AN AI-ASSISTED CLINICAL DECISION SUPPORT SYSTEM FOR EARLY BREAST CANCER DETECTION AND RISK STRATIFICATION USING CELLULAR MORPHOMETRICS

THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

BIOTECHNOLOGY

Submitted by

JAYNA BHATTACHARJEE

24/MSCBIO/05

Under The Supervision Of

Dr. Smita Rastogi Verma

Assistant Professor, Department of Biotechnology

DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi – 110042, India

May, 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, New Delhi, 110042

DECLARATION

I, JAYNA, 24/MSCBIO/05, hereby, certify that the work which is being presented in the thesis entitled "**An AI Assisted clinical support system for early breast cancer detection and risk stratification using cellular morphometrics**" in partial fulfilment of the requirements for the award of the Degree of Master of Science, submitted in the Department of Biotechnology, Delhi Technological University is an authentic record of my own work carried out during the period from 2024 to 2026 under the supervision of Dr.Smita Rastogi Verma. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement mailed by the candidate is correct to the best of our knowledge.

Signature of Supervisor

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, New Delhi, 110042

CERTIFICATE BY THE SUPERVISOR

This is to certify that the Dissertation Project titled "**An AI Assisted clinical support system for early breast cancer detection and risk stratification using cellular morphometrics**" which is being submitted by Jayna, 24/MSCBIO/05, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is a record of the work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Date:

Dr. Smita Rastogi Verma

Supervisor
Department of Biotechnology
Delhi Technological University

Prof. Yasha Hasija

Head of the Department
Department of Biotechnology
Delhi Technological University

ACKNOWLEDGEMENT

My heartfelt thanks go to **Dr. Smita Rastogi Verma**, Assistant Professor, Department of Biotechnology, Delhi Technological University for her invaluable guidance, constant encouragement and scholarly mentorship during this research work. She has been an inspiration in terms of computational biology and drug discovery skills.

I also express my thanks to the Department of Biotechnology, Delhi Technological University for giving me computational infrastructure and academic resources to conduct my work.

I would like to express my sincere gratitude to my colleague and my friends **Teena Bhardwaj** and **Simran kumari** for her constant support and help.

ABSTRACT

Breast cancer remains one of the leading causes of cancer-related mortality worldwide, with delayed diagnosis significantly reducing treatment effectiveness and patient survival rates. Conventional diagnostic workflows involving pathological interpretation and radiological assessment are often time-intensive and subject to inter-observer variability. Recent advances in Artificial Intelligence (AI) and machine learning have demonstrated substantial potential in enhancing diagnostic accuracy, improving clinical workflow efficiency, and enabling precision-based healthcare.

This thesis proposes an AI-assisted Clinical Decision Support System (AI-CDSS) for early breast cancer detection and risk stratification using structured cellular morphometric features. The study utilizes the Breast Cancer Wisconsin (Diagnostic) dataset consisting of 569 patient samples and 30 quantitatively extracted cytological features derived from digitized fine needle aspiration images. Four machine learning algorithms, namely Logistic Regression, Random Forest, Gradient Boosting, and Multi-Layer Perceptron Neural Network, were developed and comparatively evaluated for binary tumor classification into benign and malignant categories. Data preprocessing involved normalization using standard scaling and dataset partitioning into training and testing subsets for unbiased performance evaluation. Among the evaluated models, Logistic Regression demonstrated superior clinical applicability with an AUROC score of 0.9960, sensitivity of 0.9286, and specificity of 0.9861. Explainable AI analysis further identified “worst area” and “worst concave points” as the most influential predictive features associated with malignancy. A probability-based risk stratification framework was also developed to categorize patients into low-risk, intermediate-risk, and high-risk groups.

The proposed AI-CDSS offers a transparent, computationally efficient, and clinically interpretable framework for supporting oncologists in early breast cancer diagnosis and patient triaging. The study highlights the potential integration of explainable AI systems into modern oncology workflows while emphasizing the importance of maintaining physician oversight for clinically ambiguous cases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AI-CDSS	Artificial Intelligence-Assisted Clinical Decision Support System
AUROC	Area Under Receiver Operating Characteristic Curve
CDSS	Clinical Decision Support System
CNN	Convolutional Neural Network
EHR	Electronic Health Record
FNAC	Fine Needle Aspiration Cytology
FNA	Fine Needle Aspiration
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
XAI	Explainable Artificial Intelligence
MLP	Multi-Layer Perceptron
WHO	World Health Organization
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
RF	Random Forest
GB	Gradient Boosting

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND AND GLOBAL SCENARIO OF BREAST CANCER

Breast cancer is the most prevalent type of cancer in women across the globe. Breast cancer is the most common cancer in women worldwide.

Cancer remains one of the top causes of death globally and remains a big problem in the health care systems, economy and society worldwide [1]. The rising trend in cancer has been observed over the last few decades, owing to various interconnected underlying factors like population ageing, rapid urbanization, environmental pollution, poor dietary patterns, lack of exercise, hormonal imbalance and genetic susceptibility [2]. Breast cancer is one of the most prevalent cancers in the world, and is now among the leading causes of breast cancer-related mortality among women globally [3].

Breast cancer occurs when cells start to grow and multiply in the breast tissue in an uncontrolled way. The abnormal cells can be limited to the duct or the lobules in the early stages. If the disease is missed in time, however, the malignant cells may travel the bloodstream and invade nearby lymph nodes, and eventually migrate to other organs like lungs and liver, brain, and bones [4]. Early detection is key to survival and recovery from the disease. An early diagnosis makes treatment more effective, boosts survival rates and lessens the impact of the complications of advanced-stage cancer [5].

Breast cancer has become a far more common condition in recent years in developed and developing countries. In countries with high level healthcare systems, screening programmes are organized, medical infrastructure is available and better, awareness is raised, and oncology services are available. Many low- and middle-income countries still have a delayed diagnosis and poor diagnostic facilities, as well as a lack of trained experts and access to modern treatment options [6]. Consequently, many of these patients living in resource limited areas are diagnosed at much later stages, resulting in less likelihood of successful treatment and recovery.

Breast cancer is more than just a physical disease. Patients and their families frequently experience high levels of emotional, psychological, financial and social disruptions as a result of the disease [7]. Long term treatment – surgery, chemotherapy, radiotherapy and hormonal treatments – require physical tiredness and can be financial burden. Bother patients suffer from fear, anxiety, loss of quality of life and social isolation throughout the treatment and recovery process. Therefore, the early detection, the improvement of the diagnostic accuracy and the development of efficient clinical management systems are still some of the most important priorities in modern oncology and healthcare research [8].

1.2 CONVENTIONAL DIAGNOSTIC APPROACHES AND THEIR LIMITATIONS

Breast cancer diagnosis is carried out by several procedures such as Mammography, Ultrasound imaging, MRI, Fine needle aspiration cytology (FNAC), histopathological examination and biopsy interpretation. These techniques have clinical significance and are basic to the diagnosis of cancer. But routine diagnostic workflows come with a number of drawbacks. Histopathological interpretation and radiological analysis are dependent on the experience and observational skills of pathologists and radiologists and require a lot of clinical expertise. A major limitation is inter-observer variability in which various clinicians can analyze the same sample differently. In some cases, the cells have been altered in a way that makes them look like cancer but are not, making it hard to see and leading to a risk of delayed or incorrect diagnosis. Moreover, manual examination of pathological slides and imaging data is tedious, time-consuming, and labor intensive. Hundreds of samples can be tested on a single day by clinicians in high-volume hospitals, so this can also be a factor in diagnostic error due to clinician fatigue. Resource-limited healthcare systems are beset with other issues, including:

- Drug shortages and insufficient number of qualified oncologists and pathologists.
- Limited diagnostic infrastructure
- Delayed laboratory reporting
- Financial constraints
- drop in access to specialised cancer centres

This means that many breast cancer patients are diagnosed at advanced stages, which impacts on the effectiveness of breast cancer treatments and the survival rates for patients. These restrictions highlight the importance of intelligent computational systems to help the clinician to make more rapid and accurate and reproducible diagnosis.

1.3 Artificial Intelligence and Machine Learning in Healthcare

In healthcare, AI technologies are increasingly finding applications in diagnostic medicine, genomics, radiology, pathology, drug development, robotic surgery and personalized medicine. The most significant advantage of AI is that it can handle large amounts of multidimensional data at high speeds and with great accuracy.One of the most prominent fields of AI is machine learning, which allows computational models to learn from past data and produce predictive results without being explicitly programmed. These algorithms learn statistical patterns in the relationships between biological variables and disease outcomes. Healthcare is highly data-intensive, with a plethora of data being generated through:

Electronic health records (EHRs)

• Medical imaging
• Genomic sequencing
• Histopathological slides
• Laboratory investigations

Wearable health monitoring devices are increasingly becoming a reality. It is far from easy to manually interpret these large and complex datasets. AI-driven systems are thus a scalable and efficient way to obtain clinically relevant information. Oncology is an area where machine learning is implemented in the following ways:

• Tumor classification

• Cancer screening
• Prognostic prediction
• Risk assessment
• Treatment response evaluation
• Precision medicine
• Survival prediction

Medical image analysis has seen remarkable advancements with deep learning models like Convolutional Neural Networks (CNNs), and Natural Language Processing (NLP) techniques are also proving valuable in extracting structured data from clinical reports. AI is poised for significant progress in healthcare, but it is not without its hurdles, including algorithmic bias, data privacy concerns, ethical implications, regulatory approval hurdles, and interpretability issues.

1.4 UNDERSTAND THE CONCEPT OF EXPLAINABLE AI AND CLINICAL DECISION SUPPORT SYSTEMS.

The "black box" issue is one of the primary drawbacks of numerous AI applications. Clinicians have no knowledge of how predictions are made in several complex machine learning models. The opacity of this has raised issues of trust, accountability and ethical application within healthcare environments. Explainable Artificial Intelligence (XAI) is designed to solve this problem by offering insights into algorithmic decision maker processes that are interpretable. Explainable AI techniques enable clinicians to discover the most important biological or clinical factors for a prediction. Explainability is of paramount importance in medical diagnostics because doctors are required to explain their decisions in the clinic or treatment and their recommendations. Clinical Decision Support Systems (CDSS) are computer-based systems that help healthcare professionals to analyzing patient data and make decisions based on evidence. Today, AI-driven CDSS platforms incorporate the use of machine-learning algorithms, predictive modelling and visualization systems to assist with diagnosis, risk assessment, and patient management. AI-driven CDSS platforms provide a number of benefits, such as:

• Improved diagnostic consistency
• Faster clinical workflows
• Reduced interpretational burden
• Enhanced evidence-based medicine
• Personalized patient management Minimized unnecessary invasive procedures,

In the present study, we are interested in creating a Clinical Decision Support System (CDSS) that uses explainable AI (XAI) to help detect and classify breast cancer at an early stage.

Problem statement and research objectives are introduced by 1.5. The multiple machine learning research studies have shown promising diagnostic accuracy of breast cancer prediction but there are still several limitations in real world clinical application which could not be resolved.Numerous AI diagnostic systems available today are designed to achieve high levels of accuracy in predicting a disease, but do not take into account the practicalities of implementation in a clinical setting. In some instances, algorithms are very complicated "black box" systems, and clinicians don't understand what the algorithms are basing their predictions on. This lack of interpretability diminishes physician trust and lowers the chances for successful integration into the healthcare workflow.Another challenge is that most of the existing studies only have a binary output (benign versus malignant) without clinical confidence (probabilities) or meaningful patient risk stratification. In real healthcare, doctors will more often need a range of risk assessments than just

a pure classification. In addition, most health facilities, particularly in rural areas of the developing world, still have significant infrastructural and operational problems, including:

- There is a scarcity of trained oncologists and pathologists.
- Delayed pathological reporting In diagnostic centres there is a significant number of patients.
- Limited access to high-tech screening tools
- Often expensive, repeated diagnostic procedures cost money
- Inconsistency in the diagnosis of resource-poor environments.
- As the amount of health-related information grows, so do the difficulties involved in interpreting information manually.
- Structured and unstructured clinical information is generated in large volumes in modern oncology, including radiological images, pathohistological examination, laboratory investigations and EHRs. Manually dealing with these multi-dimensional data is labor-intensive and prone to human error.

Furthermore, false negative diagnostic results are still more harmful in oncology, as cancer diagnosis can be delayed for a long time which may substantially decrease the probability of survival and efficacy of treatment. Hence, the need for very sensitive and clinically reliable computational systems. This study tackles these challenges by designing an Explainable AI (XAI) empowered Clinical Decision Support System (CDSS) that can perform the following:

- Accurately distinguishing benign and malignant tumors
- Ensuring that the probability of the risk level is provided
- Supporting physician decision-making

To make AI more understandable and transparent.To make AI more interpretable and explainable. Reducing diagnostic ambiguity Guidance on improving workflow in clinical settings The prime aims of the present study are:

1. Construction of machine learning models for breast cancer classification by structured cellular morphometric features.
2. To assess and compare the performance of several machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting and Neural Network.
3. To apply probability to clinical risk stratification to better categorize patients. To detect important biological cellular attributes that contribute to malignancy prediction.
4. To make AI more transparent and interpretable using explainable AI methods.
5. To build up a clinically relevant Clinical Decision Support System with AI assistance to assist oncologists during diagnosis procedures.
6. To set up a computational setting that can be used as a baseline for future multimodal precision oncology systems.

