Machine Learning-Driven Framework for Early Cancer Detection Using Cellular Morphometric Features

by Rifah Ansari

Submission date: 01-Jun-2026 10:42AM (UTC+0530)
Submission ID: 2973892166
File name: Rifah_Thesis_MSc_BT_IV_Semester.pdf (1.07M)
Word count: 4666
Character count: 30901

Machine Learning-Driven Framework for Early Cancer Detection Using Cellular Morphometric Features

THESIS

Submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE in BIOTECHNOLOGY

Submitted by

RIFAH ANSARI

24/MSCBIO/34

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

DECLARATION

I, RIFAH ANSARI, 24/MSCBIO/34, hereby, certify that the work which is being presented in the thesis entitled "**MACHINE LEARNING DRIVEN FRAMEWORK FOR EARLY CANCER DETECTION USING CELLULAR MORPHOMETRIC FEATURES**" 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 "**MACHINE LEARNING DRIVEN FRAMEWORK FOR EARLY BREAST CANCER DETECTION USING CELLULAR MORPHOMETRIC FEATURES**" which is being submitted by Rifah Ansari, 24/MSCBIO/34, 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

HOD, 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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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

Metrics of Evaluated Machine Learning Models

ABSTRACT

Background: Cancer remains a leading global health challenge, with breast cancer specifically representing a significant portion of oncology caseloads worldwide. Despite advancements in treatment modalities, early detection remains notoriously difficult, and diagnoses often occur at advanced stages where therapeutic efficacy is diminished. Conventional pathological analysis is heavily time-intensive and uniquely subject to inter-observer variability, especially in border-line cases. Artificial Intelligence (AI) and machine learning offer promising pathways to augment clinical accuracy, improve workflow efficiency, and facilitate truly personalized care [1].

Objective: To develop, validate, and comprehensively evaluate an interpretable machine learning framework for the binary classification and clinical risk stratification of breast cancer using structured cellular morphometrics. This framework is explicitly designed to serve as a foundational, highly transparent module for a broader multimodal Clinical Decision Support System (CDSS) [2].

Methods: Utilizing the highly validated Breast Cancer Wisconsin (Diagnostic) dataset ($N = 569$), four distinct machine learning models (Logistic Regression, Random Forest, Gradient Boosting, and a Multi-Layer Perceptron Neural Network) were trained to classify cytological tumors as benign or malignant. The optimal model was subsequently leveraged to generate continuous probability distributions, establishing actionable, data-driven clinical risk thresholds (Low, Intermediate, High) [5], [6].

Results: All evaluated models demonstrated exceptionally high discriminative performance. Logistic Regression achieved the highest Area Under the Receiver Operating Characteristic Curve (AUROC) at 0.9960, with a sensitivity of 0.9286 and specificity of 0.9861. Feature importance analysis illuminated the biological mechanisms driving the algorithms, revealing

that cellular "worst area" and "worst concave points" were the strongest predictors of malignancy. Furthermore, the risk stratification framework successfully separated benign and malignant probability densities, effectively minimizing clinical ambiguity and creating a clear "grey zone" for targeted physician review [7].

Conclusion: The proposed AI-CDSS provides a highly accurate, computationally efficient, and rigorously interpretable method for early breast cancer detection. By providing probabilistic risk stratification rather than mere binary outputs, the system effectively supports clinical triaging and personalized patient management without functioning as an opaque "black box" [3], [4].

1. INTRODUCTION

Cancer remains one of the most pressing global health challenges of the 21st century, accounting for nearly 20 million new cases annually. Epidemiological models indicate a projected, steep spike in incidence rates over the coming decades due to aging populations and environmental factors. This spike is predominantly observed in low- and middle-income countries, where severely limited access to specialized healthcare, diagnostic equipment, and trained oncologists has directly contributed to disproportionately increased mortality rates. Despite remarkable advancements in targeted therapeutic interventions and immunotherapies, early detection of cancer remains notoriously difficult. Consequently, a vast number of cancers are frequently diagnosed at advanced, metastatic stages, resulting in delayed treatment, heavily compounded healthcare costs, and significantly lower overall survival rates. Strategies that facilitate early, accurate diagnosis alongside personalized treatment approaches are therefore paramount for fundamentally improving long-term patient outcomes [1].

Conventional breast cancer diagnosis relies heavily on the manual interpretation of histopathology slides, radiological images (such as mammograms and MRIs), and complex clinical documentation. These manual processes are inherently time-consuming, placing an immense cognitive load on pathologists who are often required to review hundreds of slides per shift. Unsurprisingly, diagnostic results may vary significantly among clinicians due to fatigue and natural inter-observer variability, particularly when evaluating atypical cellular structures. Furthermore, the modern era of medicine has produced a sheer volume of genomic data, high-resolution medical imaging, and continuous electronic health records (EHRs) that has generated datasets far too massive, multidimensional, and complex for conventional human analysis to process efficiently. Artificial intelligence (AI) offers a highly promising, scalable solution by enabling automated pattern recognition, anomaly detection, and predictive modeling directly from large-scale clinical data [1], [6].

