-

CURA CANCER GENOMICS

Integrated Expert Systems for Cancer Recognition and Treatment

BY

IHAB ESSAM, AHMED ABDALLAH, AMANI BASHER, MUHAMMAD KAMAL, HATEM SAYED, KHALED MOHAMED, AHMED SHA'BAN, AND MOHAMED ZAKARIA

UNDER SUPERVISION OF

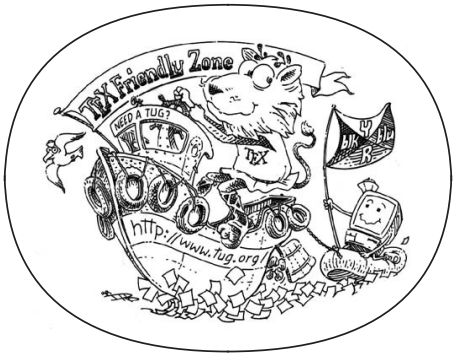
PROFESSOR ALAA ZAGHLOUL, AND T.A. NEHAL A. MOHAMED

(refer to **??** for more information)

[January 14, 2019 at 17:51 – classicthesis v4.6]

CURA CANCER GENOMICS

-



-

Jan 2019 – classicthesis v4.6

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

THEORETICAL BASIS

1 INTRODUCTION

An integrated collection of expert systems, with the aim of the identification, recognition, diagnosis, patient follow-up and help in the treatment journey for three main types of cancer: **brain**, **breast** and **gastric (stomach)** cancer, is proposed to serve both oncololgists who seek to automate a part of the general practicioner's job and use a system that will help along the journey of diagnosis, follow-up and treatment, and patients who want their questions answered starting from the very first phases before the confirmation of the illness to later phases of treatment (e.g: chemotherapy), as well as keeping up with how their health is doing so far in an easily accessible way. The system is planned to be available for both oncologists and patients through multiple applications in a way that will make it possible for both parties to access crucial data in the right times.

MOTIVATION

By the end of 2018, an estimated 1,735,350 new cases of cancer had been diagnosed in the United States alone with an estimated 609,640 people might have died as a result of the disease. By observing different cases from 2011 to 2015, an estimated 0.00439% (around 439 per 100,000) of men and women are diagnosed with cancer each year, while during the same period, the estimated number of deaths resulting from cancer was 163 per 100,000. Death as a result of cancer is slightly higher in men with an esti-

mated 197 per 100,000 compared to an estimated 140 per 100,000 for women.

According to data collected in the period of 2013-2015, it is estimated that 38.4% of both men and women will get diagnosed with cancer at some point in their lives.

Cancer has impacted societies around the world in different ways, the numbers aforementioned show approximately how impactful cancer had been and it serves as a wake up call for people who never thought of cancer as a danger they might have to anticipate at some point in their lives.

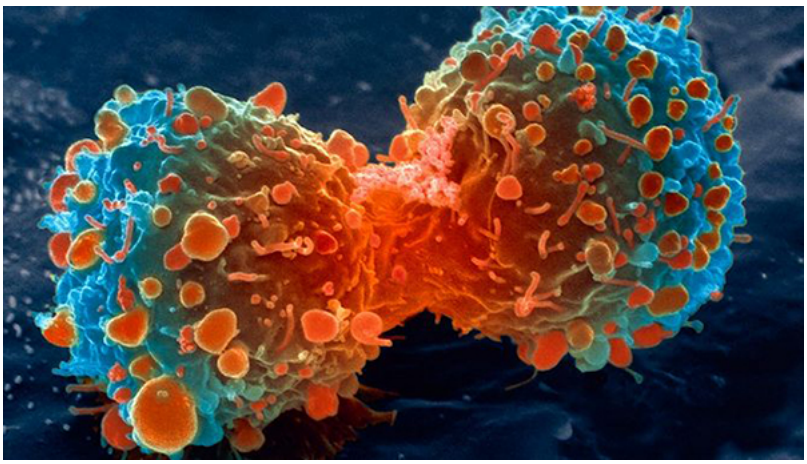


Figure 1.1: A dividing lung cancer cell.

In the most basic terms, cancer is an umbrella term for a group of related diseases that all share a common characteristic: an unstoppable division of the body's cells that spread into surrounding tissues. Cancer can basically start in any of the trillions of cells that make up the human body; normally, human cells grow and divide to form new cells, but the cells grow old or somehow get damaged, they die, and new cells take their place.

Glycation (or non-enzymatic glycosylation) is the result of a covalent bond formed between a sugar molecule (i.e: glucose or fructose) to a protein or

lipid molecule, without the controlling action of an enzyme. Glycation may either occur inside the body (endogenous glycation) or outside the body (exogenous glycation).

Excogenous glycation refers to the formation of **Advanced Glycation Endproducts (AGEs)** when sugars are cooked with proteins and fats; typically temperatures over 248 $°F$ ($\sim$ 120$°C$) greatly spur the reactions, that being said, lower temperatures with longer (or slow) cooking also promote the formulation. Such compounds are absorbed during digestion with about 10% efficiency; browning effects are evidence of pre-formed glycations, for example: sugar is often an ingredient in many french fries and baked good recipes to promote browning. Glycation may contribute to the formation of acrylamide a potential carcinogen, during cooking.

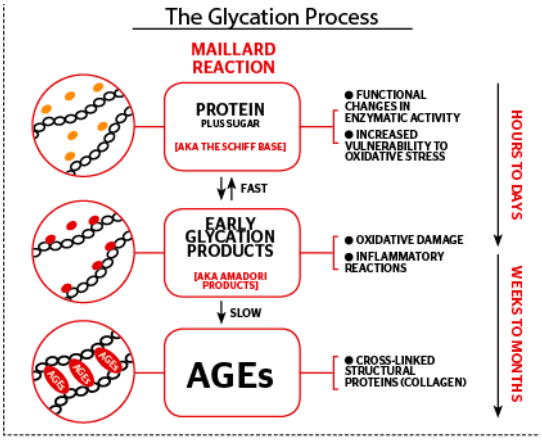


Figure 1.2: The Glycation Process

Endogenous glycations describe the type of gylcations occuring mainly in the bloodstream to a small percentage of the absorbed simple sugars (e.g: glucose, fructose and galactose). The effects of fructose were formely drastically underestimated due to the use of inaccurate assay techniques, but it turned out to have effects that are approximately ten times worse

than the effects of glucose which is considered the primary body fuel. Some AGEs are benign, but some are much more effective than the sugars they are derived of. AGEs contribute to the development of cancer, along with a massive list of other diseases including cardiovascular diseases, usually when covalent bonds are formed between the sugars and collagens. The list goes on to include Alzheimer's disease as a result of the formation of amyloid proteins as a side-product of AGEs and deafness due to demyelination. Long-lived cells, such as nerves, long-lasting proteins (such as crystallins of the lens and cornea), and DNA may accumulate substantial damage over time. Damage by glycation results in the stiffness of collagen in the blood vessel walls, usually leading to diabetes.

A telomere is a region at the end of a chromosome, often called the tail of the chromosome. Telomeres usually contain repetitive nucloetide sequences to protect the end of the chromosome from deterioration or from fusion with neighbouring chromosomes. During the process of chromosome replication, the enzymes that duplicate DNA can not continue their duplication all the way to the end of the chromosome, so each time the end of the DNA is shortened, this is mainly contributed to the synthesis of Okazaki fragments which require RNA primers attaching ahead on the lagging strand. Basically, telomeres are truncated during cell divsion; their presence protects genes contained in the chromosome from being truncated instead. The telomeres themselves are protected by shelterin proteins along with the RNA that telomeric DNA encodes (known as TERRA). When telomeres get too short, the cell can no longer divide

resulting in it being inactive or dead; this shortening process is associated with aging, cancer and death.

When cancer develops, the orderly process of cell division is broken down; cells become more and more abnormal, damaged cells survive when they are supposed to die, and new cells are formed when they are not needed, these new cells divide continuously without stopping and may form tumours.

Cancerous tumours are malignant, meaning they can spread to nearby tissues, and as tumours grow, some cancer cells break off and travel to distant places in the body through the blood or the lymph system and form new tumours in places other than the origin of the tumour. In contrast with malignant tumours, benign tumours do not spread to nearby areas, and while they can be quite large, once they are removed they usually do not grow back; most benign tumours are safe except the ones that grow in the brain.

Cancer cells are not as specialised as normal cells, meaning that normal cells mature into very distinct cell types while cancerous ones do not, that is why cancerous cells keep dividing in an unstoppable fashion. That is in addition to the fact that cancerous cells are programmed to ignore signals that tell cells to stop dividing, a process called programmed cell death, or apoptosis, which the body utilises to get rid of unncessary cells.

The microenvironment, or the area surrounding and feeding a tumour could be hugely influenced by cancer cells that may be able to affect the normal cells, molecules and blood vessels. As an example of such an effect, cancer cells could prompt normal cells to form blood vessels that supply the tumours with oxygen and nutrients which they need to grow. Cancer

cells can dodge the immune system and specialised cells that protect the body from infections and other conditions.

Being a genetic disease, cancer is caused by changes to genes that control the way our cells function, especially how they divide and grow. Such genetic changes could either be inherited from the patient's parents or from errors during the patient's lifetime that occur as cells divide or as a result from DNA damage caused by certain environmental exposures, e.g: tobacco smoking, consumption of unrefined sugars and grains, radiations such as UV rays from the sun.

SYSTEM DESCRIPTION

The system is composed of several subsystems that interact with and complement one another to help patients recognise the symptoms of cancer early on, follow their progress and have their questions about chemotherapy answered, while helping doctors to organise and easily access patient's records, support in the diagnosis, treatment and postoperative chemotherapy. The following is a description of each of the proposed subsystems.

Subsystem 1: Cura General Parctitioner

The first subsystem is aimed at the patient before the official diagnosis by a doctor. Cura General Practicioner is a rule-based expert system built with the support of natural language processing to help interact with the patients to figure out if the symptoms they are experiencing can potentially be diagnosed as cancer or not.

If the system suspects the patient is in serious need of help because they may have one of the three types

of cancer aforementioned, they are forwarded to the second subsystem and are referred to an oncologist on the system, otherwise they are recommended basic cautionary tips and may be recommended with the addresses of nearby doctors specialising in the suspected illnesses they may have. The result of the session the patient has with the system is a detailed report sent to the suggested oncologist before progressing to the second subsystem.

