A data pipeline is considered to be a series of steps or procedures that processes raw data from various data sources to a format version of the data for the sake of analysis or data modeling (*What Is a Data Pipeline | IBM*, n.d.). In the case of medical images and the machine learning algorithm to classify tumors, a batch processing pipeline will be used to stream new cases overnight to renew and improve the model to better perform within the hospitals. This model will receive stored medical data within a file or server for the sake of retraining the model through the use of TensorFlow. The OSEMN methodology for building pipelines will be used to further develop and complete the pipeline for the machine learning model.

The first step to the developing the pipeline is to develop the way that one obtains the data. As mentioned before, the pipeline will depend on batch processing for furthering the model accuracy at the desired hospitals. These updates will occur periodically and will extract raw data from different hospital servers and their affiliates, anonymize the data and finally retrain the model. The data extraction aspect of medical imaging is well established through the DICOM imaging standard. The DICOM imaging standard is well established enough to assume that most if not all hospitals and medical devices use the DICOM standard to create medical images. After extracting the data from servers, one will firsst have to work on anonymizing the data.

The DICOM Standard allows for a file type that contains not only image data, but also header data. As part of the Scrubbing or Data Cleanup, one will have to access the header data within the DICOM file to remove any identifying information (such as the name of the patient). For this, one must follow the HIPAA Privacy Rule De-Identification Methods that provide two methodologies (Office for Civil Rights (OCR, 2012)). The Expert Determination is the first of the two methods and mainly involves the application of statistical methods to remove important and personally identifiable information (Office for Civil Rights (OCR, 2012)). This method functions by having an expert and an identifier apply certain principles and statistical techniques to determine whether certain information will pose a risk to identify the patient (*DE-IDENTIFICATION Reference HIPAA Expert Determination De-Identification Method*, n.d.). Once a procedure to anonymize patients based on the data provided is done, the expert creates documentation and provides results from the anonymization (*DE-IDENTIFICATION Reference HIPAA Expert Determination De-Identification Method*, n.d.). This method is not viable for developing an automated data pipeline as the method depends solely on the expert to determine the fields may be high risk or low risk. Hence the second method, the Safe Harbor method, will be used to meet HIPAA standards. The Safe Harbor method functions by removing any potentially identifiable data from the data set (Office for Civil Rights (OCR, 2012)). The following fields are the kind that get removed from data sets based on the HIPAA Privacy Rules, Names, Geographic locations smaller than states, dates (with the exception of year), telephone numbers, vehicle identifiers, fax numbers, device identifiers, emails, URLs, social security numbers, IP addressess, medical record ids or numbers, biometric identifiers, health plan beneficiary numbers, photo Ids, account numbers, and finally certificates (Office for Civil Rights (OCR, 2012)). None of this field pose any potential risk to the performance of the model. At worst, there may be some biometric identifiers that could improve the performance of the model, but due to the anonymization standard by the HIPAA association, this may have to be removed from the model.

Once the data has been anonymized, one can move on to further scrubbing the data. The next key step is to further reduce the number of dimensions within the data set through the use of PCA and heatmaps. Heatmaps will first be used by converting textual data to categorical data and in turn using the correlation matrix which will tell the Pearson correlation coefficient of each variable present. This will allow one to remove the variables with low Pearson coefficients and leave variables with high correlation. Now that the number of features have been narrowed down, one can further narrow down the number of features through the use of Principle Component Analysis (or PCA). This procedure can be done by first standardizing the categorical values, then proceding to calculate the covariance matrix

and finally calculating the eigen values (Jaadi, 2019). Once the most important features have been found, one will explore certain patterns of the processed data.

With the key features extracted from the header data found within the DICOM file, one can now observe certain key patterns that may lead to diagnoses in tumors. One of the key features that will need to be explored is ethnicity. Although hospitals do not discriminate, it is clear that there are differences in survival rate based on race andn ethnicity (Ellis et al., 2018). For this reason, such a pattern will be explored throughout the data set and if found, then the data will be balanced to remove any bias that could be introduced unto the data set. The intent and purpose of the removal of this pattern from the dataset is to prevent the model from misdiagnosing certain ethnicity while under the impression that certain tumors for some ethnicities is larger than others. This is the main cause for the lower survival rate between races, as hispanics and african americans are usually diagnosed within the later stages as opposed to asians or caucasians (Ellis et al., 2018). Other key features that will be explored to balance the data will be sex, type of cancer, state, education level and more. Once the data has been fully explored and balanced, one can enter the modelling phase.

The modelling phase is where the data pipeline aspect will finish and the application will begin. The processed data will then be used to further train a machine learning model designed and fine tuned to predict malignancy and the tumor stage based on the image data using three convolutional layers paired with average pooling layers, and a flattening layer to convert the image into a vector that will then be appended to the vector containing the categorical values from the header. This vector will then pass through several linear (or dense) layers to provide a prediction of the malignancy of the tumor and the stage. Then the application will display the singular images with the diagnosis written at the bottom right of the image. The use will then be able to observe the predictions for interpretation. Alternatively, a dashboard will be provided containing a summary of the predictions based on the type of cancer images used for predictions and the stage related to them. The interpretation of the data will be mainly up to the user. A radiologist will be able to judge whether the prediction provided is true or not based on the statistical values provided of the model and the image itself. To further supplement and assist in the interpretation of the data, the app will also used the data before the flattening layer is reached to highlight the key feature of the image using a separate model structure.