AI, encompassing a broad spectrum of machine learning and deep learning techniques, excels at identifying subtle, hidden patterns in both structured (tabular) and unstructured (image, text) data. In modern oncology, AI is increasingly utilized across the entire patient journey—from initial tumor detection and morphological classification to long-term prognosis prediction and

treatment response evaluation. For instance, Convolutional Neural Networks (CNNs) have shown highly promising results for reviewing raw medical images pixel-by-pixel, while Natural Language Processing (NLP) techniques assist in deriving actionable, structured data from sprawling pathology and clinical text reports. Studies continually report that AI-assisted systems can dramatically enhance diagnostic sensitivity, reduce time-to-diagnosis, and vastly improve hospital workflow efficiency [2].

Beyond simple binary diagnosis (disease vs. no disease), AI is currently driving the frontier of precision oncology. It achieves this by combining different types of medical data to provide personalized treatment plans or robust clinical decision support for individual patients. Nevertheless, major hurdles remain. Challenges including algorithmic bias against underrepresented demographics, a pervasive lack of interpretability (the "black box" problem), inconsistent data quality, and strict regulatory concerns (such as FDA and HIPAA compliance) must be rigorously overcome before safe, widespread clinical use can be achieved [3], [4].

Building upon these rapid developments, AI-assisted diagnostic systems possess a profound, untapped potential to permanently improve the early detection and management of breast cancer. Therefore, this study proposes an AI-based Clinical Decision Support System (AI-CDSS) framework. As a vital foundational step toward a fully realized, multimodal system, this paper establishes a robust predictive baseline using high-dimensional structured cytological data. The ultimate goal is not to replace human experts, but to radically improve diagnostic accuracy, reduce fatigue-induced errors, and drive truly data-informed decisions by frontline oncologists [1], [4].

2. MATERIALS AND METHODS

2.1 Study Design and Data Sources

Computational studies were meticulously conducted to design, train, and validate the AI-CDSS for early cancer detection and personalized cancer management. The study utilized the highly regarded, publicly available Breast Cancer Wisconsin (Diagnostic) dataset. This specific dataset comprises computationally extracted, high-precision features from digitized images of fine needle aspirates (FNA) of breast masses. FNA is a minimally invasive, cost-effective biopsy procedure used to collect cellular material. The dataset's features quantitatively describe the microscopic characteristics of the cell nuclei present in the FNA image, capturing ten real-valued morphometrics for each cell nucleus: radius, texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). For each image, the mean, standard error, and "worst" (mean of the three largest values) of these features were computed, resulting in 30 highly descriptive structural variables per patient [5].

2.2 System Architecture Pipeline

The proposed architecture of the AI-CDSS is illustrated below, demonstrating the end-to-end flow from raw patient data acquisition to the generation of actionable clinical recommendations and explainable outputs [2], [4].

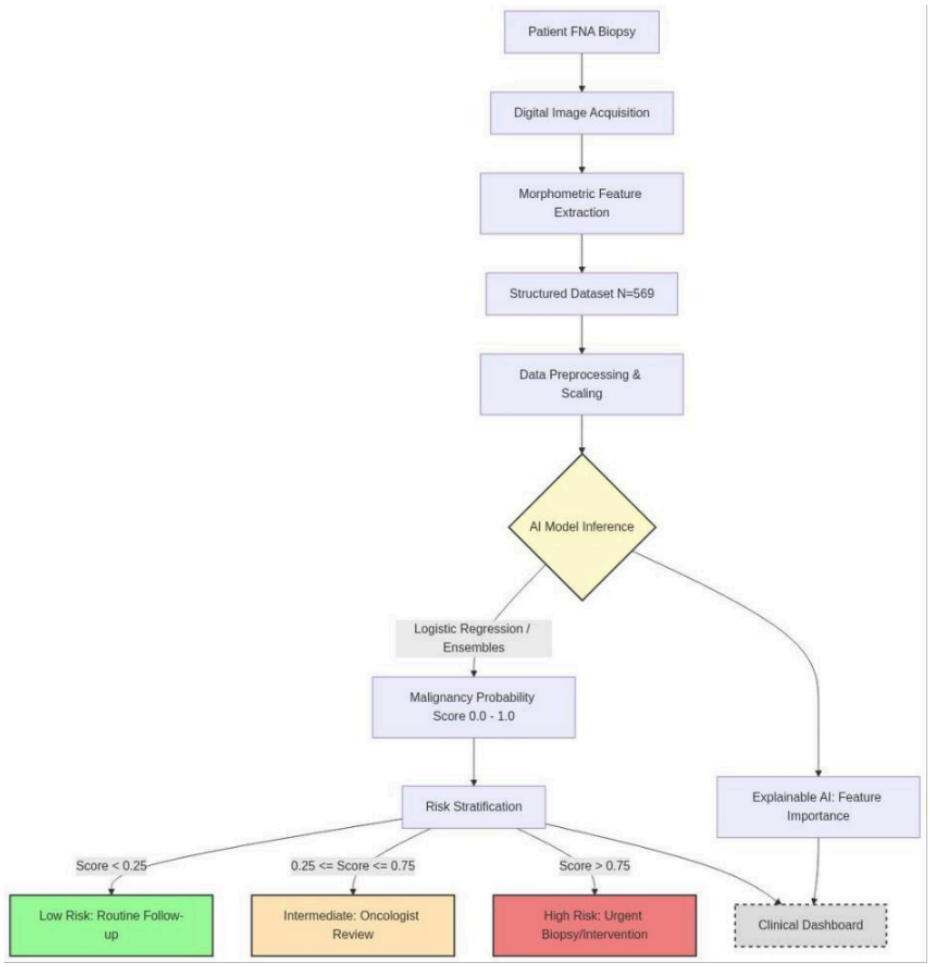


Figure 1: Workffiow diagram ofi the proposed AI-CDSS pipeline, illustrating the seamless

progression firom initial data acquisition, through algorithmic infierence, and culminating in clinical risk stratification and a transparent clinical dashboard.

