTOPHER CHEPKEN

JULY 2016

A research project report submitted in partial fulfillment of the requirements of the Degree of Master of Science in Information Technology Management at the University of Nairobi.

DECLARATION

This project, as presented in this report, is my original work and has not been presented for any other award in any other University.

Student: WANYONYI PETER SIMEON Registration Number: P54/66259/2013

Signature: .... Date: ....

This project has been submitted as a partial fulfillment of requirements for the Masters in Information Technology Management of the University of Nairobi with my approval as the University supervisor.

Supervisor: DR. CHRISTOPHER CHEPKEN

Signature: .... Date: ....

ACKNOWLEDGEMENT

I express my uttermost gratitude to Jehovah God Almighty for everything has been done by His help and provision. He is my rock and fortress.

I am also thankful to my supervisor, Dr. Christopher Chepken, for his patience in leading me through the research process. I am deeply indebted to you for the help, suggestions and encouragement I received during the research and when writing this report.

To my wife, Rasoa, your support for this work is immeasurable.

To my son, Israel, you are such an inspiration, assuredly your will go further than this.

Finally, to my Mom, Rose Luseno you laid the foundation and I am forever grateful.

Contents

ABSTRACT

Background

The use of Clinical decision support systems helps improve medical diagnosis and also minimize diagnostic errors. Older diagnosis systems have proved cumbersome to use and avail limited success in identifying the correct diagnosis in complicated cases like breast cancer at early stages.

Objectives

The objectives are to design, develop, and assess a clinical decision support system that offers a suite of services for early detection of breast cancer.

Methods

The CDSS prototype was developed based on cased based reasoning and fuzzy logic artificial intelligence technologies. The functionalities of the CDSS were developed iteratively through requirement- development cycles using enterprise-grade software-engineering methodology. Within each cycle, the acquisition of clinical knowledge was done by a health informatics engineer and a team of oncologists. The research involved 50 case records at St. Francis Mission hospital, Kasarani whose final diagnosis had already been ascertained as breast cancer. The patient symptoms from the records were manually entered in to the system so as to determine how often the CDSS would suggest the correct diagnosis. In addition to this, the speed at which data entry could be done and results recovered were evaluated.

Results

The clinical decision support system suggested the correct diagnosis in 48 of the 50 cases (96%). Manual data entry took less than a minute while results were provided within 2–3 seconds.

Conclusions

The CDSS prototype suggested the correct diagnosis in almost all of these complex cases during testing and evaluation. The prototype therefore merits evaluation in more natural settings and clinical practice.

LIST OF TABLES

LIST OF FIGURES

Chapter One

INTRODUCTION

1.1 Background of the study

Cancer causes more deaths than HIV, TB and Malaria combined globally. However, 70% of the global Cancer burden is in Low and Middle Income economies (LMICs) like Kenya where it is the 3rd highest cause of morbidity after infectious and cardiovascular diseases. Breast cancer is the most prevalent cancer among Kenyan women, and constitutes a major public health problem. Breast cancer alone contributes to 23.3 % of cancer deaths in Kenya. Several factors can help predict an individual's risk of developing cancer and these include: weight (obesity), high-risk habits (smoking, heavy drinking), exposure to environmental pollutants, family history, age, menstruation, breast tissue, and exposure to previous chest radiation, exercise and continuous use of oral contraceptives. (Mutuma and Korir, 2006; WHO and IARC, 2008).

The epidemiology of breast cancer is complex and several risk factors have so far been established. These are majorly associated with family history, age, menstruation, breast tissue and exercise (Magoha, 2000). American Cancer Society shows more modifiable risk factors, which can be changed. These modifiable risks include obesity, workplace exposure and diet.

According to (KNCCS, 2011) the disease cannot be eradicated but its effects can be significantly reduced if effective measures are put in place to control risk factors, *detect cases early* and offer good care to those with the disease.

So fare Artificial intelligence algorithms have been used successfully recently to learn cancer patterns and provide early signal of likely infection. In this study, two AI tools were used i.e. case based reasoning and fuzzy logic to establish the likelihood of breast cancer given the symptoms. Policy makers, medical practitioners (physicians), and patients can use these patterns to provide essential input into the rational planning of cancer control programs.

Personalized and predictive medicine has been picking momentum in the recent past due to the use of knowledge management technologies in disease prediction and prognosis. This movement towards predictive medicine is important, not only for patients (in terms of lifestyle and quality-of-life decisions) but also for physicians (in making treatment decisions) and health economists and policy planners (in implementing large scale cancer prevention or cancer treatment policies) (Cruz, J. A. and Wishart, D. S, 2006).

1.2 Problem Definition

30% of cancers are curable if detected early; 30% of cancers are treatable with prolonged survival if detected early; 30% of cancer patients can be provided with adequate symptom management and palliative care (Cruz, J. A. and Wishart, D. S, 2006). An ideal situation would be a breast cancer free continent but the reality is that this disease is becoming common and leading to more deaths year after year especially among women. The vast medical data on breast cancer symptoms can be used to help detect the disease before it is too late.

Medical dictionary defines diagnosis in two ways; (1) as the determination of the nature of cause of a disease. (2) a concise technical description of the cause, nature, or manifestation of a condition, situation or problem? This process is the most important of all other health care processes as it guides the rest of the treatment process. Unfortunately, this process is facing challenges in the developing countries for it is done manually, thus it depends on the ability of the medical provider to remember the disease with which the symptoms match with (L. Stefano Nardini, Germano Bettoncelli, Vincenzo Lamberti and Patrizio Soverina, 2006). With the very many people who visit the health facilities every day, and they have to undergo the process, many patients end up being diagnosed with the disease they are not suffering from. Cases have been reported of confused patient diagnosis ending up with wrong patient's diagnosis identification. Either, the diagnostician may get overwhelmed by the many patients and decides to make their work easy by not doing all diagnoses required, and this may lead to misdiagnosis. This accounts for the many cases of people who visit hospitals, get treated but never get well. In lieu of this, an expert system that can create a level ground for all experts in the medical field is worth consideration (S.S., Smita, S., Sushil & M.S., Ali, 2013).

This research study proposes a clinical decision support system that can be used to diagnose breast cancer based on the report symptoms hence assisting the clinician to make an informed decision.

1.3 *Project Objective*

The main objective of this study was to design, develop, and test a CDSS prototype for the early detection of breast cancer.

Specific Objectives are:

i. Identify AI tools appropriate for developing CDSS for breast cancer.

ii. Develop a CDSS prototype for clinicians and for breast cancer patients making prognostic assessments, using the particular characteristics of the individual patient.

iii. Assess the performance of the CDSS by making it available to a hospital and using it to diagnose newly diagnosed breast cancer cases.

1.4 *Research Questions*

The following research questions were used to meet the above objectives;

i. Which AI tools are appropriate for developing a breast cancer CDSS?

ii. How a CDSS is developed using AI tools?

iii. How can the performance of the CDDS be measured?

1.5 *Justification*

The traditional method of medical diagnosis misses out on early diagnosis. This means patients are put on medication when the real disease has not been identified. These diseases are diagnosed at the later stages when curing them is almost impossible or very expensive (Intrahealth International, 2012). The late diagnosis especially of breast cancer necessitated this research.

1.6 Scope of the Study

This study focused on cases of breast cancer that have been reported or diagnosed at the St. Francis Community Hospital, Kasarani in Nairobi County.

Chapter Two

LITERATURE REVIEW

Research organizations and companies have gone a long way in research and development of clinical diagnosis support systems. However, most of these systems have not been very successful in developing countries. The failure could be attributed to various reasons but majorly developers' lack deep understanding of the medical field and the diagnosis process (C. Abouzahr and T. Boerma, 2010).

The diagnostic decisions taken by medical experts depend upon familiarity, experience, expertise, knowledge, capability and perception of the medical scientist. As the complexity of the system increases, it is not easy to follow a particular path of diagnosis without any mistake.

AI tools have been used to imitate the operations of medical doctors, in this research, Fuzzy logic and case based reasoning are used in developing a prototype that can imitate a doctor`s operation. Fuzzy logic and case based reasoning presents powerful reasoning methods that can handle uncertainties and imprecision. An aggregation of the knowledge, observation and experience of medical experts serves as the backbone of a fuzzy models based medical diagnostic system (Rana & Sedamkar, 2013).

2.1 Introduction to breast cancer

Breast cancer usually appears in the ducts that transport milk to the nipple and the lobules of glands that produce milk. Breast cancer is one of the most widespread cancers (West V. Ensemble et al 2005). According to figures from the Kenya National Cancer Institute, one woman in eight will be afflicted with breast cancer in her life. Breast cancer recognized at an early stage can usually be treated (Sherring and Varsha, 2009). Therefore, patients suspicious of breast cancer should seek medical attention at the earliest time possible. In this way the treatment outcome will usually be more favorable for the patient and the healthcare service (Sherman DW, Haber J, Noll Hoskins C, Budin WC, Maislin G, ShuklaS, et al. 2012). It is significant to note that due to improved detection methods, the number of cancer patients has increased; nevertheless the mortality rate has fallen. Sudarshan in his study shows that early detection can lead to an 85% chance for survival compared to 10% for late detection (Sudarshan, 2001). Breast cancer is the most commonly diagnosed cancer in females accounting for about 30% of cancers.

It also occurs in males; however, the frequency is significantly lower in the latter group (Baider and Andritsch 2004). According to National Cancer Institute in the US, during 2003‑2007, the average mortal age in LMICs for breast cancer was 68 years. The number of deaths due to breast cancer varies with different age brackets as shown in table 1 below;

Table 1. Approximate percentages of deaths due to breast cancer according to age group.

Age years	Percentage breast cancer deaths (%)
< 20	0.0
20 - 34	0.9
35 - 44	6.0
45 - 54	15.0
55 - 64	20.8
65 - 74	19.7
75 - 84	22.6
>85	15.1

Note: adapted from the American Cancer Society. (2009, November 9). Breast Cancer. Atlanta, GA: American Cancer Society.

Table 2. Approximate breast cancer percentages diagnosis according to age group

Age in years	Percentage diagnosis
<20	0
20 - 34	1.9
35 - 44	10.2
45 - 54	22.6
55 - 64	24.4
65 - 74	19.7
75 - 84	15.5
>85	5.6

Note: adapted from the American Cancer Society. (2009, November 9). Breast Cancer. Atlanta, GA: American Cancer Society.

2.2 Breast cancer Risk Factors

Unlike many diseases, breast cancer does not have a single cause. Instead, it may result from the interaction of multiple factors that range from genetic characteristics to personal lifestyle. The term risk factor is used to refer to anything that is associated with an increased chance of developing breast cancer (Quinn M and Babb P, 2002).

