**Reliable detection of breast cancer using Machine learning algorithms Gray-Level Co-occurrence Matrix (GLCM)** and **Local Binary Pattern (LBP)**.

**Chapter 1**

**Introduction:**

Despite decades of research and advances in technology across many fields, cancer remains a formidable challenge. Significant progress has been made in diagnostic techniques and treatment approaches, but important gaps still exist particularly in the fight against cancer. According to the World Health Organization (WHO), breast cancer is both the most common cancer and the leading cause of cancer‐related death among women worldwide.

According to recent studies, breast cancer accounts for more than 29 % of all new cancer cases in the USA and over 24 % in Europe. However, early detection remains a major global challenge. The quality of diagnostic images and the expertise of medical personnel are critical to identifying tumours at a treatable stage. Unfortunately, many deaths result from late diagnoses, when available treatments are far less effective.

To this day, breast cancer remains a significant global health challenge, particularly among women. However, one major positive development is the critical role of mammography in detection and diagnosis. Despite the challenges, mammographic imaging has proven essential in assessing a patient’s condition, determining whether the cancer is in an early or critical stage, and guiding timely treatment decisions to prevent further spread. Early detection continues to have a profound impact on improving treatment outcomes and overall patient health.

Although mammograms play an important role in the early detection of breast cancer, they have proven to be insufficient on their own, even today. A wide range of advanced techniques and imaging technologies are currently used to detect abnormal breast tissue; however, these methods still have limitations and cannot guarantee consistent early diagnosis. While these technologies have contributed to earlier treatments and improved survival rates, the medical community continues to seek more reliable solutions. As the fields of Artificial Intelligence (AI) and Machine Learning (ML) continue to evolve and advance across many sectors, including healthcare, there is growing interest in leveraging these technologies to develop more accurate, efficient, and accessible tools for the early detection and diagnosis of breast cancer.

Despite the considerable progress made with Artificial Intelligence (AI) and Machine Learning (ML) in recent years, several key challenges still persist. One major issue is the heterogeneity of medical images, which plays a critical role in the development of protocols and in the effectiveness of feature extraction techniques. While technology has certainly made the detection process faster and more accessible, concerns remain particularly among medical professionals about the reliability of results generated by AI systems. A significant factor contributing to this distrust is the "black box" nature of many AI models; the internal decision-making processes and data extraction methods are often opaque and lack transparency. This lack of explainability makes it difficult for doctors to fully understand or verify how certain diagnostic conclusions are reached, leading to hesitation in fully adopting complex AI-based diagnostic tools in clinical practice.

One of the key challenges in developing effective AI models for cancer detection lies in several critical factors. First, the size of the dataset is essential—larger datasets generally lead to more reliable and generalizable models. Equally important is the balance within the dataset, particularly in the case of breast cancer, where there must be an adequate and proportional representation of both benign and malignant cases. This balance is crucial for training models that can accurately differentiate between the two categories. Additionally, the quality and reliability of the annotations used in both the dataset and the model training process play a vital role. Poor or inconsistent labelling can significantly impact the model’s performance, leading to misclassifications and reduced trust in the system's results. Therefore, maintaining high standards in dataset preparation and annotation is fundamental for building accurate, trustworthy, and clinically useful AI systems.

Despite the presence of many challenges in this field, researchers have proposed various solutions to improve the reliability and performance of AI models. One common approach involves refining and adjusting the datasets, ensuring accurate labelling and consistent annotations to support model training. To address classification tasks in breast cancer detection, researchers have increasingly turned to advanced techniques such as Convolutional Neural Networks (CNNs). CNNs are widely recognized as powerful end-to-end classifiers that operate in multiple stages, aiming to achieve an optimal balance between accuracy, interpretability, and efficiency. However, despite their proven effectiveness, CNNs remain highly complex, requiring a well-designed and carefully structured architecture to perform reliably. In addition to CNNs, other traditional machine learning methods—such as Support Vector Machines (SVM), K-Nearest Neighbours (K-NN), Naïve Bayes, and Random Forests—are also commonly used. While these methods are considered reliable, they too present significant challenges in terms of complexity, parameter tuning, and scalability, especially when dealing with large and diverse medical imaging datasets.