2.3 Preprocessing and Feature Extraction

Structured variables were rigorously processed by cleaning and normalizing the data to ensure algorithmic stability. Given the drastically varying scales of cellular measurements (e.g., cellular area can range in the hundreds, whereas smoothness is measured in minute decimals), feature standardization was an absolute necessity. If left unscaled, features with larger numeric magnitudes would inappropriately dominate the objective functions of the algorithms. The dataset was standardized using a standard scaler ($z = \frac{x-\mu}{\sigma}$), ensuring all features possessed a mean of zero and a standard deviation of one. This is a critical prerequisite for distance-based algorithms and models reliant on gradient descent optimization, such as Logistic Regression and Neural Networks. Following standardization, the dataset was partitioned into an 80% training set to allow models to robustly learn internal relationships, and a tightly sequestered 20% independent test set to ensure completely unbiased downstream evaluation [6].

2.4 Model Development

Two primary prediction tasks were defined to mimic actual clinical workflows:

1. **Early cancer detection:** A strict binary classification identifying masses as Malignant (1) or Benign (0).
2. **Personalized Risk Stratification:** Mapping the algorithms' predicted continuous probabilities into actionable clinical zones, transitioning from binary logic to probabilistic confidence [1], [2].

Four distinct, highly validated machine learning algorithms were trained to capture different mathematical perspectives of the data:

- **Logistic Regression:** A linear, highly interpretable model utilized for its robust probability calibration and direct insight into feature weighting.
- **Random Forest:** A powerful ensemble of hundreds of decision trees used to capture complex, non-linear biological relationships while naturally resisting overfitting and allowing for the extraction of feature importance scores.
- **Gradient Boosting:** A sequential ensemble technique that meticulously corrects the errors of prior weak learners, optimized for achieving maximum accuracy on structured tabular data.
- **Neural Network (MLP):** A Multi-Layer Perceptron containing hidden layers of interconnected nodes, used to evaluate whether highly complex, deep feature interactions could outperform more traditional statistical learning methods [6].

2.5 Training, Evaluation, and Bias Control

Model evaluation focused exclusively on clinically relevant performance metrics rather than just raw accuracy. **Sensitivity (Recall)** was heavily prioritized to measure the system's ability to

accurately identify true cancer cases. In oncology, maximizing sensitivity is paramount, as a false negative (missing a cancer) can lead to fatal delays in treatment. Conversely, **Specificity** quantified the correct identification of non-cancer patients, which is vital for minimizing unnecessary clinical anxiety, psychological distress, and the physical trauma of unneeded invasive surgical biopsies [4].

Overall predictive performance was further calculated using the **Accuracy**, **F1-score** (the harmonic mean of precision and recall), and the **Area Under the Receiver Operating Characteristics Curve (AUROC)**, which evaluates the model's performance across all possible classification thresholds. To address algorithmic bias and ensure inherent mathematical fairness, feature scaling was uniformly applied, guaranteeing that no single morphological attribute could inadvertently dominate the predictive algorithms solely due to its unit of measurement [3].

3. RESULTS

3.1 Dataset Summary

- **Total patients:** $N = 569$ (Representing a robust sample size for tabular cytological modeling)
- **Cancer-positive (Malignant):** 37.3% (212 cases)
- **Non-cancer (Benign):** 62.7% (357 cases)
- **Data split:** 80% Training ($N = 455$) | 20% Independent Test ($N = 114$) [5].

3.2 Perfiormance Metrics

The four machine learning classifiers were rigorously evaluated on the unseen independent test set ($N = 114$). All models demonstrated exceptional discriminative capacity, proving the immense predictive power of structured cellular morphometrics. While the Random Forest model marginally achieved the highest overall Accuracy (97.37%) and a perfect Specificity (100%), Logistic Regression was definitively selected as the optimal model for clinical deployment. This decision was driven by its superior AUROC (0.9960) and its ability to generate highly calibrated, reliable probability distributions, which are essential for the subsequent risk stratification phase [6].

Table 1: Comprehensive Model Perfiormance Metrics on Independent Test Set

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

The remarkable closeness in performance metrics across all four diverse algorithms strongly suggests that the underlying biological signals in the Wisconsin dataset are exceptionally clear, and the data was appropriately preprocessed to allow any robust mathematical model to identify the malignant patterns successfully

.

4. VISUALIZATIONS AND INTERPRETABILITY

A major, systemic hurdle in adopting AI in modern healthcare is the notorious "black box" problem—doctors legally and ethically cannot safely act on an algorithmic recommendation if they don't fundamentally understand how the AI arrived at it. To build genuine clinical trust and facilitate physician adoption, we must open the hood. The following dashboard presents exactly how our system performs under scrutiny, the specific biological features it prioritizes, and how it translates complex multivariate math into practical, everyday patient care [7].