Risk factors are a matter of probability. They influence an individual`s odd of developing the disease. That`s not the same thing as actually causing the disease to occur. According to Kathleen and Morgan, breast cancer risk factors have been grouped in to three: (i) Established risk factors - For this class of risk factors, there is clear scientific evidence linking the factors with breast cancer risk. These risk factors include Female gender, age, previous breast cancer, Benign breast disease, family history of breast cancer, Early age at menarche, late age at menopause, late age at first full-term pregnancy, obesity(postmenopausal), High-dose exposure to ionizing radiation early in life. (ii)Speculated Risk factors - Speculated risk factors for breast cancer are those for which there is some scientific support but not enough to be considered conclusive. Speculated risk factors include never having been pregnant, having only one pregnancy rather than many, not breast feeding after pregnancy, use of oral contraceptives, alcohol consumption, tobacco smoking, breast augmentation, low intake of phytoestrogens. (iii) Unsupported Risk factors - are primarily myths or misconceptions rather than true risk factors, this include Obesity (premenopausal), exposure to low-dose ionizing radiation in midlife, high intake of phytoestrogens, large breast size, antiperspirants(Kathleen and Morgan, 200).

2.3 Breast Cancer Diagnosis

Early detection is the key factor for successful breast cancer treatment. Diagnostic techniques such as clinical examination, ultrasound, mammography, magnetic resonance and thermography are employed to detect and provide an accurate diagnosis of breast cancer (Araujo 2009). Mammography is considered the most favored test; (Acharya 2005) nevertheless, it is not effective for fibrocystic diagnosis in dense or surgically implanted breasts. Dense breasts are common in women around 40 years of age. Mammographic examination presents the risk of ionizing radiation and the discomfort of compression. In the search for other techniques, thermography has emerged as a potential method to complement mammography and improve the efficiency of overall detection Thermography is a noninvasive, economic, rapid diagnostic method that does not touch the patient and does not inflict pain (Kapooret al 2010). Additionally it is risk-free and does not emit ionizing radiation (Wang et al 2009). Thermography is simple and is based on quantification of the surface temperature of the body measuring infrared radiation emitted by human skin (Acharya 2005).

In 1982, the US Food and Drug Administration (FDA) approved InfraRed Thermography (IRT) as an adjunct tool for the diagnosis of breast cancer. Kiran in a recent review presented a comparative study of IRT and other imaging techniques for breast screening and have concluded that IRT provides additional functional information on the thermal and vascular condition of the tissues (Kiran, 2012). Ng presented an excellent review of IRT as a non-invasive breast tumor detection modality, where he described in detail the basic methodology, standard practices, image capture and image analysis. According to Ng, an abnormal breast thermogram was indicative of significant biological risk. Tumors generally have an increased blood supply and an increased metabolic rate which leads to localized high temperature spots over those areas, rendering them relatively easy to detect by IRT (NG, 2003). Apart from passive breast imaging, cold stimulation-based imaging procedures are also in practice. Blood vessels produced by cancerous tumors are simple endothelial tubes devoid of a muscular layer. Such blood vessels fail to constrict in response to sympathetic stimulus like a sudden cold stress and show a hyper thermic pattern due to vasodilatation. Deng and Liu have shown that induced evaporation enhances thermographic contrast in cases of tumors that are underneath the skin (Deng and Liu 2005).

Although, the use of thermal cameras dates back to the 1960s, new research began in the late 1990s. The initial work was centered on the identification of cancer by imaging. Physicians diagnosed cancer by thermal imaging, after which the illness was classified via further image analysis (Jennifer 2012). It should be pointed out that this method was not suitable for the population in general as it presented a rather laborious task. The sheer volume of images limited the possibility of efficient revision and diagnosis. An alternative method was asymmetric thermal imaging; however, the low associated image quality limited its success (Jennifer 2012). Qi and Kuruganti found out that Intelligent identification by software based on detachment of all of the unique parts (thermal red parts) of the image showed the extracted parts, however it required revision by a physician. The system could not specifically extract tumor images, but just extracted the thermal part (Qi and Kuruganti 2003). Keyserlingk and associates published a retrospective study that reviewed the relative ability of clinical examinations, mammography, and infrared imaging in an attempt to detect 100 new cases of ductal carcinoma in situ, stage I and 2 breast cancers. Results from the study found that the sensitivity for clinical examination alone was 61%, mammography alone was 66%, and infrared imaging alone was 83%. When suspicious and equivocal mammograms were combined, the sensitivity was increased to 85%. A sensitivity of 95% was found when suspicious, equivocal mammograms were combined with abnormal infrared images. However, when clinical examination, mammography, and infrared images were combined a sensitivity of 98% was reached (Keyserlignk, Ahlgren, et al1998). In their paper, Quek et al shows that FNN complements breast thermography in various ways. The combination of breast thermography and FNN gives rise to more consistent results than merely using breast thermography. Whether it is cancer detection, tumor classification or breast cancer diagnosis (multi-class problem), FNN outperforms conventional methods, showing the strength of complementary learning in the recognition task. FNN assists physicians in distinct diagnostic tasks by providing a relatively accurate decision support tool, which could potentially enhance patient outcome. FNN not only gives superior results compared with conventional methods, but it also offers intuitive positive and negative fuzzy rules to explain its reasoning process.

2.4 Computational Intelligence application in breast cancer

Computational Intelligence enables, through intelligent techniques some of them inspired by nature, the development of intelligent systems that imitate aspects of human behavior, such as: learning, perception, reasoning, evolution and adaptation (Engelbrech, 2007). Some examples of Computational Intelligence techniques are: Artificial Neural Networks, case based reasoning, biological neuron-inspired technique (*Aruna, 2011*); Evolutionary Computation, inspired by biological evolution (*Mohamed and Hegazy, 2011*); Expert Systems, inspired by inference process (*Anagnostopoulos, 2006*); and Fuzzy Logic, inspired by language processing.

Late diagnosis of breast cancer in MLICs has triggered intensive studies and research on the application of these intelligent techniques in solving the problem. Intelligent methods such as Fuzzy logic and case based reasoning have been intensively used in developing clinical decision support systems (Ryua, Chandrasekaranb, & Jacobc, 2007).

2.5 Fuzzy logic

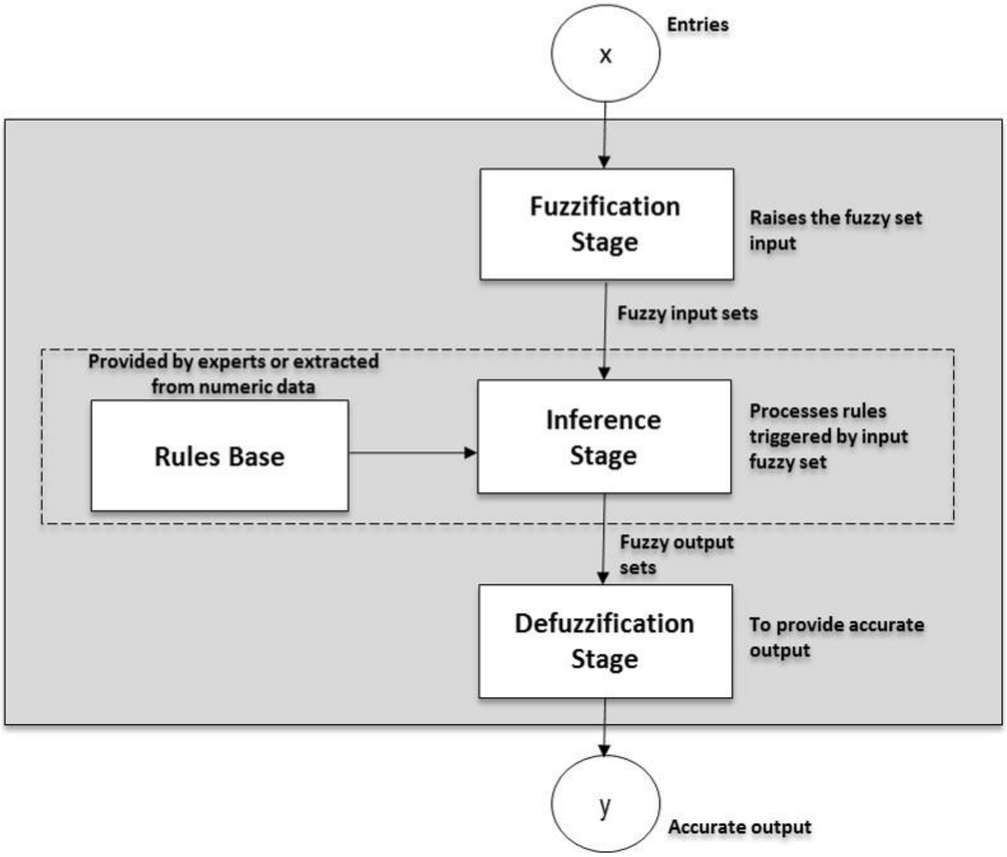
The fuzzy systems theory is a formal approach that aims to address the modeling, representation, reasoning and the inaccurate information procedure as a troubleshooting strategy (Dubois & Prade, 2001). Introduced in 1965 by Zadeh, the fuzzy set theory is a tool to model the imprecision and ambiguity that arises in complex systems (Zadeh, 1965), and it was created from the combination of the concepts of classical logic and groupings of Łukasiewicz et al defining degrees of relevance (Łukasiewicz et al, 1970, 87-88).

A fuzzy set differs from a classic set to assign to each element a value in the unit interval (*Chen H, Yang B, Liu J, 2011*). Specifically, a fuzzy set is defined as a function A of a set x, called universe of discourse, to (*Chen H, Yang B, Liu J, 2011*). The function A is referred to as a membership function, and the value A(x) represents the degree of relevance – or compatibility – of the element x with the concept represented by all the fuzzy set. Thus, the fuzzy logic proposed by Zadeh provides a mathematical model for the processing of inaccurate or vague information and concepts, intending to make computers carry out inferences as people (*Zadeh, 1979*).

The fuzzy processing is generally composed of: Rules Base (provided by specialists or extracted from numerical data); Fuzzification Stage (it activates the rules from a set of precise entries);

Inference Stage (determines how rules are enabled); Defuzzification Stage (it provides precise output, generating a fuzzy set of output), as illustrated in Figure 1.

Figure 1. Structure of a Fuzzy System Process.


x
Entries
Fuzzification Stage
Raises the fuzzy set input
Fuzzy input sets
Provided by experts or extracted from numeric data
Rules Base
Inference Stage
Processes rules triggered by input fuzzy set
Fuzzy output sets
Defuzzification Stage
To provide accurate output
y
Accurate output

Reprinted from Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis by Ana MG Guerreiro and Adrião D Dória Neto, 2012.

Fuzzy set theory has successfully been applied in handling uncertainties in various application domains (Jang, Sun, and Mizutani, 1997) including medical domain because of its ability to handle imprecise values. Inexact medical entities can be defined using fuzzy sets. The fuzzy sets explain fuzziness existing in a human thinking process using fuzzy values instead of using a crisp or binary value. Use of fuzzy logic in medical informatics begun in the early 1970s. In fuzzy CBR, fuzzy sets can be used in similarity measure (Bonissone and Cheetham 1998; Dvir, Langholz and Schneider 1999; Wang 1997). A discussion about the relationship between the similarity concept and several other uncertainty formalisms including fuzzy sets can be found in (Richter 2006). In the proposed CDSS, fuzzy set theory is used for matching similarities between existing cases and a current case to model imprecise expert's knowledge in the psychophysiological domain. It matches cases in terms of degrees of similarities (Aamodt and Plaza, 2001) between attribute values of previous cases and a new case.

