Report - PRML Major Project

Histopathological Cancer Detection

- Marri Bharadwaj (B21CS045)
Mohit Kumawat (B21CS046)
Shrashti Saraswat (B21CS081)

Abstract

This paper reports our experience with building a CNN model for predicting and labelling the images as normal (benign) or cancer. The train dataset contains about 220, 025 images and their classes/labels, which helped us train our CNN model on it. We have primarily used convolutional neural networks for this binary classification problem; but prior to this performed Random Sampling, Stratified Sampling as well as Dimensionality Reduction and compared their influence on our CNN model.

I. Introduction

Cancer is a complex and heterogeneous disease characterised by the uncontrolled growth and spread of abnormal cells. It can affect any part of the body and has the potential to be life-threatening if not diagnosed and treated early. To detect these in patients' samples, pathologists use different staining techniques to identify cancer in tissue samples. But manual cancer identification is inefficient because it is time-consuming, prone to human error, and limited by the observer's expertise. Thus, the aim of our project is to determine if a tumour is benign or malignant to improve accuracy and save labour costs.

Dataset

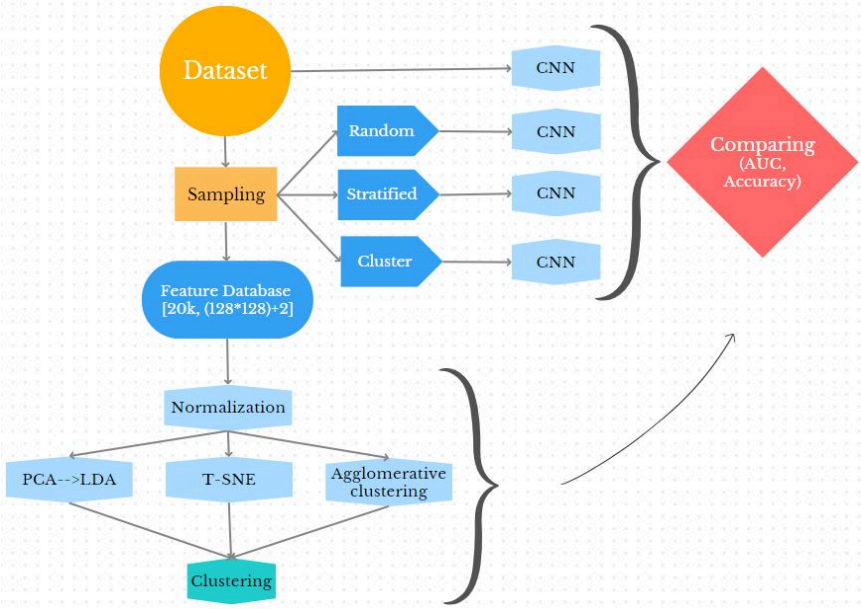
Cancer Images dataset: The dataset provided was the *train_labels.csv* dataset. It contains the images of the stained tissue samples along with their class labels as '0' for normal or benign and '1' for cancerous. The dataset consists of 220025 rows, where each row represents an image with its class.

II. Methodology

Overview

In this binary classification problem, we have chiefly used Convolutional Neural Networks (CNN) which applying the following techniques prior to it:

- CNN on complete dataset and on sampled dataset
  - Simple Random Sampling
  - Stratified Random Sampling
  - Cluster Sampling
- Dimensionality Reduction using -
  - PCA
  - LDA
  - t-SNE
  - Agglomerative Clustering