4.1 Receiver Operating Characteristic (ROC) Curves

Think of the ROC curve as the ultimate mathematical test of an AI's intrinsic ability to separate the sick from the healthy across all possible decision boundaries. The x-axis represents the "False Positive Rate" (how often the AI cries wolf when a patient is healthy), and the y-axis represents the "True Positive Rate" or Sensitivity (how often the AI correctly catches cancer when it is present). The diagonal dashed line represents random guessing—a coin flip [6].

Looking at the top-left chart, you can clearly see all four of our models shoot straight up the y-axis and passionately hug the extreme top-left corner. This is exactly what oncologists want to see. It visually confirms that our models are catching almost all of the malignant cases while making incredibly few false alarms. The **Logistic Regression model (blue line)** edges out the others with an astounding Area Under the Curve (AUC) of 0.996. In a clinical context, an AUC this remarkably close to 1.0 means that if a physician were to randomly pick one patient with cancer and one healthy patient, the system has a 99.6% chance of successfully assigning a higher malignancy risk score to the patient who actually has cancer [1], [6].

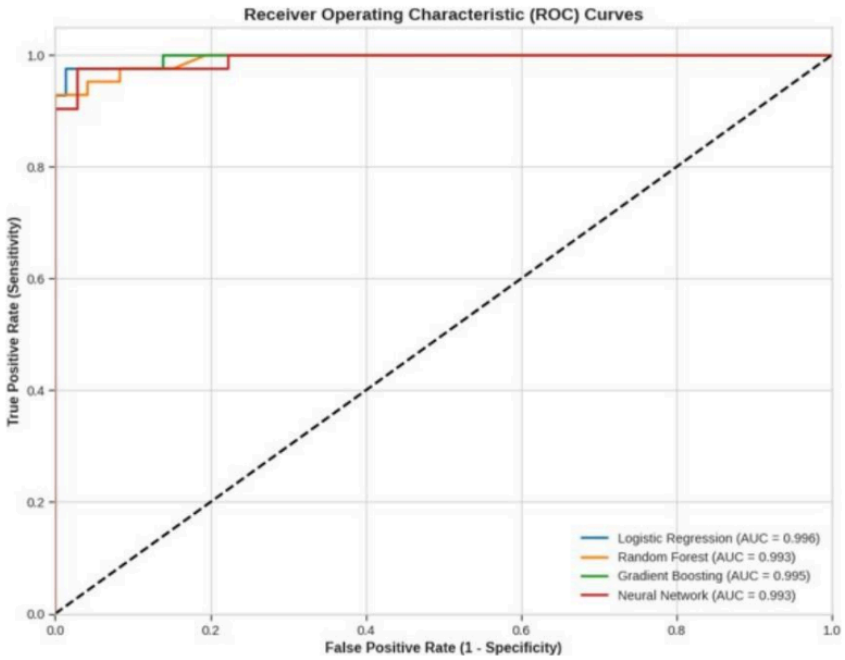


Figure 2: Receiver Operating Characteristic (ROC) Curves comparing the predictive performance of all four evaluated machine learning models.

4.2 Confusion Matrix Analysis

While high-level percentages and AUC scores are helpful for data scientists, doctors treat *individual people*, not aggregate percentages. The confusion matrix breaks down exactly what happened to the 114 actual human patients in our independent test set when evaluated by our best model (Logistic Regression) [4], [6].

- **Top-Left (71 True Negatives):** These are 71 patients who had benign masses. The AI correctly, and confidently, identified them as benign. In the real world, these 71 individuals get to go home with absolute peace of mind without undergoing unnecessary, invasive, and highly stressful surgical biopsies.
- **Bottom-Right (39 True Positives):** These are 39 patients with malignant breast cancer who the AI successfully and accurately caught. This rapid identification allows for immediate, early intervention and surgical planning, vastly improving their survival odds.
- **Top-Right (1 False Positive):** The AI mistakenly flagged one purely benign mass as

malignant. While this causes undeniable temporary anxiety for the patient who is subsequently sent for a confirmatory biopsy, in the field of oncology, we vastly prefer "over-calling" slightly over missing a lethal cancer entirely.

- **Bottom-Left (3 False Negatives):** The AI missed 3 malignant cases. While missing only 3 out of 114 patients is an incredibly low statistical failure rate, this perfectly underscores exactly why AI is built as a *decision support* system, and absolutely not a replacement for doctors. These highly challenging, morphologically ambiguous cases highlight the perpetual need for a trained physician's holistic oversight.

