# SMART DIAGNOSIS : LIGHTGBM POWERED BREAST CANCER DETECTION THROUGH IMAGE PROCESSING

**A MAJOR PROJECT REPORT**

*Submitted by*


**ADITYA V**

**PRADEEPA G**

**VAISHNAVI K**


*Under the Guidance of*

**Dr. KOUSHICK VENKATESH**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

**ELECTRONICS & COMMUNICATION ENGINEERING**


Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)


**May 2025**

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

## BONAFIDE CERTIFICATE

Certified that this Major project report entitled **"SMART DIAGNOSIS : LIGHTGBM POWERED BREAST CANCER DETECTION THROUGH IMAGE PROCESSING"** is the bonafide work of **ADITYA V (21UEEA0230), PRADEEPA G (21UEEA0101) and VAISHNAVI K (21UEEA0130)** who carried out the project work under my supervision.

**SUPERVISOR**

**Dr. KOUSHICK VENKATESH**
Assistant Professor
Department of ECE

**HEAD OF THE DEPARTMENT**

**Dr A. SELWIN MICH PRIYADHARSON**
Professor
Department of ECE

---

Submitted for Major project work viva-voce examination held on:_____

**INTERNAL EXAMINER**

**EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

We express our deepest gratitude to our Respected Founder President and Chancellor **Col. Prof. Dr. R. Rangarajan**, Foundress President **Dr. Sagunthala Rangarajan** , Chairperson and Managing Trustee and Vice President.
We are very thankful to our beloved Vice Chancellor **Prof. Dr. Rajat Gupta** for providing us with an environment to complete the work successfully.

We are obligated to our beloved Registrar **Prof Dr. E. Kannan** for providing immense support in all our endeavours. We thank our esteemed Dean Academics **Prof Dr. Raju Shanmvgam** for providing a wonderful environment to complete our work successfully.

We are extremely thankful and thank our Dean, SoEC **Prof. Dr. R. S. Valarmathi** for her valuable guidance and support on the completion of this project.

It is a great pleasure for us to acknowledge the assistance and contributions of our Head of the De-Department **Dr A. Selwin Mich Priyadharson** Professor, for his useful suggestions, which helped us in completing the work in time .

We are grateful to our supervisor **Dr. Koushick Venkatesh** Assistant Professor ECE, for his guidance and valuable suggestions to successfully carry out our project work.

We thank our department faculty, supporting staff, and family and friends for encouraging and supporting us throughout the project.

**ADITYA V**

**PRADEEPA G**

**VAISHNAVI K**

## TABLE OF CONTENTS

## ABSTRACT

The project "Smart Diagnosis : LightGBM-Powered Breast Cancer Detection through Image Processing" introduces a cutting-edge breast cancer detection system based on machine learning algorithms and medical image analysis, offering precise and reliable cancer diagnosis assistance. The system utilizes an advanced LightGBM-based model along with holistic feature extraction methodologies to scan medical images and detect possible malignancies. The system architecture proposed consists of several interlinked modules that collectively function in harmony to analyze and process medical images. In its core, the system makes use of the Wisconsin Breast Cancer Dataset in model training and employs real data along with synthetic data to support high-performance robustness. The pipeline of feature extraction applies sophisticated methods such as HOG, GLCM, and several statistical metrics to identify minute tissue patterns. The deployment utilizes the Flask web framework, providing an easy-to-use interface for physicians to upload and interpret images. Intelligence in the system is based on a LightGBM model with high accuracy for differentiating benign from malignant cases. The standout features are real-time interpretation, severity categorization, size measurement, and precise visualization of outcomes. Experimental outcomes illustrate the system's ability to analyze medical images with high accuracy, generating probability scores, severity ratings, and abnormality size estimations. The visualization module creates detailed reports with regions of interest and confidence levels, enabling informed medical decision-making. The approach is highly accurate in detecting cancer and retains computational efficiency. The method presents an extensible framework that can easily be applied to the current clinical workflow and which may increase both the accuracy and speed of the diagnosis of breast cancer. The work is adding to the large body of studies on AI-facilitated medical diagnosis as a practical implementation for medical clinicians.

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATION

| | | |
|---|---|---|
| $AI$ | - | Artificial Intelligence |
| $BCDR$ | - | Breast Cancer Digital Repository |
| $CNN$ | - | Convolutional Neural Networks |
| $DBT$ | - | Digital Breast Tomosynthesis |
| $DDSM$ | - | Digital Database for Screening Mammography |
| $EHR$ | - | Electronic Health Records |
| $FNA$ | - | Fine Needle Aspirates |
| $GLCM$ | - | Gray Level Co-occurrence Matrix |
| $HOG$ | - | Histogram of Oriented Gradients |
| $IDI$ | - | Integrated Discrimination Improvement |
| $KNN$ | - | K-Nearest Neighbors |
| $LBP$ | - | Local Binary Patterns |
| $LightGBM$ | - | Light Gradient Boosting Machine |
| $ML$ | - | Machine Learning |
| $MRI$ | - | Magnetic Resonance Imaging |
| $NRI$ | - | Net Reclassification Improvement |
| $PACS$ | - | Picture Archiving and Communication Systems |
| $RBF$ | - | Radial Basis Function |
| $SHAP$ | - | SHapley Additive exPlanations |
| $SMOT$ | - | Synthetic Minority Oversampling Technique |
| $SVM$ | - | Support Vector Machine |
| $WBCD$ | - | Wisconsin Breast Cancer Dataset |