Cura GP is composed of a set of interrelated subsystems, the principal of which are:

1. the reasoner: a rule based expert consultant that forms the basis of the system and is the main contributor to the medical report sent to the oncologist.
2. the natural language processor: a component that helps interpret user messages and makes the communication experience a little bit more friendly and easier for the patient.
3. the interviewer: the user interface of the system which combines modern UI technologies with both the NLP and Rule Based System components to achieve the most optimal results.

Overview of the Problem Domain

Cura GP is designed to assist patients in the recognition of the symptoms they might be experiencing and how they might impact their chance of treatment. Cancer is a general term describing a variety of related diseases having different prognoses and natural histories.

The domain of this particular expert system is not exactly typical; it is mainly concerned with the symptoms of the three types of cancer aforementioned:

brain, breast and gastric cancers, along with all possibly related illnesses; for example in the case of gastric cancer the system also deals with the symptoms of gastritis, ulcerative colitis and possibly related gastric diseases. The resources used to help articulate and describe such a domain include various researches detailed in the reference section of this book.

Overview of the NLP Brain

The brain of the natural language reasoner is developed using Rivescript, an excerpt of the brain is shown as follows:

Listing 1.1: Excerpt of Cura GP Brain

```
! var name=cura
! array bodypart= breast brain stomach
! array paintype = burning crampy dull Sharp
! array painspan= sudden ongoing episodic Steady
! array painlocation= lower abdomen|upper abdomen|middle abdomen
! array paintrigger= stress|drinking alcohol|eating certain food|
    Coughing or other jarring movements
! array painrelief=Antacids|changing position|drinking more water
< begin
+Describe how you are feeling right now{topic=pain}
>topic pain
+[*](@bodypart)[*]
- in which area does it hurt <set pp=<star>>
+[*](@painlocation)[*]
-Describe how is the pain like <set pl=<star>>
```

Cura CBR

Cura CBR is designed with the purpose of assisting oncologists in the follow-up and analysis of the cases of different patients based on currently used therapy protocols. Its aim is to extend our knowledge of cancer at the molecular and cellular levels, both theoretical and empirical, to achieve a system with a proper

knowledge base that will be as knowledgeable as the current technology allows.

Cura CBR system is proposed to be built based on research on hybrid neurosymbolic models, and motivations for such an approach are:

1. Cognitive processes are heterogenous – a large pool of representations and techniques are used.
2. A proper intelligent system can benefit greatly from the use of a combination of different techniques, since a single technique can not do everything.

Classical neurosymbolism deals with the integration of neural networks and rule based expert systems. The approach we are using is different as it utilises the same techniques to combine neural networks with a combination of rule and case based reasoning techniques.

Neural networks will form the basis for the entirety of the case-based system including the indexing, retrieval, adaptation and learning subsystems.

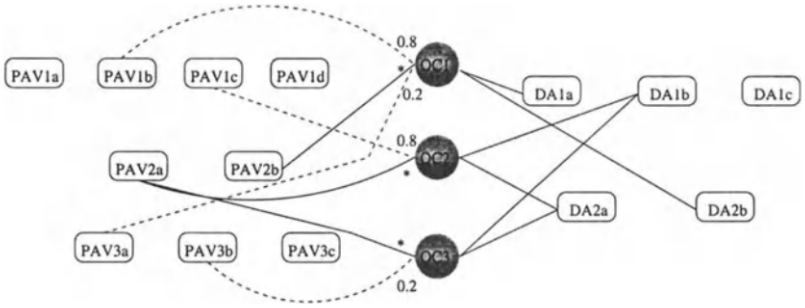


Figure 1.3: Neural Networks for CBR

As shown in figure 1.3, a connectionist implementation of the CBR is proposed, three types of neuron are used:

1. *PAV* neurons each represent an input attribute.

2. OC neurons correspond to stored cases as the hidden layer neurons.

3. DAC each correspond to the neurons describing the decision features.

Weighted connections exist between the input and hidden layers; a weight corresponds to the importance of the attribute for the determination of the input case; when a new case is introduced to the system, the hidden layer is activated by the weighted sum of the input signals, if the sum is above the threshold then the output layer will be activated.

- The input layer contains a neuron for each attribute value.
- the hidden layer contains m groups of neurons. Each group has l neurons and neuron A_{ik} has n_i connections from the neurons $a_{i1}, \cdots, a_{in_i}$in the input layer. Weights connecting $A_{ik}$to neuron a_{ik} to neuron a_{ij} are given by:

$$W(A_{ik})_j = P(A_i = a_{ij}|C = c_k) \tag{1.1}$$

$W(A_{ik})_j$ is the probability that the attribute A_{ik} takes the value a_{ij} given that the observed patient matches a previous patient (or case) c_k. The activation of A_{ik} is achieved by:

$$S(A_{ik}) = \sum_{j=1}^{n} W(A_{ik})_j \times S(a_{ij}) \tag{1.2}$$

- The output layer contains l neurons and each correspond to a previous case c_k. The activation

is computed by:

$$S(c_k) = \theta_k \times \prod_{i=1}^{m} S(A_{ik}) \tag{1.3}$$

where $\theta_k = P(C = c_k)$ is a constant stocked in the neuron c_k.

This approach of forward propagation allows to perform retrieval of similar cases; the hidden layer corresponds to similarity measures. As for adaptation, the probabilities can be retro-propagated which allows memorised cases to contribute to the adaptation process. The retro-propagation implies updating activations in the hidden layer as follows:

$$S(A_{ik}) = \frac{S(c_k)}{S(A_{ik})}$$

Activations of input layer neurons are updated using the following formula:

$$S(a_{ij}) = S(a_{ij}) \times \sum_{k=1}^{l} W(a_{ij})_k \times S(A_{ik})$$

where $W(a_{ij})_k$ is the weight from A_{ik} to a_{ij},with a value of $P(a_{ij}|c_k)$.

SYSTEM REQUIREMENTS

Functional Requirements

1. **Interface Requirements:**
   a) All interfaces shall be easy to use.
   b) NLP shall be implemented in the first subsystem to aid users in their description of their symptoms.

2. Cases shall be implemented in such a way that appropriate cases can be easily and correctly retrieved by oncologists.
3. Patient reports resulting from the general practitioner shall be sent in a proper form to the oncologist prior to the CBR subsystem.
4. The system shall be flexible as to accomodate patient follow-ups and visits.

Non-Functional Requirements

1. **Usability Requirements:**
   a) Minimising impact of errors
   b) Adapting to user needs
   c) Improved human-computer interaction through better UX.
   d) Speed
2. **Efficiency Requirements:**
   a) Performance
      i. Transaction processing
      ii. User/Event response time
      iii. Screen refresh time
   b) Space: the site requirement spaces is good enough to handle website
3. **Dependability Requirement**s:
   Users shall enjoy all their privacy rights restricting unauthorised access to their records and data.
4. **Security Requirements:**
   The implementation and use of top-notch security algorithms to protect user's sensitive and personal data.

5. **Scalability Requirement**:
   The system shall be scalable to a very flexible degree in its various subsystems where oncologists are encouraged by means of an economic game to add their data sets, and by the natural way the CBR system learns over time from follow-ups with patients along their treatment journey.

6. **Operational Requirements:**
   System servers shall be up and running around the clock to help both oncologists and patients access crucial data at all times.

7. **Development Requirements:**
   The system shall be flexible and extremely dynamic.

8. **Implementation Requirements:**

   a) Parallel processing across server and client.

   b) Support for different languages.

   c) Cross-platform support.

   d) Database integrity.

9. **Fault Tolerance:**
   The system shall be able to continue operating properly in the event of the failure of one or more of its main components. A high-availability approach shall be used.

10. **Backup**

11. **Interoperability:**
    The different components of the system shall be interoperable with one another, and with other systems as well, through a properly developed API.

12. **Testability:**
    Different components of the system shall be subjected to different kinds of tests from unit tests to integration tests.

13. **Integrability:**
    The different components of the system shall be integrated easily.

ADVANTAGES AND DRAWBACKS

Advantages:

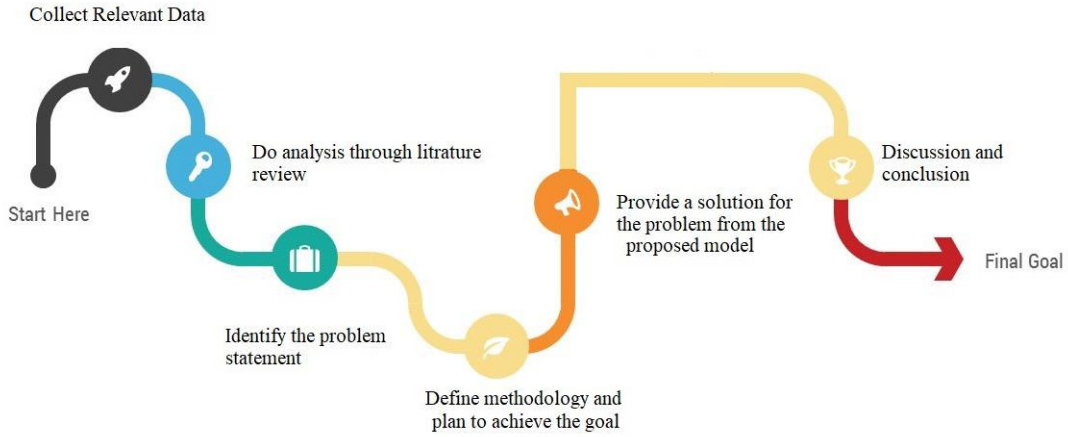
1. Easily accessible for patients and oncologists.
2. CBR learns more about different patients and cases over time.
3. The system is fed with more information over time, hence the system will develop further as the knowledge about cancer accumulates and develops over time.
4. The system can still reason with missing information about the patient by piecing together data about similar patients.
5. Oncologists can record measures of success or failure of previously suggested solutions which the system uses to know what caused those failures and predict future failures.
6. Economic games are employed to incentivise oncologists to share data sets that are relevant to cancer research.

Drawbacks:

1. The system does not cover all known types of cancer.

2. The system might not be convenient for patients throughout the entire recognition and treatment process.
3. The system might not be as efficient right from the beginning, but this is solvable with the progression of time.

RESEARCH METHODOLOGY



Research Strategy

The research done in this project is rather an applied one, based on numerous previous academic and industrial research spanning across numerous fields including bioinformatics, mathematical biology, blockchain among others. As such, our research took the form of a new improvised and revised version of a collection of previous researches in the aforementioned fields.

Goal

In an endeavour to help oncologists do a better job managing, analysing, treating and following up with their patients' health, we set out to develop and implement a proper set of tools built based on years of research in fields like mathematical biology, bioinfo-

matics, computer science, artificial intelligence and medical research by dozens of passionate scientists in all of the aforementioned fields. And thus to realise this mission we set out a goal for CuraCG is to extend our knowledge of cancer at the molecular and cellular level, both theoretical and empirical, to achieve a system with a proper knowledge base that will be as knowledgeable as the current technology allows.

