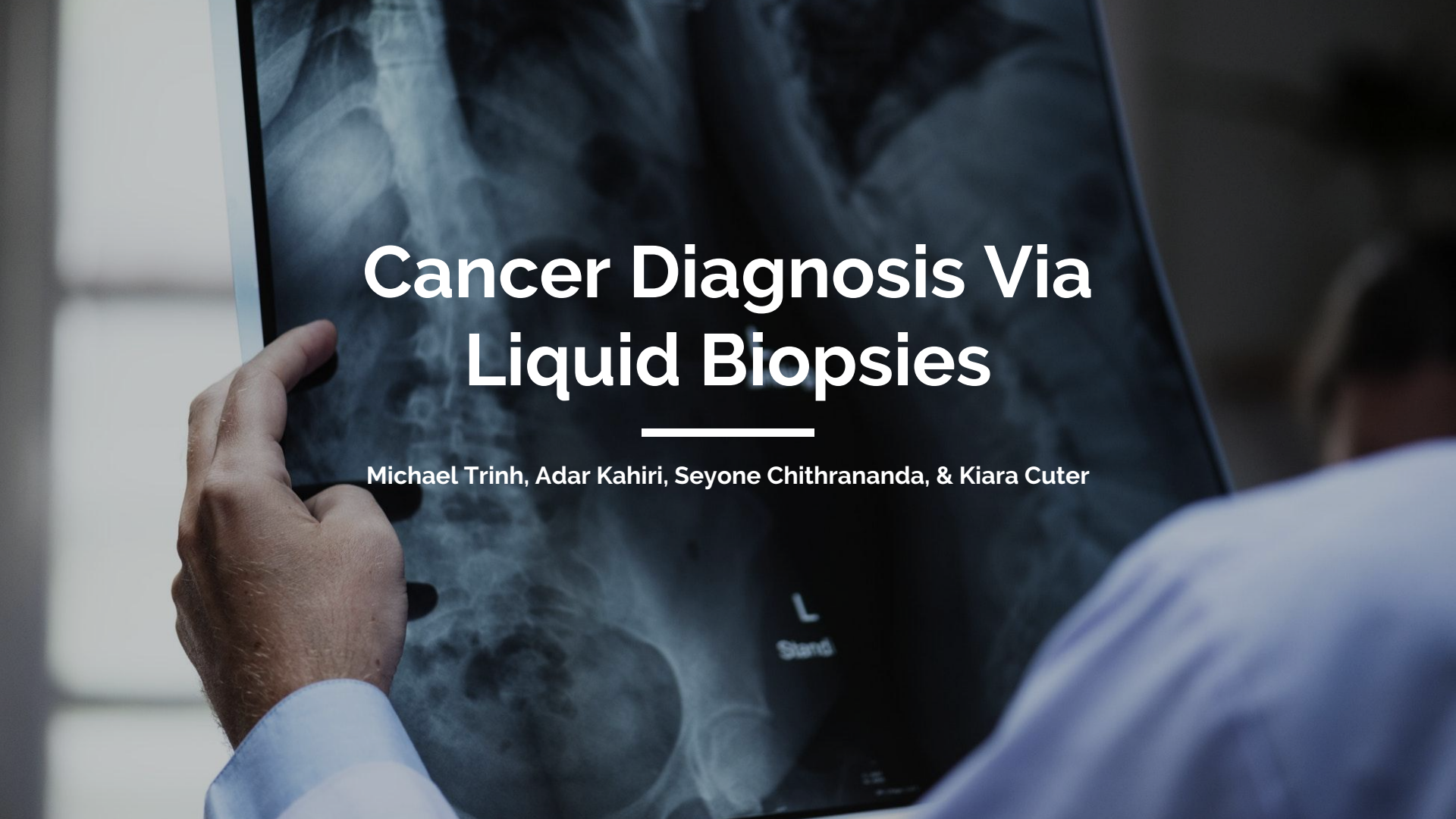
