1 BACKGROUND AND RELATED WORK

This project draws upon several diverse fields including bioinformatics, computer science, blockchain, nudge behavioral 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 approach used to recognize and treat cancer, along with motivational incentives for doctors and specialists to willingly share their relative data sets.

BEHAVIORAL 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 behavior 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:** behavioral 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 application, 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 incentivize oncologists to share their data sets whether, 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 incentivie program employs blockchain research and techniques built on behavioral 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 recognizing 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 categorized 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, specialized algorithms are used to recognize the patterns in a process called unsupervised learning.

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. 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 the data and therefore reduces the computational cost of analyzing new data. Then, the problem of how to classify on 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-organization 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 generalizations produced by the KSVM.

CANCER

Cancer refers to uncontrollable cells that continuously row 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 trough a process called programmed cell death but sometimes cancerous cells can bypass this process to grow, divide and sprad 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 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)

Sample of the blood is tested to measure the amount of various type 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 maybe 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 canceror and human chorionic gonadotropin (HCG) for germ cell tumors, 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 analyzed 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 specially important in diagnosis blood

cancers. Bone marrow is a spongy material inside the larger bones in the human body in which blood cells are produced. Analyzing a sample could help reveal the problem.
Bone marrow is the spongy material inside some of your larger bones where blood cells are produced. Analyzing a sample of bone marrow may reveal what's causing your blood problem. A bone marrow biopsy could help diagnose malignant tumours that traveled to the bone marrow.

2. **Endoscopic biopsy:** using a thin, flexible tube with a light on its end to see the 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 to a tumour in the occipital lobe or temporal lobe of the cerebrum.
    a) 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.
11. Inability to look upward can be because of a pineal gland tumor.
12. Lactation and altered menstrual periods in women are linked tos a pituitary tumor.
13. Difficulty swallowing, facial weakness or double vision is a symptom of a tumor 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 a thickening in the breast.

Cancer Statistics

Cancer Type	Estimated Cases	Estimated D
Bladder	81,190	17,240
Breast (Female – Male)	266,120 – 2,550	40,920 –
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.

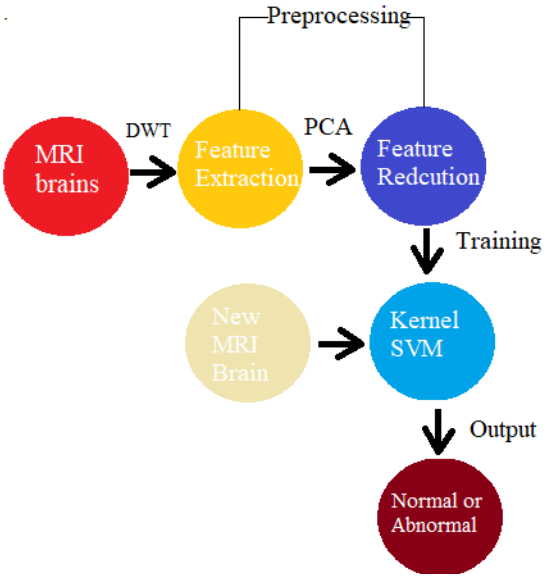

Figure 1.1: Methodology

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 analyze 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

$$W_{\Psi}(a,b) = \int_{-\infty}^{\infty} x(t)\Psi_{a,b}(t)\,dt \tag{1.1}$$

where

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi(\frac{t-a}{b}) \tag{1.2}$$

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

$$ca_{j,k}(n) = DS[\sum_n x(n) g_j^*(n - 2^j k)] \tag{1.3}$$

$$cd_{j,k}(n) = DS[\sum_n x(n) h_j^*(n - 2^j k)] \tag{1.4}$$

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 fundamental 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 sub-band (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 components of the image. As the level of decomposition increased, compacter but coarser 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 utilized 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 utilized 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 orthogonalizes the components of the input vectors so that 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 to have zero mean and unity variance before performing PCA. 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 categorization, bioinformatics and computer vision;

2. have few tunable parameters; and

3. training often involves convex quadratic optimization.

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

Givenn 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.

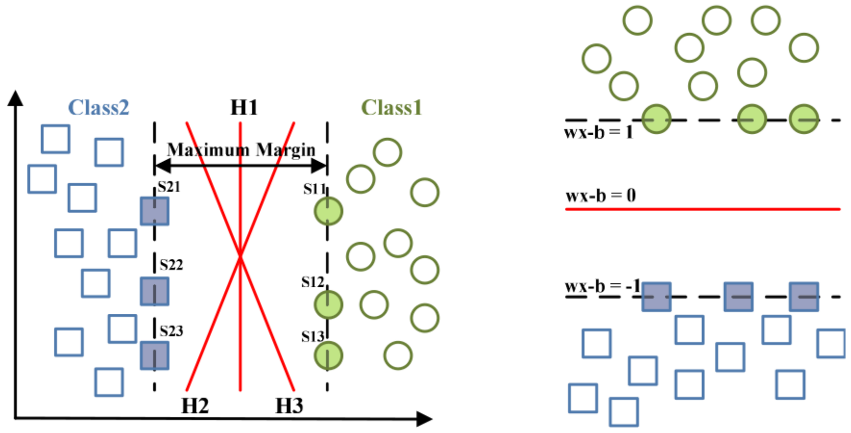

Figure 1.2: Maximum Margin Hyperplane

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 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 separate 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 generalize 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.

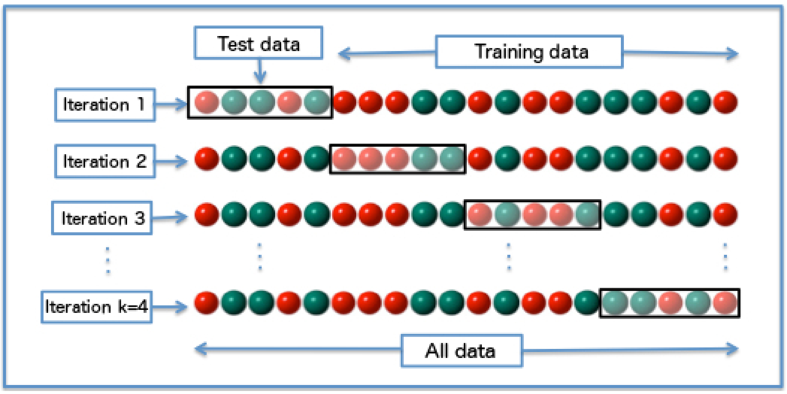

Figure 1.3: K-Fold Cross Valdiation

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 remainging 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 application NLP-based chatbot whose purpose is to help patients diagnose what disease they might have by posing some questions about symptoms so it can reach a conclusion of what disease you may have by analyzing your answers.

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

Care Zone can manage medications and doctor's instructions. Managing prescriptions, medications and health in one place, ant the application makes sure you always have an up to date list of your or your loved one's 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 organize 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. Organized.

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 UX.

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 organization to help patients deal with those side effects by finding out what's causing them and what might help minimize discomfort. The application links to the developers (Pearl point) online archive where the user can find several supports resources.