

Cancer, our global community's incurable, undiagnosable and life-threatening disease, continues to plague humanity, targeting the human body in a multitude of different ways. Scientists are baffled. How is it that our own bodies can malfunction and cause uncontrolled cell growth? How can we diagnose this disease before cancer cells start spreading throughout the body using lymphatic and circulatory systems? Does the distinct appearance of a tumour help classify whether or not its cells are cancerous and prone to metastasis? From previous research, benign tumours are visibly smoother with clear borders, compared to more harmful malignant tumours, which grow faster and therefore have less distinct borders. Can we use data to identify a correlation between appearance of a tumour and the likelihood of the tumour spreading throughout the body?

HIGHLIGHT questions that may require additional data or would benefit from:

Q: Time series analysis: Are there certain ages/ years for a human with a typical life span when certain types of cancers are more likely to occur? How has the pattern of the age of cancer patients changed over time, and can we predict future trends based on historical data?

- Q: Hotspot analysis: Can we identify "hotspots" of cancer development where multiple tumors occur with similar traits in close proximity to one another? This could be useful for researchers to target their treatments more effectively.

Q: Spatial analysis: Are there certain demographics of people that are more likely to get cancer? Are there patterns or trends within cancer data that can tell us something about the demographics and geography of tumors?

- Q: Tumor classification: Can we use machine learning and data science techniques to automatically predict the type of tumors based on features constituting of: location within the body, sex, age, prior health complications, etc. This could be vital in developing specific treatments and recognition devices.

- Q: Classification-Factor relationship analysis: Are there trends between types of tumors and their visual characteristics? Are certain tumors correlated with specific masses or growth rate?

- Q: Does an individual's race/ethnicity determine what type of cancer or existence of cancer?

- Q: Environmental analysis: Are there certain environmental factors that dictate whether a cancer is cancerous or not?

-Q: How do socioeconomic and racial disparities impact the treatment and outcome of cancer patients?

-Q: How do different cancer types react to different drugs?

-Q: How impactful are early screening and detection methods in treating patients?

EDA LIFE CYCLE:

1. Gather data from UCI Breast Cancer Dataset (1995)
 - a. <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>
2. Clean data
 - a. Remove any null values in dataset
 - b. Remove unnecessary data/columns
3. EDA
 - a. Use OneHotEncoder on variables (smoothness, size, concavity, etc)
 - b. Use logistic regression to predict whether smoothness of a tumour indicates that it is benign + whether or not size and compactness give an indication of the type of tumour a patient may have.
 - c. Use decision trees to predict whether smoothness of a tumour indicates that it is benign + whether or not size and compactness give an indication of the type of tumour a patient may have.
 - d. Visualize data using .corr() and matplotlib