INTELLIGENT MAMMOGRAPHY DATABASE MANAGEMENT SYSTEM FOR A COMPUTER AIDED BREAST CANCER DETECTION AND DIAGNOSIS

Isaac Adusei, Ognjen Kuljaca, Kwabena Agyepong

Alcorn State University, System Research Institute

{iadusei, okuljaca, kwabena}@alcorn.edu

ABSTRACT

To efficiently and intelligently archive, retrieve, and analyze digitized mammogram images in breast cancer detection, a robust knowledge based system needs to be developed to provide decision support to radiologists and researchers. A Mammography database is therefore designed for the development of a knowledge base to support the detection and diagnosis as well as research in mammography. The system also has implemented Computer Aided Detection (CAD) and Diagnosis (CADx) information system of digitized mammogram images to aid in breast cancer detection. In this paper, we will outline the architectural design for the CAD/CADx system and then focus primarily on the design and modeling of the mammography database. The database design combined two standards, the Breast Imaging Reporting and Data System (BI-RADS) by American College of Radiology and the Facility Oncology Registry Data Standards (FORDS) by the Commission on Cancer standards. The complete system is in final development and debugging phase.

KEYWORDS

Breast Cancer Detection, Knowledge-Based System, Decision Support System, Computer Aided Detection, Computer Aided Diagnosis, Mammography, Database Management System.

1. INTRODUCTION

In recent years several new CAD systems appeared in markets in USA and Europe. Also, there are several databases that are available offering access to digitized mammograms for training or information purposes. Probably the best known and oldest one of them is DDSM [4]. DDSM is open data used by many researchers. It does contain some data following available images, but by far not as much as it is available in FORDS [1] and BI-RADS systems [2]. DDSM does not have CAD abilities or search engine. The Mammographic Image Analysis Society (MIAS) database [7] is organized in UK and available to researchers in downsampled form for free and in full a modest fee. Database has only images and malignancy decision, but no other data or search capabilities. AMDI, Indexed Atlas of Digital Mammograms [8] is web based digital mammograms database that allows user to upload and download images, provide statistical comparison methods and content based retrieval of images. However, there are no CAD capabilities. IRMA [9], Image Retrieval in Medical Applications is a project aimed towards development of content based image retrieval systems, but not CAD.

There are few CAD systems in market, but none of them offers broad image search or database capabilities.

The CAD/CADx System is a Computer Aided Detection (CAD) and Diagnosis (CADx) System with a content-based searchable database and a belief network developed to support mammography in clinical practice as well as in research. It is made up of 3 main components (A 3-Tier Distributed System):

i. **SQL Database:** Hosts cancer patient's mammograms, treatments, and other patient records (metadata).
ii. **Applications Engine:** Are server applications for Calcification, Mass, and Architecture Distortion detection.
iii. **Client Applications:** Has four client applications at the moment. System is built in such way that it is possible to add more client applications to the system (both web-based and standalone applications).

The system is distributed with the business logic layer as the applications engine. The Data Access layer is obviously the SQL Database and the Presentation layer hosts the client applications. The database design is capable of providing information on cancer registry and radiology mammography imaging and reporting. It also has the functionalities such as the provision of images and information for the scientists to develop CAD algorithms, allowing radiologists to retrieve and annotate images, and provide other information for statistical analysis too.

To make the entire range of computing devices of the system work together and to have user information automatically updated and synchronized, we have decided to use the Microsoft DOT NET platform for the system development. The programming language we have chosen is C Sharp but C++ is also acceptable.

2. CAD/CADx SYSTEM ARCHITECTURE

The 3-tier distributed CAD/CADx system which comprises various complex components (shown in Fig. 1) is broadly defined with these complexities:

- Application Engine:
  - Calcification Algorithm
  - Architectural Distortion Algorithm
  - Mass Algorithm
- Database
- Application Manager
- Client Applications:
  - Radiologist Application
  - Data Entry (Cancer Registry Data) Application
  - Scanner Application
  - Research Applications

Before we outline to concept of our database management system, we will give a short overview of the other 3 components of the CAD/CADx system namely the "Application Engine", "Client Applications", and "Application Manger".