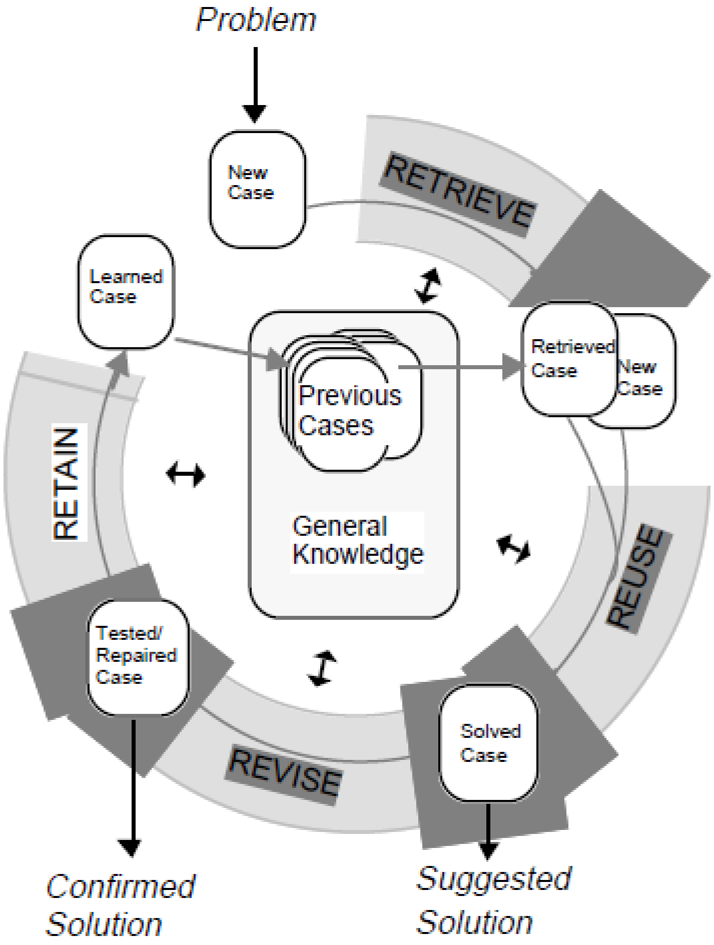
Decisions taken by medical experts during diagnosis depend on the experts experience, knowledge and familiarity with medical domain. Most of the diagnosis does not follow a particular path making the process error prone. Fuzzy logic implements a powerful reasoning method that can easily and comfortably deal with ambiguities, uncertainties and imprecision in the course of diagnosis. An aggregation of the knowledge, observation and experience of medical experts serves as the foundation of fuzzy based medical diagnosis system. The novel methodologies are presented for physician's decisions in medical informatics, medical problems solving and for the assessment of treatment planning decision process in diseases and therapies (Djam et al 2008).

Fuzzy logic systems and expert systems are used in handling complex and difficult tasks, however, fuzzy logic`s ability to handle ambiguity gives it advantage over the expert systems. To effectively handle ambiguities, linguistic rules are used to emulate human operation and assist make decisions. The ability to make decisions in fuzzy logic is time saving and minimizes need for human engagement (Tsoukalas, and Uhrig, 2003).

2.6 Case-Based Reasoning

A case-based reasoning (CBR) (Aamodt and Plaza 1994; Watson 1997) method can work in a way close to human reasoning e.g. solves a new problem applying previous experiences. A clinician/doctor may start his/her practice with some initial experience (solved cases), then try to utilize this past experience to solve a new problem and simultaneously increases his/her case base. So, this method is getting increasing attention from the medical domain since it is a reasoning process that also is medically accepted. CBR has shown to be successful in a number of different medical applications (Nilsson and Sollenborn 2004). Aamodt and Plaza has introduced a life cycle of CBR (Aamodt and Plaza 1994) with four main steps as shown in Figure 2. Retrieve, Reuse, Revise and Retain present key

Figure 2. The CBR cycle



Reprinted from The case based reasoning technique by Aamodt and Plaza , 1994.

In the retrieval step, for any new problem, the system tries to retrieve the most similar case(s) by matching previous cases from a case base. If it finds any suitable case that is close to a current problem then the solution is reused. A clinician may revise the selected case with solution and retain this solution along with the new problem into the case base. The CBR method in the proposed system is used to suggest recommendations for diagnosis of breast cancer-related disorder for a new case by retrieving and matching previously solved similar problems from the case base (BEGUM, S., 2007).

Chapter Three

RESEARCH METHODOLOGY

3.1 Research Design

There are two fundamental research approaches: qualitative and quantitative approaches. Despite the ongoing debate, recent development in research methodologies suggest that the two approaches should be integrated in comprehensive research designs in order to improve research rigor and address several of the epistemological and methodological criticisms (Kelle, 2006; Olsen, 2004).

This study applied both qualitative and quantitative approaches in order to satisfactorily answer the research questions.

This research design outlines: the data collection; data processing; system design and implementation; evaluation.

3.2 Data Collection

Data collection for this project was carried out at St. Francis Mission hospital Kasarani in Nairobi County, Kenya. The data was collected using primary data collection tools which involved direct interview of an oncologist.

3.3 Processing and classification of data – Fuzzy Method

Application of fuzzy logic and case based reasoning brings greater benefits (like expert knowledge acquisition, rules base generation, process automation and pre diagnosis greater precision) and satisfactory results, in addition to dealing with modeling, representation, the reasoning and the inaccurate information procedure as a troubleshooting strategy (Barro and Marin, 2002)).

Thus, the implementation of the intervention and control actions in the intelligent method developed, uses fuzzy logic and CBR since it enables to capture the experts' knowledge, as well as the appropriate treatment to fuzzy situations inherent in the problem classifying the likelihood of breast cancer presence (BEGUM, S., 2007).

The algorithm developed to assist the creation of fuzzy system applied to the medical field is presented next.

Algorithm: establishment of fuzzy system applied to the medical area

```
Step 1: Definition
        > Identify the problem
Step 2: Medical knowledge acquisition
        > Obtain technical information from one or more medical specialists
        > Extract data and information from gold pattern databases (with diagnosis confirmed)
        > Obtain information in technical literature available
Step 3: Fuzzification stage
        > Define entry membership functions and their fuzzy rules
Step 4: Rules base
        > Define fuzzy rules covering all possibilities
Step 5: Inference Stage
        > Reporting observations to fuzzy sets
        > Evaluate each case for all fuzzy rules
        > Combine the information from the defined fuzzy rules
Step 6: Defuzzification stage
        > Define membership functions and output sets
        > Define the defuzzification function
Step 7: Results verification
        > Ask results are satisfactory?
If answer = “No”
        > Return to Step 2
If answer = “Yes”
Finalize
```

This way, the definition of Fuzzy Method to assist in the diagnosis of breast cancer and its stages (Fuzzification Stage, Rules Base, Inference Stage and Defuzzification Stage) are listed below and instantiated through the system implemented.

3.4 Fuzzification

This is the first step in the process of fuzzy inferencing. Fuzzification applies a membership function to determine the degree of membership to a fuzzy set. This is done by selecting input parameters into the horizontal axis and projecting vertically to the upper boundary of membership function to determine the degree of membership.

Development of the CDSS was preceded by the design of the fuzzy set for all the relevant input variables. This is illustrated in the four (1 to 4) equations below. On the basis of domain experts' knowledge, the input and output parameters selected for this research were described with four linguistic variables (minor, moderate, severe and very severe). The range of the fuzzy value for each linguistic is shown in table 1 below:

Linguistic Variables - Fuzzy Values

Mild $0.1 < x \leq 0.3$

Moderate $0.3 < x \leq 0.45$

Severe $0.45 < x \leq 0.7$

Very Severe $0.7 < x \leq 1.0$

After the declaration of the linguistic variables, the raw data is transformed with the help of the functions specified in the equations below. In this process, a triangular membership function evaluates the linguistic variables and the degree of membership (between 0 and 1) identified (Djam, Wajiga, Kimbi , & Blamah, 2011).

These formulas are determined by aid of an oncologist.

$$\mu_{\text{mild}}(x) = \begin{cases} 00 & \text{If } x \leq 0.1 \\ \frac{x-0.1}{0.2} & \text{If } 0.1<x\leq 0.2 \\ \frac{x-0.2}{0.3} & \text{If } 0.2<x\leq 0.3 \\ 0 & \text{If } x>0.3 \end{cases} \qquad (1)$$

$$\mu_{\text{moderate}}(x) = \begin{cases} 0 & \text{If } x < 0.3 \\ \frac{x-0.1}{0.2} & \text{If } 0.3\leq x\leq 0.45 \\ \frac{x-0.45}{0.15} & \text{If } 0.45<x\leq 0.6 \\ 0 & \text{If } x>0.45 \end{cases} \qquad (2)$$

$$\mu_{\text{severe}}(x) = \begin{cases} 0 & \text{If } x < 0.5 \\ \frac{x-0.6}{0.2} & \text{If } 0.6\leq x\leq 0.8 \\ \frac{x-0.7}{0.3} & \text{If } 0.7<x\leq 0.8 \\ 0 & \text{If } x>0.7 \end{cases} \qquad (3)$$

$$\mu_{\text{very severe}}(x) = \begin{cases} 0 & \text{If } x < 0.8 \\ \frac{x-0.1}{0.2} & \text{If } 0.8\leq x\leq 1.0 \\ \frac{0.2-x}{0.1} & \text{If } 0.9<x\leq 1.0 \\ 0 & \text{If } x\leq 1.0 \end{cases} \qquad (4)$$

The next step in the fuzzification process is the development of fuzzy rules. The fuzzy rules for this research were developed with the assistance of domain experts (two oncologists). The knowledge-base has many fuzzy rules designed with the aid of combination theory: only the valid rules were chosen by the domain experts (Murat Karabatak and M. Cevdet, 2009). A rule is said to fire if any of the precedence parameters (mild, moderate, severe, very severe) evaluate to true (1); otherwise, if all the parameters evaluate to false (0), it does not fire (Barro & Marin, 2002).