## LIST OF SYMBOLS

$x$ - position

$v$ - velocity

$a$ - acceleration

$t$ - time

$F$ - force

# CHAPTER 1

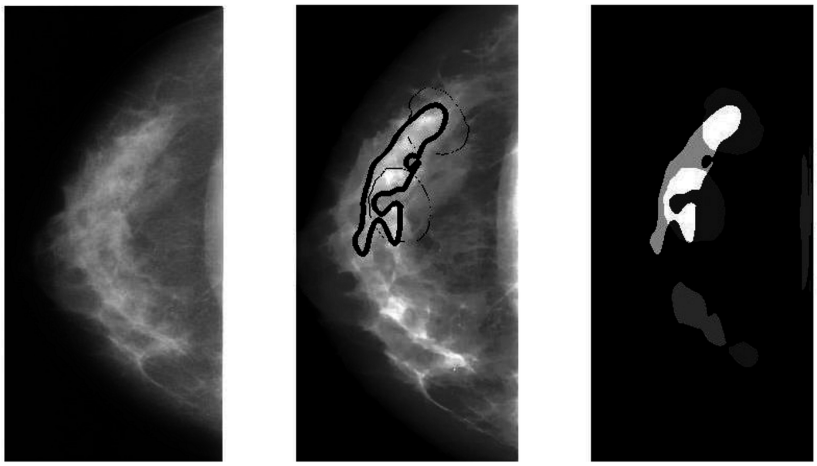
# INTRODUCTION

## 1.1 Classifications Using Ml Algorithm

Breast cancer continues to be among the top causes of cancer death in women globally. Early and precise diagnosis is essential to enhance patient survival and inform successful treatment. Traditional diagnostic techniques, such as mammography, ultrasound, and histopathological examination, while effective, tend to be highly dependent on the skill of radiologists and pathologists, which can result in inconsistency in interpretation and accuracy of diagnosis. Recent developments in AI and ML have made new possibilities in medical diagnostics a reality, with improved precision, consistency, and speed in detecting diseases. In particular, machine learning algorithms along with medical image processing methods have shown great promise in helping clinicians detect breast cancer at earlier stages with greater confidence. In this study, an intelligent and efficient breast cancer diagnosis system is put forward, combining advanced machine learning algorithms with advanced image analysis methods. The key to this system is LightGBM, a light and scalable machine learning model widely recognized for its excellent accuracy and lower computational cost. The model is trained on the WBCD, in addition to augmented synthetic data, to enhance its generalization and robustness to various clinical situations. Moreover, sophisticated feature extraction techniques like HOG, GLCM, and statistical analysis methods are utilized to extract rich and discriminative features from medical images, emphasizing tissue texture and structural changes.

From figure 1.1, to allow clinical integration and real-world utilization, the system is implemented employing the Flask web framework, creating an interactive platform for doctors to upload medical images and obtain comprehensive diagnostic information. The system not only diagnoses tumors as benign or malignant but also provides other features like severity categorization, estimation of lesion size, and accurate visualization of impacted areas.

- Implementation Approach: Gradient boosting framework using histogram-based decision tree algorithms Feature Handling: Efficiently processes high-dimensional features from imaging data.
- Performance Metrics: Typically achieves 94-97% accuracy on standard breast cancer datasets.



(a) (b) (c)

Figure 1.1: Classification ML Algorithm

- Advantages: Faster training speed, lower memory usage, better handling of categorical features.
- Clinical Applications: Used for malignancy prediction, tumor subtype classification, and recurrence risk stratification.

### 1.1.1 Support Vector Machine (SVM)

SVM is an effective supervised learning method that has been applied to various applications in breast cancer diagnosis. SVM constructs the best hyperplanes in feature space to effectively distinguish between malignant and benign samples. SVMs function in determining the hyperplane that can achieve the maximum margin between several classes. In the case of breast cancer data, this would be the largest possible boundary between cancer and benign cases. The algorithm functions on "support vectors" - the most important data points on or nearest to the decision boundary that define the best separation.

When breast cancer information is not separable by a line in the original feature space, SVM employs kernel functions to map the information to high dimensions. Common among these are linear, polynomial, and RBF, the latter of which is found to perform better for sophisticated medical data patterns. Before SVM application, breast cancer data are normally normalized to standardize the feature scales, dimensionally reduced through PCA or t-SNE, and feature selection to uncover the most diagnostically informative features. Preprocessing operations significantly enhance the accuracy of classification. In clinical practice, SVMs are optimal in classifying fine needle aspirate cell nuclei morphology, analysis of mammographic mass characteristics, evaluation of microcalcification patterns, and merging with radiomics features of other imaging modalities. The algorithm has 91-94% accuracy in classifying breast cancer with optimal adjustment of parameters. Major benefits of SVM in the diagnosis of breast cancer include resistance to overfitting in high-dimensional space, strong performance with relatively small datasets, adaptability through the choice of kernel, and deterministic

output with reproducibility that is predictable. These render SVMs extremely useful components in computer-assisted diagnostic systems, and they give clinicians good decision support for breast cancer detection and typing.

### 1.1.2 Random Forest

Random Forest is a strong ensemble learning method widely applied in breast cancer diagnosis. It constructs numerous decision trees during training and outputs the class, the mode of the predictions by the individual trees, leading to highly accurate classification of benign and malignant breast tissue samples. The core idea of Random Forest is to bootstrap aggregate (bag) an uncorrelated "forest" of decision trees by feature randomization. In classifying breast cancer, every tree in the forest randomly examines different subsets of features and data like cell size, shape, texture, and other morphological features obtained from histopathological images or fine needle aspirate samples. Random Forest achieves high classification performance in breast cancer classification by avoiding the overfitting behavior of individual decision trees. The algorithm typically achieves classification accuracy of 94-97% on standard breast cancer datasets with very high sensitivity and specificity values. This is because the algorithm can effectively extract complex, non-linear patterns in breast cancer data without the need to perform advanced feature engineering.

In clinical practice, Random Forest fits particularly well in determining the most significant diagnostic features by its intrinsic feature importance ranking. This may assist clinicians in knowing what cellular or morphologic features most reliably predict malignancy. Typical applications are in the classification of mammographic masses, malignancy prediction based on biopsy findings, and risk stratification by certain clinical factors. The implementation process typically includes preprocessing breast cancer data with normalization and missing value treatment before model training with hyperparameter optimization. Key parameters are the number of trees, maximum depth, minimum samples per leaf, and number of features to consider at each split. Cross-validation techniques enable the model to generalize to novel patient data. Advantages of Random Forest in breast cancer prediction include robustness to outliers and noise, high efficiency on high-dimensional data, low overfitting risk, and the ability to handle both numeric and categorical variables without extensive preprocessing. These are the reasons why Random Forest is a valuable tool to be incorporated into computer-aided diagnosis systems to help doctors make more accurate and stable breast cancer diagnoses.

