# Detection and Classification of cancer cells in MRI Images

## Problem Statement

An accurate classification of human cancer, including its primary site, is important for better understanding of cancer and effective therapeutic strategies development. The available big data of somatic mutations provides a great opportunity to investigate cancer classification using machine learning. In this research primary sites classification using machine learning and somatic mutation data is proposed. Here, in this research the patterns exploration of 1,760,846 somatic mutations identified from 230,255 cancer patients along with gene function information using support vector machine is proposed. In this a multiclass classification experiment over the 17 tumor sites using the gene symbol, somatic mutation, chromosome, and gene functional pathway as predictors for 6,751 subjects is to be performed. Adding the information of mutation and chromosome will improve the result. Among the predictable primary tumor sites, the prediction of five primary sites (large intestine, liver, skin, pancreas, and lung) could achieve the performance with more in *F*-measure. expected outcome of this research is that more accurate result will be generated for Detection and Classification of cancer cells in MRI Images.

## Background

Cancer is a complex disease, which is driven by the combination of genetic, environmental, and lifestyle factors. Among these factors, the combination of multiple genes driving cancer development varies considerably among cancer types and patients. During the past decade, investigation of mutations at both large-scale and specific loci has been made in order to increase our knowledge of the molecular heterogeneity in this complex disease. Notably, several large-scale, network-based cancer genome projects have generated multidimensional and genome-wide data. These projects include The Cancer Genome Atlas (TCGA) , Welcome Trust Sanger Institute's Cancer Genome Project , and the International Cancer Genome Consortium (ICGC) . These projects have dramatically advanced cancer research, especially in its genetics and genomics. A cancer somatic mutation landscape, primarily focusing on nucleotide change patterns (e.g., C->T) and mutation signatures in the cancer genomes, has been released to the community. Among these achievements, some have been translated into molecular diagnosis, better prognosis, and new targeted therapies. For example, the germline mutations in *BRCA1* and *BRCA2* confer high risks to breast and ovarian cancers. Their genotyping is used to determine susceptibility to breast and ovarian cancer. To monitor the treatment, the increased expression level of circulating tumor marker, human epidermal growth factor receptor 2 (HER2), is used to determine the treatment of a monoclonal antibody trastuzumab in breast cancer. However, cancer is strongly heterogeneous, and the cancer classification is a critical first step in the further investigation of the pathology of cancer and the development of effective treatments.

For cancer classification, the fundamental method is mainly based on the cell of origin or their histological types. During the last two decades, molecular profiling has been unveiled for classification of cancer types and subtypes, as well as assessment of heterogeneity of cancer samples However, as other data integration schemes, it presents a big challenge to develop an effective and comprehensive method for cancer classification.

## Methodology

### Step 1: Data collection and dataset preparation

This will involve collection of images from COSMIC database and preprocessing them, and extracting features.

### Step 2: Developing a classification based Model using machine learning and somatic mutation data

In this step a classification model is developed using machine learning and somatic mutation data for Detection and Classification of cancer cells in MRI Images.

**Step 3: Training and experimentation on datasets**

The classification based Model using machine learning and somatic mutation data will be trained on the COSMIC datasets to do the prediction accurately.

**Step 4: Deployment and analysis on real life scenario**

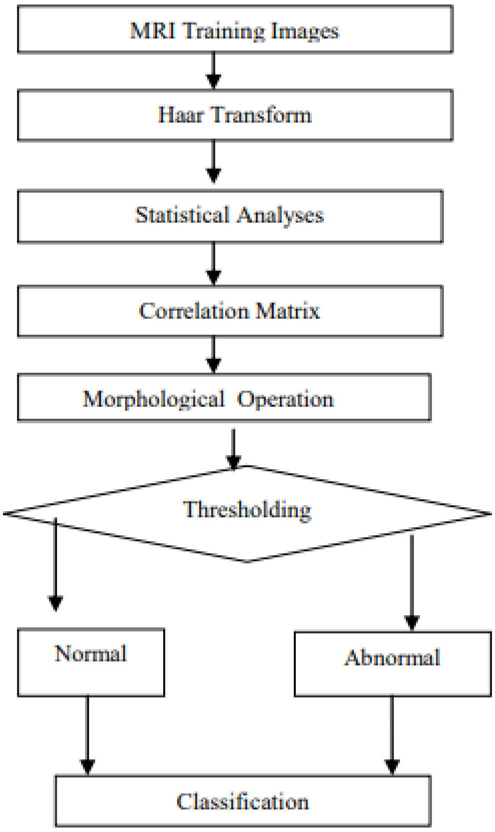

Figure 1 Block diagram of proposed system for Detection and Classification of cancer cells in MRI Images [Rani, Neha, and ShardaVashisth. "Brain Tumor Detection and Classification with Feed Forward Back-Prop Neural Network." *arXiv preprint arXiv:1706.06411* (2017).]

The trained and tested classification based model will be deployed in a real-life scenario for further analysis where Detection and Classification of cancer cells in MRI Images will be leveraged for further improvement in the methodology.

**Experimental Design**

**Dataset**

The COSMIC database is established to collect, store, and display somatic mutations and related information extracted from the primary literature on human cancers as well as those identified from cancer genome projects. The COSMIC data provides a consistent view of histology and tissue ontology with the mutation information. We downloaded the data from COSMIC website on April 18, 2014. The downloaded data contained 990,529 samples, 25,660 genes, 1,292,597 coding mutations, 1,528,225 noncoding variations, and 11,330 references.

**Evaluation Measures**

Evaluation is measured in terms of accuracy by using microaverage and macroaverage methods to report the accuracy, Precision Recall and F-measure.The performance of prediction on each primary tumor site by precision, recall, and F-measure are calculated as follows:

$$\mathrm{Accuracy}=\sum \mathrm{yiTP(yi)}\sum \mathrm{yiPred(yi)},$$
$$\mathrm{Precision(yi)}=\mathrm{TP(yi)Pred(yi)},$$
$$\mathrm{Recall(yi)}=\mathrm{TP(yi)True(yi)},$$
$$\mathrm{F\text{-}measure(yi)}=2\ast\mathrm{Precision(yi)}\ast\mathrm{Recall(yi)Precision(yi)}+\mathrm{Recall(yi)},$$

**Software and Hardware Requirements**

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, and libraries such as OpenCV, Tensorflow, and Keras will be utilized for this process. Training will be conducted on NVIDIA GPUs for training the classification based model for Detection and Classification of cancer cells in MRI Images