Exploratory Data Analysis and Pre-Processing

- We installed the necessary library packages and functions such as OpenCV, PyTorch and Sklearn, which would be used for the upcoming images' processing and CNN and sampling tasks; as well as for visualisation.
- We have directly imported the dataset into Google Colab from Kaggle. We have used the 'train_labels.csv' dataset to access the images and train our model. The dataset contains 220, 025 images with their respective labels as 0: 'No Cancer' or 1: 'Cancer'.
- Instead of training on the entire dataset, we have sampled 10,000 instances from each of the two classes and combined them to form our training dataset. We have then created our own dataset class. Using this class, we have created our dataset while applying transformations on the training dataset.
- We have applied various sampling techniques and data reduction techniques to our this training dataset such as Stratified, Simple and Cluster Sampling and PCA & LDA. We have then defined our own CNN class, and shifted to GPU for faster and more-efficient processing. We have specified BCE Loss as our criterion and Adam as our optimiser.
- We have then trained our CNN model on the differently sampled and transformed training dataset. During this training process, we have chosen our best model based on the loss in accuracy. At last, we have similarly transformed and created our testing dataset from the available 'test' dataset and classified its stained samples as 'Normal' and 'Cancer'.

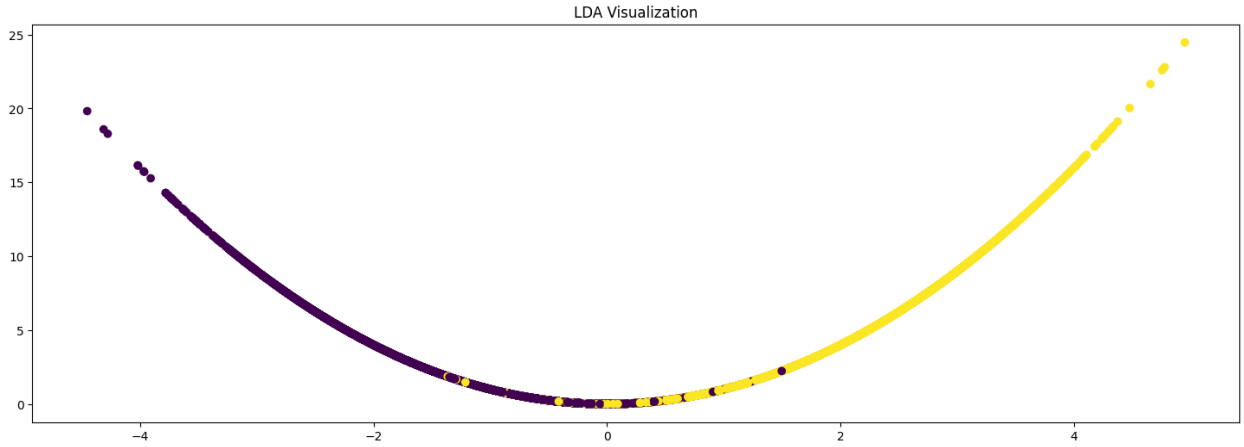
Implementation of Sampling Techniques & Dimensionality Techniques

We applied three types of sampling- ***Simple Random Sampling, Stratified Random Sampling and Cluster Sampling,*** in addition to ***Dimensionality Reduction methods: PCA and LDA***; as well as trained the complete dataset in the ***absence of any sampling or dimensionality reduction.***

- *Simple Random Sampling:* In simple random sampling, a subset of data is randomly selected from a larger dataset without any specific criteria. Each data point has an equal chance of being selected, and the sampling process is done without any external bias. Simple random sampling can be done with or without replacement.

- *Stratified Random Sampling:* Stratified random sampling involves dividing the population into homogeneous subgroups and then selecting a sample from each stratum in proportion to the population size. By ensuring that the sample is drawn from each stratum in proportion to its size, stratified random sampling reduces the sampling error and improves the precision of the estimates obtained from the sample.
- *Cluster Sampling:* Cluster sampling randomly selects clusters or groups of individuals from a population instead of randomly selecting individuals. These clusters are groups that share common characteristics. The individuals within the selected clusters are then sampled to form the final sample.
- *Convolutional Neural Networks:* CNN on a complete dataset, no sampling technique is used. The entire dataset is used for training and validation. The model is trained on the complete dataset in multiple epochs until the validation loss becomes stable. The dataset is divided into training, validation, and testing sets in a certain ratio. The training set is used to train the model, while the validation set is used to check the performance of the model during training.