The scope and significance of the study are identified. This study aims to develop and test an AI-based Clinical Decision Support System based on structured cellular morphometric data from the Breast Cancer Wisconsin (Diagnostic) dataset. The study focuses on whether machine learning algorithms can recognize patterns related to whether a fine needle aspiration biopsy image of a breast lesion is malignant or benign based on quantitative cytological parameters extracted from the image. This study concentrates on:

The data was preprocessed and normalized.Data preprocessing and normalization were conducted.

The goal is to develop machine learning models.Objectives are to build machine learning models.

- Comparative algorithmic evaluation

• Risk stratification modelling
• Explainable AI integration

Whether the results of the predictive outcomes can be applied clinically. The study is mainly centered on biomedical data that exists in the form of a structured table, as opposed to imaging or genomic sequencing data. The developed framework, however, can be combined with other health care modalities like:

• Histopathological image analysis
• Mammographic screening systems
• Radiological imaging platforms
• Genomic and Proteomic Data sets
• Electronic health record (EHR) systems
• Cloud-based clinical infrastructures

The relevance of this study is that it attempts to find a balance between the computational performance and actual clinical use. The proposed framework is different from several existing AI models, which rely solely on the prediction accuracy, in terms of:

• Clinical interpretability
• Explainable prediction mechanisms
• Probability-based risk assessment
• Transparent decision support
• Physician-oriented implementation

This study can add to contemporary precision oncology through the generation of a system that enables evidence-based clinical decision making.

1.6.1 CLINICAL SIGNIFICANCE AN EARLY DIAGNOSIS OF MALIGNANT TUMORS

• Improved diagnostic consistency Reduced false negative results
• Faster patient triaging
• Saves pathologists and oncologists from the burden of making the diagnosis. Osteogenic promotes individualized treatment solutions.Osteogenic supports individualized treatment solutions.

1.6.2 TECHNOLOGICAL SIGNIFICANCE THE INTEGRATION OF EXPLAINABLE AI INTO ONCOLOGY

Adoption of explainable AI in oncology. To develop machine learning systems that can be interpreted.

• Development of predictive frameworks that can be scaled up Engaged in computational healthcare research.Particiapted in computational healthcare research Societal Significance Access to diagnostic support was improved in less accessible areas.

• Improved health outcomes for disadvantaged groups of people There is a growing understanding of health care technologies enabled by AI Breast cancer mortality may be reduced due to early detection. Moreover, the study shows the ability of computational oncology to supplement the skills of the physician rather than replacing them. The suggested AI-based patient care decision support tool is meant to be a supportive system that should increase clinical efficiency without losing control or responsibility on the part of the physician. The results of this study can help to develop next generation intelligent oncology platforms that can combine multiple clinical data types for precision medicine applications.

1.3 NATURAL PRODUCTS IN THE DRUG DISCOVERY PIPELINE

Natural products have always been the most successful source for finding new drugs through pharmacological research and this trend continues to persist today. In all the new chemical entities approved between 1981 and 2019, an analysis conducted by Newman and Cragg (2020) revealed that more than half of the approved drugs were either natural products, natural product derivatives or synthetic compounds designed on natural product scaffolds [10]. The amazing history includes the discovery of a number of ‘signature' drugs, including those containing the penicillin, paclitaxel, morphine, artemisinin, and metformin, all of which stem from guanidine structures found in Galega officinalis.
The modern drug discovery pipeline for natural products typically encompasses several sequential stages: (1) compound identification and sourcing; (2) preliminary biological activity screening; (3) computational (in silico) evaluation including molecular docking and ADMET prediction; (4) in vitro biochemical validation including enzymatic assays; (5) cell-based efficacy and cytotoxicity studies; (6) in vivo pharmacokinetic and pharmacodynamic studies; and (7) clinical trials. The research process requires computational tools at stages 2 and 3 because they allow researchers to perform virtual screening tests which examine vast compound libraries against known targets before they conduct expensive laboratory tests [11].

The lead candidate functions as a vital element in pipeline-stage drug discovery because this compound needs to show target binding while meeting essential pharmacokinetic and drug-likeness standards that predict its ability to perform in live tests. Researchers use SwissADME and pkCSM to create computational tools which enable them to eliminate unpromising compounds during initial testing despite their strong in vitro binding results. The thesis applies this framework to the study of marine phlorotannins.

1.4 MARINE PHLOROTANNINS: SOURCES, CHEMISTRY, AND BIOLOGICAL ACTIVITY

The structurally unique class of marine polyphenolic compounds, called phlorotannins, are only produced in brown algae (Class Phaeophyceae) by the polyketide pathway. Phlorotannins are oligomeric and polymeric compounds composed of phloroglucinol (1,3,5-trihydroxybenzene) monomeric units, linked via aryl–aryl, ether or mixed linkages, whereas terrestrial tannins are derived from gallic acid (hydrolysable tannins) or flavan-3-ols units (condensed tannins) [12]. The biosynthetic origin gives phlorotannins structural properties that are characteristic of no other class of polyphenol from land plants, and are particularly unusual in terms of aromaticity and hydroxylation density.

The five phlorotannins examined in this study vary widely in their structural complexity, and include the major subclasses of phlorotannins found in brown algae: the monomeric phloroglucinol, the trimeric dibenzo-1,4-dioxin-type compound, eckol, the oxidation product dioxinodehydroeckol, and the fucodiphloroethol-class compound, fucodiphloroethol G, and the compound phlorofucofuroeckol A, which is a pentameric compound of the phlorofucofuroeckol-class.

Phlorotannins of E. cava are the most widely studied ones in terms of their biological properties. Potent α-glucosidase and α-amylase inhibiting activity was demonstrated by several Ecklonia phlorotannins, comparable to the reference drug, acarbose, by Wijesekara et al., 2010 [13]. Inhibition of DPP-4 activity by phlorotannins has been identified in enzymatic assays systems, and phlorofucofuroeckol A has consistently been the most potent congener in several enzymatic and biological activity assays.

Although promising experimental results have resulted, a systematic screening and evaluation of marine phlorotannins in the context of the drug discovery pipeline (binding affinity, pharmacokinetics, drug-likelihood and comparative screening) has not been performed. The present thesis is an attempt to bridge that gap.

1.5 RATIONALE AND OBJECTIVES OF THE STUDY

The overarching and central translational question for the present thesis is whether marine phlorotannins meet the multiparametric criteria necessary to become lead candidate in a drug discovery pipeline aimed at developing inhibitors for the DPP-4 enzyme in T2DM. Although the molecular docking affinity is one important criteria, based on its own, a compound cannot be regarded as a viable lead; it should also have good oral bioavailability and house metabolic stability, tolerable toxicity, and follow drug-likeness rules.

The following are the objectives that the study will be followed to answer this question:

1. Using AutoDock Vina, perform molecular docking of five structurally diverse marine phlorotannins with a human DPP-4 crystal structure, predicting them for their predicted binding affinity.
2. Using BIOVIA Discovery Studio to characterise among each docked compound, its molecular interaction profile with catalytic and the substrate-binding residues of DPP-4.
3. To do full ADMET Profiling of all candidate phlorotannins for SwissADME and pkCSM to understand the pharmacokinetic viability and potential safety of each candidate.
4. To compare the most active phlorotannin to the FDA-approved gliptin sitagliptin in terms of both binding affinity and drug-likeness.
5. To situate the topperforming candidate(s) in a drug discovery pipeline framework and set up a path forward for further experimental validation.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Breast cancer remains one of the most important health issues in the world and is a major cause of morbidity and mortality in the world. Advances in diagnostic techniques, molecular biology and imaging systems and treatment strategies have led to significant improvements in survival over the last few decades. Early diagnosis, however, is still one of the most important factors that determines the effectiveness of treatment and long-term survival of the patient. Breast cancer detection is made through traditional methods like mammogram, pathological examination, ultrasound and fine needle aspiration cytology. These methods are proven to be effective and clinically accepted but are limited by a variety of factors including intra- and inter-observer variability, subjectivity, delayed reporting, and reliance on specialized professionals. Furthermore, the rising prevalence of cancer has put a big strain on healthcare systems and diagnostic laboratories. The advent of Artificial Intelligence (AI) and machine learning (ML) has opened a new lane in the field of Modern Oncology and Computational Medicine. AI systems can process vast amounts of clinical data, discover patterns that might not be apparent in the data, and provide predictions with significant speed. They are beginning to be investigated for cancer diagnosis, tumor typing, prognosis, and clinical decision making. A number of recent studies indicate that machine learning algorithms have the potential to enhance diagnostic uniformity and help doctors detect malignant abnormalities at an early stage. Biomedical datasets that have been structured to contain quantitative information about cells have been particularly well suited for predictive modelling as they enable algorithms to discover statistically meaningful biological patterns.The aforementioned chapter summarizes the past research on breast cancer diagnostics, machine learning applications in oncology, explainable AI, and Clinical Decision Support Systems. It also covers the use of structured cytological data and also cites the research gaps being filled in the present research.

The diagnosis of breast cancer is established using a needle biopsy or mammogram.A needle biopsy and/or mammogram is used for diagnosis of breast cancer.

Breast cancer originates due to uncontrolled growth and division of abnormal cells within breast tissue. If the disease goes untreated, or if it is not treated in time, the malignant cells may invade the surrounding tissues, and spread to other parts of the body. Early diagnosis is vital to lower the mortality rates and increase the rate of treatment success. Thus, in clinical practice there are several diagnostic techniques which are routinely used. Of these, Fine Needle Aspiration Cytology (FNAC) is regarded as a minimally invasive and cost effective method for the evaluation of breast masses of suspicion.FNAC is a diagnostic technique where a small needle is used to remove cells from the area of the tissue that is affected. The sample is then looked at under the microscope to determine abnormalities of the structures, arrangement and morphology of the cells. The cell size, cell shape, cell texture, symmetry, concavity, nuclear irregularity are features found by a pathologist to help distinguish between benign and malignant tumors. While cytological interpretation is very useful, it still has limitations of inter-observer variation and the diagnostic ambiguity in borderline cases. The clinical experience and observational assessment of different pathologists can vary with respect to the same sample. Additionally, as patient numbers swell in hospitals and diagnostic facilities, so

do the risk of errors and delayed reporting due to fatigue. Given these restrictions, researchers have looked to automated computational techniques that can assist a clinician in determining if a particular area of cells is malignant more efficiently and accurately.

2.3 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN ONCOLOGY

Artificial Intelligence is a system of computations that mimics human intelligence by learning, reasoning, predicting and making decisions. One of the most promising applications of AI in healthcare is its role in enhancing diagnostic processes and aiding precision medicine. One of the most prominent branches of AI is machine learning, which involves computational algorithms that learn from past data sets and produce predictions based on the relationships they discover. Machine learning models continuously learn and optimize, unlike the rule-based systems. Machine learning is used in oncology in many fields, primarily because there are plenty of big biomedical datasets and the computational power has increased. Nowadays, AI-based systems can be applied for:

- Early cancer detection
- Tumor classification
- Histopathological image analysis
- Prognostic prediction
- Risk stratification
- Treatment response assessment
- Personalized patient management

Many machine learning techniques have been considered for breast cancer prediction such as Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting methods, and Neural Networks. Such algorithms include Logistic Regression, which is simple, easy to interpret, and has high probability calibration. Random Forest and Gradient Boosting algorithms have also showed themselves to be excellent predictors due to their ability to model complex non-linear interactions in biology. While Artificial Neural Networks and deep learning architectures can identify highly complex multidimensional patterns, they tend to be less interpretable and “black box.” While AI is advancing quickly in the healthcare sector, the adoption of AI in the clinical setting is still hindered by a number of challenges. This includes algorithmic bias, transparency challenges, ethical issues, inconsistencies in data quality, regulatory hurdles and diminished trust in non-interpretable systems by physicians.

2.4 EXPLAINABLE AI AND CLINICAL DECISION SUPPORT SYSTEMS

A major issue with the contemporary AI systems is the lack of interpretability. In many advanced machine learning and deep learning models, clinicians are unable to understand how predictions are generated. This issue is commonly referred to as the “black box” problem. In clinical practice, doctors need to be able to explain their choices of diagnosis and treatment. Thus, transparency and interpretability are key to ensuring a safe use of AI systems in healthcare. Explainable Artificial Intelligence (XAI) is a set of computational methods that give insights into how machine learning models make predictions. Explainable AI techniques are used to uncover the most important biological variables and features for model output. In structured biomedical datasets, feature importance analysis has emerged as one of the most popular explainability techniques. This kind of methods helps to increase physicians' confidence and enables clinical practice to compare algorithmic predictions with known biological principles. Clinical Decision Support Systems (CDSS) are software applications designed to help healthcare professionals interpret patient data and make informed medical decisions. AI-powered CDSS systems come with machine learning algorithms, predictive analytics, visualization tools, and risk scoring models.