Source(s):

12 Startups Bringing Artificial Intelligence To Cancer Diagnostics And Therapeutics. (2016, September 15). CB Insights Research. https://www.cbinsights.com/research/ai-startups-fighting-cancer/

Bazargan, M., Lucas-Wright, A., Jones, L., Vargas, R., Vadgama, J. V., Evers-Manly, S., & Maxwell, A. E. (2015). Understanding Perceived Benefit of Early Cancer Detection: Community-Partnered Research with African American Women in South Los Angeles. *Journal of Women's Health*, *24*(9), 755–761. https://doi.org/10.1089/jwh.2014.5049

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21552

Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P., Rebbeck, T. R., & Balasubramanian, S. (2022). Early detection of cancer. *Science*, *375*(6586). https://doi.org/10.1126/science.aay9040

Digital Mammography DREAM Challenge. (n.d.). Sage Bionetworks. Retrieved September 8, 2022, from https://sagebionetworks.org/research-projects/digital-mammography-dream-challenge/

Ellis, L., Canchola, A. J., Spiegel, D., Ladabaum, U., Haile, R., & Gomez, S. L. (2018). Racial and Ethnic Disparities in Cancer Survival: The Contribution of Tumor, Sociodemographic, Institutional, and Neighborhood Characteristics. *Journal of Clinical Oncology*, *36*(1), 25–33. https://doi.org/10.1200/jco.2017.74.2049

Hale, C. (2020, January 2). *An instant 2nd opinion: Google's DeepMind AI bests doctors at breast cancer screening*. Fierce Biotech. https://www.fiercebiotech.com/medtech/instant-second-opinion-google-s-

deepmind-ai-bests-doctors-at-breast-cancer-screening#:~:text=Google%27s%20DeepMind%20team%20showed%20that

IBM Cloud Education. (2020a, July 15). *What is Machine Learning?* IBM. https://www.ibm.com/cloud/learn/machine-learning

IBM Cloud Education. (2020b, August 17). *What are Neural Networks?* Www.ibm.com; IBM. https://www.ibm.com/cloud/learn/neural-networks

Individual Heath Insurance Options. (n.d.). Www.cancer.org. https://www.cancer.org/treatment/finding-and-paying-for-treatment/understanding-health-insurance/private-insurance-options.html

Kakushadze, Z., Raghubanshi, R., & Yu, W. (2017). Estimating Cost Savings from Early Cancer Diagnosis. *Data*, *2*(3), 30. https://doi.org/10.3390/data2030030

Kaur, S., Baine, M. J., Jain, M., Sasson, A. R., & Batra, S. K. (2012). Early diagnosis of pancreatic cancer: challenges and new developments. *Biomarkers in Medicine*, *6*(5), 597–612. https://doi.org/10.2217/bmm.12.69

Lee, S. (2014). *Benefits and limitations of regular cancer screening*. Canadian Cancer Society. https://cancer.ca/en/cancer-information/find-cancer-early/screening-for-cancer/benefits-and-limitations-of-regular-cancer-screening#:~:text=Cancer%20screening%20helps%20find%20cancer%20before%20it%20spreads%20when%20it

Ross, C., & Swetlitz, I. (2017, September 5). *IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close*. STAT; STAT. https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Savage, N. (2020). How AI is improving cancer diagnostics. *Nature*, *579*, S14–S16. https://doi.org/10.1038/d41586-020-00847-2

Smith, & Smith. (2022, May 11). *8 Reasons Why Your Doctor Missed a Cancer Diagnosis*. STL.News. https://www.stl.news/8-reasons-why-your-doctor-missed-a-cancer-diagnosis/519494/

Staff, N. (2022, March 22). *Can Artificial Intelligence Help See Cancer in New Ways? - National Cancer Institute*. Www.cancer.gov. https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging

Stage at Diagnosis | Cancer Trends Progress Report. (2019). Cancer.gov. https://progressreport.cancer.gov/diagnosis/stage

The Lancet. (2010). Late-stage cancer detection in the USA is costing lives. *The Lancet*, *376*(9756), 1873. https://doi.org/10.1016/s0140-6736(10)62195-2

Ultrasound for Cancer. (n.d.). Www.cancer.org. https://www.cancer.org/treatment/understanding-your-diagnosis/tests/ultrasound-for-cancer.html#:~:text=Ultrasound%20cannot%20tell%20whether%20a

Virnig, B. A., Baxter, N. N., Habermann, E. B., Feldman, R. D., & Bradley, C. J. (2009). A Matter Of Race: Early-Versus Late-Stage Cancer Diagnosis. *Health Affairs*, *28*(1), 160–168. https://doi.org/10.1377/hlthaff.28.1.160

Xie, Y., Bagby, T. R., Cohen, M., & Forrest, M. L. (2009). Drug delivery to the lymphatic system: importance in future cancer diagnosis and therapies. *Expert Opinion on Drug Delivery*, *6*(8), 785–792. https://doi.org/10.1517/17425240903085128

Zhang, C., Zhang, C., Wang, Q., Li, Z., Lin, J., & Wang, H. (2020). Differences in Stage of Cancer at Diagnosis, Treatment, and Survival by Race and Ethnicity Among Leading Cancer Types. *JAMA Network Open*, *3*(4), e202950. https://doi.org/10.1001/jamanetworkopen.2020.2950