Table 3. **Fuzzy Rule Base for breast cancer**

Rule no	A lump in the armpit	Changes in breast skin, shape or size	bone pain	nausea	loss of appetite	Jaundice	weight loss	fluid around the lungs	shortness of breath	muscle weakness	cough	headache	Conclusion
1	mild	mild	mild	mild	mild	mild	mild	mild	mild	mild	mild	severe	mild
2	mild	mild	mild	mild	mild	mild	mild	mild	mild	mild	severe	mild	mild
3	moderate	moderate	moderate	severe	severe	moderate	moderate	moderate	mild	mild	mild	mild	moderate
4	moderate	severe	mild	mild	e	severe	mild	moderate	severe	mild	moderate	moderate	severe
5	mild	severe	severe	moderate	mild	severe	moderate	mild	moderate	moderate	moderate	mild	severe
6	mild	mild	severe	moderate	mild	mild	mild	severe	mild	mild	mild	mild	severe

7	severe	mild	moderate	mild	moderate	mild		severe	moderate	mild	moderate	moderate	very severe
8	very severe	moderate	moderate	moderate	mild	moderate	mild	moderate	mild	moderate	mild	moderate	very severe
9	mild	mild	moderate	mild	mild	mild	mild	mild	mild	mild	severe	mild	moderate
10	mild	moderate	mild	mild	mild	mild	mild	mild	mild	mild	severe	mild	moderate
11	mild	mild	mild	mild	mild	mild	mild	mild	mild	mild	severe	mild	mild
12	moderate	moderate	moderate	severe	severe	moderate	moderate	moderate	mild	mild	mild	mild	moderate

Note: adapted from Medical Diagnosis System Using Fuzzy Logic by J.B. Awotunde1, O.E. Matiluko and O.W Fatai, 2014.

3.5 Inference

This is the process of making conclusions from existing data. A human expert reasoning is modeled in to a knowledge processor known as the inference engine. In the notion of fuzzy logic, the process of mapping an input set to an output using the theory of fuzzy sets is called fuzzy inference (Tsoukalas & Uhrig, 1993).

There are two inference techniques:

Forward chaining:

In this strategy, the starting point is a set of known facts, the strategy then attempts to derive new facts using the available rules whose premises match the known facts. This process is repeated until either a goal state is reached or no further rules have premises that match the known or derived facts (Ahmed, Sherif & Ahmed, 2011).

Backward chaining:

This strategy tries to put together all the relevant information so as to prove a hypothesis. The Mamdani Inference type is the fuzzy inference mechanism employed in this research. The rules in the knowledge-base are used by the fuzzy inference engine so as derive conclusion.

The inference engine used to implement this CDSS will use the forward chaining mechanism to search the knowledge for the breast cancer symptoms. The inference engine technique employed in this research is the Root Sum Square (RSS).

RSS is given by the formula in equation (6):

$$\sqrt{\textstyle\sum} R = \sqrt{}\ (R_1^2 + R_2^2 + R_3^2 + \ldots\ldots + R_n^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Where R12 + R22 + R32+……+ Rn2 are strength values (truth values) of different rules which share the same conclusion. Whenever, the inference engine receives user queries, the decision making process is initiated. The inference engine then generates the weights for the inputs.

In this stage, the entries in this case; symptoms were analyzed to generate the fuzzy output set with its respective compatibility degree. The CDSS developed used the fuzzy model proposed by Mamdani (Mamdani 1974), in which the activation function of each rule is enabled and the system of inference determines the degree of compatibility of the rules premise contained in the rules base. After this, it determines which rules are enabled and applies them to the output membership function, remaining just linking all output nebulous sets activated (and their respective degrees of compatibility) into a single Output Set (OS). This OS represents all results (diagnosis) that are acceptable for the input set, each with its compatibility level. Each case was also assessed, at this stage, for all fuzzy rules and the combination of information was carried out from the rules already defined in the Rules Base.

3.6 Defuzzification

Most real life situations, solid values are needed however the inference engine will usually output a fuzzy set (Miller & Sittig, 2011). This stage was used to generate a single numeric value, from all possible values contained in the fuzzy set obtained in the inference stage, to generate the diagnosis. The defuzzifier undertakes a translation of the inference engine output to give a firm output. The set of symptoms and the level of effect to the patient (fuzzy set) was input to the defuzzification process to output a single number as the output.

There are three common defuzzification techniques used widely:

- Max criterion,
- Center-of gravity
- Mean of maxima.

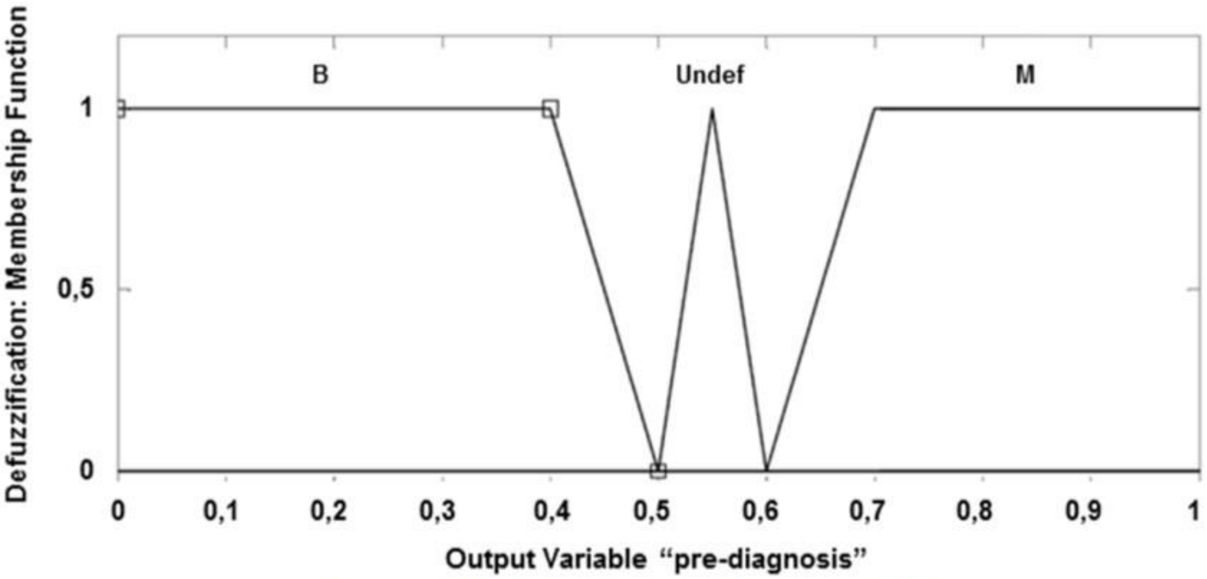
The max criterion is the most common technique because of its simplicity to implement. This technique produces the point at which the possibility distribution of the action reaches a maximum value (Tsoukalas, & R.E. Uhrig, 2003).

$$\mathrm{CoG}\ (Y^n) = \frac{\sum \mu_y(x_i)x_j}{\sum \mu_y\, x_i} \quad \text{...}(5)$$

Where $\mu_y x_{i\,=}$ me

$x_{i=}$ center

The Pre-Diagnosis (PD) membership function, Defuzzification, is composed of "Benign", "Undefined" and "Malignant", represented linguistically as BPD, UndefPD and MPD, respectively, representing the tracks [≤ 0.5; 0.5 – 0.6; and ≥ 0.6], as output set below and illustrated in Figure 3



Defuzzification: output variable "pre-diag", with domain = [0 – 1], where:
B (B_{PD}, Benign pre-diagnosis) = trapezoidal function {(-0.5; 0); (0; 1); (0.4; 1) (0.5; 0)};
M (M_{PD}, Malignant pre-diagnosis) = trapezoidal function {(0.6; 0); (0.7; 1); (1; 1) (1.5; 0)};
Undef ($Undef_{PD}$, undefined pre-diagnosis) = triangular function {(0.5; 0); (0.55; 1) (0.6; 0)}.

Figure 3
Defuzzification **Membership Function.**

Output Set (OS):

PDbenign(BPD)≤0.5→BPD={(−0.5;0),(0;1),(0.4;1),(0.5;0)};

PDundefined(UndefPD)≥0.5e≤0.6→UndefPD={(0.5;0),(0.55;1),(0.6;0)};

PD malignant(MPD)≥0.6→MPD={(0.6;0),(0.7;1),(1;1),(1.5;0)}.

3.7 Post-processing

In post-processing, the result, in the form of malignant or benign pre-diagnosis, is stored on the apache server in a MySQL database. The saved result is then made available on the screen using PHP and MySQL.

3.8 Development of the CDSS

Development of the CDSS involved closely working with the clinicians and oncologist at the St. Francis Mission Hospital Kasarani. Implementation tools included the following

- Use of a PHP framework known as YII, which uses the MVC model.
  - This was used for the development of the backend system

3.9 Validation

Testing of the CDSS, the object of this study was done in two phases, firstly using functional testing which is concerned with testing the functionality of a system without regard to the method of implementation was applied (DeMillo, 2007). Therefore, test cases were developed and used to test a variety of different implementations of the system. The choice of test cases for functional testing of the CDSS modules and their integration is more challenging, as all system inputs and outputs must be identified and specified or predicted. Test cases for functional testing were inspired by real-world observations, but must also be derived from design specifications.

The test cases used in testing the CDSS were chosen to perform both structural and functional testing. The description of the test case appears in one column then the next column has a drop down with options Pass or Fail. At the bottom of the test script, there is a summary of the number and percentage of test cases passed or failed (see appendix II for a list of test case).

Secondly, testing was also carried out using the MATLAB R2010a (student version), due to the tools available in this application to the development of models and the rapid visualization of the results obtained in the fuzzy system.

In addition to the testing carried out at the St. Francis hospital, further tests were initially carried out using the WDBC especially on the identification of the main characteristics of the fuzzy system, such as:

- identification of the set of descriptors that provide the best results, called "best input set" (BIS);
- identification of the best set of rules (BSR); and
- Definition of what membership functions, which parameters and what defuzzification functions are most suitable for use with the BIS and the BSR.

The validation of the rules base was held in conjunction with medical professionals (two oncologists and two general practitioners), considering the fuzzy set indicators of both malignant and benign diagnosis. As a consequent action of the descriptors' relations and variability the domain [0 –1], representing the tracks [< 0.5; 0.5 – 0.6; > 0.6], was adopted to defuzzification, which is represented in linguistic terms as "Benign", "Undefined" and "Malignant", respectively, as presented in Figure 9.

After this phase, cross-validation was used for testing, in order to fine-tune the parameters of the membership functions of the CDSS. A twofold validation model was adopted. First, three databases were generated, each of them with 150 (one hundred and fifty) gold pattern clinical cases randomly extracted from WDBC. Secondly, 50 breast cancer cases previously diagnosed at the St. Francis hospital were extracted. In both the first and second case, the final diagnosis was known and confirmed.

The validations of both the knowledge gained and the results achieved were performed during the development of the CDSS and also, in the final instance, by medical specialists; oncologists and general practitioners.

In order to assure a high pass rate at regression testing on all modules which may be affected by the erroneous module is stopped until corrections can be made. While this slows the overall testing process, it does reduce the probability of diagnosing "false errors" because of erroneous data propagating through the system.

3.10 Error Handling and Recovery

Any errors in the implementation of the CDSS were logged in the BugZilla and then reported to the research team comprising the programmer and the knowledge engineer. An incident report was then created which notes the observed effect, the test conditions which lead to the error and the probable location of the error. Errors in implementation were corrected by the programmer who was responsible for completing the resolution section of the incident report noting the exact problem, its resolution, the date and time of the correction and attests to the correct operation of the module. Both the programmer and knowledge engineer must attest to the correct operation of the adjusted module.