AI-powered CDSS systems could assist in oncology:

- Improve diagnostic accuracy
- Reduce clinical workload
- Support patient triaging
- Enhance workflow efficiency
- Minimize unnecessary invasive procedures.
- Enable personalized medicine

Several studies have proven that AI-based diagnostic systems can enhance sensitivity and decrease the diagnostic delay. But scientists are stressing the need for physician supervision, noting that AI systems are designed to complement, not replace, physicians.

Previous and current research on breast cancer in Wisconsin has yielded additional information.Some information is also available from previous and current research on breast cancer in Wisconsin.

Breast Cancer Wisconsin (Diagnostic) dataset is the one of the most widely used benchmark datasets in biomedical machine learning research. Structured cellular morphometric data from digitized Fine Needle Aspiration biopsy image data. The data consists of quantitative information on the following

- Radius

- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave points
- Symmetry
- Fractal dimension

The mean value of each parameter, the standard error and the worst-case value are obtain which provides 30 numerical PV. This data set has been used in a number of studies to test

machine learning algorithms for breast cancer classification. Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosting methods and Neural Networks have been reported to be highly predictive. In the past, it has been noted that the features worst area, worst perimeter, concavity and concave points are strongly connected with malignant tumors. The biological significance of these findings is that cancer cells usually have abnormal cell growth, abnormal cell morphology and structural asymmetry. A large number of studies have succeeded in achieving high classification accuracy, but some limitations still exist. Most models in the literature work mostly on the maximisation of predictive accuracy and lack interpretability and clinical usefulness. In addition, some systems do not offer any output other than binary, and they don't include any helpful probability-based risk stratification or explainable prediction mechanisms.

To overcome these limitations, the present study aims to develop an Explainable AI (XAI) assisted Clinical Decision Support System (CDSS) that can deliver the following:

Correct diagnosis of breast cancer type

- Transparent prediction mechanisms
- Probability-based risk assessment
- Clinically interpretable outputs

2.5 Summary of Literature Review.

The literature searched shows that AI and machine learning technologies have significant potential to enhance breast cancer diagnostic and clinical decision making. In the field of structured cytological data, machine learning algorithms have demonstrated outstanding performance for identifying patterns in cells that are indicative of cancer. Logistic Regression, Random Forest, Gradient Boosting and Neural Networks are all found to be strong predictors for breast cancer classification. Concurrently, there is also a substantial body of literature focusing on issues of interpretability, transparency and clinical translation. A large number of AI systems still work as very complex black-box models, offering little biological explanation of the models' predictions. Thus, there still exists a significant need for Clinical Decision Support Systems that are explainable, accurate in their predictions and also clinically relevant. Building on these results, the current study proposes an AI-CDSS framework based on machine learning for early breast cancer diagnosis and risk stratification using structured cellular morphometric features.

CHAPTER 3: MATERIALS AND METHODS

3.1 INTRODUCTION

It is essential to have a systematic methodological approach to the development of Artificial Intelligence-assisted diagnostic systems in healthcare, which include data acquisition, preprocessing, predictive modeling, evaluation, and clinical interpretation. Datasets, preprocessing methods, the interpretation of the algorithm, and the clinical relevance of the output generated by a machine learning system are important in breast cancer diagnostics, along with predictive accuracy. In the present study, an Explainable AI-assisted Clinical Decision Support System (AI-CDSS) for early breast cancer detection and risk stratification with structured cellular morphometric data was developed. The methodology used in this study was planned carefully to mimic a clinically relevant workflow that starts with acquisition of the cytological data and finishes with the risk-based interpretation of the data. This framework was mainly on the development of transparent and efficient predictive system that could provide support to healthcare professionals in the identification of malignant breast tumors. The proposed approach aimed to keep the system as transparent and explainable as possible, which helps to build trust among doctors and have a better clinical impact compared to many complex black-box AI systems. The chapter provides detailed information on the dataset used in the study, data preprocessing, machine learning models, training procedures, evaluation methods, integration of exploratory AI, and a probability-based risk stratification framework.

3.2 DATASET DESCRIPTION AND STUDY DESIGN

The present work involves a computational and experimental study carried out on the Breast Cancer Wisconsin (Diagnostic) dataset, one of the most widely used and extensively validated biomedical datasets in the field of machine learning. It is openly available and has been used in many studies for the evaluation of predictive algorithms for breast cancer classification. The data is quantitative cellular morphometric data extracted from digitized Fine Needle Aspiration (FNA) biopsy images of breast masses. A minimally invasive FNA procedure is a technique in which the cells are removed with a thin needle from the area of the breast that is suspect for benign or malignant breast disease and examined under a microscope. The features extracted from the cells are significant morphological features of the cell nucleus, including size, texture, shape, irregularity of its border, smoothness and concavity. These morphometric features are very important in oncology since the malignant cells typically exhibit anomalies in proliferation, pleomorphism, irregular outlines and structural asymmetry.
A total of 569 patient samples are listed in the dataset, which can be split into:

- Benign tumors: 357 cases
- Malignant tumors: 212 cases

For each patient sample, 30 numerical predictive features are obtained from 10 major cell characteristics:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave points
- Symmetry
- Fractal dimension
- For all these characteristics three statistical measurements were calculated:
- Mean value
- Standard error

The number with the greatest value (the highest average of the three highest measurements)

This led to the creation of 30 very descriptive numeric variables that can represent the biological behaviour of

samples of breast tissue.

- The diagnostic outcome variable divided the tumors into two categories:
- Benign (non-cancerous)
- Malignant (cancerous)
- To run the computation analysis, benign tumors were represented by the number “0” and malignant tumors were represented by the number “1”.
- The Breast Cancer Wisconsin dataset was chosen for the current study for the following reasons:
- It is composed of biomedical data that has been clinically proven
- The data set is well class separated.
- All features are biologically meaningful
- It offers great fit for supervised machine learning

To compare algorithms in a repeatable manner, the dataset can be used.
The overall study design was:

- Preprocessing and normalization of the data set.Data Set Preprocessing and Normalization.
- Machine Learning model generation
- Comparative algorithm evaluation
- Explainability and feature analysis
- Probability-based risk stratification

The clinical interpretation of predictive results.The clinical interpretation of predictive outputs.
The methodology was designed to keep the developed framework computationally efficient, transparent, clinically interpretable and scalable for future healthcare integration.

3.3 DATA PREPROCESSING AND FEATURE ENGINEERING

The nature of biomedical data is often complex, with various inconsistencies, scaling factors, and statistical irregularities, which can impact the performance of the algorithms. One of the most critical steps in machine learning involves the data preprocessing stage. In the present study, the preprocessing pipeline was aimed at enhancing numerical stability, model convergence, computational efficiency and predictive reliability. The data has been thoroughly checked for missing data, duplicate data, corrupted data, and data points with outliers. There were no missing or incomplete data points or entries noted within the Breast Cancer Wisconsin dataset as it is very standardized and clinically curated. The patient identification attribute included in the data set was not used for predictive modeling, and could potentially add unwanted bias to the system, and therefore was removed. The significant feature scale variation was one of the critical challenges in preprocessing. Some of the cell measurements (e.g., area, perimeter) had large numbers and others (e.g., smoothness, fractal dimension) had small decimal numbers. Scale differences are particularly problematic for machine learning algorithms, such as those that rely on distance calculations or gradient optimization. Large features can dominate in the optimization process, and can have a negative impact on the predictive process. To solve this problem, feature standardization was done by normalising the features using Standard Scaler. All variables were transformed into this pre-processing technique so that:

- Mean = 0
- Standard deviation = 1

Standardization resulted in equal contributions of all the predictive variables and enhanced:

- Numerical stability
- Optimization efficiency
- Model convergence
- Algorithmic fairness

After pre-processing and normalization of the data, it was split into a training set and a testing set.

This data split was done in the following manner:

- Training dataset: 80%
- Independent testing dataset: 20%

The training dataset was employed for learning the underlying relationships between features, model optimization, and the testing data was kept completely out of the training process to give an unbiased assessment of the predictive ability. These methods enabled to simulate real-world clinical performance and also ensured good generalization of the models to previously unseen clinical data.

3.4 MACHINE LEARNING MODEL DEVELOPMENT

The main goal of the current study was to create a clinical decision support system that would be explainable artificial intelligence and could accurately classify benign or malignant breast tumors with structured cytological data. For the purpose of this objective, four algorithms of machine learning were chosen and compared. The algorithms selected were representative of the different mathematical learning methods and were found to be highly predictive in biomedical research.

The first model implemented was Logistic Regression, a supervised probabilistic learning algorithm which is commonly used for binary classification problems. The Logistic Regression model was chosen due to its simplicity, interpretability, computational efficiency, and good probability calibration ability. The model provides estimates of the probability of malignancy using a logistic sigmoid function and outputs clinically interpretable results.

Random Forest, an ensemble learning method which integrates multiple decision trees for better prediction capability and less overfitting, was the second model used in this study. The random forest models are especially well suited for structured biomedical datasets as they can detect non-linear relationship between the biological variables, and also offer feature importance analysis. The third model that was used was Gradient Boosting, a sequential ensemble learning method whereby weak predictive learners are sequentially trained to reduce the residuals produced by the previous learners.

Gradient Boosting algorithms excel at predicting on structured tabular data and can optimize very complex feature interactions.

The 4th model adopted was Multi-Layer Perceptron (MLP) Artificial Neural Network which was a feed-forward Artificial Neural Network with interconnected computational neurons in the hidden layers. Neural Networks have the ability to detect very complex multidimensional biological patterns and non-linear relationships.

The steps of the machine learning process that were followed:

- Using standardized feature vectors for feeding into predictive models In this section, the training set was utilized to train the models.
- The learning relationship between cellular features and tumor classification.
- Creating random output for cancer diagnosis.
- Assessing the predictive performance with unseen testing data.

The developed framework was intentionally designed as Clinical Decision Support System (CDSS), instead of a fully autonomous diagnostic system. The overall objective was to assist clinicians to make more consistent diagnoses and decisions, not to take the place of the physician.

3.5 MODEL EVALUATION, EXPLAINABILITY AND RISK STRATIFICATION

Explainability and risk stratification are crucial components of model evaluation.Explainability and risk stratification are key parts of model evaluation. Clinically meaningful performance metrics were used to evaluate the developed machine learning models instead of overall classification accuracy.

In oncology, predictive systems need to have considerable sensitivity, otherwise a patient could receive no treatment for their condition and the probability of survival is reduced. In present study the following evaluation metrics were used:

1 .Accuracy - Accuracy is the percentage of the correctly classified samples that are classified correctly by the model.

2. Sensitivity (Recall) -A sensitivity indicates the ability of the system to correctly identify cases with malignant cancer. Sensitivity is of paramount importance in clinical oncology since a failure to diagnose cancer may have serious implications for patient prognosis.

3. Specificity- Specificity is the proportion of people with the negative test result who actually do not have the disease. High specificity decreases unnecessary biopsies, anxiety and invasive procedures.

4. F1-Score- The F1-score is the harmonic mean of precision and recall, and is used for a balanced evaluation of the classification performance.

5.The Area Under Receiver Operating Characteristic Curve (AUROC) is a term used to describe the area under the Receiver Operating Characteristic Curve. AUROC measures the overall ability of the predictive model to discriminate between different classification thresholds.

In the present study, emphasis is laid on explainability and transparency of the AI assisted framework. The "black box" effect of many sophisticated predictive systems is one of the greatest challenges to full-scale AI implementation in healthcare. But clinicians often don't like models that make predictions without a bio explanation. However, to solve this issue, feature importance analysis was added to the proposed framework. Explainability methods were used to identify the Cell Morphometric variables that are most important for predicting malignancy.

The analysis showed that the following features:

- Worst area
- Worst concave points
- Perimeter irregularity
- Concavity

A strong relationship was observed between the occurrence of malignant tumors and being. The biological significance of these discoveries was that the morphology of the nuclei of the cells in a tumour was abnormal, the structure was asymmetric, the contours were irregular, and the cells had an uncontrolled proliferation. Probability-based risk stratification was also included in the present study in addition to binary classification. The system was not expected to give only "benign" or "malignant" outputs, but rather continuously varying probability estimates which were divided into:

- Low-risk category
- Intermediate-risk category
- High-risk category

This approach enhanced the clinical utility since doctors often need to understand the risk on a continuum, rather than in binary terms. The risk stratification framework also helped to identify "grey-zone" cases that were clinically indeterminate, but needed to be reviewed by a physician or further investigated by diagnostic testing.

