Data Science in Breast Cancer Diagnosis
Abigail Granrud Data 100

Almost 280,000 people in America will be diagnosed with some form of breast cancer this year, making it one of the most prevalent types of cancer for women. According to the American Cancer Society, about 42,000 people are predicted to die due to breast cancer in 2020. An early diagnosis of breast cancer can greatly increase the prognosis and change of survival in a patient. Correct classification of the tumor type can also prevent patients from undergoing unnecessary treatment, and early examination can ensure that breast cancer patients are placed in the correct treatment category for their illness. Machine learning offers unique advantages in the screening and detection of critical features in complex data sets, and is often used in breast cancer pattern classification and forecast modeling.

To create a dataset from a patient to be used in the machine learning technique, doctors use fluid samples from patients with prominent breast masses. At the University Of Wisconsin Hospital at Madison, Wisconsin, Dr. Wohlberg uses a computer program called Xcyt, which takes a digital scan of the information and is able to analyze its cytological features. The program computes ten values for each nucleus that it sees in the scan, including the radius, texture, compactness, and symmetry. This information can now be processed through computers to determine whether the mass is malignant or benign. Each of the ten variables is assigned a “point value.” The more points that the mass receives, the more likely it is to be malignant. A less symmetrical nucleus or a larger radius would acquire more points for the mass, making it more likely to be malignant. After the initial scan is taken and the Xcyt program has assigned values to each nucleus, the data is processed. Dr. Wohlberg uses Spyder to interpret the data and determine how many of the tested masses were malignant. In his particular data set, 357 of the patients were identified as benign, and 212 were labeled malignant out of the 569 person data set. Dr. Wohlberg also uses the Label Encoder in Python to categorize this data, describing the patients in a more useful way, such as noting their age, race, country, or SES.

It can be useful to use computer programs such as Xcyt to scan the nuclei of each patients’ mass, but the trends that show when the data is interpreted become even more critical in determining the treatment each patient should get. When a program is used to determine whether each mass is malignant or benign, it assigns value, or importance to each factor that can be identified. After trials that indicate which factors of the nucleus are most important to the determination of malignance, doctors can see trends and know exactly what to look for in patients. Access to a large library of data

improves prognosis for all patients, not just those whose masses were scanned by computer programs. Collection of data from breast cancer patients brings doctors and researchers more insight into the trends followed by this relatively unpredictable disease, and can allow more patients to be treated properly.

As time progresses, more doctors will have access to more data about patients with breast cancer, and will be able to more accurately diagnose and treat their patients. Doctors and researchers worldwide are working to collect more data to make diagnoses more predictable. A lot of progress has been made in the past years with breast cancer, but medicine still has a long way to go. Hopefully, using data science, doctors will be able to save more lives of breast cancer patients.