

In the world of medical science, predictive modeling shines as a ray of hope. It opens new paths for diagnosing and comprehending breast cancer. Through the power of statistical methods, machine learning, and artificial intelligence, researchers and doctors can now foresee disease outcomes, categorize patient risks, and personalize treatment options with remarkable accuracy. These predictive models are like sophisticated detectives, diving deep into extensive and intricate data to uncover hidden patterns and links that might otherwise go unnoticed.

There are two types of tumors that are generally classified in the cases of cancer, Benign and Malignant. Benign tumors are dormant tumors which are noncancerous and do not affect the body at this stage. Malignant tumors however, are cancerous and are classified by stages as to what severity the cancerous tumor holds.

We delve into a comparative analysis of various machine learning and deep learning models for breast cancer diagnosis, leveraging the Wisconsin Breast Cancer Dataset as our primary data source. Our methodology systematically addresses the challenges of high-dimensional data through Principal Component Analysis (PCA), which simplifies the dataset to its most informative features, thereby improving computational efficiency and model interpretability. We delve into the capabilities of Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) for their profound ability to recognize complex patterns within the data, a crucial aspect for accurately classifying cancerous tissues from medical images.

To enhance the predictive accuracy and robustness of our models, we employ a variety of ensemble techniques, including AdaBoost, Bagging Classifier, and Random Forest, which integrate multiple learning algorithms to reduce overfitting and improve generalization. Our approach also incorporates AutoML tools, like the Tree-based Pipeline Optimization Tool (TPOT), to automate the selection and optimization of machine learning pipelines, ensuring that the best possible model is identified through an exhaustive search over a predefined space of algorithms and parameters.

Furthermore, we evaluate the performance of specific algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machines (SVM), alongside advanced models like MobileNetV2, EfficientNetB0, and DenseNet121, to determine their efficacy in the context of breast cancer classification. Special attention is given to Gradient Boosting algorithms like XGBoost and LightGBM for their exceptional performance on structured data, as well as the exploration of Generative Adversarial Networks (GANs) for generating synthetic medical images to augment our dataset.

This holistic approach to model evaluation and selection is aimed at identifying the most effective combination of techniques for early detection of breast cancer, contributing to the body of knowledge by not only enhancing the accuracy of diagnosis but also by improving the

understanding of how different models perform under various data conditions. Through this comparative study, we aim to delineate the strengths and limitations of each model, providing a roadmap for future research to build upon. We are doing a comparative study of this scale to identify the parameters that give us the most information about the data and its effects. Accuracy of model on the dataset helps us know which combination of parameters is most suited for helping in the early detection of breast cancer, which can help save a lot of lives as treating cancer in early stages has way higher chance of survival and a lot of the cases can be cured before this becomes lethal.

In later stages of our research, we plan to integrate Explainable Artificial Intelligence (XAI) to further enhance the interpretability and transparency of our models. XAI will allow us to uncover the underlying mechanisms of our AI-driven predictions, providing insights into how decisions are made. This advancement is pivotal for fostering trust among clinicians and patients by making the diagnostic process more understandable and justifiable, thereby bridging the gap between complex machine learning models and their real-world applications in medical diagnostics.