Chapter Four

RESULTS AND DISCUSSION

4.1 Results

The testing of all the CDSS modules generated more than 200 test cases. Full CDSS testing required more than 900 man-hours of work, that would have led to a system which performs as designed and is easily maintained. Since the implementation of these systematic methods for testing the proposed CDSS were implemented, the rate of implementation and logic errors has decreased significantly.

The evaluation of the prototype was conducted using historical patient records that have final diagnosis. 50 breast cancer cases at the St. Francis Mission Hospital Kasarani were considered, the data captured included the patient address, phone Number, First Name, Last Name, Middle Name, Date of Birth, Gender, physical examination findings, and laboratory test results. All the cases selected for the evaluation had been diagnosed at the hospital within the last three years. After entering the patient details the system was invoked to give a diagnosis. The diagnosis given by the CDSS was compared to the final diagnosis that had earlier been recorded.

Using the CDSS, the list of diagnoses suggested by the CDSS contained the correct diagnosis in 48 of the 50 cases (96%). The 2 diagnoses that were not suggested were not included in the CDSS database at the time of the evaluation.

The input set that featured the best results while using the WDBC database for evaluation has the following characteristics:

1. a) **fuzzy system:** Mamdani;
2. b) **membership functions of the entry set:** trapezoidal;
3. c) **input set** composed of 4 variables (descriptors), with the following fuzzy sets:

c.1)
AREAcomSMAREA={(184.5;0),(185;1),(748.8;1),(1000;0)}andLAAREA={(508.1;0),(2194;1),(4255;1),(4256;0)};
c.2.)
PERIMETERwithLPERI={(49.5;0),(50;1),(92.58;1),(103;0)}andLAPERI={(85.1;0),(159.8;1),(252;1),(252.5;0)};
c.3.)
UNIFORMITYwithMOUNIF={(−0.5;0),(0;1),(1.669;1),(2.6;0)}andLEUNIF={(0.65;0),(6.205;1),(12;1),(12.5;0)};and
c.4.)

HOMOGENEITYwithMOHOM={(0;0),(0.01;1),(0.1232;1),(0.19;0)}andLEHOM={(0.0295;0),(0.2168;1),(0.45;1),(0.5;0)};
d) rules base: 16 rules;
e) membership functions of the output set:
e.1)
trapezoidal for classification Benign, beingBPD={(−0.5;0),(0;1),(0.4;1),(0.5;0)};
e.2)
trapezoidal for classification Malignant, beingMPD={(0.6;0),(0.7;1),(1;1),(1.5;0)};and
e.3)
triangular for classification Undefined, beingUndefPD={(0.5;0),(0.55;1),(0.6;0)};
f) defuzzification: Centroid function;
g) output variable: 1 (result = pre-diagnosis).

The best result achieved is shown in the Diagnostic Test Assessment Matrix presented in Table 4, as well as in the Matrix of Confusion presented in Table 5.

Diagnostic test Assessment			
	GOLD PATTERN DIAGNOSIS		
FUZZY-FNA	Malignant (%)	Benign (%)	**TOTAL**
Malignant (%)	36.73	9.14	45.87
Benign (%)	0.53	53.60	54.13
TOTAL	37.26	62.74	100.00
Sensitivity = 98.59%	**Specificity = 85.43%**		

Table 4 : Diagnostic test of assessment matrix

Confusion matrix		
	GOLD PATTERN	
FUZZY-FNA	Malignant	Benign
Malignant	**0.99**	0.15
Benign	0.01	**0.85**

Table 5: Confusion matrix of the diagnostic test

It is noted in the diagnostic test assessment matrix (Table 4), that the CDSS developed presents: 98.59% sensitivity, which is the ability of a diagnostic test to identify the real positive in individuals truly ill, meaning a satisfactory percentage of hits in the pre-diagnosis of malignancies; and 85.43% specificity, which is the ability of a diagnostic test to identify the real negative in individuals truly healthy, corresponding to the correct pre-diagnosis of benign cases.

We must point out that, in the laboratory examination (biopsy) of smears obtained by FNA for identification of breast cancer, it is more important to get good results in sensitivity than in specificity (*Armitage & Berry 1994, Office for Official Publications of the European Communities 2006).* Subsequently, among the tests performed during the development of the CDSS to assist in the diagnosis of breast cancer, there were several with satisfactory results as well, but they were not selected as the best solution, having been discarded, as, for example, the test sets A, B and C, presented below.

The tests of set A were conducted from the best input set, with changes in nebulous sets (parameters) of the membership functions. In Table 6, the results of sensitivity and specificity of the same are presented. Notably test A.1 presents 99.06% sensitivity, however the medical experts found the specificity of 64.15% unsatisfactory. The tests A.8 and A.10 presented the same sensitivity of CDSS developed (98.59%), but lower specificity (84.31% and 84.87%, respectively). The other tests presented sensitivity less than 98.59% and thus were discarded.

Tests	**Sensitivity (%)**	**Specificity (%)**
CDSS developed	**98.59**	**85.43**
Test A.1 (1)	99.06	64.15
Test A.2 (2)	92.92	90.48
Test A.3 (3)	98.11	70.87
Test A.4 (4)	93.87	89.92
Test A.5 (5)	96.23	88.80
Test A.6 (6)	97.17	87.39
Test A.7 (7)	97.64	86.83
Test A.8 (8)	98.59	84.31
Test A.9 (9)	98.11	86.55
Test A.10 (10)	98.59	84.87

Table 6: Comparison of the tests presented in “TEST SET A" (changes were realized in the fuzzy sets of membership functions)

(1) changes in Test A.1: SM_{AREA} = {(184.5; 0), (185; 1), (749; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.6; 1), (95; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.67; 1), (1.87; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.123; 1), (0.143; 0)}.

(2) changes in Test A.2: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (127.1; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (3.09; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.2278; 0)}.

(3) changes in Test A.3: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (95; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (3.09; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.2278; 0)}.

(4) changes in Test A.4: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (110; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (3.09; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.2278; 0)}.

(5) changes in Test A.5: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (106; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (3.09; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.2278; 0)}.

(6) changes in Test A.6: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (106; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (3.09; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.18; 0)}.

(7) changes in Test A.7: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (106; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (2.5; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.18; 0)}.

(8) Changes in Test A.8: SM_{AREA} = {(184.5; 0), (185; 1), (748.8; 1), (800; 0)} e MO_{UNIF} = {(-0.5; 0), (0; 1), (1.669; 1), (2.5; 0)} e MO_{HOM} = {(0; 0), (0.01; 1), (0.1232; 1), (0.18; 0)}.

(9) Changes in Test A.9: SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (103.5; 0)}.

(10) Changes in Test A.10: SM_{PERI} = {(49.5; 0), (50; 1), (92.58; 1), (102.3; 0)}.

The tests of set B were conducted from the best input set, with changes in the types of membership function of the input set and, consequently, in their nebulous set (parameters). In Table 7, the results of sensitivity and specificity of the same are presented. Notably the tests B.1 and B.4 showed the same sensitivity that the CDSS developed (98.59%), but lower specificity (84.47% and 82.91%, respectively). The other tests showed sensitivity less than 98.59%, having been discarded.

Tests	**Type of membership function (after adjustments in fuzzy sets)**	**Sensitivity (%)**	**Specificity (%)**
CDSS	**trapezoidal (1)**	**98.59**	**85.43**

Tests	**Type of membership function (after adjustments in fuzzy sets)**	**Sensitivity (%)**	**Specificity (%)**
developed			
Test B.1	triangular(2)	98.59	83.47
Test B.2	gaussian2(3)	98.11	84.31
Test B.3	dsigmoidal(4)	98.11	84.59
Test B.4	polinomial zero(5)	98.59	82.91

Table 7: Comparison of the tests presented in "TEST SET B" (changes were realized in the membership functions of the entry set and its fuzzy sets)

(1) trapezoidal - function with straight lines with a flat top, resembling a truncated triangle.
(2) triangular - function with straight lines, in the form of a triangle.
(3) gaussiana2 - composed of two different gaussian curves.
(4) dsigmoidal - created from the difference between two sigmoidais functions.
(5) polinomial zero – asymmetric polynomial function, being zero at both ends, with an increase in the middle.

The C set were performed from the best input set, with changes only in the defuzzification function. Presented in Table 8, are the results of sensitivity and specificity of the same. It is worthy to note that all of the tests presented the same sensitivity that the CDSS developed (98.59%), but lower specificity.

Tests	**Defuzzification function**	**Sensitivity (%)**	**Specificity (%)**
CDSS developed	**centroid (1)**	**98.59**	**85.43**
Test C.1	bisector(2)	98.59	83.47
Test C.2	mom(3)	98.59	77.59
Test C.3	lom(4)	98.59	73.67
Test C.4	som(5)	98.59	77.59

Table 8: Comparison of tests presented in "TEST SET C" (changes were realized in the defuzzification functions)

(1) centroid - calculates the output set (OS) area center generated in the inference stage and determines its projection on the x-axis, that is the control output value.
(2) bisector - exact position that splits the output set into two equal areas.
(3) mom (Middle of Maximum) - it performs the arithmetic mean of all maximum values of the OS.
(4) lom (Largest of Maximum) - considers the greatest among all the maximum values of the OS.
(5) som (Smallest of Maximum) - considers the lowest among all the maximum values of the OS.

Thus, the results achieved by the CDSS, the object of this study, were considered satisfactory by the oncologists mainly for their high sensitivity (malignant cases hit) presented, as can be seen in Table 4.

The sensitivity of 98.59% presented by CDSS is at the same level of prominence of other works using the same dataset with other techniques such as, for example, work done by *Anagnostopoulos et al 2006*, using Probabilistic Neural Network-PNN with 31-568-2 topology. Although other works, for example, done by (*Mohamed and Hegazy, 2011)* are more accurate than the specified cdss, they use ten descriptors, while this CDSS uses only four descriptors, two of which are extracted indirectly from WDBC, which simplifies the model and streamlines processing.

4.2 Discussion

Current practice of clinical diagnosis with no respect to chronic cases is done manually. This process proves ineffective as it largely depends on the experience and the remembering ability of the individual doing the diagnosis. This has led to misdiagnoses which in some occasions cost people's lives. Furthermore, data collected on paper during the clinical process may be difficult to use for decision and policy making because of the unease of access. The proposed CDSS has implemented an electronic clinical diagnosis support system that can assist any clinician in making the decision about breast cancer irrespective of his/her experience. In addition, the proposed CDSS has implemented data viewer module where patient data can be visualized and manipulated for analysis purpose. Compared with manual methods, which require extra efforts of filling the forms and thereafter enter the data manually in computer software such as MS Excels and spreadsheet, the proposed CDSS has linked the data collection feature and data manipulation feature. Therefore, the proposed prototype may reduce difficulties in patient data collection, ensure early breast cancer diagnosis and minimizes the time lag for data to be available for usage for decision making. Patient data capture with the proposed prototype is simplified as the data is fed directly to the database through a web browser; therefore human data transcription errors can be minimized and increase data accuracy. The proposed CDSS has the capability of capturing data of text type but can be enhanced to capture, audio, video, images, in order to add more flexibility in kinds of data that can be collected about a patient. Furthermore, proposed prototype is customizable (flexible form design) and can be deployed in user defined settings. The flexibility on terminologies used in data collection forms allows easy way of setting uniformity of data formats and therefore increase coordination of different levels. The framework selected to design the proposed prototype is based on open source technologies, which allow future development with less effort, which can be affordable and manageable even in economic situations.