Objectives

1. Implement a general practitioner that helps diagnose the user by means of an interactive NLP-powered chatbot.
2. Implement an expert system that concludes the major communication link between the oncologist, the patient and his visits, along with a case-based reasoning component that creates knowledge based on cases from its casebase.
3. Allow oncologists to monitor, moderate and follow-up patient treatment and gain insights on its effectiveness.
4. Provide a collaborative platform for oncologists and scientists using Blockchain technology, incentivising them to share their datasets and vote on current ones.

Research Method - Qualitative versus Quantitative Techniques

In order to carry out research that would be beneficial to the point where it would help to achieve the objectives listed in 1.5.3, a mixed approach of both qualitative and quantitative research was ap-

Qualitative	Quantitative
The aim is providing detailed, complete descriptions.	The aim is classifying features, counting them and constructing statistical models to explain the observations.
Recommended during the early phases of research.	Recommended in later phases in the research.
The design emerges as the study unfolds.	The study is designed carefully before data is collected.
Data is in the form of words, pictures or objects.	Data is in the form of numbers and statistics.
Subjective	Objective
Time consuming, and less able to generalise to independent data	More efficient, and could be cross-validated

Table 1.2: Qualitative versus Quantitative Research

plied. Briefly, qualitative research could be the approach to use if the sample size is appropriately small, but its outcomes are not measurable nor quantifiable. Yet, it offers a complete description and analysis of the topics being examined. Quantitative research focuses on classifying features, counting them and constructing statistical models to explain the observations.

Research Questions

1. How widespread are the cancers being considered (brain, breast and stomach)?
2. How significant are the symptoms in recognising cancers in early stages?
3. How important is it to discover the cancer in an early stage?
4. What are the environmental factors affecting the considered cancers?

5. What are the words most associated to each of the considered cancers?
6. What are the features of each cancer most useful for indexing in a case-based system?
7. What are the features most interesting in an MRI/endoscopy?

2 BACKGROUND AND RELATED WORK

This project draws upon several diverse fields including bioinformatics, computer science, blockchain, nudge behavioural psychology and game theory. The purpose of this chapter is to give a general sense of how each of these areas affected the project, as well as point to similar projects that gave rise to this project.

This chapter discusses the approaches used to recognise and treat cancer, along with motivational incentives for oncologists and specialists to willingly share their relative data sets.

BEHAVIOURAL PSYCHOLOGY

Cura's system is particulary interested in nudging, in which small changes are introduced to an individual's choice set to push individuals towards certain behaviour without much change to their current routine. Usually employed techniques include

- **Anchoring Bias:** people tend to be heavily biased by the initial data points presented to them
- **Availability Bias:** people tend to overestimate the likelihood of things that are easy to remember
- **Representativeness Bias:** people tend to see patterns when there are none
- **Overestimation:** almost all individuals think they are better than average
- **Loss Aversion:** people tend to prefer avoiding losses over acquiring equivalent gains

- **Status Quo Bias:** behavioural inertia must be overcome when changing a habit
- **Framing Effect:** people react to a particular choice in different ways depending on how it is presented
- **Priming Effect:** people can be influenced to make a certain choice based on what they see or experience directly before making their choice
- **Hyperbolic Discounting:** people prefer rewards that happen sooner rather than later, even if the rewards have the same actual value

BLOCKCHAIN

Blockchain technology has been the revolution recently, and the introduction of Ethereum was especially revolutionary because it had the goal of building decentralized applications set from the beginning, as opposed to Bitcoin, that mainly focused on financial transactions. Decentralized applications, as promising as they are perceived to be, are in fact very limited and uncompetitive against traditional centralized applications.

The use of blockchain in CuraCG's projects is to create an incentive-based project to help grow the scientific community built on top of the CuraCG's subsystems. A special token named CuraToken (CT for short) is issued at a certain rate, which is to be discussed in CuraTherapy's white and yellow papers, to incentivise oncologists to share their data sets, especially the ones they built prior to them starting to use CuraCBR in order to have an evergrowing repository of datasets related to cancer and related diseases for the reasons aforementioned including better understanding of cancer on molecular and cellular levels. The incentive program employs blockchain research

and techniques built on behavioural psychology and game theory.

PATTERN RECOGNITION

Pattern recognition is a mature field that is still very fast growing and exciting. It spans different areas including computer vision, image processing, text and document analysis, and neural networks. It supported huge developments across different applications like biometrics, bioinformatics, multimedia data analysis and data science.

Pattern recognition, by definition, is the process of recognising patterns and regularities in data. It is closely related to machine learning, sometimes even the terms are used interchangeably. However, machine learning is one of the approaches used in pattern recognition.

Different approahces are used in pattern recognition and they could be categorised into two different categories: the first is done by using labeled "training" data, or in other terms: supervised learning. In the case where no labeled data are provided, specialised algorithms are used to recognise the patterns in a process called unsupervised learning, which is the second category.

MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging (MRI) is an imaging technique that produces high quality images of the anatomical structures of the human body, especially in the brain, and provides rich information for clinical diagnosis and biomedical research. The diagnostic values of MRI are greatly magnified by the automated and accurate classification of the MRI images.

WAVELET TRANSFORM

Wavelet transform is an effective tool for feature extraction from MR brain images, because it allows analysis of images at various levels of resolution due to its multi-resolution analytic property. However, this technique requires large storage and is computationally expensive. In order to reduce the feature vector dimensions and increase the discriminative power, the principal component analysis (PCA) was used. PCA is appealing since it effectively reduces the dimensionality of data and therefore reduces the computational cost of analysing new data. Then, the problem of how to classify the input data arises.

SUPPORT VECTOR MACHINES

In recent years, researchers have proposed a lot of approaches for this goal, which fall into two categories. One category is supervised classification, including support vector machine (SVM) and k- nearest neighbors (k-NN). The other category is unsupervised classification, including self-organisation feature map (SOFM) and fuzzy c-means. While all these methods achieved good results, and yet the supervised classifier performs better than unsupervised classifier in terms of classification accuracy (success classification rate). However, the classification accuracies of most existing methods were lower than 95%, so the goal of this paper is to find a more accurate method.

Our approach starts first by segmenting the photos and then using wavelet transform to extract features, then applying PCA to reduce the features extracted before submitting the reduced features to the Kernel Support Vector Machine KSVM classifer and finally

using K-fold cross validation to enhance the generalisations produced by the KSVM.

CANCER

Cancer refers to uncontrollable cells that continuously grow and invade bodily tissues. Cells may turn cancerous due to different factors including accumulation of mutations in the DNA. Examples of genetic mutations include BRCA1 and BRCA2 mutations. Cancer could also be caused by terrible lifestyle choices such as the consumption of smoked red meat (e.g: hot dogs and sausages), which are classifed as IARC Group 1 carcinogens, meaning they do cause cancer in humans. Cells can detect and repair DNA damage most of the time, and the body usually gets rid of damaged cells through a process called programmed cell death but sometimes cancerous cells can bypass this process to grow, divide and spread abnormaly in the tissues of the human body. Benign tumours are tumours that grow only in one place and usually do not grow back if treated, while malignant tumours grow in and travel to different places and hence they are more dangerous.

Cancer Detection

There are several ways by which patients can know whether or not they have cancer. Examples of which will be discussed in this section.

Complete Blood Count (CBC)

A sample of blood is tested to measure the amount of various types of blood cells. Usually this method is used to detect blood cancer if a certain blood cell type has too many or too few blood cells or by the

existence of abnormal cells. Usually, bone marrow biopsy could help confirm the diagnosis of a blood cancer.

Tumour Marker Tests

Although tumour markers are typically chemicals made by tumour produced by some normal cells, and the levels of which may be higher in noncancerous conditions. This is usually considered a problem for tumour marker tests to detect and diagnose cancer. It is rare to only use such a method alone to confirm the diagnosis of cancer.

Examples of such tests include prostate-specific antigen for prostate cancer, calcionin for medullary thyroid cancer, alpha-fetoprotein (AFP) for liver cancer, or human chorionic gonadotropin (HCG) for germ cell tumours, such as testicular cancer and ovarian cancer. That said, use of tumour marker tests is still highly controversial.

Biopsy

A test in which a sample of cells or a piece of a tissue is removed and analysed in a laboratory with the purpose of diagnosing cancer. There are different types of biopsy tests including:

1. **Bone Marrow Biopsy:** as mentioned previously this test is especially important in diagnosing blood cancers.
   Bone marrow is the spongy material inside some of your larger bones where blood cells are produced. Analysing a sample of bone marrow may reveal what's causing your blood problem. A bone marrow biopsy could help diagnose malignant tumours that travelled to the bone marrow.

2. **Endoscopic biopsy:** using a thin, flexible tube with a light on its end to see inside the body. The tube usually takes small samples of tissues for later analysis. An endoscopic biopsy could be inserted through the mouth, rectum, urinary tract or a small incision through the skin.
3. **Needle biopsy:** using a needle, cells from suspicious areas could be extracted. A needle biopsy may be more suitable for tumours that can be felt through skin such as suspicious breast lumps and enlarged lymph nodes.
4. **Skin biopsy:** involves removing cells from the surface of the body. Usually used to detect skin cancer.

Symptoms

Symptoms of Brain Cancer:

1. Severe headache whose onset is usually in the early morning.
2. Seizures, including motor seizures.
3. Changes to personality and/or memory.
4. Vomitting and nausea.
5. Fatigue.
6. Problems falling asleep
7. Problems with memory.
8. Changes to motor abilities such as walking.
9. Loss of balance is usually associated with a tumour in the cerebellum.

10. Complete or partial loss of vision is associated with a tumour in the occipital lobe or temporal lobe of the cerebrum.
11. Changes in speech, hearing, memory, or emotional state, such as aggressiveness and problems understanding or retrieving words can be attributed to a tumor in the frontal and temporal lobe of the cerebrum.
12. Inability to look upward can be because of a pineal gland tumour.
13. Lactation and altered menstrual periods in women are linked to a pituitary tumour.
14. Difficulty swallowing, facial weakness or double vision is a symptom of a tumour in the brain stem.

Symptoms of Stomach Cancer

1. Vomitting and nausea.
2. Fatigue.
3. Indigestion.
4. Loss of apetite.
5. Feeling that food is stuck in throat while eating.
6. Unexplained weight loss.
7. Diahrrea or constipation.
8. Vomitting blood in advanced stages.