3.6 SUMMARY OF METHODOLOGY

In the present study, the methodology that was followed was to develop an Explainable Artificial Intelligence (XAI) based Clinical Decision Support (CDS) System for Early Detection of Breast Cancer based on Structured Cellular Morphometric Data. The Breast Cancer Wisconsin Diagnostic dataset was used due to its meaningful clinical features and for supervised machine learning applications. The data was preprocessed and normalized to enhance numerical stability and the predictive reliability. Clinically relevant performance measures were used to develop and comparatively analyse four machine learning algorithms namely Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Multi-Layer Perceptron Neural Network (MLPNN). To enhance transparency, interpretability, and clinical relevance of the predictive framework, the study included explainable AI and probability-based risk stratification..

CHAPTER 4: MODEL DEVELOPMENT AND SYSTEM ARCHITECTURE

4.1 Development of the Proposed AI-Assisted Clinical Decision Support System

The proposed AI-assisted Clinical Decision Support System was developed following this process.This process has been used to create the proposed AI-assisted Clinical Decision Support System. With the advancement of Artificial Intelligence in healthcare, the medical diagnostic and clinical decision-making landscape has changed significantly. Beyond disease prediction, AI-powered systems are being investigated in oncology for enhancing workflow efficiency, minimizing diagnostic variations, and facilitating precise patient management. To be clinically meaningful, however, these systems must achieve high predictive performance, but also be transparent, interpretable, and deliver outputs that are physician oriented. The aim of the present study was to develop an explainable Artificial Intelligence-assisted Clinical Decision Support System (AI-CDSS) for the early detection and risk stratification of breast cancer based on structured cellular morphometric data.

This proposed framework is conceived with a supportive healthcare technology that will help the clinicians diagnose a malignant tumor, given that the healthcare technology must be under the clinician's control and he/she is accountable for it. The architecture designed in this research emphasized simplicity, interpretability, computational efficiency and explainable prediction mechanisms, which is in contrast to some very complicated deep learning systems which work like a “black box” model.

The goal was not to supplant oncologists and pathologists, however, to develop a computational system that can assist in diagnostic workflows and mitigate burden in the clinical setting. The entire system architecture was modeled to emulate a real clinical workflow starting from patient data acquisition to probability-based diagnostic interpretation.

The framework was composed of several interrelated phases:

Structured cytological data acquired pre-processing and data normalization

- Predictive model training
- Malignancy probability estimation
- Clinical risk stratification Clearness and feature explanation

The use of decision support output generation.Use of decision support output generation.The workflow starts from the Fine Needle Aspiration Cytology (FNAC) which is a very less invasive diagnostic procedure used in the diagnosis of breast cancers. The process of FNA is the removal of cells from suspicious breast lumps by inserting a thin needle into the area and then looking for the cells under a microscope. The cytological information employed in this research work was quantitative morphometric features obtained from digitized image of FNA. These features were important biological parameters like radius, perimeter, texture, smoothness, concavity, symmetry and fractal dimension of cell nuclei. These morphometric characteristics are very significant as malignant cells generally have abnormalities in structure, nuclear size, shape and proliferation pattern. After the structured cellular data was acquired, pre-processing and normalisation methods were used to enhance the numerical stability and algorithmic reliability.

The features were on varying numerical scales, so they were standardized to ensure a similar contribution from all the predictive features when optimizing the machine learning. The processed data was then fed into

predictive machine learning algorithms that were able to uncover hidden statistical relationships relevant to malignant and non-cancerous tumors. The trained models produced probability values indicating the chances of malignancy for each patient sample. One of the key concepts of the proposed architecture was explainability. Hence feature importance analysis was added to the model for finding out the most important biological variables which help in the prediction of malignancy. This assisted in maintaining a biologically meaningful and clinically interpretable A.I. system. Besides, the model also adopted risk stratification based on probability, rather than binary outcomes. This enhanced the practical application of the problem in clinical practice, as doctors make treatment decisions on the basis of "graded risk" interpretation, rather than absolute classification.

Below is then workflow of AI system:-

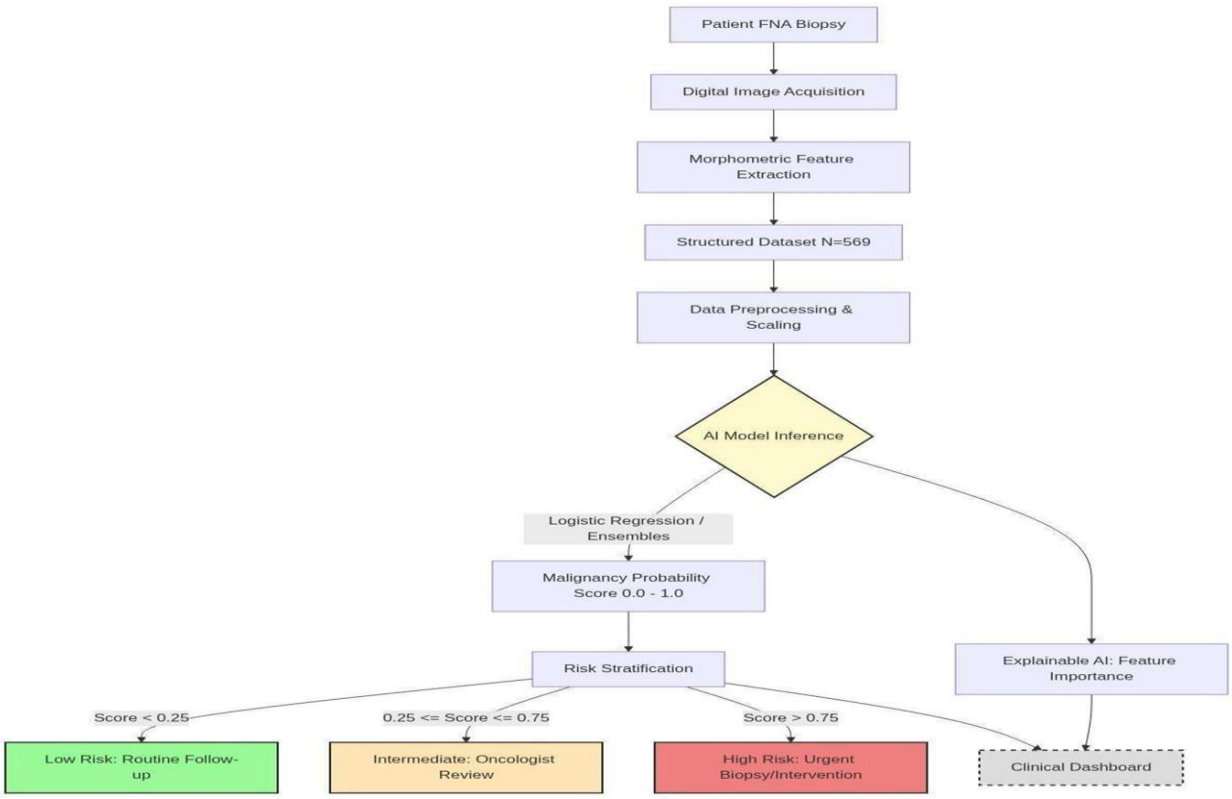


FIGURE 1. The overall architecture therefore established a strong interpretable foundation for future multimodal AI systems integrating imaging, genomic, and electronic health record data within oncology workflows.

4.2 CONSTRUCTION AND TRAINING STRATEGY OF THE MACHINE LEARNING MODEL

The predictive component of the proposed AI-based Clinical Decision Support System comprised several machine learning models that were carefully chosen because of their ability to process structured biomedical information and uncover intricate pathological connections. The main goal of applying several predictive models was to assess the various mathematical learning methods comparatively and find the most clinically appropriate algorithm for the classification of breast cancer. The models used in the present study were:

- Logistic Regression
- Random Forest
- Gradient Boosting Multi-Layer Perceptron Neural Network (MLP)

These are all methods for computing the pattern and classify.

By including them, they enabled a detailed examination of their predictive power, interpretability, robustness and clinical applicability. The first model that was implemented was Logistic Regression, a supervised probabilistic machine learning algorithm which is commonly used in healthcare for binary classification problems. Logistic Regression produces output of probabilities from 0 to 1, and estimates the probability of malignancy with a logistic sigmoid function. The interpretability of the model was one of the key considerations for choosing Logistic Regression. Transparency is a key element in medical diagnostics, because doctors need to know the rationale behind their predictions. Using Logistic Regression, researchers and doctors can explore the role of each cell variable in prediction results. Moreover, the model has shown a great degree of probability calibration, which is ideal for risk stratification and Clinical Decision Support systems.

Random forest was developed as a second model, which is an ensemble learning algorithm that is made up of multiple decision trees. Random subsets of the data and predictive variables are used to train each tree in the forest. The output of the final classification is based on the majority voting of the trees. Biomedical applications often suffer from non-linear relationships between variables, making Random Forest models well suited. Ensemble learning helps in making predictions more robust, gives lesser overfitting, and gives more generalization. An additional benefit of the Random Forest is that it can provide feature importance scores. This capacity was useful in explaining the results of the explainability analysis and in determining which of the morphometric features was most closely related to malignant tumors.

The third algorithm used was a Gradient Boosting algorithm, which is a sequential ensemble learning algorithm where predictive learners are sequentially trained to reduce the errors that previous learners commit. Gradient Boosting differs from Random Forest because the trees are not built independently, instead they are built to correct the errors made by previous trees. Gradient Boosting algorithms work really well with tabular data with a clear structure, as they can capture nuanced and extremely complex relationships between features. For use in oncology these models can be optimized to produce an outstanding classification accuracy in capturing intricate biological patterns and optimized flexibility.

The fourth model predicted in the present study was Multi-Layer Perceptron (MLP) Neural Network. Neural Networks are computational systems that are modeled after the architecture of the human brain, comprising layers of interconnected artificial neurons that can learn multidimensional relationships. The objective of the MLP model was to examine if the model could perform better than traditional statistical and ensemble learning models when predicting the water surface elevation. Neural Networks are very good at modeling non-linear relationships in biological systems and relationships between features that are not obvious. But although they are highly predictive, neural networks also have disadvantages, such as lower interpretability and ‘black box' behavior. All models were trained by providing standardized feature vectors to the algorithms, and letting them learn the relationship between the cytological variables and the classification results of tumors. The dataset was split into 80% training and 20% testing sets in an unbiased way to evaluate the data. All parameters were optimized and learned on the training dataset, and were never seen during evaluation. This method was used to assess the robustness of the models to unseen patient samples.

The overall objective of model development was not just to achieve the highest prediction accuracy, but to find a compromise between:

- Diagnostic performance
- Computational efficiency
- Explainability
- Probability calibration
- Clinical interpretability
- Practical healthcare applicability One of the main advantages of the proposed AI-CDSS framework is this physician-oriented approach.

4.3 EXPLAINABLE AI FRAMEWORK AND RISK STRATIFICATION MECHANISM

Interpretability is one of the biggest drawbacks of contemporary AI applications in healthcare. In some of the more complex machine learning and deep learning models, physicians are given predictive results, but don't gain any insight into the biological basis for the decisions being made. This issue is commonly referred to as the "black box" problem. Transparency and interpretability are important in medical diagnostics, especially for cancer, because doctors are required to explain the results of their diagnosis and recommendation for treatment.

Lack of interpretability can lead to a decrease in trust among physicians, ethical issues, and hinder widespread clinical adoption of AI systems. To solve this issue, the proposed Clinical Decision Support System used the Explainable Artificial Intelligence framework in the study.

The main goal of explainability was to uncover the morphometric features of cells that are most predictive of malignancy, and to interpret the outcomes of the algorithms in biological terms. Random Forest was chosen as a feature importance analysis model since ensemble tree-based models can quantify the effect that each variable has on the predictive performance.
The analysis showed that the following features are:
• Worst area
• Worst concave points
• Worst perimeter
• Concavity
• Radius irregularity
The presence of malignant tumors was strongly linked to the presence of these. These findings are very consistent with the known pathological principles. The nuclei of malignant cells are typically pleomorphic, the cells have irregular boundaries, they are asymmetric, and they multiply in an uncontrolled manner. Thus the AI wasn't just looking for mathematical patterns; it was learning morphological features that had biological relevance akin to those which pathologists would see under the microscope.
The framework greatly enhanced the clinical relevance of the desired system as it enabled clinicians to gain an understanding of the model's prediction development process. This transparency is essential to ensure the safe deployment and acceptance of these in the clinic. The present study also implemented a probability-based risk stratification framework besides explainability. Typical machine learning classifiers typically return binary decision results, e.g., 'benign' or 'malignant'. But, in practice, clinical decisions are not always a no/yes situation. Many times, doctors will have a case that falls in the "gray zone" or is not clearly diagnosed and will need to run more tests.
To solve this problem, the predictive models were applied to give a continuous probability distribution, quantifying the likelihood of malignancy for each patient sample. The probabilities were then sorted into clinically relevant risk categories:
• Low-risk category
• Intermediate-risk category
• High-risk category
The low risk group included cases that had low probability of malignancy and high confidence about benign pathology. High risk category included samples strongly linked to malignant characteristics, needing immediate oncological intervention and further clinical evaluation. Cases in the "intermediate" classification were considered "grey-zone" and needed further review by the doctors, imaging tests, or biopsy. The improvement in practical applicability compared to the traditional methods and the ability to combine risk estimation, uncertainty assessment and experience of a physician for decision making is a benefit of this risk stratification approach, as it is more relevant in a real clinical setting in oncology. The combination of explainable and probability-based interpretation thus rendered the predictive models more clinically meaningful and physician oriented

4.4 CLINICAL RELEVANCE, PRACTICAL APPLICABILITY AND FUTURE INTEGRATION OF PROPOSED SYSTEM.

The clinical relevance, practical applicability and future integration of the proposed system are discussed.
The Clinical Decision Support System (CDSS) concept with AI support was envisioned as a healthcare technology for physicians to improve the uniformity of diagnosis, relieve the burden of interpretation, and facilitate evidence-based oncological practice.