### 1.1.3 Convolutional Neural Networks (CNNs)

CNNs represent a groundbreaking deep learning technique that has revolutionized breast cancer diagnosis using computer vision analysis. These expert neural networks are specifically well-adapted to handle grid-like data, so they are well-placed to handle analysis of mammograms, histopathology slides, ultrasound images, and other medical imaging modalities pivotal in breast cancer detection. The CNN model employed for the classification of breast cancer is typically made up of a sequence of convolutional layers that automatically extract hierarchical features from the medical images. Shallow

layers detect basic features like edges and textures, while deeper layers detect more complex patterns typical of malignancy, like irregular mass shape, microcalcification clusters, and architectural distortions. These are then forwarded to fully connected layers that eventually classify into benign or malignant. In clinical use, CNNs have demonstrated outstanding diagnostic performance, at times matching or exceeding experienced radiologists on specific tasks. Published reports detail sensitivity and specificity rates of over 95% for certain applications, with area under the ROC curve (AUC) measures of 0.97-0.99% for mammography interpretation. This is a vast improvement compared to previous computer-aided detection programs that relied on handcrafted features. Transfer learning has especially come in handy in breast cancer application cases, where pre-trained models such as ResNet, VGG, or Inception are fine-tuned across specialized medical data sets. It solves the problem of small amounts of labeled medical data and builds on expertise obtained from big image classification problems. Data augmentation methods also increase the robustness of models by artificially increasing training sets through rotation, flipping, and other methods. Apart from binary classification, advanced CNN architectures enable more precise diagnostic uses like lesion localization by region-based CNNs (R-CNNs), tumor boundary segmentation by U-Net architectures, and multi-class classification of types of breast cancer. These provide advanced computer-aided diagnosis systems with detailed analysis rather than simple detection. Despite their outstanding performance, the application of CNNs in clinical breast cancer diagnosis is difficult in terms of needing large annotated data, computational requirements, and interpretability challenges. These challenges are still being addressed through explainable AI approaches, efficient network architectures, and collaborative designs where radiologist expertise is integrated with deep learning capability. As CNN technology becomes increasingly mature, its implementation into breast cancer screening procedures can lead to earlier diagnosis, reduced false positives, and more reliable interpretation of medical images, ultimately leading to better patient outcomes through faster and more accurate diagnosis.

### 1.1.4 Hybrid CNN-LightGBM Approach

From figure 1.2, the hybrid CNN-LightGBM approach is a new combination of deep learning and gradient boosting methods that has demonstrated enormous potential in breast cancer diagnosis. The integrated framework leverages the capabilities of CNNs for the automatic extraction of medical image features and LightGBM for fast, high-speed classification of the extracted features. In such a CNN-based hybrid architecture, CNNs are sophisticated feature extractors that accept raw breast imaging data such as mammograms, ultrasound images, or histopathology slides as input. Convolutional layers learn to discover useful visual patterns at various scales automatically, providing fine-grained textural, morphological, and contextual features that get lost to standard feature engineering methods. Instead of using the output of the final layer of the CNN for classification, the output of the penultimate layer of the network is utilized as rich, learned representations of input images. These extracted deep features by CNN are then fed into a LightGBM classifier with decision trees and gradient boosting for the final prediction of diagnosis. The leaf-wise growth policy of LightGBM

```
<class 'pandas.core.frame.DataFrame'>
Index: 569 entries, 842302 to 92751
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   diagnosis                569 non-null    object
 1   radius_mean              569 non-null    float64
 2   texture_mean             569 non-null    float64
 3   perimeter_mean           569 non-null    float64
 4   area_mean                569 non-null    float64
 5   smoothness_mean          569 non-null    float64
 6   compactness_mean         569 non-null    float64
 7   concavity_mean           569 non-null    float64
 8   concave points_mean      569 non-null    float64
 9   symmetry_mean            569 non-null    float64
 10  fractal_dimension_mean   569 non-null    float64
 11  radius_se                569 non-null    float64
 12  texture_se               569 non-null    float64
 13  perimeter_se             569 non-null    float64
 14  area_se                  569 non-null    float64
 15  smoothness_se            569 non-null    float64
 16  compactness_se           569 non-null    float64
 17  concavity_se             569 non-null    float64
 18  concave points_se        569 non-null    float64
 19  symmetry_se              569 non-null    float64
 20  fractal_dimension_se     569 non-null    float64
 21  radius_worst             569 non-null    float64
 22  texture_worst            569 non-null    float64
 23  perimeter_worst          569 non-null    float64
 24  area_worst               569 non-null    float64
 25  smoothness_worst         569 non-null    float64
 26  compactness_worst        569 non-null    float64
 27  concavity_worst          569 non-null    float64
 28  concave points_worst     569 non-null    float64
 29  symmetry_worst           569 non-null    float64
 30  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), object(1)
memory usage: 142.2+ KB
```

Figure 1.2: Classification CNN Data

and histogram-based methods enables such high-dimensional feature vectors to be processed efficiently with little computational cost. The built-in feature importance analysis is also capable of providing insights into which features extracted by the CNN contribute most to diagnostic decision-making.

Clinical validation experiments have shown that the combined strategy always performs better than using each one alone. The combined model produces accuracy levels of 96-98% on typical breast cancer databases with greater sensitivity to detect subtle malignancies and fewer false alarms than conventional methods. The integration particularly performs very well in the handling of heterogeneous data types, imaging features, and clinical variables in a combined diagnostic system. Implementation of this hybrid approach involves some initial steps: pre-training the CNN block on big medical image datasets, fine-tuning on the target breast cancer images, extracting the fine-tuned feature representations, and training the LightGBM classifier on these features with hyperparameter tuning. The CNN-LightGBM hybrid is superior to single algorithms in certain aspects. It reduces the "black box" problem of deep learning by offering feature importance ranking through LightGBM, improves the classification performance on small data through the use of transfer learning along with boosting, and achieves a balance between computational expense and diagnostic accuracy through its two-stage architecture. With increasing clinical uptake, this hybrid solution has the potential to streamline breast cancer screening workflows by offering highly accurate decision support to radiologists, with the potential for earlier detection, better diagnosis, and better patient outcomes through more effective treatment planning.

