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 doctors 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 doctors and patients using different 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 will be diagnosed in the United States alone with an estimated 609,640 people will die 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 estimated 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.

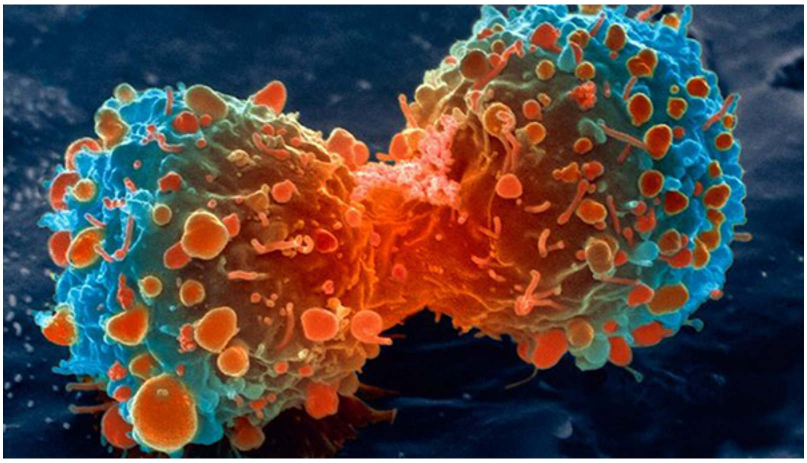

Figure 1.1: A dividing lung cancer cell.

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 sorrounding 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.

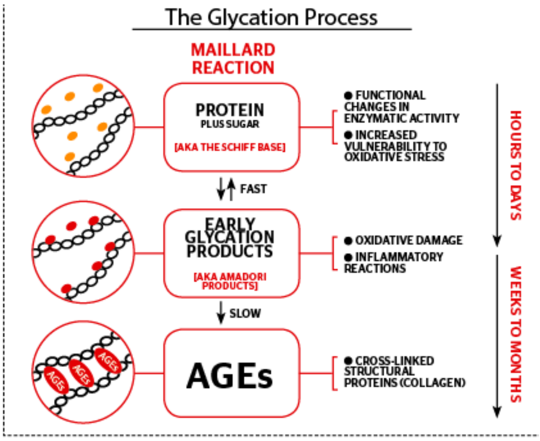

Figure 1.2: The Glycation Process

Endogenous glycations describe the type of gylcations occuring mainly in the bloodstream to a small percentae 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, 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 gylcation 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 chromose, 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 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 specialized 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 the addition to the fact that cancerous cells are programmed to ignore signals that tell cells to dividing, a process called programmed cell death, or apoptosis, which the body utilizes to get rid of unncessary cells.

The microenvironment, or the area sorrounding 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 specialized 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 four phases that interact with and complement one another to help patients to recognize the symptomps of cancer early on, follow their progress and have their questions about chemotherapy answered, whie helping doctors to organize and easily access patient's records, support in the diagnosis, treatment and postoperative chemotherapy. The following is a description of each of the proposed phases.

Phase 1: Cura General Parctitioner

The first phase in the journey is aimed at the patient before the official diagnosis by a doctor. Cura General Practicioner is a rule-based system built with the support of natural language processing to help to interact with the patients to figure out if the symptomps they are experiences 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 phase and recommended a suitable doctor who uses the system, otherwise they are recommended basic cautionary tips and may be recommended with the addresses of nearby doctors specializing 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 doctor before progressing to the second phase.

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 doctor.

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 RBS components to achieve the most optimal results

Overview of the Problem Domain

Cura GP is designed to assist patients in the recognition of the symptomps 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 symptomps of the three types of cancer aforementioned: brain, breast and gaststric cancer, along with all possibly related illnesses; for example in the case of gastric cancer the system also deals with the symptomps 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 through 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 level, both theoretical and experimental, to achieve a system with a proper knowledge base that will be as expert 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 utilizes 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 phases.

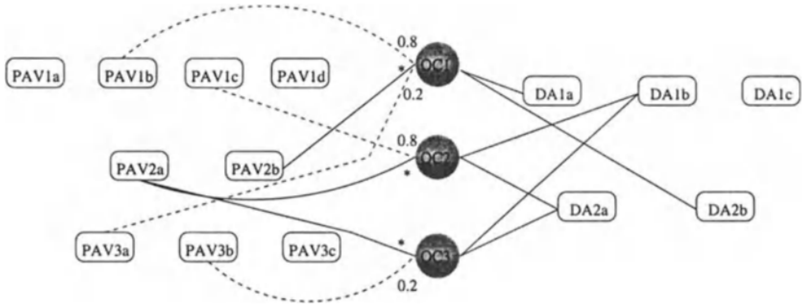

Figure 1.3: Neural Networks for CBR

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 corespond to the neurons describing the decision features.

Weighted connections exist between the input and hidden layers; the weight correspond 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 belongs 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 corresponds to a previous case c_k. The activation is computed by:

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

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 similariy measures. As for adaptation, the probabilities can be retro-propagated which allows memorized 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

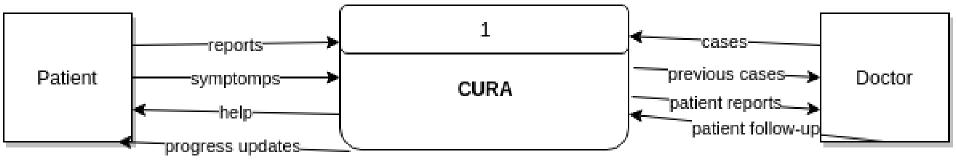

Figure 1.4: Context Diagram

1. **Interface Requirements:**
   a) All interfaces shall be easy to use.
   b) NLP shall be implemented in the first phase to aid users in their description of their symptomps.
2. Cases shall be implemented in such a way as 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 phase.
4. The system shall be flexible as to accomodate patient follow-ups and visits.

Non-Functional Requirements

1. **Usability Requirements:**
    a) Minimizing 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 unauthorized access to their records and data.
4. **Security Requirements:**
   The implementation and use of top-notch security algorithms to protect users sensitive and personal data.
5. **Scalability Requirement**:
   The system shall be scalable to a very flexible degree in its various phases, especially in the second and fourth phases, or the CBR system and the marketplace where doctors 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 doctors 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 accross 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 phase of the system shall be interoperable with one another and with other systems as well through a properly developed API.

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

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

ADVANTAGES AND DRAWBACKS

Advantages:

1. Easily accessible for patients and doctors.

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 incentivize 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

Goal

In an endeavour to help oncologists do a better job managing, analyzing, treating and following up with their patient's health, we set out to develop and implement a proper set of tools built based on years of research in fields like mathematical biology, bioinfomatics, computer science, artificial intelligence and medical research by dozens of passionate scientists in all of the aforementioned fields. And thus to realize 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 experimental, 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 ongologists 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, incentivizing them to share their datasets and vote on current ones.