Chapter Five

CONCLUSION AND RECOMMENDATIONS

This paper presents a computer-aided decision support system for diagnosing breast cancer. Our work to date features three main points, namely feature extraction from time series data, case-based reasoning, and fuzzy information processing. Feature extraction is tasked to “dig out” key characteristics from original signals to reach a concise yet sufficient description of problems. The success for this heavily relies on domain knowledge and 19 time-based features have been identified and confirmed through cooperation with domain experts. The method of case-based reasoning is employed to make recommendations for stress diagnosis by retrieving and comparing with previous similar cases in terms of features extracted. Moreover, fuzzy techniques are incorporated into our CBR system to better accommodate uncertainty in clinicians reasoning as well as imprecision in case indexes. All such ideas have been implemented and validated in a prototypical system.

Feature weighting is another important issue under investigation in our project. With available test data we have recognized that the extracted features have different importance and proper weightings for them plays a crucial role for system performance.

So far we have two sets of weight values, both of which offered acceptable system performance in evaluation.

One of such weight sets was exclusively defined by an experienced domain expert, and the other set was learnt from the case base by applying the so called discriminating power (Funk and Xiong 2007) on discretized universes of individual features. The automatic learnt weights have shown to perform sufficiently close to an expert in identifying similar cases, sufficiently good bearing in mind that different expert have different opinions and that there is no exact answer. We conjecture there would be two reasons for this inferiority. The first lies in the fact that there are merely 39 cases in the current case library and this low number of samples may degrade the reliability of weights achieved. The second and possibly more important is the lack of expert preference information in the case base. One of our future research directions will be optimization of feature weights by directly utilizing case preferences of expert as learning signals.

5.1 General implications of the study findings

The realization that computer based clinical diagnosis decision support systems are feasible in the developing world especially in ensuring early diagnosis of breast cancer has important implications for health reforms in these parts of the world. It has been noted that the use of computerized technologies is on the rise in the developing world and that the current paper-based systems are not sustainable. There is thus the need for pilot projects to adopt new models such as the prototype proposed in this research in a controlled and experimental process in the field. This will provide the needed data sets for the review of the prototype and for streamlining future research works.

5.2 Contribution to previous work

The overall contribution from this study is knowledge of how computerized clinical diagnosis decision Support systems can ensure early diagnosis of breast cancer using open source frameworks. Comparing with previous work which studied the use of CDSS such as PDA, D-xplain and manual processes (L. Tsoukalas, and R. Uhrig, 2003), this study has attempted to evaluate and test the use of emerging web based technology for clinical diagnosis. Unlike previous studies which did not draw much attention on open source technology solutions, this study has investigated the applicability of open source technologies in enhancing early breast cancer diagnosis. Electronic forms that have been implemented in the proposed prototype are user friendly and can let user fill the proper data in proper input box. The data visualization interface of the proposed prototype may help the decision makers to quickly access data. The empirical contribution from the proposed prototype is to serve as a blue print of the actual implementation of the systems that could ensure early and accurate diagnosis of breast cancer and provide the data for further analysis.

5.3 Limitations of the proposed prototype

The proposed prototype has been developed basing on the requirements from both primary and secondary sources such as literature and direct interviews, therefore it could not be deployed directly in the field rather it can be used to depict the general process of clinical diagnosis. Furthermore, the proposed prototype can only be accessed via a computer browser. The system should be enhanced to include a mobile version for easy access given that the mobile phones are prevalent in day to day communication and general use.

5.4 Recommendations for future work

The future studies in this research area could attempt to develop the proposed prototype using data and specifications from the more primary sources (actual stakeholders). Such works could also focus on enhancing health data visualization to improve the analysis process. For example, enable data visualization through mobile phone screen (mobile interface for data visualization) as this research focused on developing a web based prototype. However, a mobile version would come in hand as users can access at their convenience.

The health data analysis process could also be improved in such a way that some of the decisions to be automated can be based on the collected data. This research focused on collecting patient details and suggesting the possibility of breast cancer likelihood. However, the data fed in to the system can help provide useful reports for decision making. Furthermore, the future studies could look at the way different health data systems can be integrated to avoid duplication of data and maintain consistency in reporting. In addition to that, future work could attempt to investigate ways of coordinating different health information systems levels to avoid fragmentation of flow of information through centralization of health data centers. Moreover, there are still rooms for investigating how open source frameworks could enhance other health management services in the developing world. For example, studies are needed to evaluate the applicability of different open source software packages for health service management in the developing world. Security to health data is another area future studies can look at. The advancement of ICT increases vulnerability of the privacy and security of health data, especially sensitive health data (statistical data) which might have great impact to the health service. Future work can investigate ways of securing health data in the ubiquitous networks and other health data transmitting networks. Finally, future work should enhance the proposed prototype so as bring to a level where it can do diagnosis for diseases other than breast cancer.

REFERENCES

[1] A.E.S., Ahmed, E.B., Sherif, and A A.B.A., Ahmed, (2011). A Fuzzy Decision Support System for Management of Breast Cancer, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[2] S. Barro, R. Marin, (2002). Fuzzy Logic in Medicine.

[3] Zadeh LA (1965): Fuzzy sets. Information and control 8: 338-353, 1965.

[4] B. Faran, M. Saleem Khan, Yasir Noor, and M. Imran, (2011), Design Model of Fuzzy Logic Medical Diagnosis Control System, International Journal on Computer Science and Engineering (IJCSE), ISSN :0975-3397 Vol. 3 No. 5 May 2011 2093-2108.

[5] American Cancer Society. (2009, November 9). Breast Cancer. Atlanta, GA: American Cancer Society.

[6] Djam, X.Y., Wajiga, G. M., Kimbi Y. H. and Blamah, N.V., (2011). A Fuzzy Expert System for the Management of Cancer, International Journal of Pure and Applied Sciences and Technology ISSN 2229 – 6107, 5(2) (2011), pp. 84-108.

[7] L.H. Tsoukalas, and R.E. Uhrig. Fuzzy and neural approaches in Engineering, John Wiley & Son, Inc. (2003).

[8] L. Stefano Nardini, Germano Bettoncelli, Vincenzo Lamberti and Patrizio Soverina. The AIMAR recommendations for early diagnosis of chronic disease based on the WHO/GARD model. (2006).

[9] S.S., Smita, S., Sushil & M.S., Ali, (2013). Fuzzy Expert Systems (FES) for Medical Diagnosis, International Journal of Computer Applications (0975 – 8887) Volume 63– No.11, February 2013.

[10] Manish Rana, & Sedamkar R.R., (2013). Design of Expert System for Medical Diagnosis Using Fuzzy Logic. International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013,pg. 2914-2921 ISSN 2229-5518

[11] Kiran Reddy. Developing Reliable Clinical Diagnosis Support System Developing Personal Medical Record Application for the iPhone and web. (2012)

[12] NG, K. (2003). Structural similarity as guidance in case-based design. In: European Workshop on Case-based Reasoning, paper Presentations, 1-5 November 2013. Vol. I. University of Kaiserslautern, pp. 14-19

[13] M. Morgan, N. Mays, and W. W. Holland, "Review article Can hospital use be a measure of need for health care ?," Journal of Epidemiology and Community, vol. 41, pp. 269-274, 2007.

[14] D. M. Aanensen, D. M. Huntley, E. J. Feil, F. Al-Own, and B. G. Spratt, "EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection.," PloS one, vol. 4, no. 9, pp. 1-7, Jan. 2009.

[15] Diamond GA, Pollock BH, Work JW. Clinician Decisions and Computers. In: Shabot MM, Gardner RM, eds. Decision Support Systems in Critical Care. New York: Springer-Verlag New York, Inc., 2001:(Orthner HF, ed. Computers in Medicine;

[16] Jorgensen PC. Software Testing: A Craftman's Approach. Boca Raton, Florida: CRC Press, 2005.

[17] DeMillo RA. Software Testing and Evaluation. The Benjamin/Cummings Publishing Compayn,

[18] Inc., 2007.

[19] Miller PL, Sittig DF. The evaluation of clinical decision support systems: what is necessary versus what is interesting. Med Inf Lond 2010;15(3): 185-90.

[20] C. Abouzahr and T. Boerma, "Policy and Practice Health information systems : the foundations of public health," Bulletin of the World Health Organization, vol. 14951, no. 4, pp. 578-583, 2010.

[21] Deo Shao, "A Proposal of a Mobile Health Data Collection and Reporting System for the Developing World" 2011

[22] AAMODT, A., and E. PLAZA 2013. Case-based reasoning: Foundational issues, methodological variations and system approaches.

[23] Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, Bo von Schéele 2009, A Case-Based Decision Support System for Individual Stress Diagnosis Using Fuzzy Similarity Matching, , Computational Intelligence (CI), vol 25, nr 3, p180-195(16), Blackwell Publishing

[24] Murat Karabatak , M. Cevdet Ince 2009, An expert system for detection of breast cancer based on association rules and neural network

[25] Begum, S., m. U. Ahmed, P. Funk, N. Xiong, and B. Von Schéele. 2007. Classify and Diagnose Individual Stress Using Calibration and Fuzzy Case-Based Reasoning. In proceedings of 7th International Conference on Case-Based Reasoning, Edited by Weber and Richter, Springer, Belfast, Northern Ireland, pp. 478-491

[26] http://www.medscape.com/ 08/Aug/2009 [Medscape offers physicians, specialists, primary care GPs, and other health professionals the internet's most robust and integrated medical information and educational tools]

[27] Pople HE Jr. (1982) Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics. In: Szolovits P, ed. Artificial intelligence in medicine. AAAS Symposium Series. Boulder, CO: Westview Press. [Good illustration of mimicking human reasoning in clinical diagnosis and using of AI in healthcare]

[28] Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. (2011) Physicians' information needs: an analysis of questions posed during clinical teaching in internal medicine. Ann Intern Med. [A comprehensive study of clinicians requirements and understanding of medical knowledge in clinical diagnosis]

APPENDICES

Appendix I: Questionnaire used for qualitative study.