Symptoms of Breast Cancer

1. Change in the size of the breast.
2. Bloody nipple discharge that occurs suddenly and only in one nipple.

3. A nipple turned inward, a sore in the nipple area or any physical changes.
4. Pain in the breast that doesn't go away.
5. A lump that feels like thickening in the breast.

Cancer Statistics

Cancer Type	Estimated Cases	Estimated Deaths
Bladder	81,190	17,240
Breast (Female – Male)	266,120 – 2,550	40,920 – 480
Colon and Rectal (Combined)	140,250	50,630
Endometrial	63,230	11,350
Kidney (Renal Cell and Renal Pelvis) Cancer	65,340	14,970
Leukemia (All Types)	60,300	24,370
Lung (Including Bronchus)	234,030	154,050
Melanoma	91,270	9,320
Non-Hodgkin Lymphoma	74,680	19,910
Pancreatic	55,440	44,330
Prostate	164,690	29,430
Thyroid	53,990	2,060

MRI CLASSIFICATION PROCESS

Overall our method involves three stages:

1. **Preprocessing:** feature extraction and feature reduction.
2. SVM training and cross-validation.
3. SVM classification.

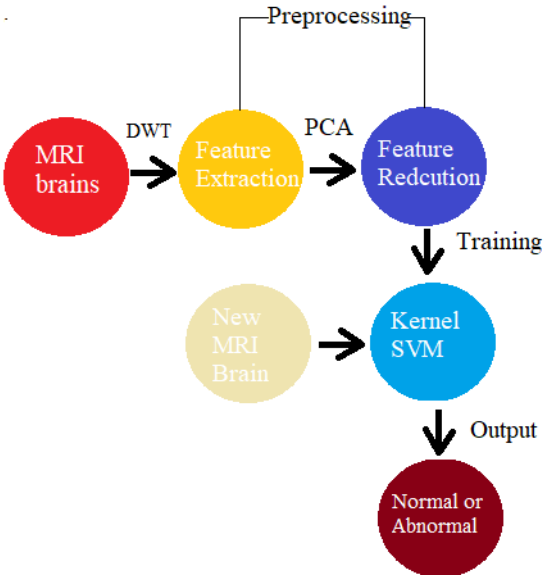


Figure 2.1: Methodology

FEATURE EXTRACTION

The most conventional tool of signal analysis is Fourier transform (FT), which breaks down a time domain signal into constituent sinusoids of different frequencies, thus, transforming the signal from time domain to frequency domain. However, FT has a serious drawback as discarding the time information of the signal. For example, analyst can not tell when a particular event took place from a Fourier spectrum. Thus, the quality of the classification decreases as time information is lost. Gabor adapted the FT to analyse only a small section of the signal at a time. The technique is called windowing or short time Fourier transform (STFT). It adds a window of particular shape to the signal. STFT can be regarded as a compromise between the time information and frequency information. It provides some information about both time and frequency domain. However, the precision of the information is limited by the size of the window. Wavelet transform (WT) represents the next logical step: a windowing technique with variable size.

Thus, it preserves both time and frequency information of the signal. The development of signal analysis is shown in Fig. 2. Another advantage of WT is that it adopts "scale" instead of traditional "frequency", namely, it does not produce a time-frequency view but a time-scale view of the signal. The time-scale view is a different way to view data, but it is a more natural and powerful way, because compared to "frequency", "scale" is commonly used in daily life. Meanwhile, "in large/small scale" is easily understood than "in high/low frequency.

DISCRETE WAVELET TRANSFORM

The discrete wavelet transform (DWT) is a powerful implementation of the WT using the dyadic scales and positions.

The fundamentals of DWT are introduced as follows. Suppose x(t) is a square-integrable function, then the continuous WT of x(t) relative to a given wavelet (t) is defined as

$$W_\Psi(a,b) = \int_{-\infty}^{\infty} x(t)\Psi_{a,b}(t)\,dt \tag{2.1}$$

where

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi(\frac{t-a}{b}) \tag{2.2}$$

Here, the wavelet $\Psi_{a,b}(t)$ is calculated from the mother wavelet (t) by translation and dilation: a is the dilation factor and b is the translation parameter (both real positive numbers). There are several different kinds of wavelets which have gained popularity throughout the development of wavelet analysis. The most important wavelet is the Harr wavelet, which is the

simplest one and often the preferred wavelet in a lot of applications.

Formula (1) can be discretised by restraining a and b to a discrete lattice ($a = 2^b$ & $a > 0$) to give the DWT, which can be expressed as follows:

$$ca_{j,k}(n) = DS[\sum_n x(n) g_j^*(n - 2^j k)] \tag{2.3}$$

$$cd_{j,k}(n) = DS[\sum_n x(n) h_j^*(n - 2^j k)] \tag{2.4}$$

Here $ca_{j,k}$ and $cd_{j,k}$ refer to the coefficients of the approximation components and the detail components, respectively. g(n) and h(n) denote for the low-pass filter and high-pass filter, respectively. j and k represent the wavelet scale and translation factors, respectively. DS operator means the downsampling. Formulas (3) and (4) are the fundamentals of wavelet decomposes. It decomposes signal x(n) into two signals, the approximation coefficients ca(n) and the detail components cd(n). This procedure is called one-level decompose.

The above decomposition process can be iterated with successive approximations being decomposed in turn, so that one signal is broken down into various levels of resolution. The whole process is called wavelet decomposition tree.

2D DWT

In case of 2D images, the DWT is applied to each dimension separately. Consequently, there are 4 subband (LL, LH, HH, and HL) images at each scale. The sub-band LL is used for next 2D DWT.

The LL subband can be regarded as the approximation component of the image, while the LH, HL, and HH subbands can be regarded as the detailed com-

ponents of the image. As the level of decomposition increased, a more compact but coarse approximation component was obtained. Thus, wavelets provide a simple hierarchical framework for interpreting the image information. In our algorithm, level-3 decomposition via Harr wavelet was utilised to extract features. The border distortion is a technique issue related to digital filter which is commonly used in the DWT. As we filter the image, the mask will extend beyond the image at the edges, so the solution is to pad the pixels outside the images. In our algorithm, symmetric padding method was utilised to calculate the boundary value.

FEATURE REDUCTION

Excessive features increase computation times and storage memory. Furthermore, they sometimes make classification more complicated, which is called the curse of dimensionality. It is required to reduce the number of features. PCA is an efficient tool to reduce the dimension of a data set consisting of a large number of interrelated variables while retaining most of the variations. It is achieved by transforming the data set to a new set of ordered variables according to their variances or importance. This technique has three effects: it orthogonalises the components of the input vectors so that they are uncorrelated with each other, it orders the resulting orthogonal components so that those with the largest variation come first, and eliminates those components contributing the least to the variation in the data set.

It should be noted that the input vectors be normalized before performing PCA to have zero mean and unity variance. The normalization is a standard procedure.

CLASSIFICATION USING KERNEL SUPPORT VECTOR MACHINE

The introduction of support vector machine (SVM) is a landmark in the field of machine learning. The advantages of SVMs include high accuracy, elegant mathematical tractability, and direct geometric interpretation. Recently, multiple improved SVMs have grown rapidly, among which the kernel SVMs are the most popular and effective. Kernel SVMs have the following advantages:

1. work very well in practice and have been remarkably successful in such diverse fields as natural language categorisation, bioinformatics and computer vision;
2. have few tunable parameters; and
3. training often involves convex quadratic optimisation.

Hence, solutions are global and usually unique, thus avoiding the convergence to local minima exhibited by other statistical learning systems, such as neural networks.

Linear SVM

Given a p-dimensional N-size training data set that can be described as follows:

$$\{(x_n, y_n)|x_n \in R^p, y_n = \{-1, 1\}\}$$

where $x_n$is a p-dimensional vector and y_n is a binary class set where -1 and 1 correspond to the classes benign and malignant.

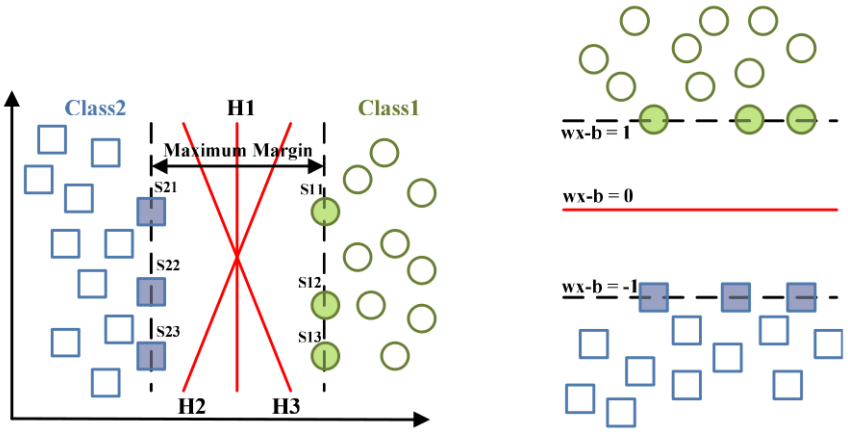


Figure 2.2: Maximum Margin Hyperplane

The maximum margin hyperplane which divides the two classes is the optimal SVM needed to solve this problem which can be of the form:

$$w \cdot x - b = 0$$

where $w \cdot x$ is the dot product of the data vector x and w is the normal vector to the hyperplane.

In order to maximize the margin between the two horizontal hyperplanes (-1, 1, i.e: benign and malignant) while still separating the data we need to define the two hyperplanes as follows:

$$w \cdot x - b = \pm 1$$

CROSS VALIDATION THROUGH K-FOLD

Cross validation is employing one of various techniques to test how the results of a statistical analysis will generalise to an independent data set. It is mainly used in an application where the goal is prediction.

In a prediction problem, the model is given a set of known data (i.e: training data set) on which training

is run, then it is given a set of first-seen data, i.e: data that was not used in estimating the model, on which the model is tested to flag issues like overfitting and selection bias.

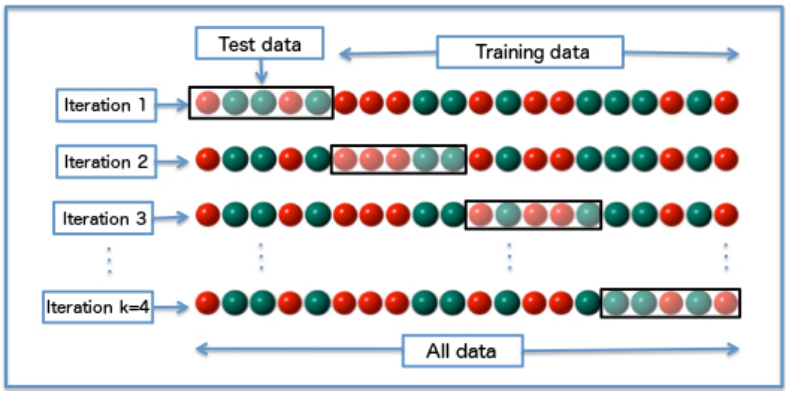


Figure 2.3: K-Fold Cross Valdiation

In K-Fold cross validation, the sample is randomly partitioned into k equal-sized subsamples, then one of the k subsamples is used as the validation data set for testing, and the remaining k-1 are used as the training data. The validation is repeated k times, with each subsamples used exactly once as the validation data set. Finally, the k results are averaged to get a single estimate.