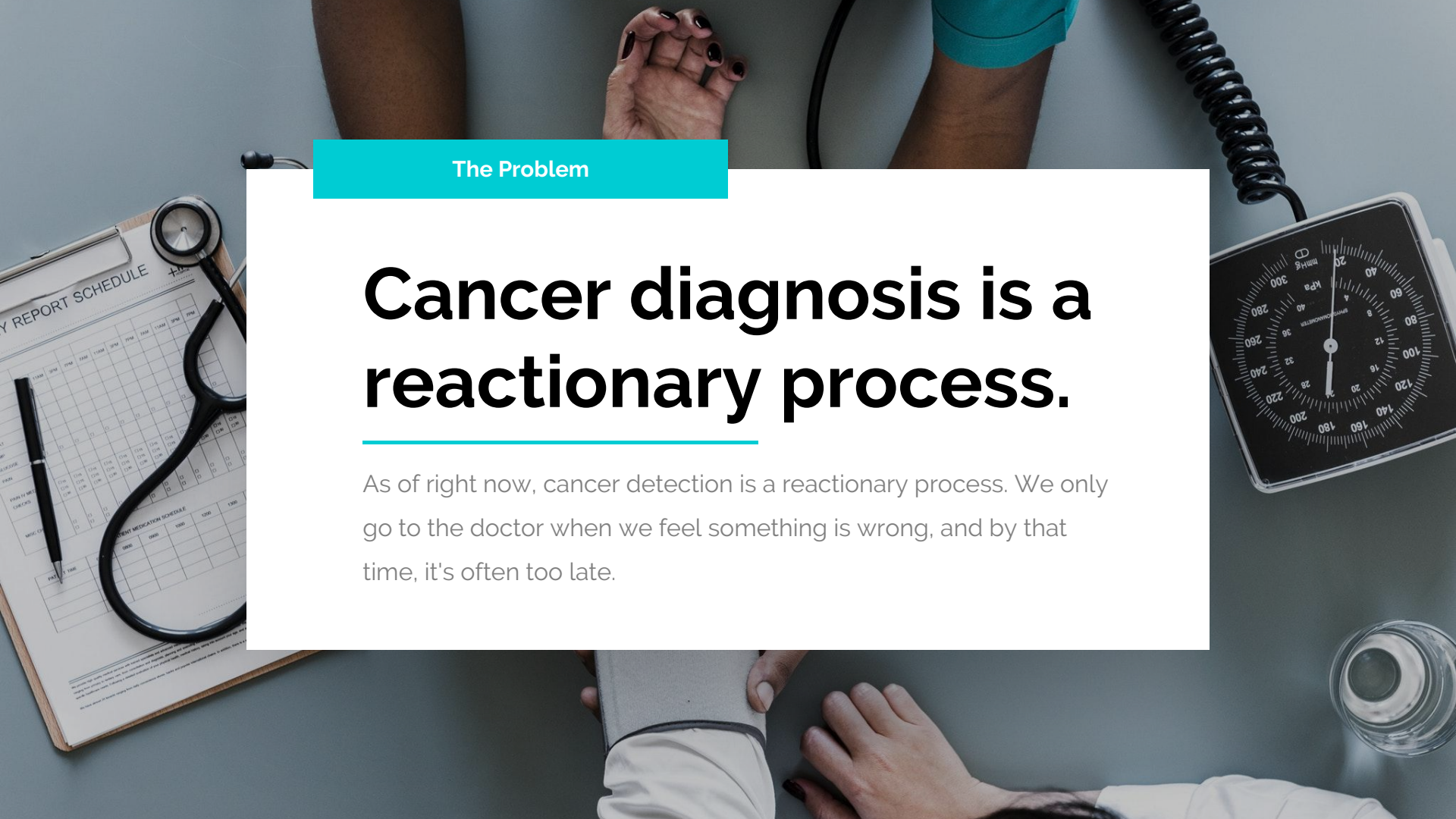