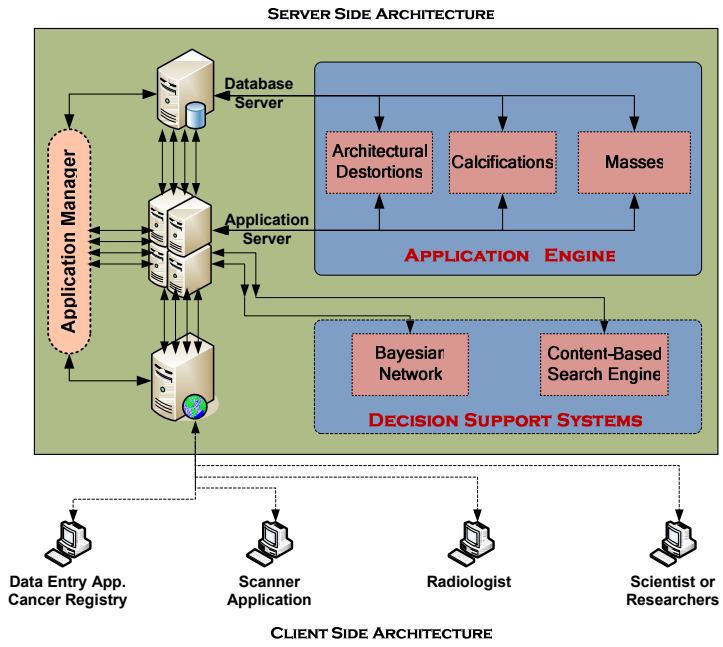

Fig. 1. CAD/CADx System Architectural Diagram

2.1. Application Engine

The application engine correlated the mammographic appearance and pathologic characteristics of different non palpable malignancies diagnosed in varied places. The common division of cancer in imaging research is in three groups based on mammographic presentation [3]: mass, calcification, and architectural distortion which form the engine of the system. They are all image processing mechanisms used to determine mammographic appearance of non palpable breast cancer that may be associated with pathologic variables having prognostic significance, which could influence clinical management.

2.2. Client Applications

This part of the system has four separate applications namely the Radiologist, Data Entry, Scanner, and Research applications; the core application being the Radiologist application. With this application, radiologist can directly scan mammograms or enter cancer registry data or information. The research application deals mainly with robust query-based interface and statistical analysis pertaining to cancer research.

With this application, a radiologist can load images (filtered to display only "TIF" files) into the image pane. Annotation of image is done by mouse. When the lesion type changes, the middle group box will show the functions related to only the newly specified lesion type.

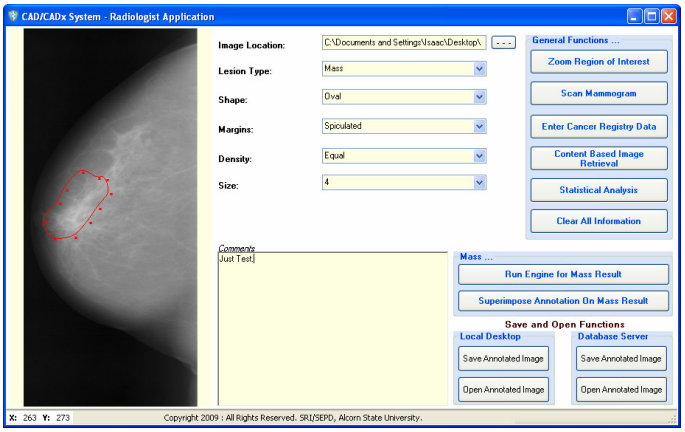

Fig. 2. CAD/CADx System – Radiologist Application

The entire form is also structured to display the parameters for the particular selected lesion (namely: Calcification, Mass and Architectural Distortion). Lesion data fields are manipulated automatically.

Clicking on the "Run Engine …" based on the lesion selection, calls one of the application engines (section III). The Data Entry application is a web-based application that allows user to input already available data from the Cancer Registry. The Scanner application is designed to automate upload of hard copy mammograms that maybe scanned into the database. For the research applications, we currently have implemented two main applications that can enhance research activities in this field. They are the "Statistical Analysis" and "Content Based Image Retrieval (CBIR)" applications. The CBIR function of the interface will take an image under investigation and search the database to retrieve similar images and display them. The initial process is to load an image browsing to the file location. By clicking on the CBIR button, it loads the CBIR window (Fig. 3). The initial image under consideration in the main window is automatically displayed in the CBIR window under the title "Original Image". The statistical analysis makes effective use of data in the database relating to individual cancer patients.