The advantages of such an approach include:

1. All observations are used both for validation and training.
2. Each subsample is used exactly once for validation.

SIMILAR WORK

ADA

ADA is a smartphone NLP-based chatbot application whose purpose is to help patients diagnose what

diseases they might have by posing some questions about symptoms so it can reach a conclusion.

Advantages:

1. Free.
2. Easy to use.
3. The plausible conclusion will be shown at the end of each session based on proper analysis of the answers provided.
4. Suggestions of actions the patient should take are presented.
5. Explanation facilities.

Disadvantages:

1. Not as flexible as might be needed. Symptoms have to be written correctly.
2. Not a very friendly user experience.

CareZone

CareZone can manage medications and doctor's instructions. Managing prescriptions, medications and health in one place, the application makes sure you always have an up-to-date list of your or your loved ones' medications, and that you're following the exact instructions given to you by doctors after diagnosis, treatment or discharge and that you always have a medication list with you. The application allows you to quickly create a detailed medications list, which is securely backed up and accessible from any mobile device or browser at the clinic, in an emergency, or any other time you need it. Reminders help

you stay on schedule. Schedule when you, or a family member, is supposed to take each dose and CareZone will send reminders to you, your family, or anyone you choose that it's time to take medications. Finally, it allows you to organise important info in one place.

Advantages:

1. Sleep journal to keep track of your sleep habits.
2. Medications journal.
3. Medication reminders.
4. Tracking options for ailments.
5. Organised.

Disadvantages

1. Not flexible when it comes to their camera feature; it does not allow you to edit the misunderstood pictorial data.
2. Medications are only scheduled on daily basis.

WebMD

WebMD Symptom Checker is a free, web-based tool that enables patients to enter their information to help them diagnose their issues whose main target audience is the general public who do not posess enough knowledge about disease symptoms.

Advantages

1. Ease of use.
2. Good user experience.
3. Surveys.
4. Explanation facilities.

Disadavantages

1. No record keeping.
2. Single rule matching system.

Miiskin Melanoma Skin Cancer

Miiskin is your personal skin monitoring app - a simple tool to assist and help you explore and monitor changes on your skin and moles. Miiskin's mission is to help you keep an eye on your skin and compare moles over time. Monitoring and detecting changes in moles can be important in finding melanoma (cancer in moles) at an early stage (Malignant Melanoma is one of the most dangerous forms of skin cancer).

Advantages:

1. Secure cloud storage for mole pictures.
2. Keeps track of moles over time.
3. Great user experience.

Disadvantages:

1. Not a free application.
2. Does not provide any information on diagnosis.
3. It only tracks moles but does nothing more.

Pearl Cancer

Everyone that's ever been through or is going through cancer treatment knows that the side effects are incredibly overwhelming. This application was made by Pearl Point Cancer Support organisation to help patients deal with those side effects by finding out what's causing them and what might help minimise

discomfort. The application links to the developers (Pearl point) online archive where the user can find several support resources.

Table 2.2

Title	Methodology	Advantages	Disadvantages	Problem Definition	Rate
Ada	A decision tree chatbot is used to take the symptoms and output potential diseases or conditions that the patient might have. A decision tree chatbot is used to take the symptoms and output potential diseases or conditions that the patient might have.	Simple concept, storing medical cases, shows information about potential diseases after diagnosis.	Text is case sensitive, requires a lot of medical knowledge by the patient, amount of questions can be boring to the user.	Provide a modern, easy to use solution for patients to diagnose and track their potential diseases and conditions.	9/10

Table 2.2

Title	Methodology	Advantages	Disadvantages	Problem Definition	Rate
Care Zone	It manages patients' prescription lists and medical conditions by means of medication lists and a schedule It also keeps a simple medical record for each patient that contains important vitals and statistics.	Modern interface, keeps the medication tracking process easier, provides graphs and visuals about patient records.	Users are unable to track medications on an hourly basis. Prescription picturing isn't accurate and has a high miss chance.	Provide a way to keep track of prescriptions and general health by means of an application.	9/10

Table 2.2

Title	Methodology	Advantages	Disadvantages	Problem Definition	Rate
WebMD Symptom Checker	It takes in the patient's basic information along with their current symptoms and conditions and medications that they take, it then shows them what symptoms they might have based on their input.	Easy interface, doesn't require expert knowledge about the domain, it suggests common symptoms. It lists potential conditions by order of their relativeness to the input.	It doesn't keep a medical record of the patient; each new session is a new patient to the system. It requires an Internet connection.	Provide a web solution for patients to diagnose their diseases and conditions.	8/10
Miiskin-Melanoma-Skin-Cancer	Provide side-by-side image viewing and comparison to easily detect and track skin and moles.	Has a clean, on-the-spot user interface, not cluttered with irrelevant features. Reminds the user when to take a new picture of their skin.	It is a premium app that has a 30-day trial only. It doesn't have an explanation facility or evaluation or analysis of the images, the users are left on their own.	Provide an application to facilitate skin and mole detection and tracking in a hope to early detect melanoma.	8/10

Table 2.2

Title	Methodology	Advantages	Disadvantages	Problem Definition	Rate
Pearl point the cancer side effects helper	Provides information and dictionaries about cancer treatment side effects, mostly a knowledge base or an information catalogue.	Has a simple, user-friendly interface. Has condition and symptom explanation to familiarise patients with what's going on with them. It allows users to browse all side effects and further investigate about them.	It doesn't keep a medical record or track previous patient conditions. Doesn't provide a satisfying feature set.	Help patients learn about the side effects of their treatment and generally educate themselves about their conditions and diseases.	-

3

REQUIREMENTS ANALYSIS AND SPECIFICATION

—

SYSTEM DOMAIN

The CuraGP web and mobile applications provide an interactive interface hat alllows patients to describe the symptomos they are experiencing for a quick consultation session with Cura's interactive general practitioner bot. The patient can choose to view a detailed report of the GP's diagnosis, which is to be sent to the patient's preferred oncologist for diagnosis confirmation, further diagnosis and treatment follow-up through Cura CBR. The patient does not have a prescribed method of question asking, the bot is programmed in such a way as to answer all questions, no matter how they are written, the only requirement is that they are written in English. In order for any user to be able to use any of the cCura products, they have to be registered on the main Cura portal. Patients can register through a specific website that grants them access to CuraGP, CuraCBR and CuraTherapy. Patients should provide their names, date of birth, country, city of residence, address, email, phone(s) for contact, and a password. Doctors/Oncologists have to request membership through the same portal. Specialized moderators handle registrations and confirm oncologist registration.