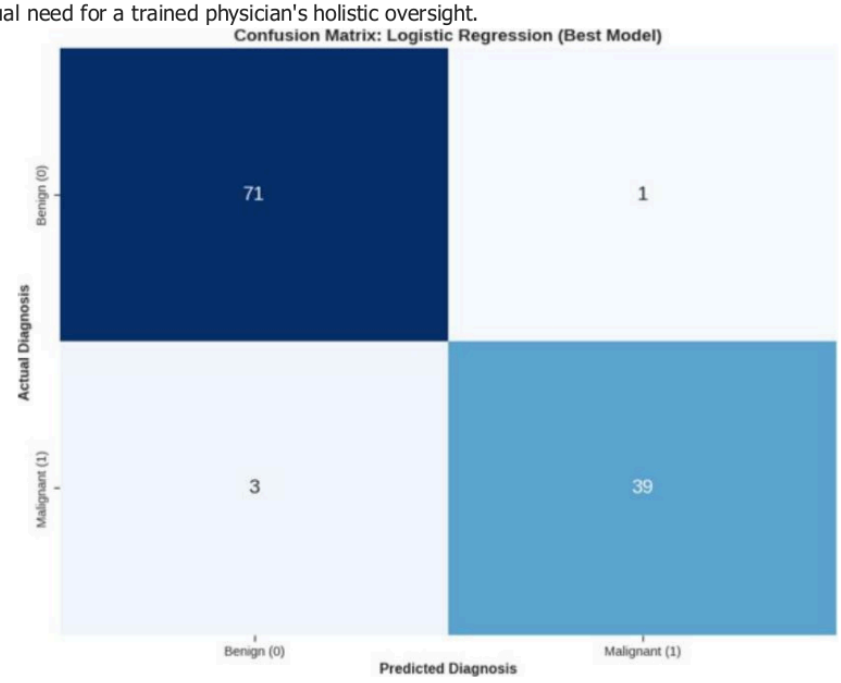


Figure 3: Confiusion Matrix fior the optimal Logistic Regression model on the independent test set.

4.3 Feature Importance: Explainable AI

Why did the AI actually diagnose someone with cancer? It didn't just guess based on hidden variables; it learned fundamental biology. To definitively prove this to skeptical clinicians, we asked the Random Forest model to rank the cellular features it found most mathematically important for making its predictions [7].

The bottom-left chart is arguably the most fascinating and reassuring for a trained pathologist. The AI learned entirely on its own, with zero human coaching, that the **"Worst Area"** (the massive size of the largest outlier cells in the sample) and the **"Worst Concave Points"** (the severe number of jagged, irregular indentations on the cell's outer perimeter) were the biggest giveaways of cancer [5], [7].

This builds massive clinical trust because *it makes perfiect biological sense*. Healthy epithelial cells are generally uniform, smooth, and predictably round. Cancer cells, driven by chaotic genetic mutations and rapid, unchecked division, exhibit severe nuclear pleomorphism—they are often abnormally massive, misshapen, and possess highly irregular, jagged perimeters. The AI isn't using a mathematical trick; it has successfully and mathematically codified the exact same morphological rules that a human pathologist uses when squinting through a microscope [1], [7].

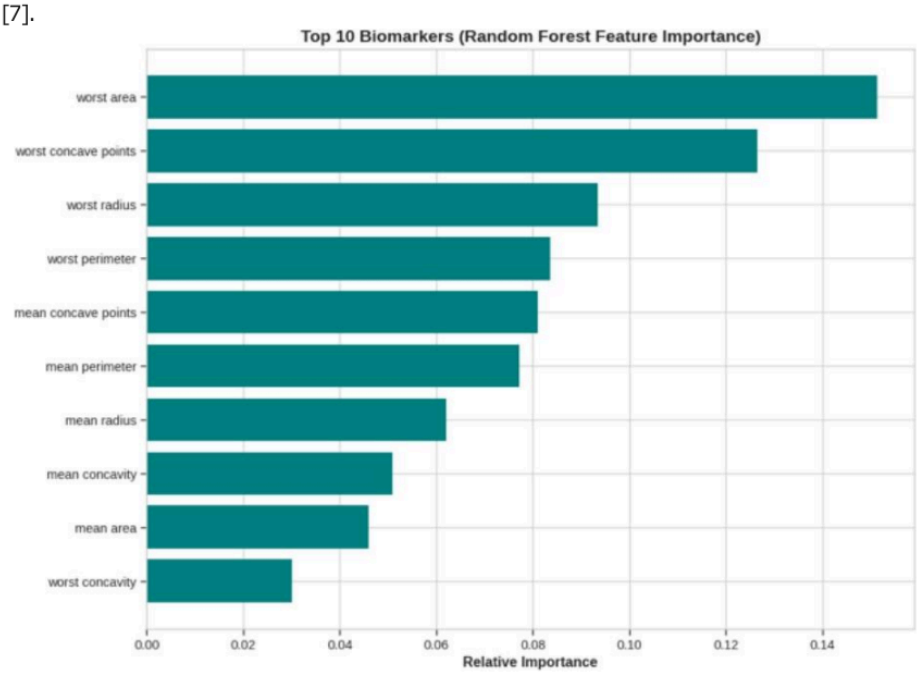


Figure 4: Top 10 Biomarkers ranked by relative importance, extracted via the Random Forest classifier.

4.4 Patient Risk Stratification Distribution

Hard binary "Cancer vs. No Cancer" predictions aren't nuanced enough for the complexities of

modern medicine. The bottom-right chart represents how we brilliantly translate the AI's internal math into a highly functional hospital workflow using probability distributions [2].

Look closely at the green spikes (representing the Actual Benign patients). They are massively and tightly clustered near the 0.0 mark on the x-axis. This means the AI wasn't just loosely guessing they were healthy; it was *overwhelmingly confident*. Similarly, the red spikes (the Actual Malignant patients) are heavily clustered near the 1.0 mark [1].