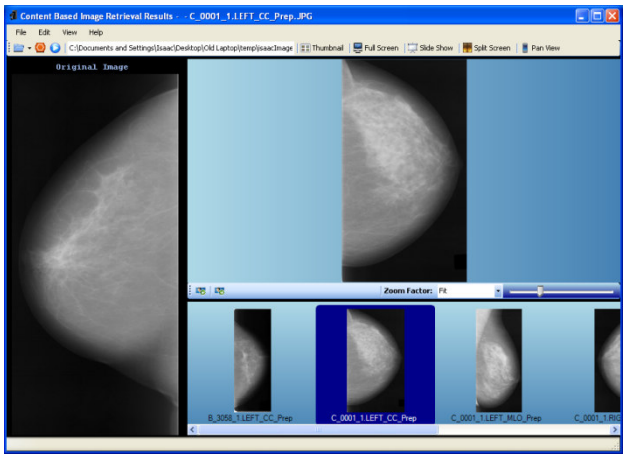

Fig. 3. Content-Based Image Retrieval Application Window

It deals with aspects of collection, analysis and interpretation of these data. Another common goal for statistical analysis tool is to investigate causality, and in particular to draw a conclusion on the effect of changes in the cancer data values of predictors or independent variables on dependent variables or response.

2.3. Application Manager

This is a lightweight multi-threaded application that runs forever when it is started. It acts like a software agents that inter-connect, retrieve and send information among the various software components of the CAD system. The entire CAD/CADx system is currently divided into separate parts (modules) which are being developed by different teams. This manager was a vital piece of software during the prototype stage of the system development. Broadly speaking, the software group and the image scientists work on different parts of the system. The fundamental role of the application manager is to interface the application engine parts with the other components of the software. The manager seeks to avoid the tedious role of re-coding the engine applications in .NET Platform compatible program like C/C++, JAVA, etc. This was very important at the prototype stage of the CAD system where a lot of changes would be made both at the engine side and the other software parts. And to do a re-coding of the engine each time a change is in effect is not only tedious but unworkable in terms of software development. A manager or a tool for the software components to use to communicate with the engine is very important at the prototype stage.

2.4. Database

The mammography database role is to give data support to the CAD/CADx system, by providing a comprehensive package of data including images, patient information, pathology, lesion types, etc., and providing the capability to retrieve these data through criteria set by end users of the system. The database system is able to store and retrieve all images and other data needed by radiologists for cancer annotation and by the scientists for research purposes. It is designed and developed as a Knowledge Base system to support detection, diagnosis and research in mammography. The

ultimate aim is to then use this database for the development of a Computer Aided Detection (CAD) and Diagnosis (CADx) system for the archiving, retrieval, and analysis of digitized mammogram images to aid in breast cancer detection.

In this section we will outline the design and modeling of this database; the main focus of the paper. The design combined two standards, the Breast Imaging Reporting and Data System (BI-RADS) by American College of Radiology and the Facility Oncology Registry Data Standards FORDS) by the Commission on Cancer standards. The current design is capable of providing information on cancer registry and radiology mammography imaging and reporting. It also has the functionalities such as the provision of images and information for the scientists to develop CAD algorithms, allowing radiologists to retrieve and annotate images, and provide other information for statistical analysis too. The database is designed in a hierarchical architecture with three different levels. This makes the entities and relationships in the database flow from general to specific. The design model makes the database more efficient and more scalable. There are 21 tables and 293 attributes in the database. Microsoft Sequel Server 2005 was chosen as the database management system for this project because it meets the security and scalability requirements.