- ***Dimensionality Reduction Techniques:***
  - ❖ *Principal Component Analysis:* PCA reduces the dimensionality of high-dimensional datasets by transforming the original variables into a new set of uncorrelated variables called principal components. It identifies the most important features that explain the maximum amount of variance in the data.
  - ❖ *Linear Discriminant Analysis:* LDA finds a suitable linear combination of features that can best separate classes or groups of data. It aims to find a projection that maximises the distance between the means of the two classes and minimises the variation within each class. LDA reduces the number of features required to classify data while preserving class discriminability.
  - ❖ *t-Distributed Stochastic Neighbour Embedding:* t-SNE works by mapping the high-dimensional data into a probability distribution in a lower-dimensional space and minimising the Kullback-Leibler divergence between the two probability distributions.
  - ❖ *Agglomerative Clustering:* It is a hierarchical clustering method that starts by treating each data point as a separate cluster and then iteratively merges the two closest clusters until a stopping criterion is reached.



III. Evaluation of Methodologies

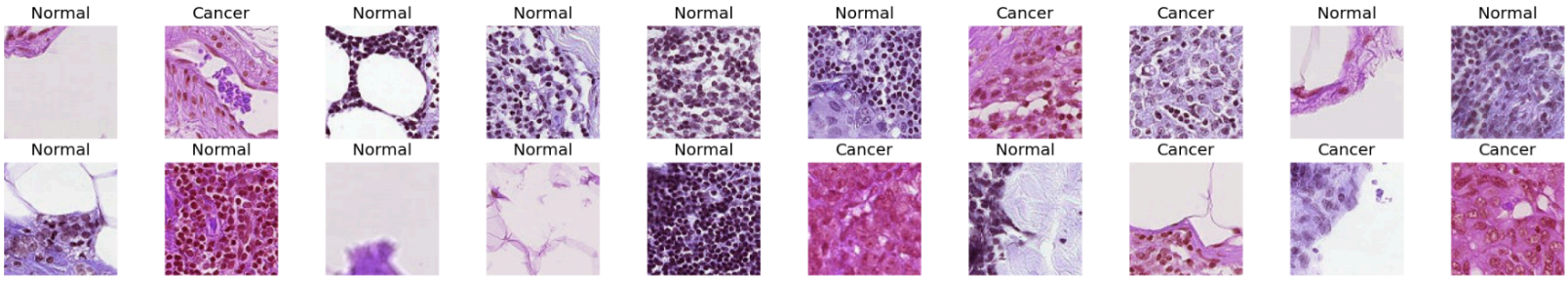
Model	Accuracy	AUC	Loss
Main CNN	95.17	0.9887	0.0759
Random sampled	92.42	0.9420	0.29164
Stratified sampled	92.31	0.9313	0.19897
Cluster sampled	93.75	0.9430	0.28363

Method	Accuracy	Final Verdict
PCA → LDA	86.64	Got two clusters at the end
T-SNE	Not Applied	No proper clusters formed
Agglomerative Clustering	50.2	No proper clusters formed

IV. Result and Analysis

From the table, CNN on the complete dataset has the highest accuracy and AUC score with least loss which is obvious considering it had trained on the complete dataset, sampling helps us reduce the massive training size to optimum value. Amongst the three sampling techniques, Cluster sampling performed the best in the metrics of accuracy and AUC while Stratified random sampling has the least loss.

Amongst the dimensionality reduction techniques, combining LDA after PCA generated a good score of accuracy; resulting in two clusters towards the end. Agglomerative Clustering had performed poorly with a binary classification equivalent of tossing a coin. Overall, we can conclude that while complete dataset training is preferable for small datasets, large ones require constructive sampling along with dimensionality reduction application.



Cluster Sampling + CNN Predicted Classes

Contribution:

Marri Bharadwaj (B21CS045)- Worked on the dataset using 3 different types of sampling and did CNN.
Mohit Kumawat (B21CS046)- Worked on dataset using dimensionality reduction techniques and did clustering.
Shrashti Saraswat (B21CS081)- Worked on complete dataset Main CNN model and did report.