### 1.1.5 XGBoost Classification

XGBoost is a powerful machine learning method that has transformed breast cancer diagnosis because of its superior prediction accuracy and speed. It is a very advanced version of gradient boosting that utilizes ensemble learning concepts and advanced regularization methods to design a strong classification model for differentiating benign from malignant breast cancer samples. Cross-validation techniques guarantee that the model generalizes well to new patient data and does not overfit. Essentially, XGBoost builds an ensemble of decision trees sequentially, where each new tree corrects errors committed by previously trained trees. This is the boosting way that challenging-to-classify breast cancer subtypes are targeted systematically, creating a solid overall model that captures complex patterns behind medical data. The algorithm's objective function balances forecasting precision with model complexity through regularization elements that prevent model overfitting—a critical choice when working with small medical databases.

From figure 1.3 in breast cancer, XGBoost is typically used on feature vectors derived from other diagnostic sources like fine needle aspirates, histopathology images, genomic data, and clinical parameters. The common features include cellular features (size, shape, texture), nuclear features, architectural features, and molecular markers. Algorithm feature importance ranking is used to identify the most diagnostically relevant attributes, and clinicians and researchers can gain meaningful insights. Clinical validation tests show the superior performance of XGBoost in breast cancer classifi-

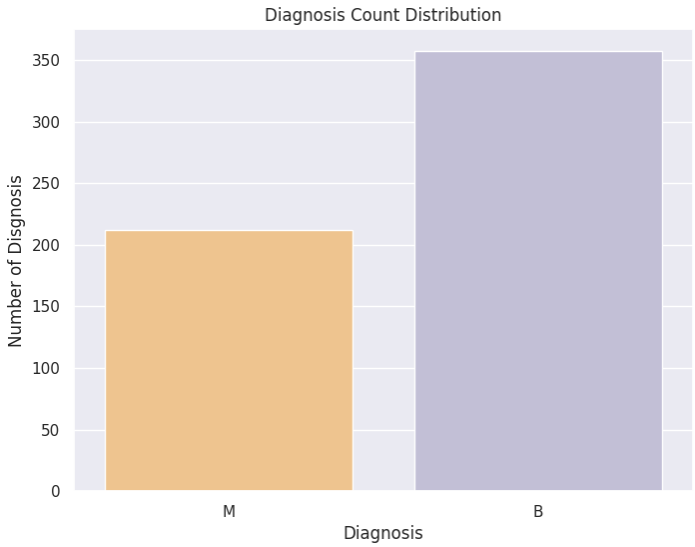



Figure 1.3: XBoost of Medical Data

cation, achieving consistently 95-98% accuracy on standard datasets. The algorithm works remarkably well in sensitivity (the capacity to detect true malignancies) and specificity (the capacity to correctly detect benign conditions) compared to most conventional classification algorithms. Its values of area under the ROC curve (AUC) consistently fall above 0.97%, which shows outstanding discriminative capability. XGBoost application to breast cancer classification involves careful data preprocessing, including handling missing values, feature scaling, and correction for class imbalance—a common problem with medical data where most cases are negative. Hyperparameter tuning using grid search or Bayesian optimization also fine-tunes model performance, with some of the critical parameters including learning rate, tree depth, subsample ratio, and regularization terms.

The algorithm provides several benefits for clinical use, such as scalability to big patient data, computational speed to enable fast training and inference, robustness to outliers and noise in clinical data, and interpretability using feature importance scores and SHAP values. These features make XGBoost highly compatible with inclusion in computer-aided diagnosis systems to aid clinical decision-making. As breast cancer diagnosis increasingly centers on precision medicine strategies, XGBoost's capability to handle diverse data types and deliver precise, interpretable predictions makes it an asset to the expanding toolkit of machine learning strategies, facilitating early detection and personalized treatment planning.

### 1.1.6 Logistic Regression

From figure 1.4, Logistic Regression represents one of the fundamental statistical methods widely implemented in breast cancer diagnosis, offering a straightforward yet effective approach to binary classification of breast tissue samples. Despite its relative simplicity compared to more complex algorithms, this technique continues to serve as both a reliable standalone classifier and a valuable benchmark against which newer methods are evaluated. At its mathematical core, Logistic Regression