2.4.1. Design Stages – Database

The mammography database has gone through a number of phases in the requirement and analysis stage. In the first design stage, some features from publicly available databases, such as the University of South Florida Digital Database for Screening Mammography (DDSM) [4] were adopted. It was redesigned after meetings with radiologists from the University of Mississippi Medical Center (UMMC). At this stage, a standard widely used by radiologists in cancer detection, Breast Imaging Reporting and Data System (BI-RADS) by American College of Radiology was adopted. After the Facility Oncology Registry Data Standards by the Commission on Cancer (FORDS), the standard used by the cancer registry at UMMC, was introduced, the design of the mammography database was again fine-tuned. The final mammography database was then developed with the cancer registry standard FORDS combined with the radiologist standard BI-RADS, radiologist clinical cancer detection and staging, and the specifications of the film scanner. The design is able to provide the capability for the database system to store and retrieve data for cancer registration, annotation, statistic analysis, research, teaching, etc. The advantages of this design are:

a. The design has been able to combine cancer registry data with the clinical radiology image and reporting categories. In hospitals, these two systems are usually separated from each other.
b. The design provides the capability to list and categorize the images and pathology according to lesion type. The Picture Archive Communication System (PACS system) widely available on the market today does not have this functionality.

This database design is also robust and scalable enough to be extended to other types of cancer detection.

2.4.2. DBMS and System Specifications

The mammography database adopts the relational database design principles. The Microsoft SQL Server 2005 is used as the database management system for the mammography database.

2.4.3. Architecture of the Database

The database design architecture is a hierarchical one. The design encompasses different standards and data from different resources, with three levels being designed into the architecture. Interfaces exist between the different levels. At the top level is general information collected from the cancer registry. This part includes the information on patient, cancer identification, co morbidities, staging and treatment, etc. Since the information held in this part is general, it can be extended to fit other types of cancer diagnosis. At the middle level is the information on breast cancer, reporting, pathology and scanning. At this level the information becomes more specific and narrowed down to each case of breast cancer, while, it is still general compared to information held in the next level. At the lowest level in this hierarchy are the entities related to image and pathology of three lesion types, architecture distortion, calcification and mass. Each case has four scanned images. The information held at this level is specific to each image, such as laterality, view, shape, distribution, boundary of annotation, number of abnormalities, etc. Fig.4 shows the overall architecture of the database. The hierarchical architecture makes the entities and relationships in the database flow from general to specific. The "IS A" parent and child design model makes the database more efficient and more scalable. Currently there are 21 tables and 293 attributes in the database. Fig. 5 depicts the entities and relationships in the mammography database.

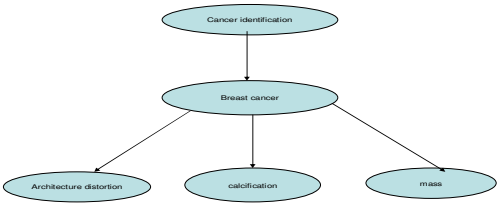

Fig.4. Hierarchical architectural Design of the database.

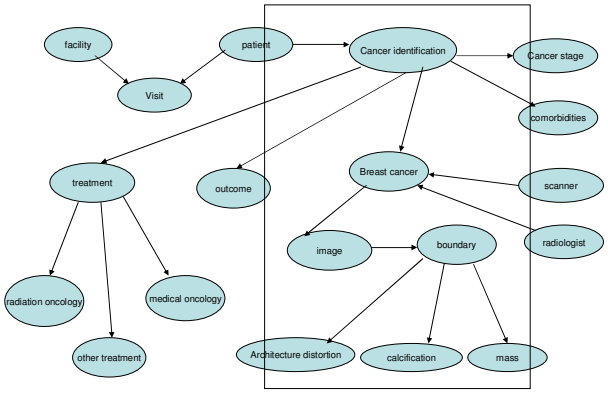

Fig. 5. Overview of the entities and relationships in the mammography database

2.4.4. Entity-Relationship Diagrams and Design Models

The mammography database design followed the general steps and diagram models of database design. The whole design process went through the conceptual design, logical design and physical design. At last, tables were created in the MS SQL Server 2005, the DBMS chosen for this project.

2.4.5. Conceptual Design

Database design usually starts with conceptual model. At this level, the process is software and hardware independent. The details of the actual physical implementation of the database do not need to be considered. The conceptual model represents the views of data, relationships between data in the real world. At this stage in our database design, all significant entities and attributes were included. Candidate keys were identified, but they did not explicitly include a complete scheme of identity.