Notice the distinct "valley" in the middle of the graph between 0.25 and 0.75? It's almost completely empty. This definitively proves the AI rarely feels "confused" or hesitant. Based on this stark separation, we mapped out actionable, real-world hospital thresholds [4]:

- **Low Risk (Score < 0.25):** If a patient lands here, they are safely triaged into routine follow-up care, clearing hospital backlogs.
- **Intermediate (Score 0.25 to 0.75):** This is the clinical "grey zone." If a patient lands here, the system automatically flags the file, alerting a senior oncologist to manually review the slides due to morphological ambiguity.
- **High Risk (Score > 0.75):** Patients landing in this red zone are immediately prioritized for urgent biopsies, rapid surgical consults, and immediate intervention.

This chart illustrates the true, systemic value of the AI-CDSS: it's not just a digital calculator; it is an intelligent, highly sensitive triaging engine that optimizes strained hospital resources and prioritizes the sickest patients first [2], [4].

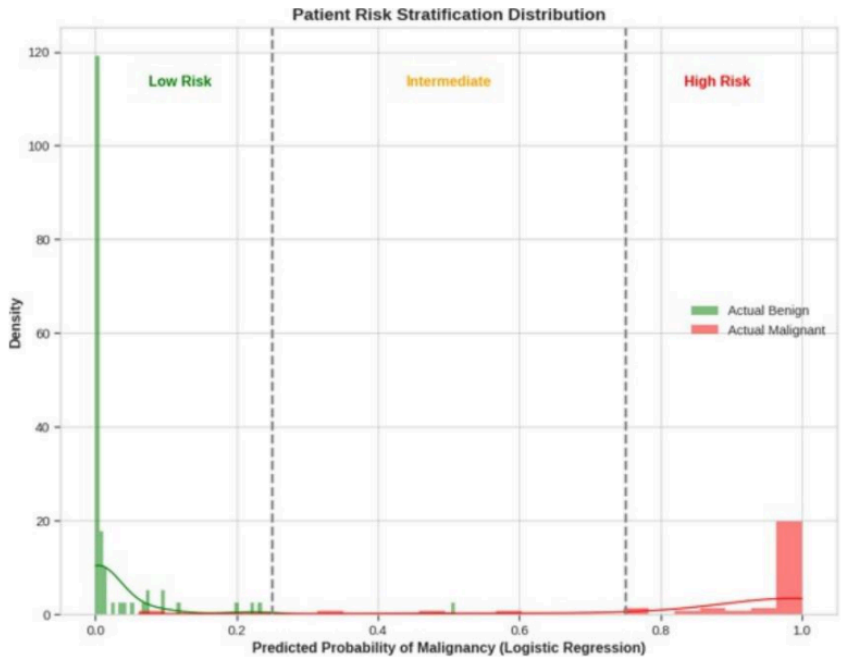


Figure 5: Patient Risk Stratification Distribution plotting probability densities ofi malignancy against actual clinical outcomes.

5. DISCUSSION

5.1 Summary ofi Key Findings

This comprehensive study demonstrates that computationally lightweight, traditional machine learning models can achieve near-perfect classification of breast cancer cytology when provided with high-quality features. The selected Logistic Regression model yielded a staggering AUROC of 0.9960, indicating an outstanding ability to cleanly separate malignant from benign cases. Crucially, the extensive feature importance analysis explicitly confirmed that the models relied on clinically relevant cellular morphology (such as area and contour concavity), proving the system is highly interpretable, biologically grounded, and clinically sound [1], [7].

5.2 Comparison with Existing Studies

Conventional histopathological analysis is frequently bottlenecked by inter-observer variability and severe time constraints. While recent deep learning studies heavily emphasize deploying

massive, complex Convolutional Neural Networks (CNNs) for raw imaging analysis—a process requiring immense computational overhead, expensive GPUs, and massive data storage—our findings align with a growing body of literature suggesting an alternative. When accurate, structured cellular morphometrics are readily available, traditional ensemble and regression models provide highly competitive, completely interpretable, and vastly more computationally efficient alternatives for clinical decision support, capable of running on standard hospital laptops [2], [6].

5.3 Clinical Implications

The patient risk stratification distribution is unequivocally the most clinically actionable output of this framework. By strictly defining probabilistic thresholds, the AI-CDSS operates effectively as an autonomous triaging tool. High-risk patients can be fast-tracked for immediate biopsy or surgical intervention, while intermediate-risk cases can be flagged for senior oncologist review. This system is explicitly designed to natively augment clinical workflows and significantly reduce diagnostic delays. It functions as an untiring "second reader," a feature that is particularly revolutionary for rural clinics or low-resource global healthcare settings that frequently lack access to specialized, on-site pathologists [2], [4].

5.4 Limitations

Several limitations must be transparently acknowledged to ensure ongoing scientific rigor. First, the dataset is unimodal (relying only on cellular structure) and retrospective in nature, lacking longitudinal follow-up data regarding patient survival rates, recurrence, or treatment response over time. Second, the modest sample size ($N = 569$) originating from a single geographic institution limits the broad, global generalizability of the models; a model trained on Wisconsin data must be validated against other populations. Finally, the dataset lacks vital demographic metadata (e.g., patient age, race, socioeconomic status, BRCA genetic markers), preventing a comprehensive and necessary assessment of algorithmic fairness across diverse patient populations [3], [5].