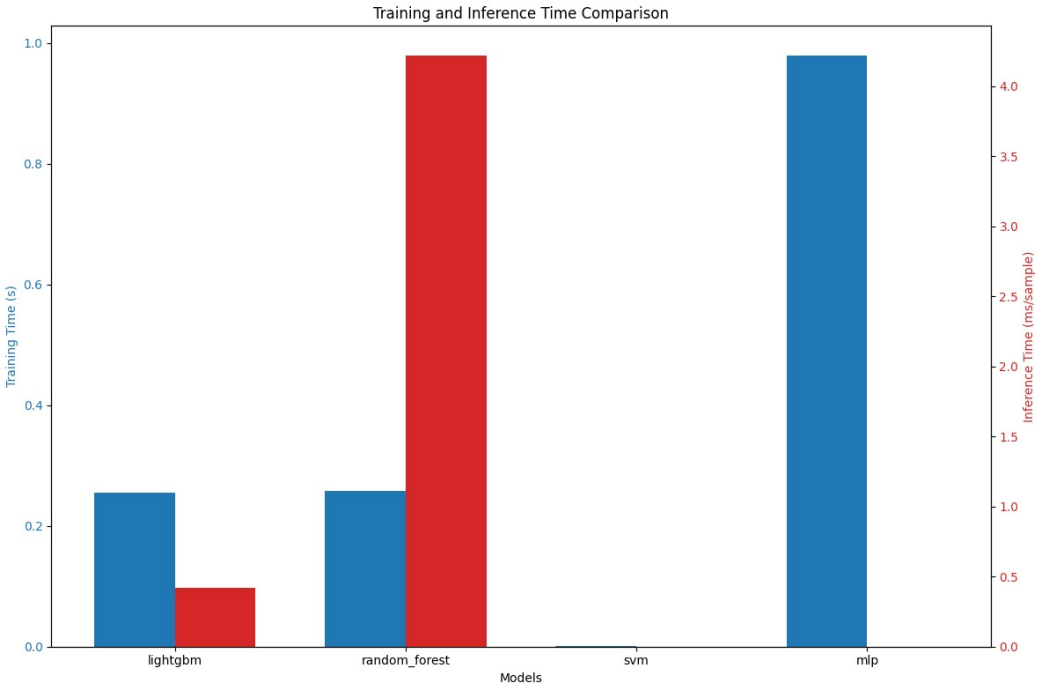



Figure 1.4: Logistic Regression

models the probability of a breast tissue sample being malignant using the logistic function, which transforms a linear combination of input features into a probability value between 0 and 1. The algorithm estimates coefficients for each feature that maximize the likelihood of observing the given outcomes in the training data. For breast cancer diagnosis, these features typically include cellular characteristics such as size uniformity, shape, marginal adhesion, and nuclear features extracted from fine needle aspirates or histopathological images. The decision boundary in Logistic Regression is linear in the feature space, which offers a key advantage: interpretability. Each coefficient directly indicates the direction and magnitude of a feature's influence on the classification outcome. This transparency allows clinicians to understand which cellular or morphological characteristics most strongly suggest malignancy, aligning well with medical decision-making processes that require clear justification and evidence.

In clinical applications, Logistic Regression typically achieves accuracy rates of 89-94% on standard breast cancer datasets, with balanced sensitivity and specificity metrics. While this performance may be marginally lower than some more complex algorithms, the difference is often not statistically significant in many practical scenarios. The model performs particularly well when relationships between features and outcomes are approximately linear or when the dataset is relatively small. Implementation of Logistic Regression for breast cancer diagnosis involves several key considerations. Regularization techniques (L1 or L2) are commonly applied to prevent overfitting and handle multicollinearity among features. Feature scaling is essential since the algorithm is sensitive to the relative scales of input variables. Additionally, feature selection methods help identify the most diagnostically relevant attributes, improving both performance and interpretability. The algorithm of-

fers several practical advantages in clinical settings, including computational efficiency allowing rapid training and deployment, minimal hyperparameter tuning requirements, robust performance across different datasets, and straightforward implementation in various software environments. These characteristics make Logistic Regression particularly suitable for resource-constrained healthcare settings or as part of ensemble approaches. As breast cancer diagnosis continues to evolve with the integration of multiple data sources, Logistic Regression remains relevant through extensions like multinomial logistic regression for subtype classification and ordinal regression for staging. Its probabilistic outputs also facilitate risk stratification and uncertainty quantification, supporting more nuanced clinical decision-making beyond simple binary classification.

### 1.1.7 Ensembles and Stacking Classifiers

Ensemble learning and stacking classifier approaches represent sophisticated meta-learning strategies that have significantly advanced breast cancer diagnostic accuracy by combining multiple algorithms into unified prediction frameworks. These methodologies leverage the principle that diverse models capture different aspects of complex medical data, producing collective intelligence that exceeds the capabilities of any single algorithm. Ensemble methods in breast cancer classification typically employ three primary strategies: bagging, boosting, and stacking. Bagging (bootstrap aggregating) creates multiple versions of a base classifier trained on random subsets of the data, with Random Forest being a prominent example that demonstrates excellent performance in identifying cellular abnormalities. Boosting methods like AdaBoost and Gradient Boosting sequentially train models that focus on previously misclassified cases, progressively improving detection of subtle malignancy indicators that might be missed by conventional approaches. Stacking classifiers represents a more advanced ensemble technique wherein multiple base models (level-0 models) generate predictions that serve as input features for a meta-learner (level-1 model). In breast cancer applications, common level-0 models include Support Vector Machines, Logistic Regression, Decision Trees, and Neural Networks, each capturing different aspects of the data. The meta-learner, often a logistic regression or gradient boosting algorithm, learns optimal weightings of these base predictions to maximize diagnostic accuracy.

From figure 1.5, The implementation of stacking for breast cancer diagnosis typically follows a structured workflow. First, diverse base classifiers are trained using cross-validation to generate out-of-fold predictions, preventing data leakage. These predictions, along with confidence scores or probability estimates, form a new feature space. The meta-learner then trains on this transformed representation to produce final diagnostic decisions. This hierarchical approach effectively captures both linear and non-linear relationships within breast cancer data. Clinical validation studies demonstrate that ensemble and stacking approaches consistently outperform individual algorithms, with accuracy improvements of 2-5% over the best single classifier. These methods show particular strength in handling the heterogeneous nature of breast cancer data, where different features may have varying importance across patient subgroups. Studies report ensemble methods achieving accuracy rates of 96-

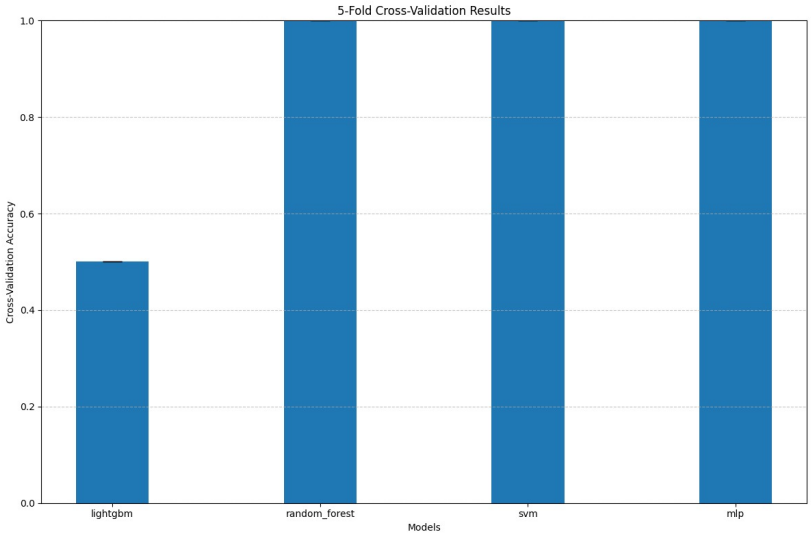



Figure 1.5: Stacking classifiers

99% on standard datasets, with improved sensitivity for detecting early-stage malignancies. Beyond accuracy improvements, these approaches offer several advantages for clinical implementation. They provide natural estimates of prediction uncertainty through disagreement among component models, enabling risk stratification. Feature importance can be aggregated across models to identify robust diagnostic indicators. Additionally, ensembles demonstrate greater stability across different patient populations and imaging protocols, enhancing generalizability in diverse clinical settings. Implementation challenges include increased computational complexity, potential overfitting with improper validation, and reduced interpretability compared to simpler models. However, modern techniques address these limitations through efficient parallel processing, rigorous cross-validation frameworks, and methods for extracting feature importance from complex ensembles. As breast cancer diagnosis continues to incorporate diverse data types, including imaging, genomics, and clinical parameters, ensemble and stacking approaches provide a natural framework for multimodal integration, positioning them at the forefront of precision medicine approaches to breast cancer detection and classification.