2.4.6. Logical Design

Fig. 6 shows the ER diagram for the logical design of the mammography database. It contains the full population of entities and attributes. Logical data types were defined and identities were selected. Propagation of identifiers as foreign keys was explicitly represented in this model. The relational database does not offer direct support for many-to-many relationships. At this stage many-to-many relationships were resolved into associative entities. For example, in the relationship between the Patient and Facility entities, one patient may go to multiple facilities and one facility may have multiple patients. Thus, the relationship between patient and facility is many-to-many. To solve this problem, an intermediate table, *Visit Table* is created.

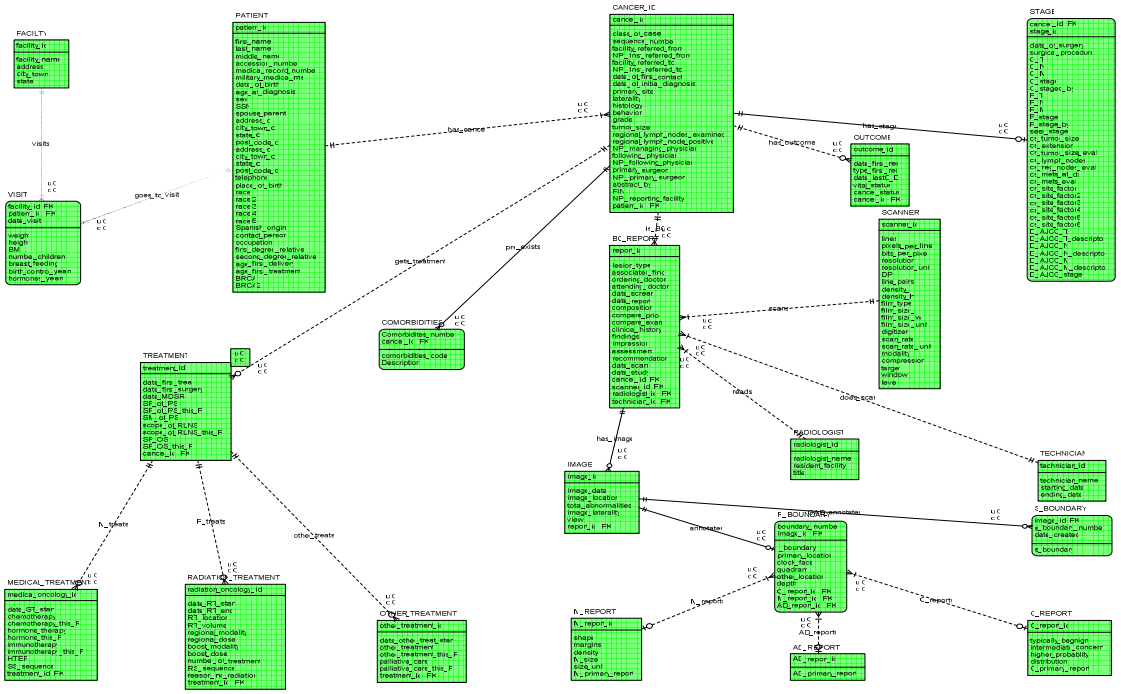

Fig. 6. IDEF1X paradigm for the logical model of the mammography database

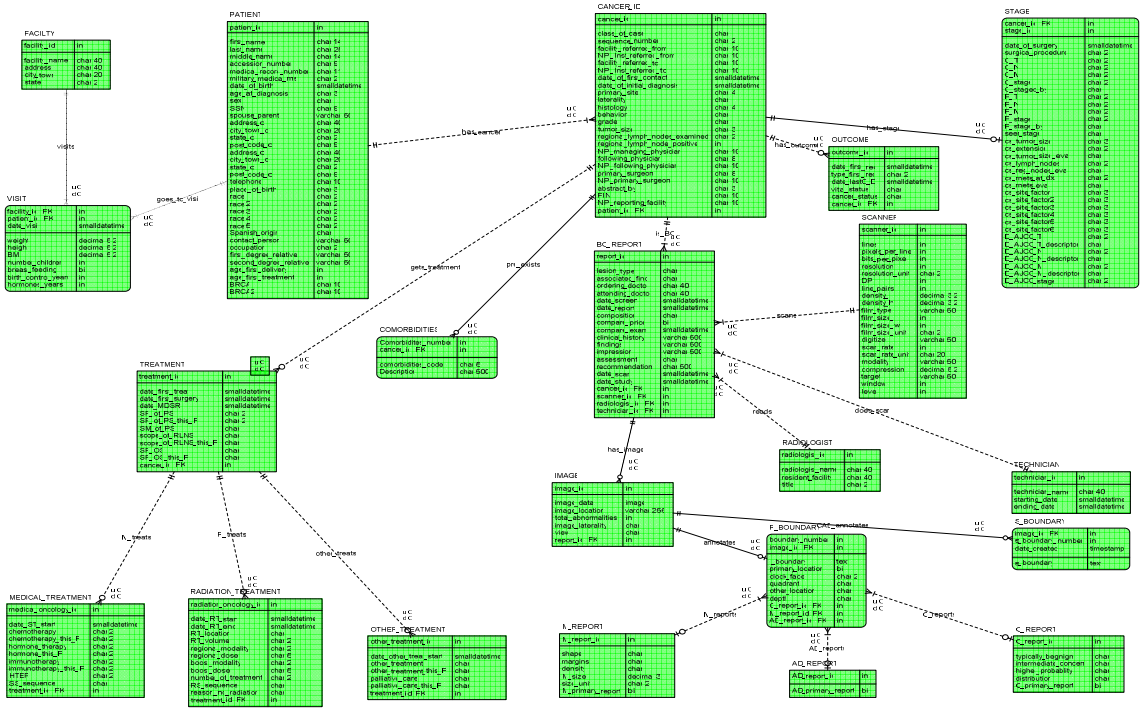

Fig. 7. IDEF1X paradigm for the physical model of the mammography database

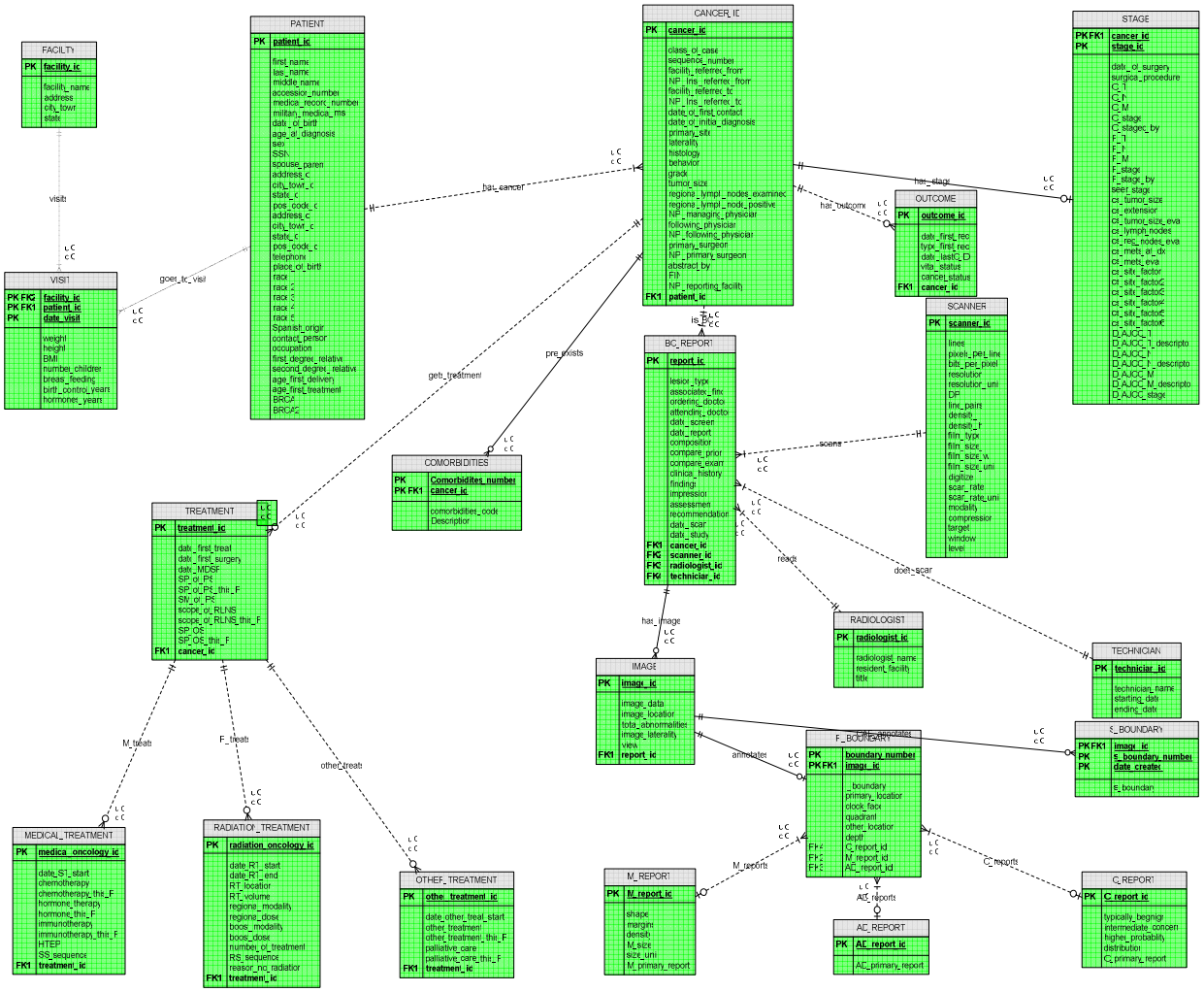