Introduction

This questionnaire is prepared by Peter Simeon Wanyonyi a student undertaking Masters of Science course in Information Technology Management (ITM) at the University of Nairobi. With the guidance of my supervisor Dr. Chris Chepken, I need to conduct a qualitative study to understand challenges in clinical diagnosis in the developing world and also to evaluate the proposed prototype for clinical diagnosis. Your participation in this research study will be highly appreciated. Your response to this questionnaire will be treated with confidentiality. Your response is very important to me as this will lead to the final write-up of my thesis. The questionnaire should not take too long to complete. Thank you very much for your time. Please do not hesitate to contact me with the email below if you have any questions concerning this questionnaire.

Regards

Peter Wanyonyi

Email: wanyos2005@gmail.com

Questions:

Multiple choice you may highlight the answers.

General questions

1. County: ....

2. Name of Medical facility: ....

3. Which position do you hold? ....

4. How is diagnosis done in your medical facility?.....................
    a. Manual methods
    b. Computer software
    c. I don’t know

5. To what extent is computer technology applied in your medical facility? ....
    a. Wide b. Intermediate c. Low

6. Do you think computer software applications could improve clinical diagnosis, health data collection and reporting process? ....
    a. Yes b. No
7. Does your organization use any computer software? ....
    a. Yes b. No
8. If yes to (7) above, please give the name of the software?................
9. Does your medical facility have stable internet connection?..................

    a. Yes b. No c. I do not know

Evaluation of the proposed prototype

10. Is the health information systems computerized in your medical facility? ....

    a. Yes b. No

11. What are the main challenges that face health information systems in your facility/county?

12. What kinds of data are reported from health facility to the management levels (secondary facilities)?

13. What is the frequency of reporting health data (e.g. monthly, weekly, quarterly, yearly)?

14. How long does it take to collect patient data and undertake diagnosis using the current method?
    a. Within 1 hour b. Within 2 hours c. more than 1 day d. specify other

15. Who are the most common users of collected data (e.g. doctors, patients, government)?

16. A web based system has been designed to capture and assist in clinical diagnosis; do you think this could improve the current way of capturing patient information and ensure early diagnosis?

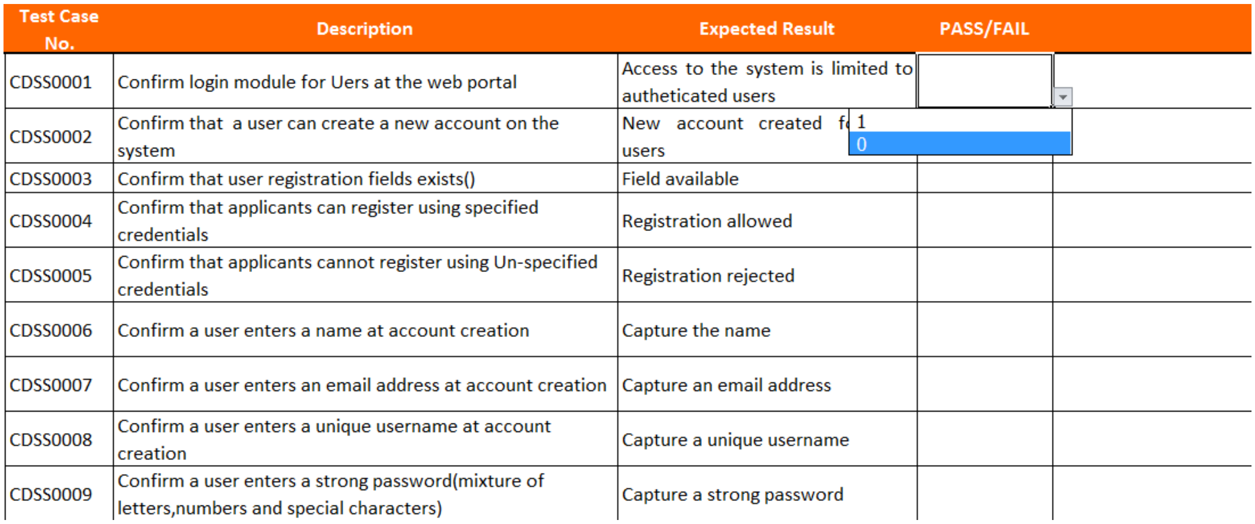
17. The proposed prototype offers health data mapping and simple graphs for analysis. What other features that could be included in future to improve analysis process?

18. The proposed prototype involves only two main stakeholders (health statisticians and health managers). What other stakeholders do you think should be involved?

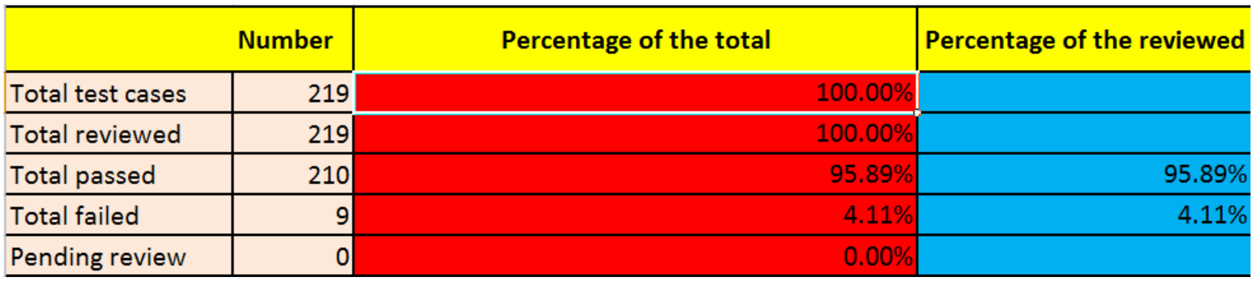
19. The proposed prototype has been developed in the PHP yii platform. Do you think that this platform is now common in your country and could be proper platform for developing community applications?

20. Any other comments or suggestion.

Appendix II: Sample test cases for prototype testing



Test Case No.	Description	Expected Result	PASS/FAIL	
CDSS0001	Confirm login module for Uers at the web portal	Access to the system is limited to autheticated users		
CDSS0002	Confirm that a user can create a new account on the system	New account created f users	1 0	
CDSS0003	Confirm that user registration fields exists()	Field available		
CDSS0004	Confirm that applicants can register using specified credentials	Registration allowed		
CDSS0005	Confirm that applicants cannot register using Un-specified credentials	Registration rejected		
CDSS0006	Confirm a user enters a name at account creation	Capture the name		
CDSS0007	Confirm a user enters an email address at account creation	Capture an email address		
CDSS0008	Confirm a user enters a unique username at account creation	Capture a unique username		
CDSS0009	Confirm a user enters a strong password(mixture of letters,numbers and special characters)	Capture a strong password		



	Number	Percentage of the total	Percentage of the reviewed
Total test cases	219	100.00%	
Total reviewed	219	100.00%	
Total passed	210	95.89%	95.89%
Total failed	9	4.11%	4.11%
Pending review	0	0.00%	

Appendix III: Proposed prototype description

The researcher has designed a CDSS for clinical diagnosis decision support that will serve health practitioners and health policy makers in their works. The proposed CDSS has focus on determining the likely hood of a patient suffering from breast cancer. The patient symptoms are fed in to the CDSS prototype. The CDSS will then search its knowledge base to see if such a set of symptoms have ever been reported and if yes what the final outcome was. In the case of a final outcome being breast cancer, the set of symptoms is stored for reuse in the future. The data captured will provide an efficient reporting on breast cancer for purposes of decision making. The prototype has been deployed in Google cloud and tested. Below are sample screenshots showing the main functionality of the application.

How to run the prototype

The prototype is web based, and is accessible through any web browser on a computer with an internet connection. The only constraint is that the CDSS does not work with versions of internet explorer earlier than I.E 8.0 The URL for the prototype is:

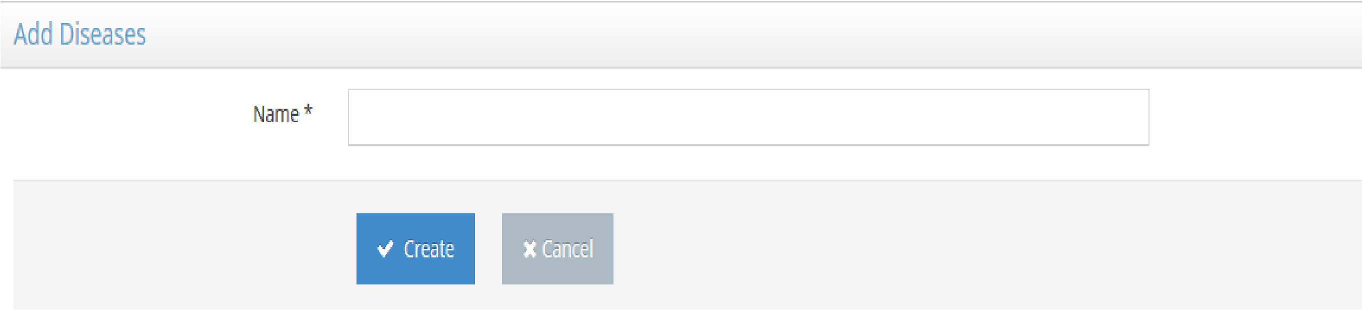
<https://www.galaxyggroup.com/cdss>

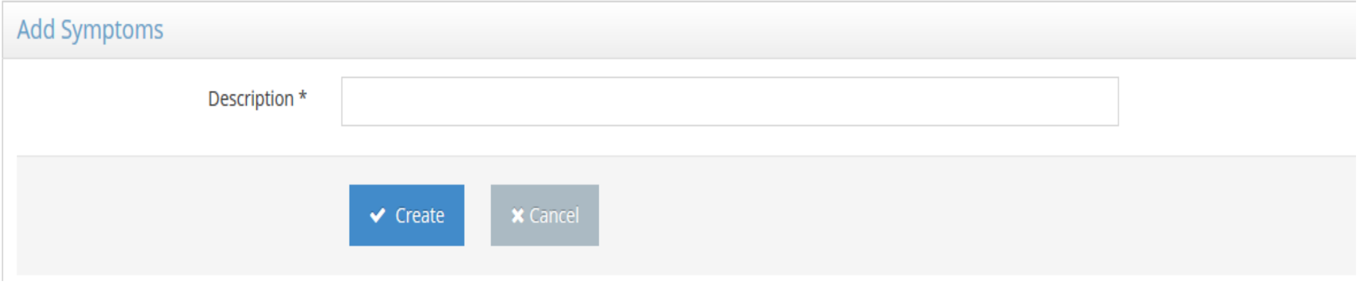
Client Module

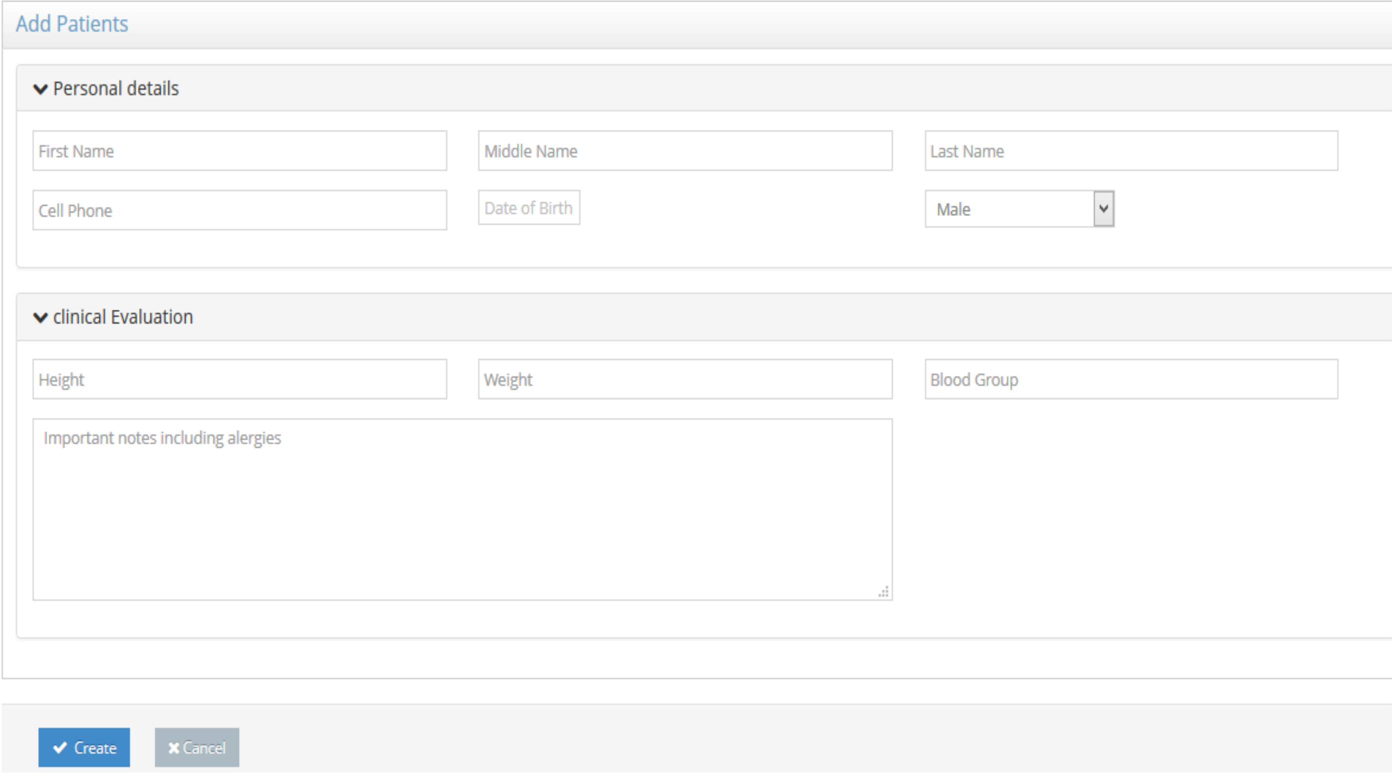
The client module has been developed to collect data about the patient. This module has been developed on PHP`s YII framework on the server side. The module`s forms are developed are developed on html 5 and CSS3 used for styling. The system is designed with the custom functionalities of getting blank forms from the web server to a computer screen using a web browser and also filling the forms and sending the forms to the server. Form validation has been done both on the client and server side. The testing of this module has been done on PHP version 5.6.8.

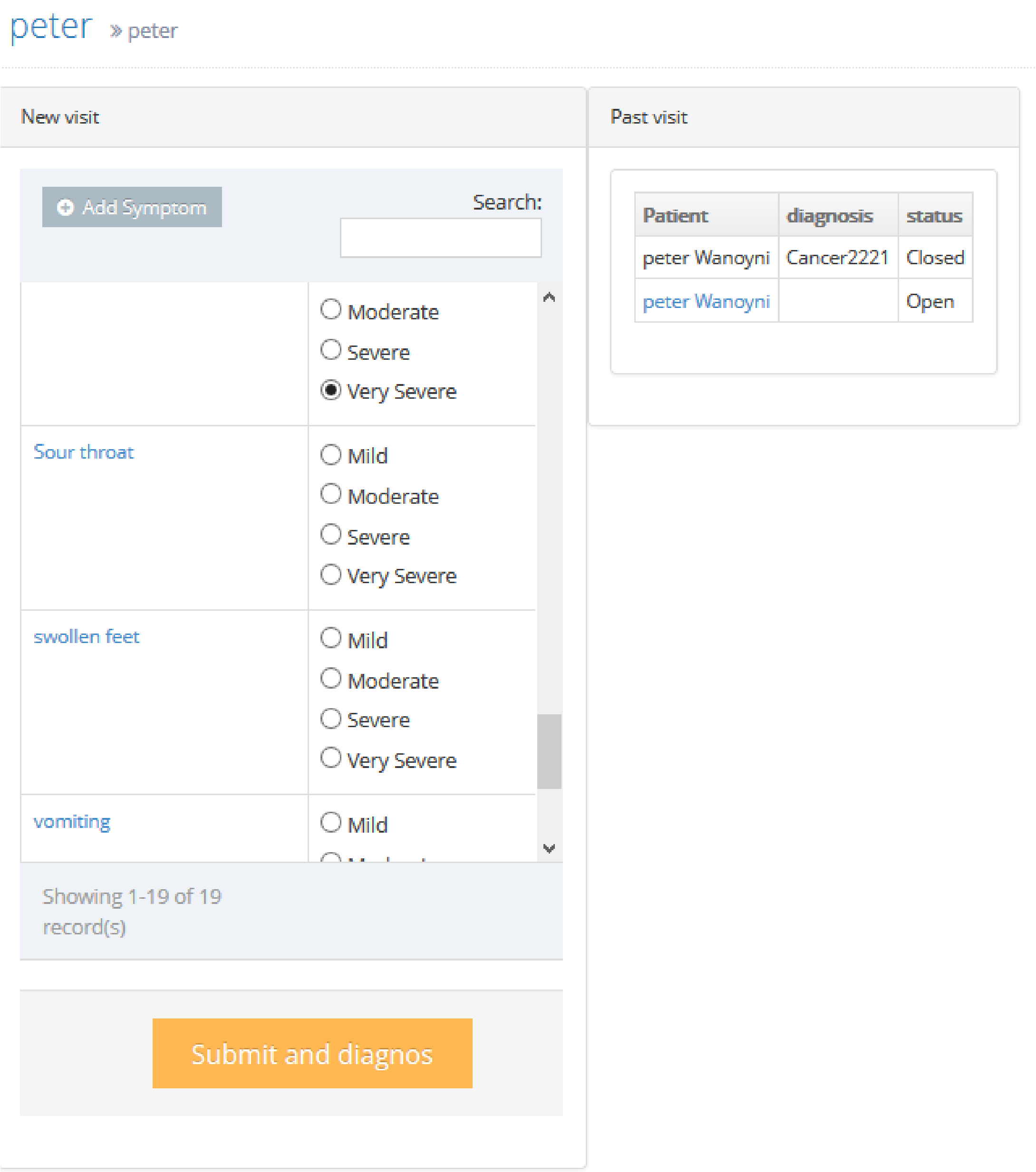
Data gathering forms

The forms are majorly used to collect data from the user or about a patient. The main forms created in this prototype are used in collecting patient details, symptoms, diseases and user details. The patient details form is the major form collecting all the information asked to a patient when he visits a hospital. The terminologies used to label data elements in our forms were found from the sample form reviewed in this study (see Appendix I and Appendix II). However, the terminologies are flexible depending on the kind of data needed to be captured; for example, data related to medical records and diseases. The data elements that are defined in forms are automatically created in the database when the form is submitted. The forms are designed using sublime text editor which offers flexible visual form design. The forms are created and uploaded to the administration module of the prototype. The users then can access these forms through the client module on a computer. The filled data can then be accessed via the administration module where the admin can visualize and process data for management purposes such as decision and policy making. The flexibility of the form design allows the prototype to work with various terminologies (data elements) and allow integration data from different sources (different primary health facilities). The flexibility on terminologies used in data collection forms allows easy way of setting uniformity of data formats and therefore increase coordination of different levels of health information system. Therefore, the proposed prototype will be used to capture health data of various types depending on the demand. Figure 6 .2 shows the sample forms as viewed on a web browser.









○ Moderate
○ Severe
● Very Severe

Sour throat
○ Mild
○ Moderate
○ Severe
○ Very Severe

swollen feet
○ Mild
○ Moderate
○ Severe
○ Very Severe

vomiting
○ Mild

Showing 1-19 of 19 record(s)

Patient	diagnosis	status
peter Wanoyni	Cancer2221	Closed
peter Wanoyni		Open

Figure 7: Sample forms

Form Manager

This is the main window, which contains links to perform different operations on forms. The operations include, getting blank forms from the server, filling blank forms and submitting the forms to the server. Figure 6.3 shows the screenshot of the form manager that is displayed in a web browser.

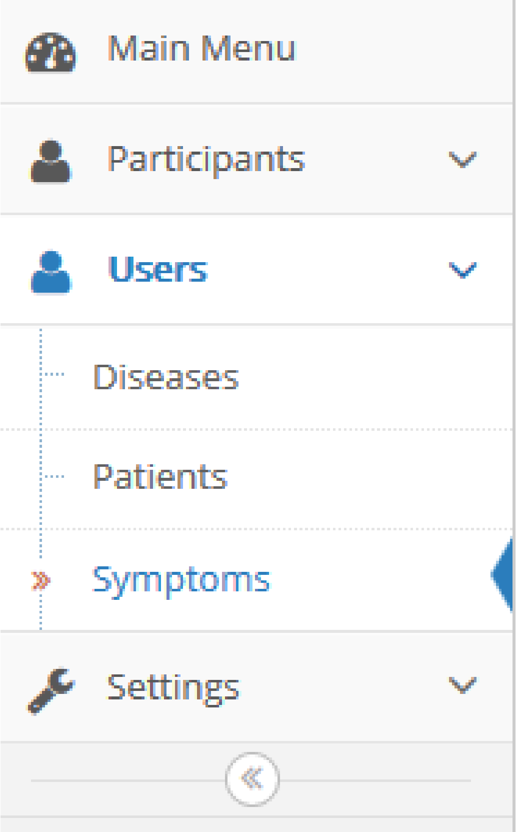


Figure 6.3: A sample screenshot of the form manager menu.

Administration Module

In developing administration module, the apache web server and the MySQL database bundled in Xampp 5.6.8 was used. YII framework provides an advanced module called ‘backendModule’ for developing the admin module of a web application. This module can be customized to function as a fully supported web application.

Backend Module

This is web application that provides interfaces to manage data collection forms and the collected data. The operations on management of data are visualization of data through maps and simple graphs, customizable data filters that allow the generation of custom reports. Furthermore, collected data can be exported to Comma Separated Values (CSV) format and visualized in other applications such as MS excel27. This web interface can be accessed through HTML web browsers. In this prototype, we have tested this application in both computer and phone HTML browsers. It was observed that the phone HTML browser seems to have limitations on the visualization of graphs and charts.