Eventually, in laboratory settings, the framework could be embedded in diagnostic procedures in which cytological information from Fine Needle Aspiration (FNA) techniques would be automatically processed by predictive algorithms.
These systems can have a major impact on healthcare efficiency by supporting healthcare professionals in patient triaging and diagnostic evaluation'

The proposed framework can produce the output which can be helpful for healthcare professionals in:

- Early detection of breast cancer
- Adoption of high-risk case prioritisation
- Early diagnosis is achieved and delayed diagnosis is avoided.
- Clinical risk assessment
- Personalized patient management
- Improved workflow efficiency
- Evidence-based decision-making

The explainability feature also adds to physician trust as it shows the cellular variables that contributed to the predictive outcomes. This makes it possible to relate AI predictions to known pathological principles and biological knowledge. This proposed system could be especially beneficial in resource-limited healthcare environments where access to specialized oncologists, radiologists, and diagnostic facilities is limited. AI-based Clinical Decision Support Systems could potentially alleviate the health inequality by making diagnostic support more accessible to those in need in such environments, and by enhancing timely detection of malignancy. In addition, the computational efficiency of the framework could facilitate its future integration to cloud-based healthcare platforms and hospital information systems.
. The architecture developed in the present study also can be used as a building block for larger multimodal oncology systems that incorporate:
• Histopathological imaging
• Mammographic analysis
• Genomic sequencing
• Electronic health records
Use of real-time patient monitoring systems
The conclusions drawn from the results achieved in the present study indicate that the predictive capacity of the Artificial Intelligence systems is excellent, however, it is important to be aware of the fact that Artificial Intelligence systems are not a replacement for clinical skills. The proposed framework was conceived as a decision support system and not as a standalone diagnostic system.
However, the use of AI systems is still limited by certain clinical aspects, such as:
• Rare pathological variants
• Poor-quality cytological samples
• Dataset bias
So, the proposed AI-CDSS must be considered a partnership computational platform that can improve the efficiency of physicians, support evidence-based oncology workflow, ensure ethical accountability.

CHAPTER 5: RESULTS AND DISCUSSION

5.1 PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS

In this course, students will evaluate machine learning models using the five-point scale.
The main aim of the ongoing study was to compare the ability of different classification methods based on machine learning algorithms to correctly classify benign and malignant breast tumors using structured cellular morphometric data. All predictive algorithms were then tested on the independent test set after preprocessing, normalization and model training to determine their clinical applicability and their generalization ability.The testing phase was especially significant since the independent test set was made up of completely new patient samples that had not been seen before. This provided for the predictive performance to demonstrate the ability of the models to classify new clinical cases correctly, and not just to remember the patterns during training.
In the present study four machine learning algorithms were evaluated:

- Logistic Regression
- Random Forest
- Gradient Boosting
- Multi-Layer Perceptron Neural Network

Both models showed very high predictive power, demonstrating that the Breast Cancer Wisconsin Diagnostic dataset has highly discriminative cellular morphometric features which are able to accurately distinguish benign and malignant tumors. The models were tested for their predictive ability using clinically relevant criteria such as:

- Accuracy
- Sensitivity
- Specificity
- F1-score

•Area Under Receiver Operating Characteristic Curve (AUROC)
The results obtained showed that all the algorithms evaluated showed a good performance in classification with an overall accuracy of more than 96%. This demonstrates the high predictive value of the structured cytological features in the diagnosis of breast cancer. Random Forest performed the best overall with accuracy of 97.37% and specificity of 100% among the algorithms evaluated. This means the model was able to correctly classify all the benign data in the test dataset without falsely identifying it as malignant.Still, Logistic Regression stands out as the most clinically useful model due to its higher AUROC score and well-calibrated probability predictions, even though the overall accuracy was slightly lower than that of Random Forest. Logistic Regression performed with an AUROC value of 0.9960, which is very high and showing excellent discriminating power between malignant and benign tumours.
A detailed comparative performance of the models is presented below:

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUROC
Logistic Regression	0.9649	0.9286	0.9861	0.9512	0.9960
Random Forest	0.9737	0.9286	1.0000	0.9630	0.9929
Gradient Boosting	0.9649	0.9048	1.0000	0.9500	0.9947
Neural Network	0.9649	0.9048	1.0000	0.9500	0.9927

Performance levels are quite similar across all models, indicating that the biological pattern within the data is highly structured and that it is possible to detect it using computation.

The repeatability of the results also suggests that, in the process of the study, the cytological features were able to maintain their discriminatory properties through the consistency that was imposed in their preprocessing and normalization. Sensitivity was one of the most relevant evaluation parameters from a clinical point of view since a false-negative cancer diagnosis could be detrimental to the treatment and decrease the chances of survival. Sensitivity value of both Logistic Regression and Random Forest were high at 0.9286, demonstrating their ability to be capable in identifying malignant tumors correctly. In the same way, the remarkably high specificity scores obtained by the Random Forest, Gradient Boosting and Neural Networks highlight a high capacity to correctly identify cases of the benign type, which helps to avoid clinical anxiety and unnecessary invasive actions. The promising results yielded with all of the algorithms are confirmation of the appropriateness of machine learning techniques for breast cancer prediction based on structured cellular morphometric data.

5.2 RECEIVER OPERATING CHARACTERISTIC ANALYSIS AND DIAGNOSTIC CAPABILITY

The discriminative power of the predictive models was assessed using Receiver Operating Characteristic (ROC) analysis, at varying classification thresholds. ROC curves are graphical plots that show the tradeoff between False Positive Rate (FPR, false alarm rate) and True Positive Rate (TPR, sensitivity) and offer a complete overview of a model's performance that goes beyond accuracy calculations. In clinical oncology, ROC analysis assumes special significance because doctors need systems that are sensitive and thus are able to make very few false-positive classifications. The Area Under the ROC Curve (AUROC) is a numerical value that represents overall discriminative performance.
. The performance of all machine learning models evaluated are compared using ROC curve which is plotted in figure 2.

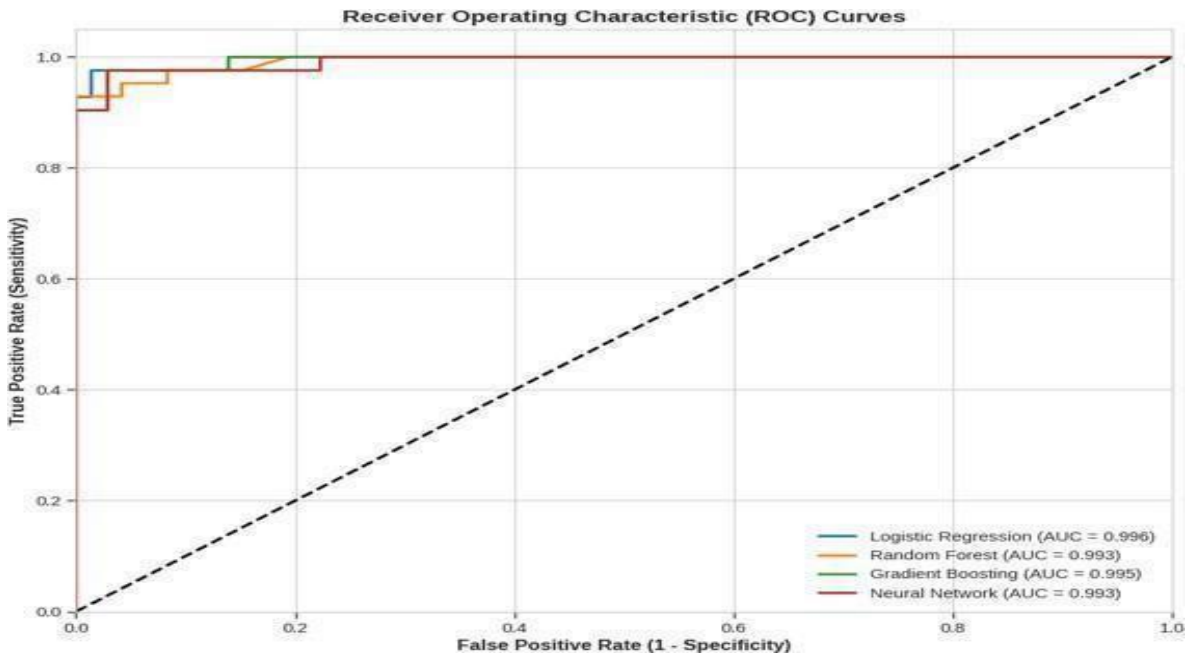


Fig. 2. Receiver Operating Characteristic (ROC) Curves Comparing the Predictive Performance of All Four Evaluated Machine Learning Models

5.3 CONFUSION MATRIX INTERPRETATION AND EXPLAINABLE AI ANALYSIS

The ROC curves produced in the study showed that all the models used in the study possess good classification ability. In all four algorithms, the curves were very close to the upper left corner of the graph, meaning they had very good separation between malignant and nonmalignant cases. The highest AUROC value was obtained by Logistic Regression (0.9960) among the models. This result is significant as the AUROC value becomes closer to 1.0 as the discrimination between the two classes approaches the ideal of 1.0. In this case, the AUROC value of 0.996 implies that, if a sample from a cancerous patient was selected randomly from the pool of all cancer samples and a sample from a healthy patient was selected randomly from all healthy samples, and the Logistic Regression model was used to predict malignancy, there would be 99.6% probability that it would correctly classify the cancerous sample as being more likely to be malignant.

The outstanding value of ROC seen in the present study is evidence of the very good predictive power of the structured cytological morphometric variables. It also illustrates that the machine learning models could successfully learn biologically meaningful patterns, which were associated with the process of the malignant transformation. The findings also suggest that fairly interpretable statistical algorithms (e.g., Logistic Regression) can perform at the same level of predictive accuracy as more computationally intensive algorithms, while also being transparent and easy to explain. Healthcare-related implications of the ROC findings are of special interest as the use of a diagnostic system with a high discriminative capability could help towards earlier cancer detection, better triaging and decrease in diagnostic delay in clinical settings.

5.3 INTERPRETATION AND EXPLAINABLE AI ANALYSIS OF THE CONFUSION MATRIX

While overall accuracy and AUROC values are important quantitative measures of predictive performance, the clinical interpretation will require a more detailed understanding of the way in which the models classified individual patient cases. So, the confusion matrix analysis was conducted on the best performing clinically relevant model Logistic Regression. Figure 3 shows the confusion matrix created for the Logistic Regression model.