Cancer Diagnosis Via Liquid Biopsies

Michael Trinh, Adar Kahiri, Seyone Chithrananda, & Kiara Cuter

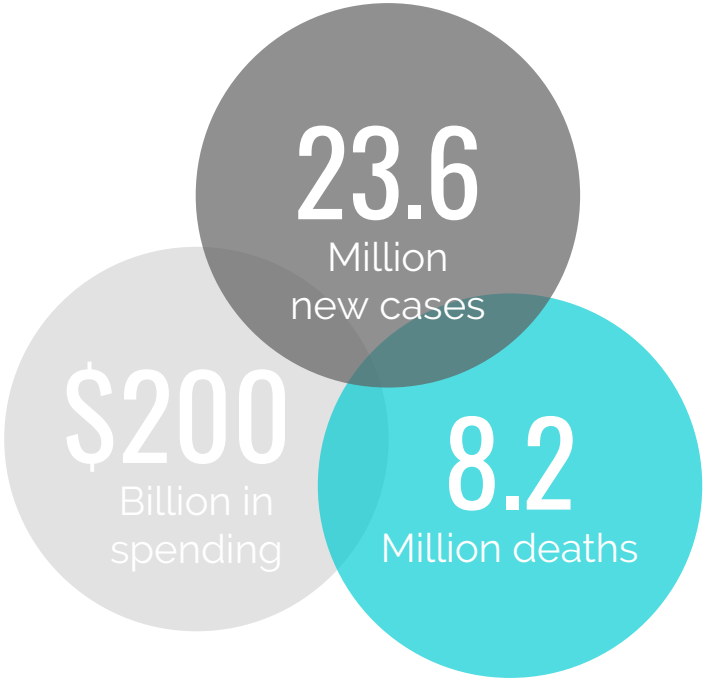


The Problem

Cancer diagnosis is a reactionary process.

As of right now, cancer detection is a reactionary process. We only go to the doctor when we feel something is wrong, and by that time, it's often too late.

The Facts



High death rate

Over **8.2 million people** die of cancer each year due to the inaccessibility of appropriate detection procedures and treatments. A quarter of cancer patients are dead in 6 months due to late diagnosis.

Cancer is expensive

The U.S. market for oncology therapeutic medicines is expected to reach as much as **$100 billion by 2022**. The global market is expected to reach as much as **$200 billion**.

The problem will only get worse

The number of new cancer diagnoses per year is expected to rise to **23.6 million** by 2030.

Death is preventable

When cancer is diagnosed at stage 1, the survival rate for 5 years is **almost 100%.** When it is diagnosed at stage 4, the survival rate is **22%.**

The Solution

Integrating precautionary cancer testing into routine medical checkups through liquid biopsies.

We can analyze biomarkers in the blood with machine learning to detect cancer in its early stages, before it has the chance to do any harm.

Machine Learning Pipeline

Gather Data

Right now, there isn't a lot of data on cfDNA that we could use for this project. Our first order of business is to gather this data from blood tests (from private and public sources). For this, we need a lab.

Preprocess the Data

Next we need to process the data in a way that's useful to our models. This includes the methylation patterns, beta values, and the location of certain CpG sites.

Detect Presence of Cancer

Analyze the methylation patterns of cell-free DNA (cfDNA) in the blood with a Support Vector Machine to detect the presence of cancer.

Determine Location of Cancer

Analyze the methylation state of CpG sites in cfDNA, extract beta-values, and determine the tumour of origin from these beta values and their locations.

Determine Severity of Cancer

Analyze the concentration of cancerous cfDNA in the blood to determine the stage of the cancer.

Gathering Data

Lack of Useful Data

Right now, public data on methylation patterns in ctDNA (cell-free tumour DNA) are relatively scarce and have questionable quality. In order to maximize the accuracy of our pipeline, we would largely need to use private datasets.

Public Datasets

Existing datasets include (https://ega-archive.org/datasets/EGAD00001003168) and (https://ega-archive.org/datasets/EGAD00001004317). However, these two datasets only have 19 samples total, which is not nearly enough.

Private Datasets

Clearly, publicly available datasets would not suffice. In order to accurately train every model in our pipeline, we would need access to private datasets owned by labs or other institutions like universities and hospitals.