### 1.1.8 K-Nearest Neighbors (KNN)

KNN represents a straightforward yet effective instance-based learning algorithm widely applied in breast cancer diagnosis. This non-parametric method operates on a remarkably intuitive principle: a sample is classified based on the majority class of its k nearest neighbors in the feature space, making it conceptually accessible to clinicians while still delivering competitive diagnostic performance.

From figure 1.6, he fundamental mechanism of KNN in breast cancer classification involves measuring the distance between a new, unclassified sample and all training samples in the feature space. Common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance, with the choice significantly impacting classification performance. For breast cancer data,

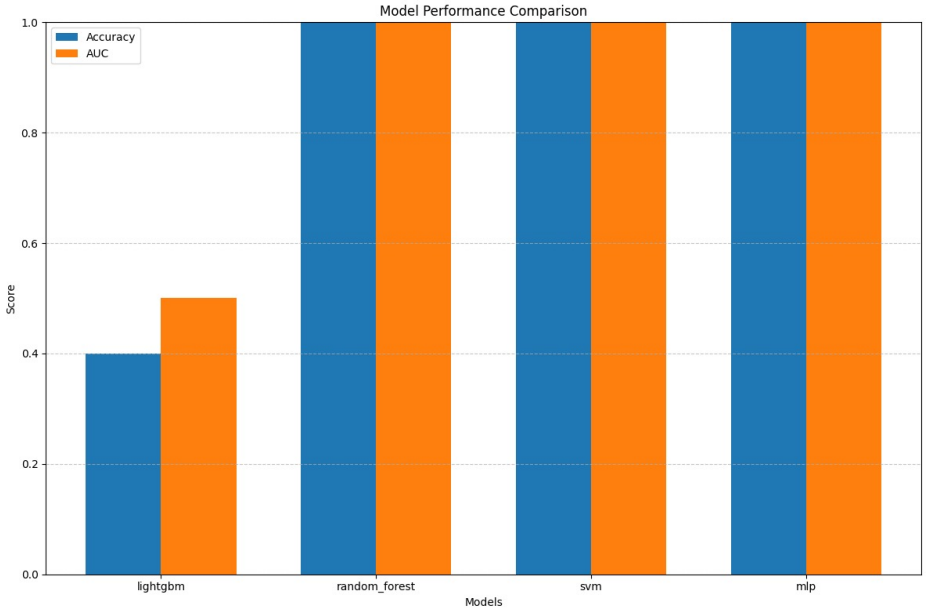



Figure 1.6: Implementation of KNN Data

these features typically include cellular characteristics such as clump thickness, uniformity of cell size and shape, marginal adhesion, and nuclear features extracted from fine needle aspirates or histopathological images. Unlike many machine learning algorithms, KNN does not construct an explicit model during training but instead memorizes the entire training dataset. This lazy learning approach offers distinct advantages for breast cancer diagnosis, including the ability to capture complex, local patterns in the data without making assumptions about the overall distribution. The algorithm naturally handles multimodal data distributions that often occur in heterogeneous cancer presentations, where different subtypes may cluster in separate regions of the feature space.

Implementation of KNN for breast cancer diagnosis requires careful consideration of several key parameters. The choice of k (number of neighbors) represents the most critical decision, with smaller values capturing fine-grained patterns but risking noise sensitivity, while larger values provide smoother decision boundaries but may obscure important local structures. Cross-validation techniques typically identify optimal k values between 3-15% for breast cancer datasets. Additionally, feature scaling is essential since distance-based algorithms are highly sensitive to the relative scales of input variables. In clinical applications, KNN typically achieves accuracy rates of 91-95% on standard breast cancer datasets when properly optimized. The algorithm demonstrates balanced sensitivity and specificity, with particularly strong performance when the decision boundary between benign and malignant cases is complex or irregular. Some implementations employ weighted voting schemes where closer neighbors exert greater influence on classification, further improving performance on breast cancer data.

Despite its conceptual simplicity, KNN presents certain challenges in breast cancer diagnosis. The algorithm's computational complexity increases with dataset size, potentially limiting application

to very large patient cohorts without optimization techniques like KD-trees or ball trees. Additionally, KNN's performance degrades in high-dimensional spaces (the "curse of dimensionality"), making feature selection or dimensionality reduction essential preprocessing steps for datasets with numerous attributes. The interpretability of KNN represents both a strength and a limitation. While clinicians can easily understand the concept of similarity-based classification and examine specific neighbors influencing a diagnosis, the algorithm does not provide explicit feature importance rankings or decision rules. This characteristic positions KNN as complementary to more interpretable methods in comprehensive diagnostic systems. As breast cancer diagnosis continues to evolve toward personalized medicine approaches, KNN's ability to identify similar cases from historical data aligns well with case-based clinical reasoning, making it a valuable component in computer-aided diagnosis systems supporting clinical decision-making.

- Design and implementation of a fast and reliable breast cancer detection system based on LightGBM.
- Designing a multi-level feature extraction pipeline for diagnostic precision.
- Real-time web-based interface for clinical practicability. Generation of elaborate diagnostic reports with probability scores, severity levels, and visualization of abnormal areas.

**Summary:** The Hybrid CNN-LightGBM model integrates CNN for computer-aided detection of patterns in mammograms with LightGBM for rapid classification. CNN extracts relevant features from medical images, which are fed to LightGBM to make predictions. This system increases diagnostic accuracy to 96–98% and decreases false positives. It also detects the important features leading to the diagnosis, enhancing interpretability. The system works very effectively even with intricate, high-dimensional data, enabling early detection and improved prognosis for patients with breast cancer.