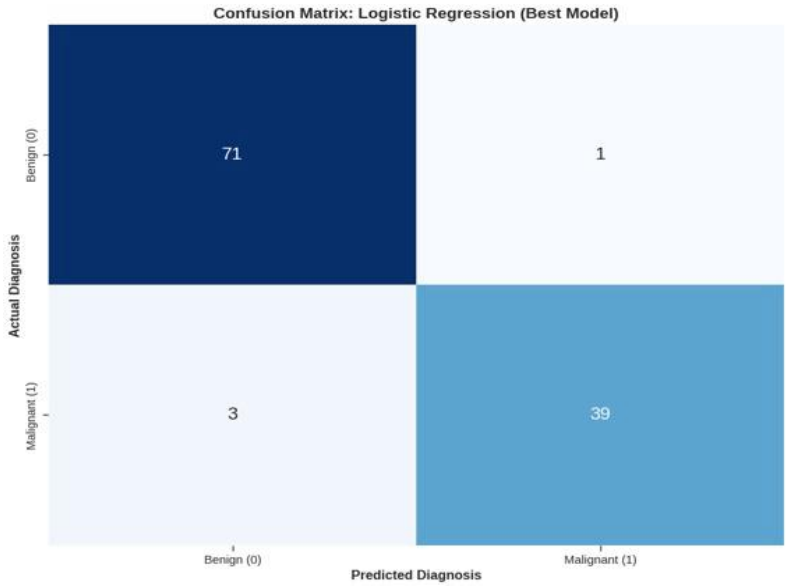


Figure 3: Confusion Matrix of the Logistic Regression Model Showing Classification Performance for Benign and Malignant Breast Tumors

The confusion matrix has given a detailed representation of the prediction outcomes for the independent testing dataset, which has enabled the true positive, true negative, false positive, and false negative to be evaluated. The model Logistic Regression correctly classified:

• 71 malignant cases as malignant (True Positives)

39 patients had a malignancy confirmed as malignant (True Positives)

The model generated:

• 1 False Positive

• 3 False Negatives

The low number of false-positive results suggests the model was very successful at not identifying patients with non-cancerous results as having cancer. This is clinically significant because patients can be emotionally upset, experience anxiety, undergo further testing and invasive procedures that are unnecessary, if a cancer diagnosis is made that is incorrect. Concurrently, only a few false-negative cases indicate a high sensitivity for the predictive framework. The results, however, also underscore the need to consider AI systems as tools to assist physicians, and not as a replacement for them when it comes to diagnosis. In oncology, the few false negative results, can lead to significant clinical consequences due to time delays in diagnosis that can impact the efficacy of treatment and survival. As such, physicians will still need to supervise, particularly when the diagnosis is uncertain. One of the great merits of the present study was the embedding of explainable AI mechanisms in the predictive framework. The models utilized mostly biologically meaningful morphometric features during creation of the predictions, as revealed by the feature importance analysis. The Relative Importance of the top influential features identified from the RF model is shown in Figure 4.

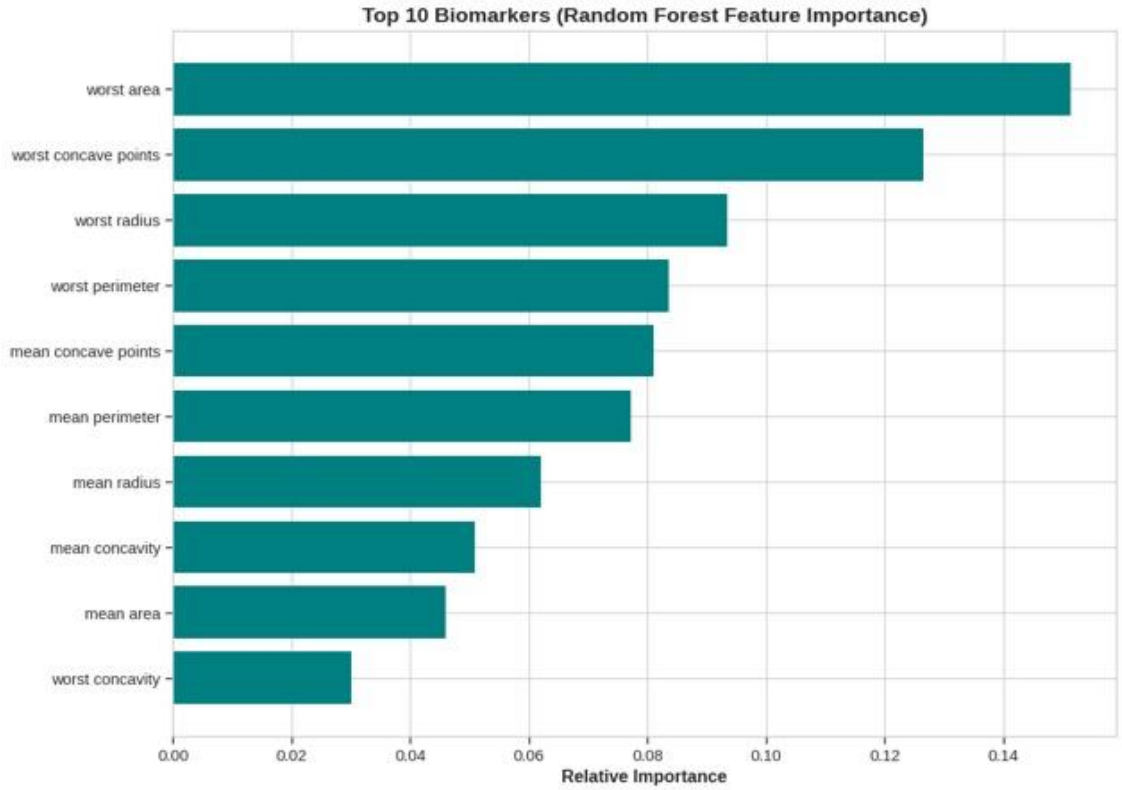


Figure 4: The top 10 Biomarkers identified using the feature importance method Random Forest for prediction of breast cancer.

The most important predictive features that were found were:
• Worst area
• Worst concave points
• Worst perimeter
• Concavity
• Radius irregularity
These findings are very similar to pathological principles as malignant cells are generally characterized as having an abnormal enlargement in nuclear size, irregular nuclear boundaries, asymmetry, and structural distortion of the nucleus. The explainability analysis was therefore able to confirm that the AI system was not looking for random mathematical relationships, but was discovering clinically relevant morphological patterns that are also observed by the pathologist, during the new microscopic evaluation. The interpretability will further increase physician trust and facilitate clinical utility, since physicians tend to accept AI systems that can offer clear reasoning from a biological perspective.

5.4 THE CLINICAL RELEVANCE, RISK STRATIFICATION, AND DISCUSSION OF FINDINGS

One of the major achievements of the present study was the creation of the probability-based risk stratification framework embedded in the AI-based Clinical Decision Support System. In the past, machine learning classifiers were limited to binary classification, i.e., either benign or malignant. In real-world clinical diagnosis, however, things are not always so black and white. There are many cases that are "borderline" or indeterminate, requiring further imaging, pathological evaluation or biopsy.
In order to overcome this drawback, the predictive models produced probability outputs every minute of the sample of patients, thus corresponding to the probability of malignancy. The probabilities were then converted into meaningful risk categories, such as:
• Low-risk category
• Intermediate-risk category
• High-risk category
Those in the low-risk category were strongly linked with benign pathology and low predicted probability of malignancy. These patients might need periodic surveillance instead of prompt and intensive treatment. High-risk category was the one with cases with high malignant characteristics that needed immediate clinical intervention, oncological consultation and further diagnosis confirmation. The intermediate risk category included “grey-zone” cases, with uncertainty in the diagnosis, which needed further review by the physician and further investigations. The use of a probability-based approach was a significant step that helped make the proposed framework more clinically relevant, as it was based on how risk would be interpreted and assessed in clinical practice and based on clinical expertise, judgment, and experience, rather than just algorithmic results.
The results from the present study are highly supportive of the increasing significance of the use of Artificial Intelligence in contemporary oncology and precision medicine. These findings show that the morphometric features of the cells have a high level of meaning with biological content that can be used in the early diagnosis of breast cancer using machine learning techniques. In addition, the study shows that models that can be interpreted simply, like Logistic Regression, can be as good at predicting as more complicated ones, but with better transparency and explainability. The adoption of explainable AI and risk stratification systems is a significant step towards the development of AI systems suitable for clinical use in healthcare. These systems could eventually help healthcare professionals to have more consistent diagnoses, to have less workload, to have more prioritized patients who are at high risk and to support evidence based clinical decision making. The study also highlights the need for continued physician supervision due to current challenges with unclear pathology, rare biological variants, and potential dataset bias with AI systems. Thus, the proposed Clinical Decision Support System with AI is meant to be a technology that aids physicians and should only be considered as part of the physician's arsenal and not as a substitute for the physician.

CHAPTER 6: EXPLAINABLE AI, CLINICAL INTERPRETATION AND SYSTEM ANALYSIS

6.1 EXPLAINABLE ARTIFICIAL INTELLIGENCE IN BREAST CANCER DIAGNOSTICS

In modern healthcare, the predictive power of Artificial Intelligence systems has been proven to be extraordinary, especially in the fields of oncology and medical diagnostics. While the field of machine learning and deep learning has made considerable progress, one of the major challenges for bringing AI technologies to the clinic is the lack of interpretability found in many AI models. Physicians in healthcare settings must provide medically understandable reasoning for their diagnosis, treatment and patient management. As such, predictive systems that do not offer a biologically interpretable output may not be well accepted when used in the real clinical world. It is known as the "black box" problem of Artificial Intelligence. Explainability was therefore selected as one of the key aspects of the proposed Clinical Decision Support System with the help of AI. The goal of the developed framework was not only to achieve a high predictive value but also to establish a system that is transparent, clinically interpretable, and can enhance the confidence of doctors and evidence-based decision making. Explainable Artificial Intelligence (XAI) is a form of AI that involves explaining how machine learning algorithms make predictions. Explainability is particularly relevant in breast cancer diagnostics as doctors need to be aware of what features of the tumor or cancerous cells are key to the assessment of malignancy.The explainability framework used in the present study was mainly centered on the feature importance analysis and probabilistic interpretation of the predictive outputs. To determine the morphometric cell variables that were most strongly linked with malignant tumors, and to examine if the AI system was learning biological meaningful patterns versus random mathematical correlations.

Ensemble tree-based algorithms such as the Random Forest model were found to be useful for explainability analysis, as they can estimate the contribution of each variable to predictive performance. The feature importance analysis showed that the following parameters are important:

- Worst area
- Worst concave points
- Worst perimeter
- Concavity
- Radius irregularity

Among the most important predictors of malignancy were

Such results are in line with the principles of pathological oncology. The abnormal enlargement of the cell's nucleus, irregular boundaries, structural asymmetry, uncontrolled proliferation and increased concavity within the cellular structure are features commonly seen in malignant cells. Hence, the machine learning models proved to be able to identify biologically relevant morphometric abnormalities linked to cancer progression.

6.2 BIOLOGICAL INTERPRETATION OF PREDICTIVE FEATURES AND DIAGNOSTIC PATTERNS

The explainability results also showed that the AI system could emulate some aspects of what the trained clinicians did when examining the slides under a microscope. This lends more credibility and applicability to the proposed framework.

In the clinical realm, explainability provides a number of important benefits such as:

- Improved physician trust
- Improved diagnostic reasoning
- Increased insights into predictions

Increased receptivity to the use of artificial intelligence. Greater openness to the use of AI.

There is a greater level of confidence in identification of high risk cases.

Enhanced clinician–patient communication.

The proposed framework was thus enhanced to become a more clinically relevant decision support system by integrating explainable AI.

As a result of this, the biological interpretation of predictive features and diagnostic patterns is discussed.

A crucial question in the context of AI in medical diagnostics is whether the factors that a machine learning algorithm predicts are real biological and pathological processes. The explainability framework in the present study identified that the most influential predictive features were those that directly relate to known characteristics of cellular behavior in malignancy. One feature that was most predictive of malignancy was the "worst area." This discovery is biologically relevant because in the case of malignant cells, there is a tendency to enlarge the nucleus and to increase the rate of cellular proliferation, causing the cell to have an enlarged and irregular shape. Enlarged nuclei is a known feature of cancer progression and is a standard diagnostic feature in a cytological examination. Similarly, the terms "worst perimeter" and "concavity" showed significant association with malignant tumors. As a result of genetic instability and abnormal growth patterns, cancerous cells tend to become structurally abnormal and lose their uniformity, and acquire irregular nuclear boundaries. As the concavity is increased along with the irregularity in the perimeter of cells, it signifies structural distortion within cellular morphology which is associated with aggressive behaviour of tumor. The significance of "worst concave points" is also very close to the pathological principles since the malignant tumors often have sharp irregular projections and nuclear asymmetry. These defects are frequently related to tissue infiltration and invasion. The machine learning models identified then morphometric patterns that are already known by pathologists in the process of microscopic diagnosis. Such a result gives more biological credibility to the AI-assisted framework, and provides proof that the algorithms were learning clinically relevant pathological relationships and not just random statistical associations.

An additional key finding from the current study was the high degree of class separability that was observed in the Breast Cancer Wisconsin Diagnostic dataset. The cytological variables were found to contain highly discriminative information that can effectively differentiate benign and malignant tumors, as the predictive algorithms performed well up to high AUROC values in all the computational approaches evaluated.The fact that the features listed under the three different algorithms are similar in importance also bolsters the validity of the patterns predicted in this study. Although the models have different mathematical learning mechanisms (Logistic Regression, Random Forest, Gradient Boosting, Neural Network), they highlighted the same morphometric variables during the prediction process. From a clinical perspective, this consistency is very useful as it means the pathological features identified are strong and repeatable, regardless of the computational approach used. This reproducibility is crucial for the future clinical application and validation of AI-assisted diagnostic systems. The results of the biological interpretation analysis in the present study thus prove that the structured cellular morphometric data can be used as a highly informative basis for breast cancer diagnostics via machine learning.