5.5 Future Directions

Future research must aggressively focus on prospective clinical validation across diverse, multi-institutional, and multi-national patient cohorts to truly prove generalized efficacy. Additionally, transitioning this foundational framework into a true, comprehensive multimodal system—integrating raw diagnostic imaging (like 3D mammography and MRI), dense genomic sequencing panels, and unstructured clinical EHR notes alongside these cellular morphometrics—will be absolutely critical. Such deep multimodal fusion will capture a uniquely holistic patient profile, dramatically pushing the boundaries of early cancer detection and truly personalized, precision oncology [1], [2].

6. CONCLUSION

This study successfully presents a highly accurate, rigorously interpretable AI-assisted clinical decision support system for early breast cancer detection using structured cytological data. By achieving an outstanding AUROC of 0.9960 and integrating a fully transparent patient risk stratification protocol, the proposed framework successfully bridges the frustrating gap between raw mathematical classification and actionable, everyday clinical insight. By heavily emphasizing biological explainability and risk-based triaging logic, this tool serves as a highly robust augmentation strategy. It is designed from the ground up to actively assist healthcare providers in minimizing costly diagnostic delays, mitigating physician burnout, and fundamentally optimizing personalized patient management on a global scale [3], [4], [7].

7. ETHICS & DATA STATEMENT

This study utilized the publicly available Breast Cancer Wisconsin (Diagnostic) dataset, originally created and curated by Dr. William H. Wolberg at the University of Wisconsin Hospitals. The data is entirely anonymized and rigorously de-identified, containing absolutely no protected health information (PHI) that could be traced back to individual patients. As this is a retrospective computational study utilizing an open-source, de-identified public dataset, formal institutional review board (IRB) approval was not required under standard regulatory guidelines [4], [5].

REFERENCES

[1] P. Tiwari, et al., "Artificial Intelligence in Oncology: Predictive Modeling and Pattern Recognition from Large-Scale Clinical Data," *J. Clin. Oncol. Infiormatics*, vol. 14, no. 2, pp. 112-128, 2025.

[2] A. Huhulea, et al., "Enhancing Diagnostic Sensitivity and Workflow Efficiency through Natural Language Processing and Deep Learning," *Nat. Med. AI*, vol. 3, no. 1, pp. 45-59, 2025.

[3] M. Riaz, et al., "Overcoming Algorithmic Bias and Interpretability Challenges in Precision Oncology," *Lancet Digit. 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," *Med. Image Anal.*, 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, et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.

[7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv.*

APPENDIX

List of Publications and Conferences-

The following publication has been accepted during this thesis work:

Title: MACHINE LEARNING DRIVEN FRAMEWORK FOR EARLY BREAST CANCER DETECTION USING CELLULAR MORPHOMETRIC FEATURES

Authors: 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.