CLASS DIAGRAM

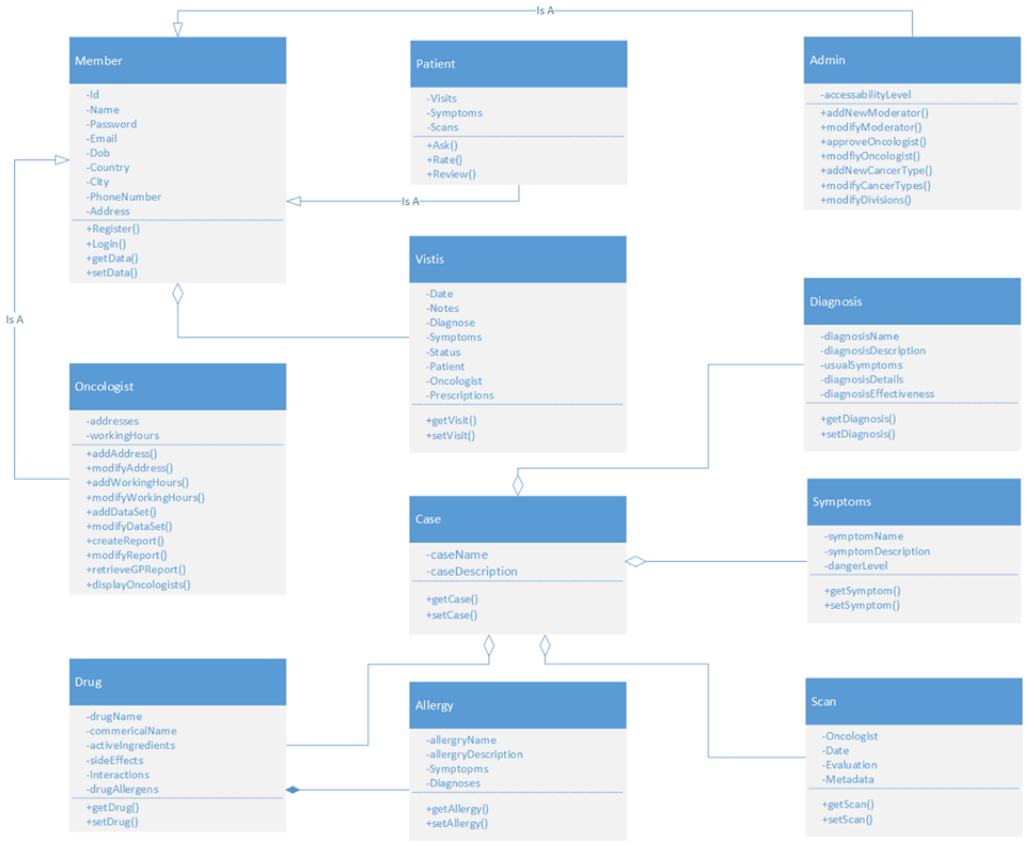


Figure 3.1: Class diagram

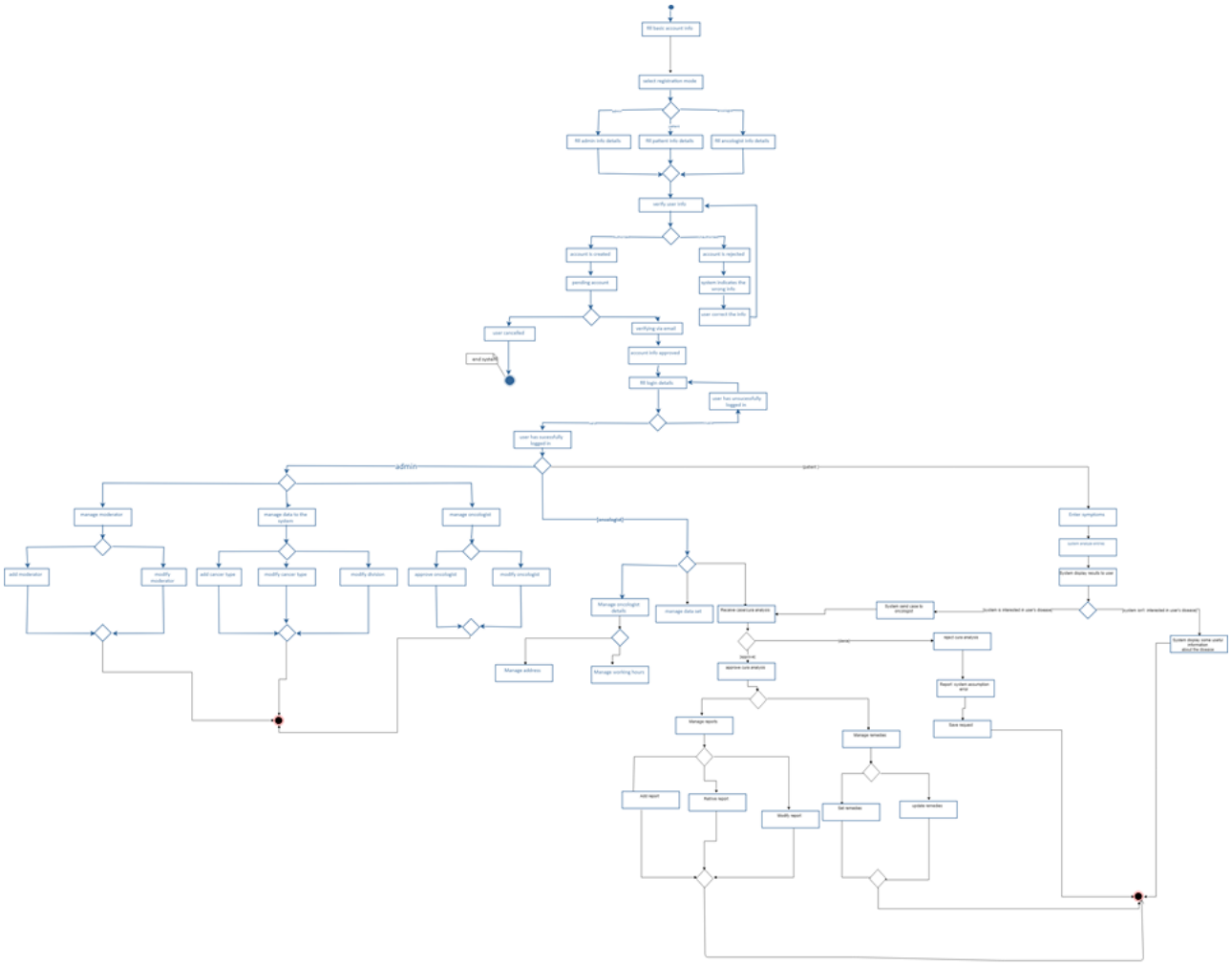


Figure 3.2: Activity Diagram

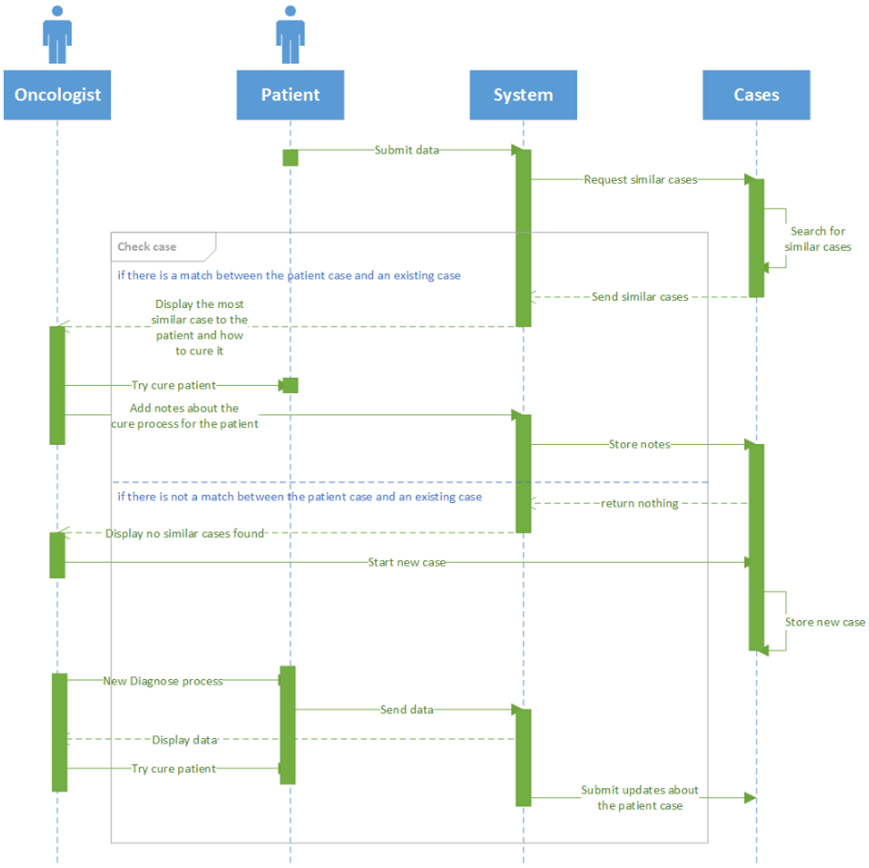


Figure 3.3: Sequence diagram 1

ACTIVITY DIAGRAM

SEQUENCCE DIAGRAMS

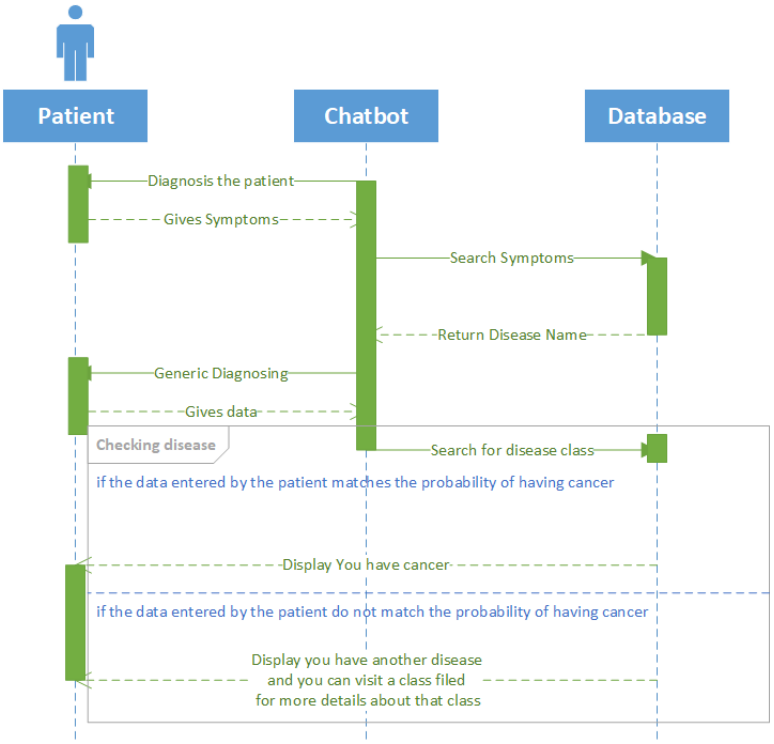


Figure 3.4: Sequence diagram 2

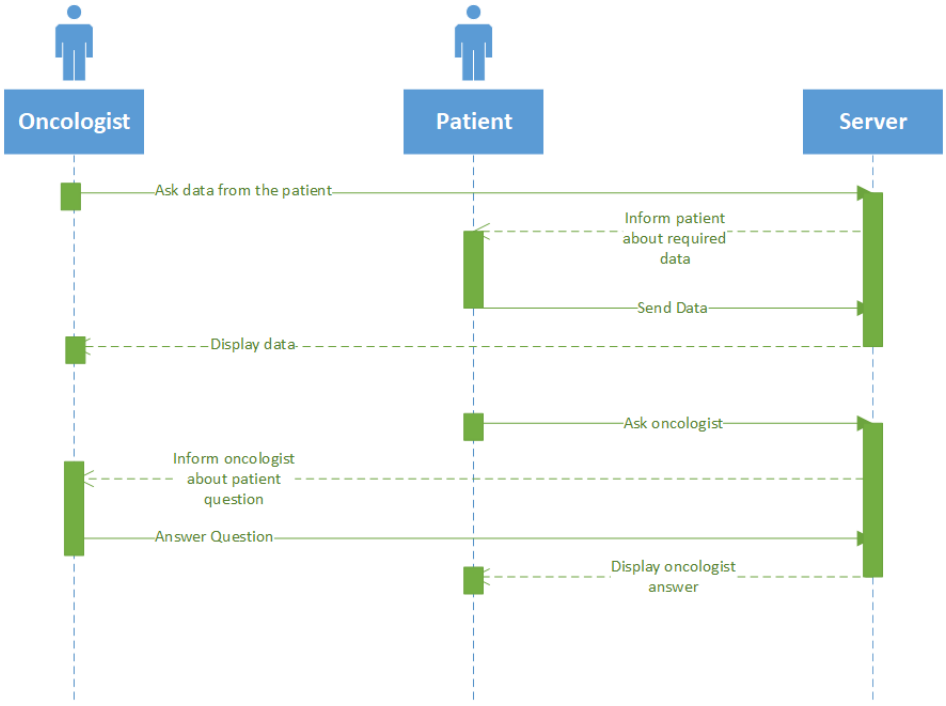


Figure 3.5: Sequence diagram 3

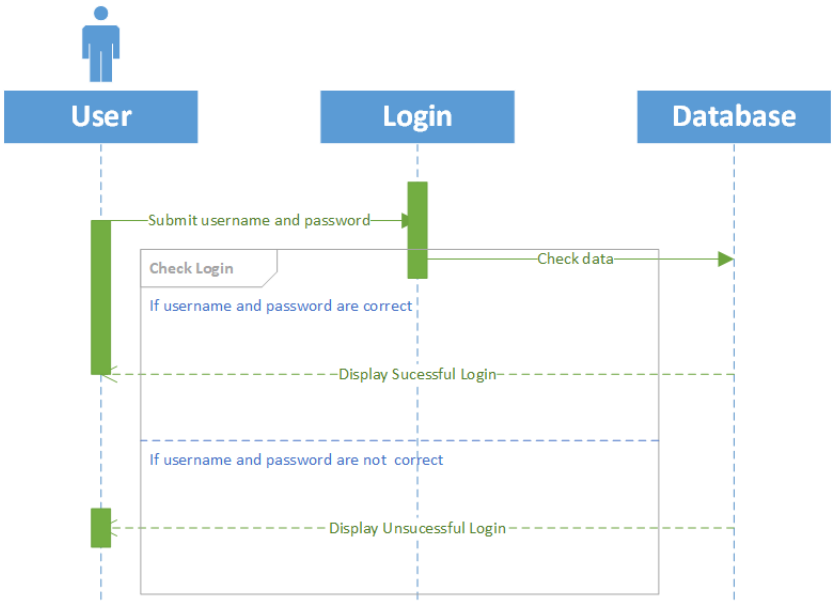


Figure 3.6: Sequence diagram 4

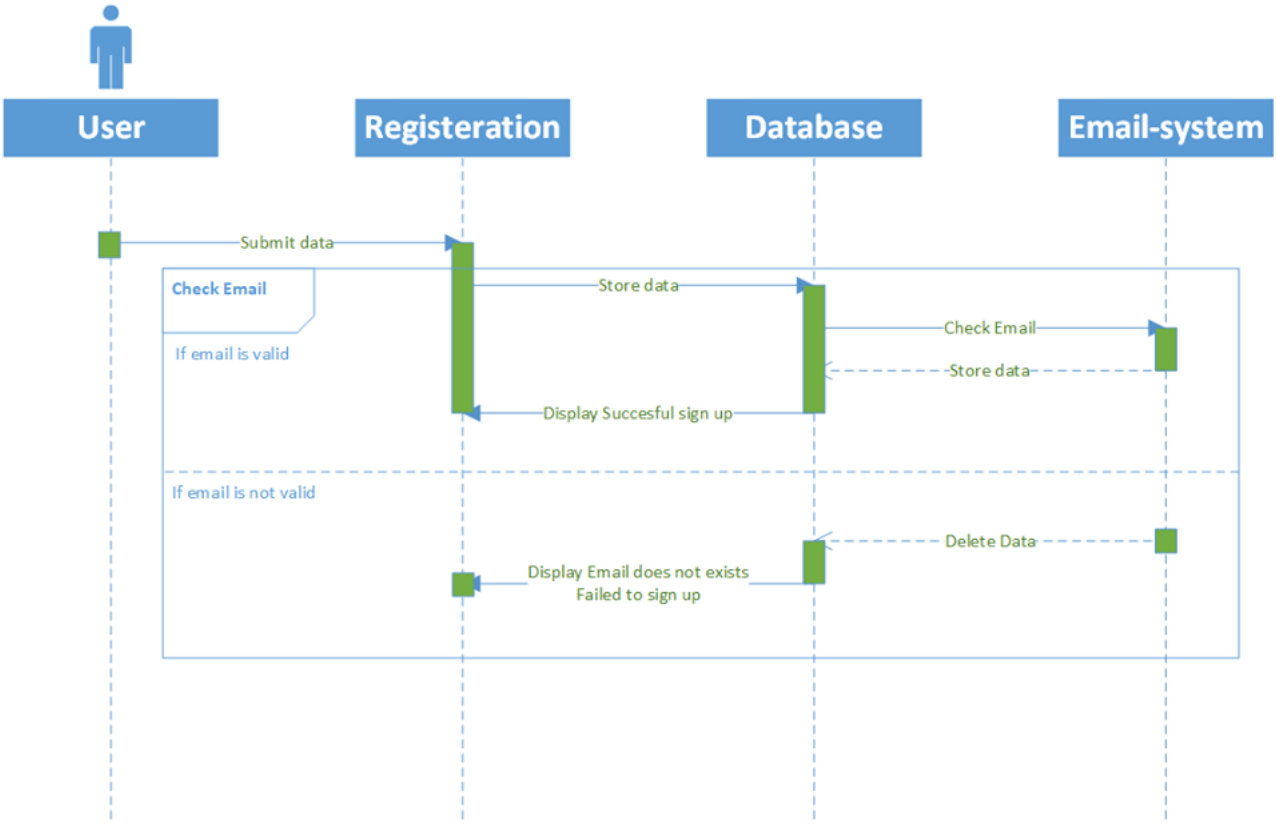


Figure 3.7: Use case diagram 5

USE CASE DIAGRAMS

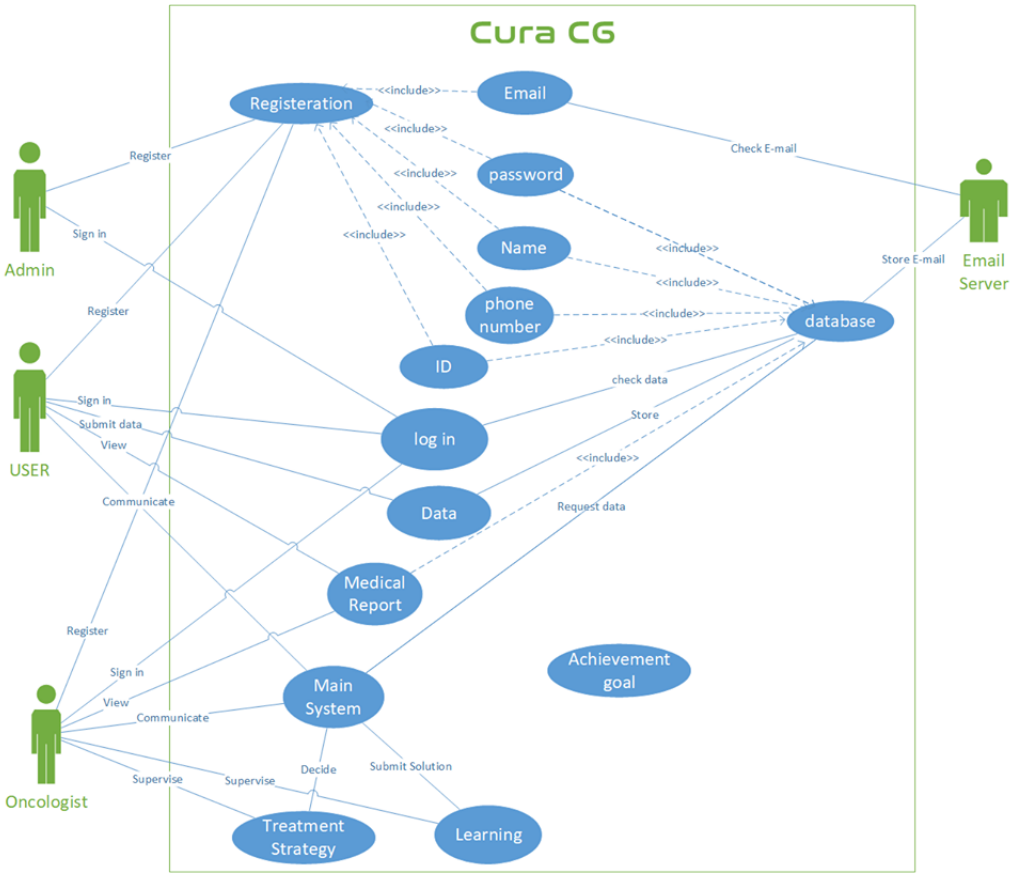


Figure 3.8: Use case diagram 1

CONTEXT DIAGRAM

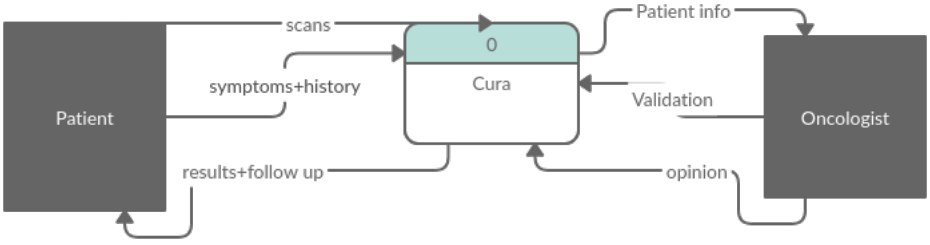
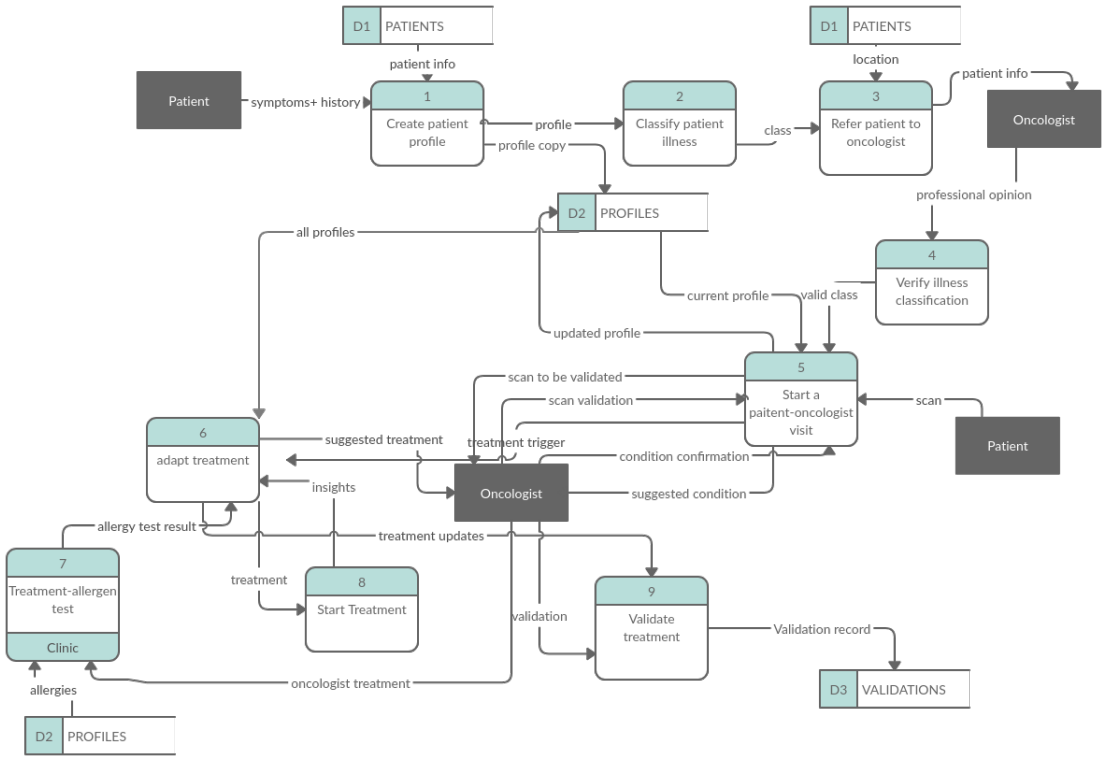


Figure 3.9: Context diagram

DFD LEVEL 0



This diagram can be summarized as follows:

The patient submits their symptoms while chatting with the Cura General Pracititioner along with their medical history which is requested when finalizing the general practitioner's session if the GP suspects the patient has one of the cancers the system is concerned with. The system then classifies the patient's symptoms along with their entire profile, which includes their medical history, according to previous cases stored in the Cura Case Based Reasoning system to confirm whether or not they have the suspected cancer. Only then the system refers the patient to a nearby oncologist specialized in that illness and uses Cura CBR to continue the diagnosis and then treatment. The oncologist has to verify the CuraGP classification of the illness as a final confirmation before the patient is admitted to the system.