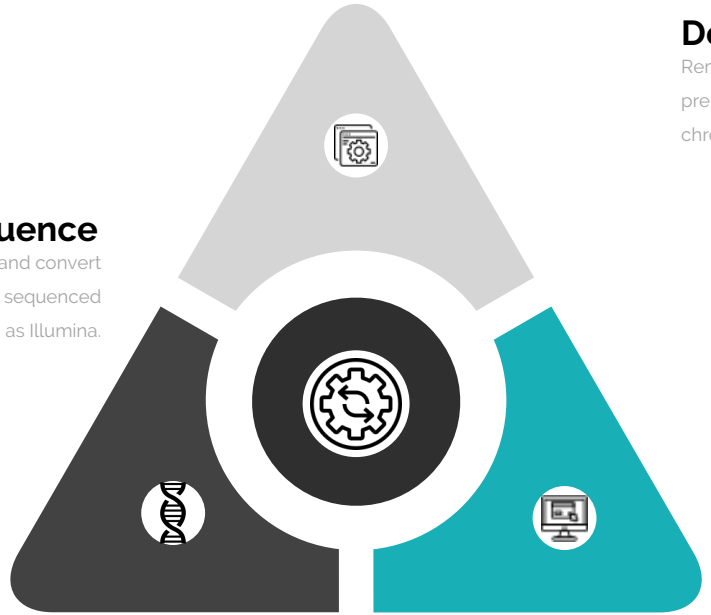
Preprocessing the Data

De-identify and Transform

Remove the identification for the cfDNA profile, and preprocess feature vectors by removing sex chromosomes, poor quality features and more.

Extract cfDNA and pair-end sequence

Extract cfDNA from plasma samples using isolation kits, and convert them into libraries. Finally,, sequences are paired-end sequenced in-lab or using platforms such as Illumina.



Normalize features

Normalize the features to take into account factors such as the strand's length, read depth, and guanine-cytosine content. (CpG)

Our Solution

A three-stage process.

Determine the Severity of the Cancer

Analyze the concentration of cancerous cfDNA in the blood to determine the stage of the cancer. Different kinds of cancer tend to have varying concentrations, but generally, the higher the concentration, the more developed the cancer. Once the type of cancer is known, it is fairly easy to estimate its stage.

Determine Tumour of Origin

Our Naive Bayes algorithm would use the beta values for each CpG site, as well as the ratio of CpG sites that are methylated to those that aren't. This would result in a model with several classes for different cancer types, such as lung, breast, colorectal, and more. The Naive Bayes algorithm would then determine the most probable class (the type of cancer) based on the features it is given

Detect Cancer

We plan to use a Support Vector Machine that would essentially take in a mix of labelled tumorous and non-tumorous cFDNA profiles as well as the beta values of each CpG site and assign a weight to each one. It would separate the labelled data points by whether the profile is cancerous or not, and establish a hyperplane separating the data into two classes

Impact

Billions of Dollars Saved

According to CNBC, global spending on cancer medicines is projected to reach $150 billion by 2020. Early diagnosis of cancer means less drugs will need to be used to treat it.

Millions of Lives Saved

According to the WHO, 9,555,027 people died of cancer in 2018. Approximately 40% of all cancer patients can't seek effective treatment because they are diagnosed whilst in the later stages of cancer. Early diagnosis could save a significant portion of these people.

Making Accurate Diagnosis Accessible

According to WHO, about 70% of all cancer deaths occur in low- and middle-income countries. Part of this problem is the fact that people in these countries don't have access to proper testing. Our method could change that.

Personal Note

We'd like to thank the judges and organizers for putting together this amazing opportunity for us to submit our project to the Google AI Impact Challenge. We hope that through this project our team can be a major impact towards cancer diagnostics and grow the use of liquid biopsy testing towards the future! We'd also like to give a special thank you to The Knowledge Society, the youth accelerator we are a part of, for helping support our idea!

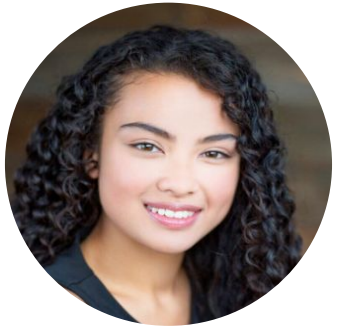
Michael, Seyone, Kiara and Adar

Seyone Chithrananda

seyonec@gmail.com

Michael Trinh Orzsag

trinh.orszag@gmail.com



Kiara Cuter

kmcuter@gmail.com

Adar Kahiri

adar.kahiri4@gmail.com