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

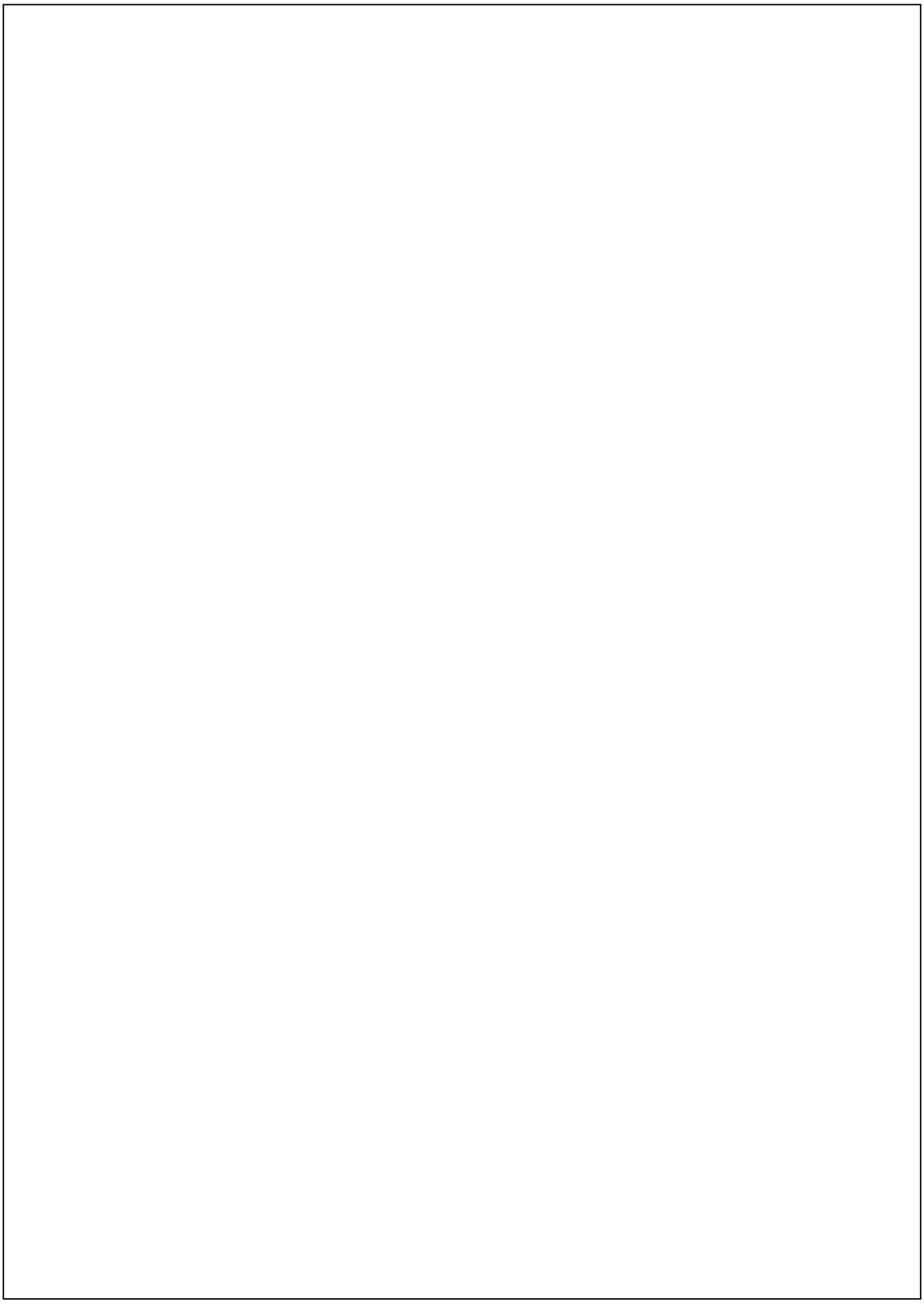
Rifah Ansari

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	K.Janani	Dr.Chirsto Ananth	Dr.Akhatov Akmal Rustamovich
Principal & Conference Chair	CEO, OSIET	Professor, Samarkand State University, Uzbekistan	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 **"MACHINE LEARNING DRIVEN FRAMEWORK FOR EARLY BREAST CANCER DETECTION USING CELLULAR MORPHPMETRIC FEATURES"**

Total Pages , Name of the Scholar **RIFAH ANSARI (24/MSCBIO/34)**

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: **10 %** Total Word Count: **8559**

C T

Jr

Quick Submit
Quick Submit
Marwadi University

Document Details

Submission ID
trn:oid:::1:3514684481

Submission Date
Mar 23, 2026, 12:56 PM GMT+5:30

Download Date
Mar 23, 2026, 12:58 PM GMT+5:30

File Name
jr.docx.pdf

File Size
372.5 KB

12 Pages
2,313 Words
13,504 Characters

turnitin Page 1 of 14 - Cover Page Submission ID trn:oid:::1:3514684481

*% 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 (i.e., our AI models may produce either false positive results or false negative results), 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.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Document Details

Submission ID
trn:oid:::3618:132448481

Submission Date
Mar 22, 2026, 11:38 AM GMT+5:30

Download Date
Mar 22, 2026, 11:41 AM GMT+5:30

File Name
Work.pdf

File Size
403.9 KB

7 Pages

2,409 Words

14,341 Characters

turnitin Page 1 of 11 - Cover Page Submission ID trn:oid:::3618:132448481

turnitin Page 2 of 11 - Integrity Overview Submission ID trn:oid:::3618:132448481

10% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

22 Not Cited or Quoted 10%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
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
5% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

turnitin Page 2 of 11 - Integrity Overview Submission ID trn:oid:::3618:132448481

turnitin Page 3 of 11 - Integrity Overview Submission ID trn:oid:::3618:132448481

Match Groups

22 Not Cited or Quoted 10%

Top Sources

6% Internet sources

Machine Learning-Driven Framework for Early Cancer Detection Using Cellular Morphometric Features

ORIGINALITY REPORT

16%	13%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Delhi Technological University Student Paper	7%
2	dspace.dtu.ac.in:8080 Internet Source	2%
3	"Proceedings of 3rd International Conference on Smart Computing and Cyber Security", Springer Science and Business Media LLC, 2024 Publication	1%
4	www.mdpi.com Internet Source	1%
5	www.techscience.com Internet Source	1%
6	ir.vistas.ac.in Internet Source	<1%
7	www.ijraset.com Internet Source	<1%
8	english.cjebm.com Internet Source	<1%
9	www.coursehero.com Internet Source	<1%

10	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Methods in Science and Technology - Volume 2", CRC Press, 2026 Publication	<1 %
11	Vasiliki Pantoula, Vasileios Mandikas, Tryfon Daras. "Enhanced Assumption-Aware Linear Discriminant Analysis for the Wisconsin Breast Cancer Dataset: A Guide to Dimensionality Reduction and Prediction with Performance Comparable to Machine Learning Methods", AppliedMath, 2026 Publication	<1 %
12	d197for5662m48.cloudfront.net Internet Source	<1 %
13	jbc.bj.uj.edu.pl Internet Source	<1 %
14	ir.lib.nchu.edu.tw Internet Source	<1 %
15	pmc.ncbi.nlm.nih.gov Internet Source	<1 %

Exclude quotes	Off	Exclude matches	< 14 words
Exclude bibliography	Off		