Once the patient is confirmed, a series of visits with the oncologist commences in which each time the oncologist asks the patient to submit scans prior to the physical session, which the system classifies according to a data set including the profiles of other patients along with their scans, this is more detailed in child diagram 5. At the end of each session there are conclusion arrived at by the system and by the oncologist, the oncologist has to verify the system's conclusion.

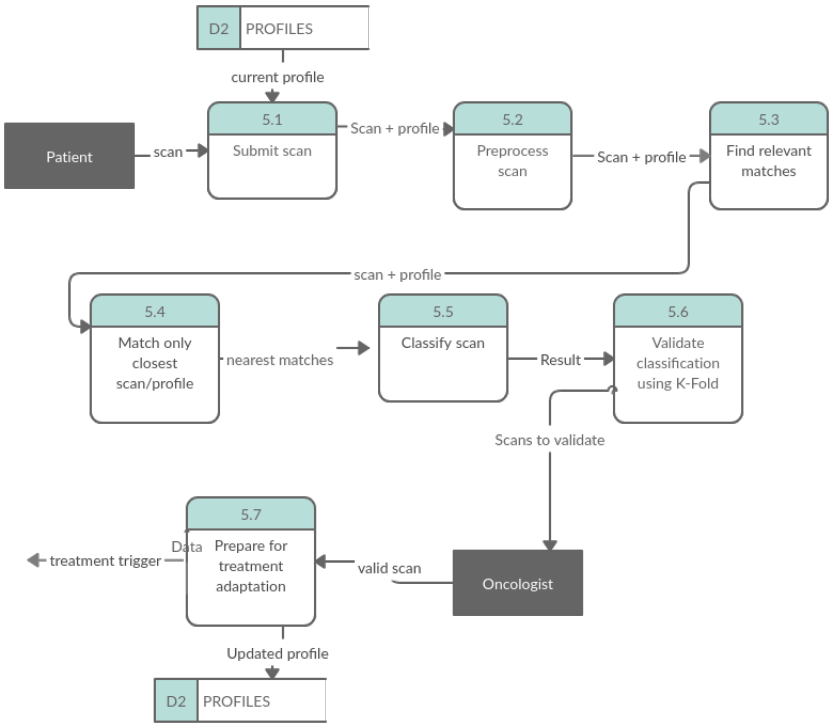
Once a treatment trigger is found, the adaptation subsystem is fired which uses a connectionist neural network to derive discriminating

features to find the most appropriate treatment for the current patient. The system triggers the oncologist to ensure no allergic reactions to the treatment will ensue by requiring the oncologist to perform an allergy test on the patient.

Just then the treatment starts, and along the road the treatment is adapted by the patient's progress, similar profiles and the oncologist's opinion until the treatment is complete. When the treatment is complete, the oncologist validates the treatment progress by completing a survey. This is done to ensure better performances in the future.

DFD CHILD DIAGRAMS

Child Diagram 5



DATA DICTIONARY

Description:	Create Patient Profile Gather all necessary data about patient from medical history to symptoms	Process ref.: 1
Inputs	Logic Summary	Outputs
P-1 symptoms + history D1-1 patient info	Once the Cura general practitioner collects all the necessary information needed to classify the patient, it immediately sends them to the classifier	1-2: profile
Physical ref.: Full description of this logic can be found in:	Part of online form through CuraGP Functional spec. section 3.7.1	

Description:	Classify Patient Illness Classify illness using training set of other patient profiles	Process ref.: 2
Inputs	Logic Summary	Outputs
1-2 profile	Required preprocessing is performed then using a connectionist neural network similar profiles are fetched and matched with the current patient's profile	2-3: class
Physical ref.: Full description of this logic can be found in:	Functional spec. section 3.7.2	

Description:	Refer Patient to Oncologist Send the patient to a preferred oncologist	Process ref.: 3
Inputs	Logic Summary	Outputs
2-3: class D1-3: patient location	After the general practioner classifies the patient's probable illness, using the class, the patient's info and preferences he is referred to an oncologist.	3-O patient info
Physical ref.: Full description of this logic can be found in:	Functional spec. section 3.7.3	

Description:	Verify illness classification Check whether the general practitioner classification is valid	Process ref.: 4
Inputs	Logic Summary	Outputs
O-4: professional opinion	Prior to starting the treatment history on the system, the doctor has to verify whether or not the patient has the cancer suggested by the GP	4-5: valid class
Physical ref.: Full description of this logic can be found in:	Functional spec. section 3.7.4	

Description:	Start a patient-oncologist visits This process handles all patient visits to the oncologist	Process ref.: 5
Inputs	Logic Summary	Outputs
4-5: valid class, D2-5: current profile, P-5: scan,O-5: scan validation,O-5: condition confirmation	The patient submits a scan required by the oncologist in each visit, the system classifies the scan before the oncologist's confirmation before the system classifies the patient's condition for the oncologist's confirmation until they find a treatment trigger. More details are in child diagram #5	5-O: scan to validate , 5-O: suggested condition , 5-D2: updated profile , 5-6: treatmetn trigger
Physical ref.:	Part of online form through CuraCBR	
Full description of this logic can be found in:	Functional spec. section 3.7.5	

	Adapt treatment	Process ref.: 6
Description:	Use previous experience to figure out treatment	
Inputs	Logic Summary	Outputs
D2-6: all profiles , 5-6: treatment trigger , O-6: treatment trigger, 8-6: insight, 7-6: allergy test result	This process handles almost all parts of the treatment; a machine learning algorithm adapts treatment and solutions throughout the patient's journey to recovery. Full operation is in child diagram 6.	6-O: suggested treatment , 6-8: treatment, 6-9: treatment updates
Physical ref.: Full description of this logic can be found in:	Adaptation subsystem in CuraCBR Functional spec. section 3.7.6	

	Treatment-allergen test	Process ref.: 7
Description:	Test the patient's allergies before starting a treatment	
Inputs	Logic Summary	Outputs
D2-7: allergies , O-7: oncologist treatment	Prior to starting a treatment, this process ensure no allergic reaction will ensue.	7-6: allergy test result
Physical ref.: Full description of this logic can be found in:	Allergy test in clinic Functional spec. section 3.7.7	

	Start treatment	Process ref.: 8
Description:	Start the treatment for a patient	
Inputs	Logic Summary	Outputs
6-8: treatment	This process follows the patient's progress throughout their treatment and logs statistics, information and insights about their progress.	8-6: insights
Physical ref.: Full description of this logic can be found in:		

	Validate Treatment	Process ref.: 9
Description:	Check the treatment as a success or not for future reference	
Inputs	Logic Summary	Outputs
6-8: treatment	This process follows the patient's progress throughout their treatment and logs statistics, information and insights about their progress.	9-D3: validation record
Physical ref.: Full description of this logic can be found in:		

	History	Data structure
Description:	A data structure representation of medical history	
	Contents	Related data flows
treatment *		p-1, 1-2, 5-d2, d2-5 d2-6
type of treatment		
date of treatment		
start date		
end date		
outcome		
oncologist*		
oncologist-name		
clinic-address * (1-)		
notes *		
note-details		
note-date		
oncologist		
		Physical Representation
		Physical record of patient

	Profile	Data structure
Description:	A data structure representation of patient profile	
	Contents	Related data flows
patient name		p-1, 1-2, 5-d2, d2-5 d2-6
patient-dob		
Email		
Location		
City		
Address		
Country		
Phone		
symptom* (1-)		
scan*		
allergy*		
		Physical Representation
		Physical record of patient

	Scan	Data structure
Description:	A data structure representation of a patient scan	
	Contents	Related data flows
Date		1-2, 1-d2, d2-6
Oncologist		O-5, 5-O, P-5
Evaluation		
Scan-Image		
Metadata		
		Physical Representation
		Scan document

	Allergy	Data structure
Description:	A data structure representation of an allergy's information	
	Contents	Related data flows
Name		D2-7, 7-6, 1-2, D2-5,
Description		5-D2, D2-6, 1-D2
Symptom*(1-)		
Diagnosis*(1-)		
		Physical Representation

	Insight	Data structure
Description:	A data structure representation of a doctor or AI insight about a scan or the patient's treatment	
	Contents	Related data flows
class		8-6
notes *		
system-notes*		
oncologist-notes*		
patient		
oncologist*(1-)		
		Physical Representation
		medical report

	treatment	Data structure
Description:	A data structure representation of a patient's treatment	
	Contents	Related data flows
patient-info		6-O, 6-8, O-7
oncologist-info* (1-)		
diagnosis* (1-)		
milestones* (1-)		
level		
validation-info*(1-)		
		Physical Representation

	symptom	Data structure
Description:	A data structure representation of the information about a particular symptom	
	Contents	Related data flows
name		P-1
description		
diagnosis* (1-)		
		Physical Representation

	Patients	Data store ref: D1
Description:	Information about the patients	
	Contents	Data flow out
patient name		D1-1, D1-3
patient-dob		
Email		
Location		
City		
Address		
Country		
Phone		
symptom* (1-)		

	Profiles	Data store ref: D2
Description:	Information about the patient progress	
Data flow in		Data flow out
1-D2, 5-D		D2-5, D2-7,
Contents		
patient name		
patient-dob		
Email		
Location		
City		
Address		
Country		
Phone		
symptom* (1-)		
scan*		
allergy*		

	Validation	Data store ref: D3
Description:	Information about the patient progress	
Contents		Data flow in
notes* (1-)		9-D3
oncologist		
level		
progress		