Classification of Cancer Cells

Team Members

Name	Email	Student ID
Chengkai Lin	linchk20@lzu.edu.cn	320200945841
Wenbo Liu	wbliu20@lzu.edu.cn	320200931121
Xuda Han	hanxd20@lzu.edu.cn	320200945781
Yifei Guo	guoyf20@lzu.edu.cn	320200930901

Dataset: **Cancer Data | Kaggle**

Introduction

In this project, we mainly focuses on classification in order to assist doctors with initial filtering on cancer types. According to reports from WHO, cancer is one of the major causes of death, which led to near one sixth (about 10,000,000) death cases in 2020. Thus, cancer classification is very meaningful since it helps in the evaluation of population-based screening and early detection efforts. Several machine learning algorithms are applied, such as *decision tree*, *SVM* and *CNN*.

Data Description

This dataset is 570 by 32, and contains information on 569 cancer cells and 30 features to determine whether the cancer cells in the dataset is benign or malignant. All but one of the columns in the dataset are numeric, with the exception being the `diagnosis` column which shows whether the tumor is benign or malignant.

Hardware and Software Involved

- **Hardware**: Computers that meets the requirements.
- **Software**: Weka and Python.

Preliminary Plan and Timeline

Things to do	Estimated Time Cost
Exploratory Analysis: Have a deeper understanding of the dataset by data cleaning, preprocessing and visualization	2 weeks
Model Building: Divide the dataset into training set and testing set, then use them to build classification models and see if they meet the requirement	4 weeks
Summary: Get all the data team members create, and see if anything needs modification or better tuning	2 weeks
Report: Prepare documents for the report and presentation	2 weeks

Files to Deliver

All critical source files and report.