Fig. 8. Relational view of the mammography database design

This new table joins two tables in the middle and encapsulates properties from both tables. It holds identifiers--primary keys that combine both tables.

2.4.7. Physical Design

At the physical design stage, the database needs to specify details of the physical implementation of data in the chosen DBMS. Fig. 4 gives the ER diagram for the physical data model of the mammography database. At this stage, data types in the logical model were converted to be SQL Server specific, the indexes were constructed, constraints were created, and the referential integrity was achieved. Fig. 5 gives the relational view of the design. Fig. 6 to Fig. 10 give more detailed views of the physical model.

2.4.8. Normalization and Choosing Primary Keys

In the design of mammography database, normalization was done extensively through to the third normal form. At the first normal form, repetitions were removed and atomicity was achieved by moving any repetition of groups of data into new tables. In the second normal form, repetition of data was further reduced with each column depending on the whole key. In the third normal form, derived data was eliminated and columns were checked to make sure that no column depended on a non-key column.

In the design, besides the general rules for database design, some special rules that are specific for the mammography database were taken into account. This was reflected especially in choosing the primary keys. Mammography database is a medical records related database.

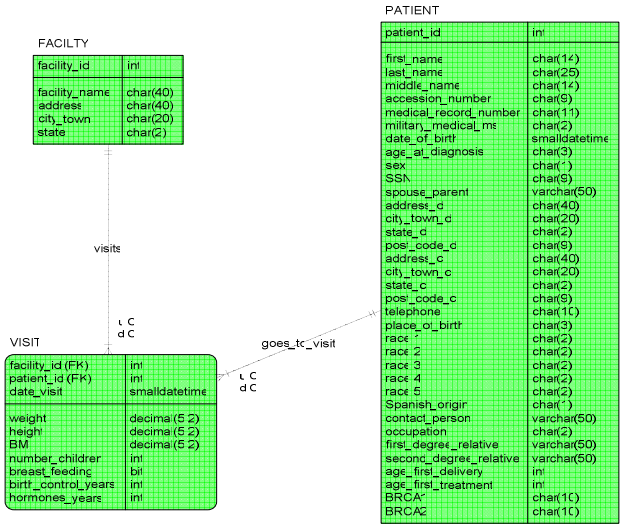

Fig. 9. Physical model for PATIENT, FACILITY, and VISIT table

It collects the attributes related to patient general information, patient's pathology, diagnosis, mammography images, etc. When collecting and managing the kind of information, confidentiality has to be given according to Food and Drugs Administration standards. So, the candidate keys that can be generally used as the primary identifiers in other database were not used in the design.