**Proposed method and similar work SCOPE!!**

The main goals of this thesis are limited in these 4 steps:

* Understanding the problem, one of the most critical issues in cancer detection is identifying the root causes of diagnostic failures and finding effective ways to overcome them. Many studies have concluded that poor clarity and quality of mammography images significantly contribute to inaccurate results. In addition, properly labelling and interpreting these images presents a major challenge, even for experienced professionals. As a first step in addressing this issue, I focused on gaining a deeper understanding of the images within the dataset. I carefully studied the image notes and metadata to better comprehend the data structure, with the aim of improving my model’s ability to accurately detect cancerous tissue. This foundational step was essential in helping the model recognize patterns and anomalies more effectively.
* Limitation of the Dataset, another major challenge lies in the limitations of the MIAS dataset. Although MIAS is a reliable and widely used dataset in many scientific studies, it is relatively small in size, containing only 324 mammography images—fewer than 100 of which are from cancer patients. This lack of sufficient data can lead to problems such as overfitting, where the model performs well on training data but poorly on unseen data. Such limitations may significantly impact the accuracy and generalizability of the final results. To address this, it is crucial to design and structure the program, code, and AI model in a way that minimizes the negative effects of limited data. Techniques such as data augmentation, regularization, and careful model validation become essential to ensure the model remains robust and reliable despite the dataset's constraints.
* Data Labelling Challenges, another significant issue is the lack of clear, structured labelling within the dataset. The MIAS dataset and its accompanying documentation do not provide pre-labelled data in a format that is directly usable for machine learning models. As a result, one of the first essential steps in the coding process is to assign appropriate labels to each image—typically, '0' for normal and '1' for cancerous cases. Although the documentation does include information about the type of condition (e.g., normal, abnormal, benign, or malignant), it does not present this information in a machine-readable or consistent format. Therefore, additional preprocessing is required to extract and standardize these labels, ensuring that the model can properly differentiate between various types of tissue and cancer states during training and evaluation.
* Method and Algorithm, for this project, the first step involved labelling all the images in the dataset based on the information provided in the documentation. Since the images are grayscale mammography scans, two specific image processing algorithms were applied during the coding phase: **Gray-Level Co-occurrence Matrix (GLCM)** and **Local Binary Pattern (LBP)**. Both of these algorithms are widely used for texture analysis in medical imaging. In this project, they were used to extract meaningful features such as the **mean** and **average intensity** from the images. These extracted features serve as the foundation for training the model to distinguish between normal and cancerous tissues. A more detailed discussion of these methods and their implementation is provided in Chapter 3.

**Structure of the work**

The remainder of this thesis is organized as follows:

**Chapter 2** presents a comprehensive literature review, covering a wide range of studies focused on reliable breast cancer detection using Artificial Intelligence (AI) and Machine Learning (ML). This chapter examines previously developed models, their methodologies, limitations, and how researchers have addressed key challenges in the field. It also explores the types of datasets used, classification techniques, data handling strategies, and evaluation methods applied in earlier works.

**Chapter 3** forms the core of this thesis, detailing the proposed methodology of the research. It outlines the problem statement, the research objectives, and the specific algorithms and techniques employed—such as GLCM and LBP. This chapter also covers the implementation process, including how the data was handled, how challenges were approached, and how the overall coding structure was developed and organized.

**Chapter 4** evaluates the effectiveness and reliability of the proposed method. It includes a critical analysis of whether the project’s objectives and requirements were successfully met. Additionally, it presents the visualization of results, allowing readers to assess whether the system functions as intended and to what extent it achieves accurate detection.

**Chapter 5** concludes the thesis by summarizing the main findings and contributions of the project. It also discusses the limitations encountered during the research and proposes directions for future work. Key issues, potential improvements, and opportunities for further exploration are highlighted to guide future studies in this area.

**Chapter 2**

**Literature Review**

In Chapter 2, the focus is primarily on building a strong foundational background for the topic. This includes identifying the key requirements for breast cancer detection and developing a clear understanding of the problem domain. The chapter reviews previous research efforts, exploring the various methods and approaches taken by different researchers, as well as the limitations they encountered. Alongside this, the chapter provides a critical examination of current methodologies and highlights potential improvements that could enhance the quality of existing techniques. By doing so, it offers a deeper insight into both the strengths and gaps in current research, laying the groundwork for the proposed approach in the following chapters.

**Understanding the problem and data**

One key issue in the mammography image is that data is not always sorted and labelled correctly so without studying the data and observing closely research cannot be very successful in overall.