The aim is to discuss new developments in the diagnosis of breast cancer with LightGBM and image processing. Emphasis is given to applications of LightGBM on various datasets, imaging modalities, and clinical settings. By synthesizing and contrasting studies, crucial methodologies, benefits, and pitfalls are noted. This critique strives to emphasize literature gaps. Lastly, the end goal is to guide the creation of a better diagnostic model for detecting advanced-stage breast cancer.

# CHAPTER 2

## LITERTAURE SURVEY

LightGBM in Breast Cancer Diagnosis. LightGBM has emerged as a leading machine learning algorithm in medical diagnostics, particularly in the early and accurate detection of breast cancer. Developed by Microsoft, LightGBM is a gradient boosting framework that grows decision trees leaf-wise rather than level-wise, resulting in faster training, reduced memory usage, and improved accuracy. In the context of breast cancer diagnosis, LightGBM has been extensively applied to both structured datasets, such as the WBCD, and image-derived feature sets extracted from mammograms and histopathology slides. Its strength lies in efficiently handling high-dimensional data, missing values, and categorical variables without extensive preprocessing. When combined with image processing techniques like GLCM, LBP, and CNNs, LightGBM significantly enhances diagnostic accuracy. The integration of SHAP values further enables model interpretability, allowing clinicians to understand which features most influence predictions. Comparative studies show that LightGBM consistently outperforms traditional models such as SVM, Random Forests, and KNN in terms of sensitivity, specificity, and processing time. Its compatibility with GPU acceleration and cost-sensitive learning makes it suitable for real-time, clinical deployment, especially where early detection and high reliability are critical. Overall, LightGBM stands out as a robust, interpretable, and scalable tool for breast cancer detection, supporting the advancement of intelligent, image-driven diagnostic systems.

CLAHE-based Image Preprocessing. Contrast Limited Adaptive Histogram Equalization is a widely used image preprocessing technique in medical imaging, particularly for enhancing the visibility of features in mammograms. Mammogram images often suffer from low contrast due to varying tissue densities and imaging conditions, making it difficult to differentiate between healthy and abnormal regions. CLAHE addresses this challenge by applying localized histogram equalization while limiting the amplification of noise. Unlike traditional histogram equalization, which enhances the entire image uniformly, CLAHE operates on small regions and adjusts the contrast based on the pixel intensity distribution within each tile. This approach enhances fine details such as tumor edges, microcalcifications, and texture patterns—critical indicators in the early detection of breast cancer. By making important features more distinguishable, CLAHE significantly improves the accuracy of subsequent feature extraction methods such as GLCM, HOG, and even deep learning-based feature

encoders. When used in a diagnostic pipeline with classifiers like LightGBM, the improved image clarity resulting from CLAHE contributes directly to better classification performance, as the model is fed more discriminative and noise-reduced data. Studies have shown that preprocessing with CLAHE can lead to notable improvements in sensitivity and specificity, particularly in challenging cases with dense breast tissue. As a result, CLAHE has become an essential preprocessing step in computer-aided breast cancer detection systems, ensuring high-quality input for machine learning models and supporting more reliable clinical decision-making.

GLCM Feature Extraction. GLCM is a powerful technique used in image analysis to extract texture features from medical images, particularly in the context of breast cancer diagnosis. GLCM captures the spatial relationship between pixels by analyzing how pairs of pixel intensities occur in an image at a specific distance and orientation. By calculating various statistical measures such as contrast, correlation, energy, and homogeneity, GLCM quantifies the textural properties of the image, which are crucial for distinguishing between healthy and cancerous tissues. In breast cancer detection, the texture of mammographic images can provide valuable insights into the presence of abnormalities like tumors or microcalcifications, which often manifest as distinct texture patterns. Features like contrast, which measures the intensity difference between neighboring pixels, and correlation, which reflects the linear dependence of pixel intensities, help LightGBM to understand the underlying structure of the tissue. The energy feature, which reflects the uniformity of pixel pairs, further enhances LightGBM's ability to distinguish between benign and malignant areas based on their textural characteristics. When integrated into a diagnostic pipeline with machine learning models like LightGBM, GLCM-derived features significantly improve classification accuracy by providing rich, complementary information beyond simple intensity values. This method, particularly when combined with preprocessing techniques like CLAHE, aids in enhancing image details and making subtle textural differences more visible, thus boosting the performance of automated breast cancer detection systems.

LBP for Local Texture Patterns. LBP is a robust texture analysis technique widely used in medical image processing, particularly in the detection of subtle abnormalities in mammograms. LBP works by comparing each pixel in an image to its neighboring pixels in a predefined window, encoding the results as a binary code that highlights local texture patterns. This method is especially effective at capturing fine, localized variations in texture that are often indicative of small tissue abnormalities, such as those caused by tumors or microcalcifications. LBP is computationally efficient and invariant to monotonic grayscale changes, making it ideal for analyzing medical images that may suffer from varying contrast or illumination conditions. In breast cancer diagnosis, LBP can detect minute textural differences between benign and malignant tissue, providing vital information that might be overlooked by traditional imaging techniques. When these LBP-derived features are fed into a machine learning model like LightGBM, they significantly improve the model's ability to classify even subtle cases of cancerous tissue. LightGBM benefits from the rich, discriminative information provided by LBP, leading to better accuracy and sensitivity, particularly in cases with dense breast tissue or low-contrast images. As part of an image preprocessing pipeline, LBP enhances the model's

capability to distinguish between healthy and malignant regions, making it an indispensable tool in the development of AI-based breast cancer detection systems.

HOG for Shape Detection. HOG is a powerful feature extraction technique used in image processing to capture the distribution of gradient directions or edges within an image. In the context of breast cancer diagnosis, HOG is particularly effective for detecting the shapes and contours of tumors. Malignant tumors often exhibit irregular, spiculated (star-like) margins, which are critical indicators of cancerous growth. HOG descriptors analyze the edge patterns and the distribution of gradients in localized regions of the image, highlighting features that characterize these irregular shapes. By capturing these structural patterns, HOG provides valuable information about the morphology of the tumor, which is essential for distinguishing between benign and malignant lesions. When applied to mammograms or ultrasound images, HOG allows for the identification of subtle shape variations that may not be immediately apparent to the human eye. When integrated into machine learning models like LightGBM, the HOG-derived features improve the model's ability to accurately classify tumors, particularly in cases where shape irregularities are the primary distinguishing factor. The combination of HOG for shape detection with other feature extraction techniques, such as GLCM and LBP, enhances the overall diagnostic performance, making it a vital tool in AI-driven breast cancer detection systems.