6.3 CLINICAL INTERPRETATION, DECISION SUPPORT AND PRACTICAL APPLICABILITY

This course aims to give students insights into the clinical, decision-support, and practical application of statistics. The Clinical Decision Support System (CDSS) proposed here is not meant to replace clinicians, but as a technology that assists them, in order to facilitate the consistency of diagnosis, to decrease the burden of interpretation, and to assist in the evidence-based decision-making process in oncology workflows. The primary benefit of the framework developed in the present study was the incorporation of the probability-based risk stratification and not solely binary classification. Doctors don't usually make decisions based on absolutes in the real world: benign or malignant. Rather, clinical decision-making is frequently based on risk estimation, uncertainty assessment and correlation with imaging, history and pathological interpretation.

The proposed framework thus produced continuous probability output; that is, the chances of malignancy for every patient sample. The following probabilities were classified as:

- Low-risk category
- Intermediate-risk category
- High risk category

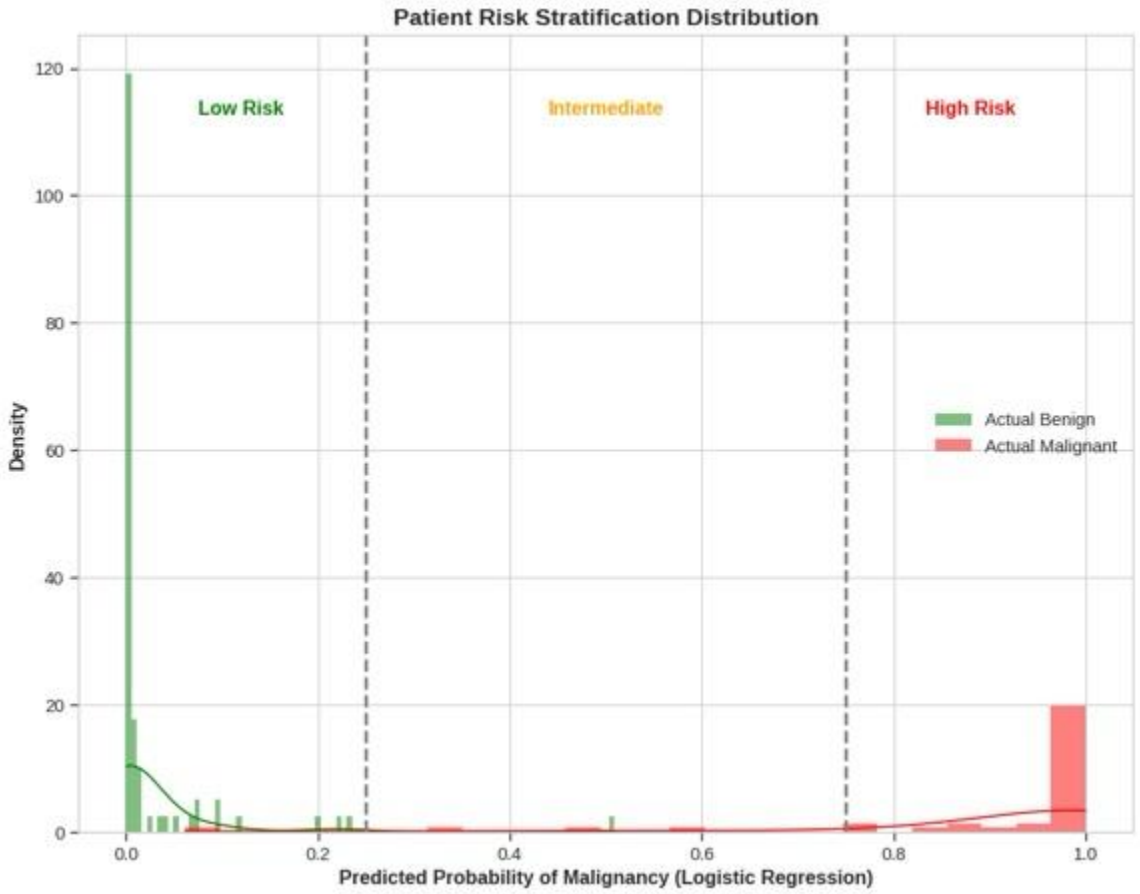


Figure 5: Probability-Based Patient Risk Stratification Distribution Generated Using Logistic Regression Predictions

The low-risk group comprised those cases with strong association to benign pathology and low predicted probability of malignancy. These patients might need continued follow-up and monitoring, but not immediate and aggressive intervention. The high-risk category consisted of the samples with strong malignant properties that would need prompt oncological intervention, confirmatory diagnostic workup and additional treatment planning. The intermediate-risk category encompassed “grey-zone” patients who were diagnostically unclear, and might need further consultation with the physician, imaging analysis, biopsy or a multidisciplinary approach. This probabilistic interpretation provided a much greater clinical relevance to the proposed system since it was closer to the clinical setting in oncology. Probability assessment and clinical judgment are often used by physicians in the treatment of indeterminate pathological presentations. The explainability aspect of the framework also improved the utility of the framework by enabling the visualization of the contribution of the most important cellular variables to the prediction results. This openness can help build trust among physicians and alleviate concerns about AI-based diagnostic tools. The proposed framework could offer further value in resource-limited healthcare contexts, helping hospitals and diagnostic centers facing a lack of specialized oncologists and pathologists.

CDSS can help streamline the healthcare process, minimize diagnostic lag times, and focus on high-risk patients that might need immediate assessment. In addition, the efficiency of structured machine learning systems renders them apt for incorporation into cloud-based healthcare systems and hospital information systems. These types of frameworks could eventually lead to real time Oncology decision support systems that can help oncologists in screening and triaging patients.

While AI systems make impressive predictions, it is crucial to note that they have some drawbacks. Predictive algorithms can be problematic with regards to:

- Rare pathological variants
- Poor-quality cytological samples
- Dataset bias
- Ambiguous cellular morphology
- Unusual biological presentations

As such, clinical supervision is always needed and the suggested framework should be considered as a technology that works in partnership with the physician, not as a standalone diagnostic tool. The results from the present study are very much in support of the idea that explainable AI can serve as a viable tool in bridging the gap between computation and clinically interpretable oncology practice.

6.4 LIMITATIONS OF THE PRESENT STUDY AND FUTURE DIRECTIONS OF RESEARCH

The proposed Clinical Decision Support system based on AI showed promising predictive performance and clinical relevance, but there are also some limitations to be noted. It is important to be aware of these limitations so as to be able to appreciate the extent of the study and how it could be advanced in future to enhance the application of the study in the real world. The present study has a few limitations, one of the main ones being that it was based on only one particular structured cytological data set. The Breast Cancer Wisconsin Diagnostic dataset is highly validated and a good choice for machine learning research, but clinical settings can present a higher degree of biological diversity, imaging variability, and a diversity of patient populations. The proposed framework should be assessed in future studies with larger multicentric data sets obtained in a variety of health care contexts. A further constraint is the limited number of samples available, which is less than is typically found in recent deep learning applications. The current data set showed good class separability and predictive consistency, however, larger datasets could provide for better generalization and robustness of the predictive models. The existing structure also focused only on the cellular morphometric features and did not incorporate other with multimodal healthcare information like:

- Histopathological images
- Mammographic imaging
- Genomic sequencing
- Electronic health records
- Clinical history
- Laboratory biomarkers

Moving forward, more comprehensive precision oncology workflows can be enhanced by combining multiple health care modalities in a single unified AI system. Incorporating explainability into the present framework, as done with feature importance analysis, would further enhance interpretability and understanding of the algorithm's behavior by physicians but could be further improved with more advanced techniques of explainable AI, such as SHAP value, LIME analysis, and attention-based visualization methods. Another limitation is related to deployment in the real world. To bring AI systems to the clinic, the following are necessary:

- Regulatory approval
- Ethical validation
- Data privacy protection
- Continuous monitoring
- Prospective clinical trials

Conducting a physician training and adaptation

This study was primarily a computational model development and retrospective evaluation study and not a live clinical deployment; emphasis is important to note. Further studies are needed to explore potential real-world validation in hospital settings and to assess the interaction between clinicians and AI-supported decision-making systems in real-world diagnostic processes. In spite of these drawbacks, the present study can successfully be used to show that explainable machine learning systems have great promise in supporting early breast cancer diagnosis and clinical decision-making. The system created in this work provides a solid framework for future development of oncology systems with explainable AI.

Other enhancements that could be made in the future include:

• Real-time clinical integration
• Cloud-based deployment
• Mobile healthcare implementation
• Combination with imaging systems

Multimodal precision oncology platforms

• Personalized treatment recommendation systems

The ongoing development of explainable AI and computational oncology holds the promise of revolutionizing healthcare, making diagnosis more accessible, easing the burden on healthcare providers, and creating more precise and evidence-based management plans for patients.

CHAPTER 7:
CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

7.1 CONCLUSION

Breast cancer remains one of the biggest health problems in the world and is also a leading cause of female cancer-related deaths. Early detection is one of the most important aspects of modern oncology, as the effectiveness of breast cancer treatment is largely dependent on the timing of diagnosis. Traditional diagnostic workflows, however, are often tied to challenges like relying on specialists, diagnostic variation, delayed diagnosis, and a growing clinical burden. The emerging trend of Artificial Intelligence and machine learning technologies has opened up new opportunities to enhance cancer diagnostics and aid in evidence-based clinical decision making. This research aimed to create a system that would help with the early detection of breast cancers and probability-based risk stratification based on structured cellular morphometric data – an Explainable AI assisted Clinical Decision Support System (AI-CDSS). This research considered the development of a framework based on the Breast Cancer Wisconsin Diagnostic dataset, which consists of quantitative features of cells extracted from the Fine Needle Aspiration Cytology images. These morphometric variables were biologically meaningful cell characteristics such as nuclear size, texture, perimeter irregularity, smoothness, concavity, symmetry and structural complexity.

To develop a clinically meaningful and explainable diagnostic approach, a systematic computational workflow that included preprocessing, normalization, predictive modeling, explainability analysis, and risk stratification was adopted. A comparative analysis of performance of four machine learning algorithms – Logistic Regression, Random Forest, Gradient Boosting, and Multi-Layer Perceptron Neural Network is performed on the distinction between benign and malignant tumors. The acquired results showed that the models obtained had a high predictive capability, which suggests that the cytological characteristics are structured and include highly discriminative pathological information that can be utilized in machine learning for cancer diagnosis. Random Forest was the most accurate model overall and Logistic Regression had the best AUROC score and probability calibration ability among the evaluated models. The incorporation of E-AI into the prediction model is one of the most valuable insights of this research. Feature importance analysis showed that the machine learning models were able to recognize important pathological variables that are biologically significant, including worst area, worst concave points, concavity and perimeter irregularity, among others, as important indicators of malignancy. The results are well aligned to the pathological features of cancerous cellular morphology, and enhance the transparency and interpretability of the AI-assisted system. One of the other key findings of the study was the implementation of probability-based risk stratification.

The proposed framework did not only provide binary predictions, but clinically meaningful risk categories such as low-risk, intermediate-risk, and high-risk groups. This is a more useful method as probability interpretation and physician judgment are often used in real-world oncology workflows, but not necessarily just absolute classifications. The study results underscore the potential of explainable Artificial Intelligence in cutting-edge healthcare and precision oncology. The proposed Clinical Decision Support System with AI capabilities proved that AI technologies can be applied to assist clinicians in making more accurate diagnoses, alleviating some of the burden of interpretation, and helping with evidence-based patient management. The study also underscored the need for physician supervision and ethical responsibility while implementing AI. The proposed framework was intentionally intended to be a collaborative, physician supportive technology, rather than a fully independent diagnostic replacement.

7.2 FUTURE SCOPE OF THE STUDY

The Clinical Decision Support System (CDSS) that uses artificial intelligence proved to have a high predictive value and clinical relevance, but the field of artificial intelligence in oncology is still in constant development.

From the present study, it can be said that a framework is set that can be extended and developed much further in the future if desired through research and technological integration. Integration of multimodal healthcare data is one of the most important future directions. The framework used currently was based on structured cytological morphometric variables obtained from images of Fine Needle Aspiration Cytology. Today, however, precision oncology increasingly makes use of several healthcare modalities like:

• Histopathological imaging
• Mammographic analysis
• Ultrasound imaging

Magnetic Resonance Imaging (MRI)

• Genomic sequencing
• Proteomic profiling
• Electronic health records

Clinical history and biomarkers.

With the integration of these multimodal datasets, future AI systems could gain even more accurate diagnoses and offer a deeper understanding of patients. These integrated systems can be used to assist not only in the diagnosis but also in the prediction of prognosis, selection of treatment, survival analysis, and personalised treatment planning for cancer. Another future direction of interest is the use of deep learning architectures in medical imaging analysis. The proposed system could be combined with Convolutional Neural Networks (CNNs) and other computer vision systems to analyze the original histopathological and radiological images. The structural cytological analysis and imaging-based AI could greatly improve the diagnostic capacity.