2.4.9. Constraints, domains and referential integrity

In the following examples, constraints are put on the primary and foreign keys, and indexes are constructed. Referential integrity is achieved through cascading actions.

2.4.10. Tables

The following screen images show the tables created in SQL Server2005 for mammography database.

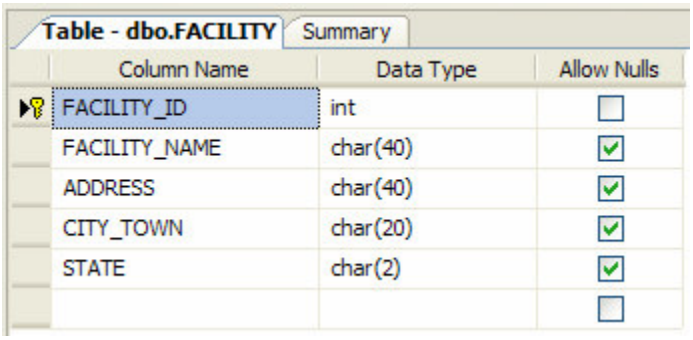

Fig. 10. Example Table - Facility

2.5. System Integration

The systems integration part of the project involves the tedious process of bringing the various part or components of the CAD/CADx system into conformity. It is the process of linking together different computing systems and software applications physically or functionally. We integrated discrete systems utilizing a variety of techniques such as computer networking, enterprise application integration, business process management or manual programming. According to [5], there are basically 3 methods of integration namely; Vertical, Star and Horizontal Integration. In the prototype stage, the horizontal integration approach was used whereby the application manager acts like a specialized subsystem is dedicated to communication between other subsystems. This approached proved to be very helpful since the application engines were still underdevelopment and a lot of changes were being made frequently. Afterwards, when the application engines were all re-coded from MATLAB to C-Sharp, we adopted the star integration approach, where each system is interconnected to each of the remaining subsystems [6].The seamless connection of the database with the client applications and the application engine was done with the .NET platform.

3. CONCLUSIONS

The CAD/CADx system under development and described here seeks to add additional complexities and functions and also improve upon the already available systems in the market today. Although the usefulness of these systems are based solely on the accuracy of the application engine algorithms to detect breast cancers, we have added additional components like high-level user interface which has many functions for both diagnosis and research tasks. Capabilities like these are all added:

i. Ability for radiologists to annotate mammograms under examination.
ii. Statistical analysis tool.
iii. Query based interface to identify and selected particular category of information from the database.
iv. Content-Based Image Retrieval ability for researchers and even radiologists to select similar mammograms of the image under investigation from the database.

Others like the Bayesian network to add more intelligence to the CAD/CADx are currently been developed. The uniqueness of CAD/CADX system under development is that encompasses features usually found in mammograms databases with features usually found in CAD systems and also adds deep knowledge database with patient metadata. System will help in diagnosis and screening as well as with research by combining features that exist today mostly as independent modules.

REFERENCES

[1]: FORDS – Faculty Oncology Registry Data Standards, Revised for 2009 and 2010; Commission on Cancer.

[2]: BI-RADS – Breast Imaging Reporting and Data System by American College of Radiology.

[3]: M. P. Sampat, M. K. Markey, A. C. Bovik, "Computer-Aided Detection and Diagnosis in Mammography", Handbook of Image and Video Processing, 2nd edition, 2005

[4]: University of South Florida Digital Mammography – URL "http://marathon.csee.usf.edu/Mammography/Database.html" as of April 2010. Digital Database for Screening Mammography – DDSM.

[5]: CIS 8020 - Systems Integration, Georgia State University.

[6]: Wikipedia Homepage, "http://www.wikipedia.com".

[7] MIAS Database, http://peipa.essex.ac.uk/info/mias.html as of May 2010.

[8] AMDI, Indexed Atlas of Digital Mammograms, "http://www.lcc.ufu.br/amdi/" as of May 2010

[9] IRMA, "http://ganymed.imib.rwth-aachen.de/irma/index_en.php" as of May 2010.