**INTRODUCTION**

Healthcare is a widespread field that focusses on betterment of health with the application of improved principles like diagnosis, treatment and maintenance. An efficient health care system can subsidize to an important part of a country's economy, development and industrialization. Healthcare is one of the most stimulating field in data analytics. Deploying a healthcare analytics suite can assist healthcare providers improve upon several areas of operation. The future of healthcare will be driven by the digital transformation and data analysis.

As with any industry, a Business Intelligence (BI) and reporting solution can significantly improve operational efficiency, reduce costs, and streamline processes through measuring and leveraging KPIs to spot opportunities and inform decision-making. The data is highly valued for improving diagnosis and can help to analyse a whole host of issues including symptoms, pharmaceuticals and dosage. Without this information, it would be significantly more stimulating for medical specialists to come to the correct choices. The business intelligence derived from data analytics provides answers in near real-time based on a huge amount of data. By uniting this info, it interprets into actionable decisions that provide patients with better healthcare.

In the field of healthcare, there is a variety of datasets that ranges from text data to huge image data files. Analysing them and reaching to valuable conclusion is always challenging in terms of technicalities and aspects of viable solution. Data integration with data-driven approaches can successfully generate results. Dirty data can quickly derail any analytics problem, especially during the case when bringing together disparate data sources. Data analytics has contributed a lot in finding optimised results in diseases like heart -failure, kidney diagnoses, laparoscopy etc. With the supervision of data analytics, helpful understandings could be withdrawn that will provide better decision-making process. As the volume of healthcare data grows there has been great impact on its critical cost, security and performance issues. Robust metadata and strong stewardship protocols makes it easier for the organisation to query their data and get answers that are being expected. The ability to query data is foundational for reporting and analysis. The most significant part is a clean and engaging data visualisation that can make it much easier for a clinician to absorb information and use it properly. In our project, the major part will be carried out for the purpose of cleaning and preprocessing. Data integration for the model building has an equal impact to generate reports that forms the basic building block for analysis and finding insights. Machine learning algorithms will play an equal role to find the accuracy and helps in minimising errors. In our case we will be using supervised machine learning models like KNN, Random Forest , Decision Trees and Linear Models. We will be working on Cancer dataset primarily focussed on breast cancer. The major objective is to mitigate the causes and provide a reasonable remedy by predictive analysis that will be helpful for society.

The breast clinical data we found is publicly available on Kaggle. So, it is following all the aspects of ethics (Consent, Clarity, Consequences, Control, Consistency). The data that has been chosen is free from bias as it has been verified. The data is transparent as it will not hamper the privacy concerns of the strata of the sample space.

The data is significant size as it has attributes that are relevant for findings. We will be considering with the image dataset to compare the insights and reach a valuable conclusion. Based upon the size of dataset, various classification and prediction models will be applied and optimised results will be generated. The volume of data that has been selected could be one of the challenges, but this can be resolved with appropriate modelling methodologies.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

**Some Risk Factors for Breast Cancer**

The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

**Age.** The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.

**Personal history of breast cancer.** A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.

**Family history of breast cancer.** A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.

**Genetic factors.** Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

**Childbearing and menstrual history.** The older a woman is when she has her first child, the greater her risk of breast cancer. Also at higher risk are:

* Women who menstruate for the first time at an early age (before 12)
* Women who go through menopause late (after age 55)
* Women who’ve never had children