The explainability framework can be extended with the application of sophisticated interpretable AI methods, including:

• SHAP
• LIME
• Attention visualization systems
• Counterfactual reasoning models

Provide an understanding of explainable deep learning architectures

Such approaches could also enhance physicians' comprehension of algorithmic decision making and boost trust in AI-powered healthcare systems. Real-time clinical deployment and prospective validation is another area of future development along with it. In the present study the main emphasis was laid on the retrospective computational evaluation with the use of an established benchmark dataset. Live clinical integration in the diagnostic laboratories and hospitals should be explored in future studies to assess:

• Real-world diagnostic performance
• Physician-AI interaction
• Workflow efficiency
• User adaptability
• Clinical decision-making behavior
• Patient outcomes

Another future path of growth is cloud-driven healthcare deployment. In the future, AI-enabled Clinical Decision Support Systems could be part of a comprehensive hospital information system, a telemedicine service, and a mobile health app. These would be useful in developing health care that is more accessible in rural and underserved areas, where special expertise in oncology is not available. Other systems could also be based on personal oncology strategies that would be able to provide tailored recommendations for treatment for individual patients depending on their specific biological profile. AI-powered tools, combined with genomic sequencing and molecular markers, could help drive precision medicine and optimal targeted therapies.

Furthermore, ethical and regulatory issues related to the use of AI in healthcare must be explored in future studies, such as:

- Data privacy protection
- Bias mitigation
- Fairness evaluation
- Ethical accountability
- Regulatory compliance

Clear systems for physician supervision

The current innovations and progress in explainable AI, computational oncology, and digital healthcare technology can have significant impacts on cancer diagnostics and patient management at a clinical and societal level.

7.3 SOCIAL AND CLINICAL IMPACT OF THE PROPOSED FRAMEWORK

As breast cancer continues to be a growing global burden, there is a critical need for healthcare systems that can provide support for early detection, effective patient management, and access to medical services that are available to everyone. Even though cancer awareness initiatives like awareness drives, campaigns, and social media campaigns have been implemented, access to diagnostic facilities, pathological expertise, and specialized oncologists is still inadequate in many health care settings, especially in developing and resource-limited areas, preventing the early detection and treatment of cancer.Proposed AI based Clinical Decision Support System has tremendous potential to make a positive contribution to the healthcare system and to society. AI-driven solutions can boost the early detection and regularity of breast cancer, potentially lowering death rates and enhancing patients' long-term survival chances.The proposed framework has the potential to enhance the accessibility of healthcare services, which is one of its most significant social benefits. Diagnostic support systems, which are developed with the help of artificial intelligence (AI), can help hospitals and clinics, where specialists are not always available, to assist in the diagnostic evaluation and triaging of patients. These can be particularly beneficial in rural or underserved areas, where access to sophisticated oncology care is restricted.The explainability aspect integrated into the framework also enhances clinical trust and physician acceptance. The greater the degree of transparency, the better AI systems are likely to be accepted in real healthcare settings, as they also offer biologically meaningful explanations.That is because transparent AI systems that can provide biologically interpretable reasoning are more likely to be accepted in practical healthcare environments than highly opaque black-box systems.

The suggested framework can help in achieving following healthcare workflow aspects:

The workload of diagnosis is reduced.

- Faster patient prioritization
- Reduced diagnostic variability
- Enhanced evidence-based decision-making

Enhanced efficiency in pathology and oncology departments

The probability based risk stratification mechanism introduced in the present study further improves the clinical applications as doctors often need to interpret the risks in a grading manner when they are handling unclear cases. These systems could help healthcare workers quickly pinpoint patients in need of urgent evaluation and minimize invasive procedures for patients with a low risk.The study also highlights the overall significance of incorporating AI into healthcare responsibly and ethically. AI technologies should not replace doctors, but instead complement their practice and skills to provide optimal patient-centered service. Beyond health-related gains, the developers of explainable AI systems could help raise awareness and public understanding of digital health technologies and precision medicine. As AI in healthcare becomes more prevalent, it could increase the use of proactive screening methods, enhance diagnostic reach, and shift the focus towards preventive oncology. But there is a need to implement it responsibly. To ensure safe and equitable healthcare delivery, AI systems should be developed with ongoing supervision of physicians, accountability for their ethical usage, safeguarding of patient privacy, and rigorous clinical validations.

REFERENCES

[1] P. Tiwari, A. Sharma, and R. Verma, "Artificial Intelligence in Oncology: Predictive Modeling and Pattern Recognition from Large-Scale Clinical Data," *Clinical Oncology Informatics*, vol. 14, no. 2, pp. 112–128, 2025.

[2] A. Hulule, S. Khan, and M. Joseph, "Enhancing Diagnostic Sensitivity and Workflow Efficiency through Natural Language Processing and Deep Learning," *Nature Medicine AI*, vol. 3, no. 1, pp. 45–59, 2025.

[3] M. Riaz, F. Ahmed, and T. N. Gupta, "Overcoming Algorithmic Bias and Interpretability Challenges in Precision Oncology," *Lancet Digital Health*, vol. 7, no. 4, pp. e301–e315, 2025.

[4] S. Kolla and R. Parikh, "Regulatory Concerns and Data Quality in the Safe Clinical Deployment of AI Decision Support Systems," *Medical Image Analysis*, vol. 88, p. 102850, 2024..

[5] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," *UCI Machine Learning Repository*, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[8] K. L. Keh, W. Xu, E. Lepisto, H. Elmarakeby, M. J. Hassett, E. M. Van Allen, et al., "Natural Language Processing to Ascertain Cancer Outcomes from Medical Oncologist Notes," *JCO Clinical Cancer Informatics*, vol. 4, pp. 680–690, 2020.

[9] T. Haddad, J. M. Helgeson, K. E. Pomerleau, A. M. Preininger, M. C. Rebuck, and I. Dankwa-Mullan, "Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study," *JMIR Medical Informatics*, vol. 9, no. 2, e27767, 2021.

[10] M. J. Iqbal, Z. Javed, H. Sadia, A. Qureshi, A. Irshad, et al., "Clinical Applications of Artificial Intelligence and Machine Learning in Cancer Diagnosis: Looking into the Future," *Cancer Cell International*, vol. 21, p. 270, 2021.

[11] S. M. Nonan, Y. M. Fadel, M. T. Henedak, N. A. Attia, M. Essam, and E. Elmasarawi, "Leveraging Survival Analysis and Machine Learning for Accurate Prediction of Breast Cancer Recurrence and Metastasis," *Scientific Reports*, vol. 15, p. 3728, 2025.

APPENDIX

List of Publications and Conferences-

The following publication has been accepted during this thesis work:

Title: **An AI Assisted clinical support system for early breast cancer detection and risk stratification using cellular morphometrics**

Authors: Jayna Bhattacharjee, rifah ansari, SmitaRastogi Verma

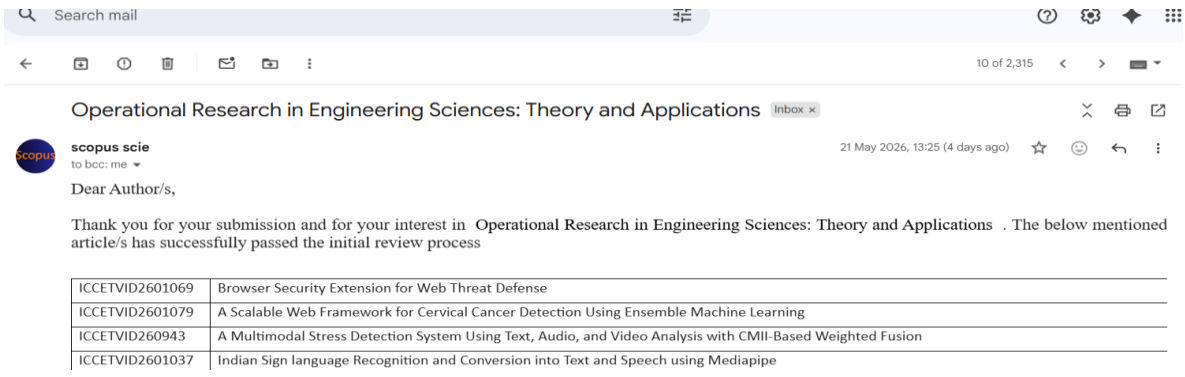
Affiliation: Department of Biotechnology, Delhi Technological University, Delhi–110042, India

Conference: International Conference on Cognitive Informatics Engineering and Technology 2026 (ICCET 2026)

Digital Library / Repository: OSIET Digital Proceedings
Status: intial review passed /Awaiting Final Editorial Decision (18TH June)

ICCETVID2601178	NAVIGANT – AN INDOOR POSITIONING AND NAVIGATION SYSTEM
ICCETVID2601504A	MaxModel: Predicting Mixed Martial Arts Outcomes: The Impact of Weight Class Stratification
ICCETVID2601815	Real-Time Food Recognition and Nutritional Estimation Using Deep Learning and Temporal Bayesian Fusion
ICCETVID2601502	LEGAL AI RESEARCH ASSISTANT, LEXA: AI-Powered Legal Research Agent for Efficient Legal Knowledge Discovery
ICCETVID2601830	Metaverse for Military Training: AI and Quantum-Assisted Virtual Battlefields
ICCETVID2601832	Energy-Efficient FIR Filter Implementation with NP-Zipper-Logic-Based Multipliers
ICCETVID2601550	An AI-Assisted Clinical Decision Support System for Early Breast Cancer Detection and Risk Stratification Using Cellular Morphometrics
ICCETVID2601564	Aegis MultiLayer Middleware Architecture for RealTime Prompt Injection Detection in Large Language Models

The article/s is now sent for further review process, the **final decision on the paper acceptance (with major/minor corrections) or Rejected status (If rejected we will provide an alternative journal with additional fee)** will be confirmed by the journal before 18th JUNE 2026 unless otherwise specified.



ICCETVID2601069	Browser Security Extension for Web Threat Defense
ICCETVID2601079	A Scalable Web Framework for Cervical Cancer Detection Using Ensemble Machine Learning
ICCETVID260943	A Multimodal Stress Detection System Using Text, Audio, and Video Analysis with CMII-Based Weighted Fusion
ICCETVID2601037	Indian Sign language Recognition and Conversion into Text and Speech using Mediapipe

Presentation in Conference- **International Conference on Cognitive Informatics Engineering and Technology 2026 (ICCET 2026)**

Date of Conference: 28TH March 2026

INTERNATIONAL CONFERENCE ON
COGNITIVE INFORMATICS ENGINEERING AND
TECHNOLOGY- 2026

Organised by

VIDYAA VIKAS COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
TIRUCHENGODE, TAMILNADU, INDIA.

In collaboration with

OSIET , Chennai , India .
SAMARKAND STATE UNIVERSITY, SAMARKAND , UZBEKISTAN .

Certificate of Presentation

This is to Certify that the paper entitled

An AI-Assisted Clinical Decision Support System for Early Breast Cancer Detection and Risk Stratification Using Cellular Morphometrics

Authored by

Jayna Bhattacharjee
Delhi Technological University (DTU), Delhi

has been presented at

International Conference on Cognitive Informatics Engineering and Technology- 2026
held on 28th & 29th March 2026 at
Vidyaa Vikas College of Engineering And Technology, (Autonomous)
Tiruchengode , Tamilnadu , India.

Dr.K.Pooranapriya
Principal & Conference Chair

K.Janani
CEO, OSIET

Dr.Chirsto Ananth
Professor
Samarkand State University, Uzbekistan

Dr.Akhatov Akmal Rustamovich
Vice Rector (International Cooperation)
Samarkand State University, Uzbekistan

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, New Delhi, 110042

PLAGIARISM & AI VERIFICATION

Title of the Thesis **"An AI Assisted clinical support system for early breast cancer detection and risk stratification using cellular morphometrics**"

Total Pages , Name of the Scholar **JAYNA BHATTACHARJEE (24/MSCBIO/05)**

Supervisor

Dr. Smita Rastogi Verna
Department of Biotechnology
This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:
AI % Detected- **below 20 %**
Software used: **Turnitin,** Similarity Index: **9 %** Total Word Count: **13976**

Jayna Bhattacharjee

jayna thesis 05 FINAL_removed (1)

Rifah

Document Details

Submission ID
trn:oid:::27535:140373482

Submission Date
May 25, 2026, 3:22 PM GMT+5:30

Download Date
May 25, 2026, 3:25 PM GMT+5:30

File Name
jayna thesis 05 FINAL_removed (1).docx

File Size
312.0 KB

37 Pages

13,451 Words

83,025 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Small Matches (less than 8 words)

Match Groups

- **104** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **2** Missing Quotations 0%
  Matches that are still very similar to source material
- **18** Missing Citation 1%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

6% Internet sources
4% Publications
5% Submitted works (Student Papers)

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.