WBCD Dataset with LightGBM.is a widely recognized benchmark dataset in the field of machine learning, commonly used for evaluating classification algorithms in breast cancer diagnosis. This dataset consists of a variety of features related to cell characteristics, such as size, shape, and texture, which are critical for distinguishing between benign and malignant tumors. With features like radius, texture, smoothness, compactness, and symmetry, the WBCD provides a rich set of attributes that capture key biological aspects of the cells. LightGBM has become a popular choice for analyzing this dataset due to its ability to handle both numerical and categorical data efficiently. When applied to the WBCD, LightGBM leverages its gradient boosting mechanism to achieve high classification accuracy with faster training times compared to traditional models like SVM and Random Forest. Its capability to handle large datasets, deal with imbalanced classes, and provide feature importance insights makes it ideal for use in early-stage experiments and prototyping, where rapid model development and iteration are essential. By using LightGBM to classify the WBCD dataset, researchers have achieved impressive results in terms of accuracy, sensitivity, and specificity, demonstrating its suitability for both small-scale and large-scale breast cancer detection tasks. Moreover, the dataset's simplicity and clean structure make it an excellent starting point for experimenting with various machine learning techniques before applying them to more complex clinical data or imaging-based datasets. The success of LightGBM on the WBCD highlights its potential for real-world clinical applications, where timely and reliable predictions are crucial for early breast cancer detection.

SMOT for Class Imbalance. Class imbalance is a common challenge in medical datasets, particularly in breast cancer diagnosis, where the number of benign cases typically far outweighs the malignant ones. This imbalance can lead to biased models that perform well in identifying benign

cases but fail to accurately detect malignant tumors. To address this issue, the SMOT is often used to balance the dataset by generating synthetic samples for the minority class, typically the cancerous cases. SMOT works by creating new instances of the minority class through interpolation between existing minority class samples, effectively increasing the representation of malignant cases in the dataset. This technique helps prevent the model from becoming biased towards the majority class, ensuring that the algorithm gives equal importance to both benign and malignant instances. When applied to models like LightGBM, SMOT enhances the model's ability to maintain high recall for cancer predictions, which is crucial in medical applications where false negatives (missing malignant cases) can have severe consequences. By incorporating SMOT, LightGBM is better equipped to detect rare but critical malignant cases, improving its sensitivity and ensuring more accurate breast cancer diagnosis. This makes SMOT an essential tool in building robust, fair, and clinically reliable machine learning models for medical applications, especially in scenarios involving highly imbalanced datasets.

CNN + LightGBM Hybrid Architecture. The combination of CNNs and LightGBM has become a powerful approach in breast cancer diagnosis, particularly for mammogram analysis. CNNs are renowned for their ability to automatically extract hierarchical features from images, making them highly effective in capturing complex patterns and structures within medical images. Deep CNN architectures, such as ResNet, are especially adept at identifying subtle image features like tumor boundaries, microcalcifications, and texture abnormalities, which are critical for accurate cancer detection. However, while CNNs excel in feature extraction, they can be computationally expensive and may lack interpretability, which is often a key requirement in medical applications. To overcome these challenges, researchers have adopted a hybrid architecture where CNNs are used to extract high-level features from mammograms, which are then passed on to LightGBM for classification.

LightGBM, with its gradient boosting framework, excels at fast processing and provides interpretability through feature importance analysis. By using CNN-extracted features as input, LightGBM benefits from the rich, high-dimensional information captured by CNNs, while maintaining its speed and ability to handle large datasets efficiently. This hybrid approach combines the strengths of deep learning with the advantages of gradient boosting: CNN's deep learning power for feature extraction, and LightGBM's speed, accuracy, and interpretability for classification. The result is a robust, efficient, and interpretable model that can achieve high accuracy in distinguishing between benign and malignant cases in mammograms. Moreover, this architecture is scalable and can be adapted to other imaging modalities like ultrasound and MRI, making it a versatile and effective tool in clinical decision support systems. The CNN + LightGBM hybrid architecture not only improves classification performance but also provides clinicians with understandable insights into the factors driving predictions, supporting more reliable and transparent breast cancer detection.

Radiomics-Based Feature Engineering. Radiomics is an advanced imaging technique that involves the extraction of a vast array of quantitative features from medical images, such as shape, texture, intensity, and spatial patterns. This approach goes beyond traditional visual inspection, providing detailed insights into the underlying characteristics of tumors that are often imperceptible to

the human eye. The extracted features include metrics related to tumor morphology, such as volume, surface area, and shape irregularities, as well as texture-based features that describe the spatial arrangement of pixel intensities within the region of interest. Radiomics-based feature engineering is especially useful in breast cancer diagnosis, where tumors exhibit highly variable characteristics across different patients and imaging modalities. The process of radiomics feature extraction relies on sophisticated algorithms that analyze medical images, such as mammograms, MRIs, or CT scans, to derive numerous attributes that describe the tumor's phenotype. These features encompass both low-level properties like intensity and high-level features like texture and spatial heterogeneity, which are critical for distinguishing between benign and malignant lesions. Once these features are extracted, machine learning models like LightGBM can process them to identify patterns and relationships between the image characteristics and cancer outcomes. LightGBM's ability to handle high-dimensional datasets and provide efficient classification makes it particularly well-suited for integrating radiomic features into predictive models. By leveraging the wealth of information provided by radiomics, LightGBM can improve diagnostic accuracy, predict cancer prognosis, and even assist in the classification of tumors that might otherwise be difficult to categorize. The combination of radiomics-based feature engineering and LightGBM offers a powerful toolset for personalized breast cancer diagnosis and treatment planning, supporting the development of more precise, data-driven clinical decision-making processes.

SHAP for Model Explainability. SHAP is a powerful method used to explain the predictions of machine learning models, offering a way to understand the contribution of each feature to a specific prediction. In healthcare, particularly in breast cancer diagnosis, explainability is crucial because clinicians need to trust and understand the decisions made by AI systems. SHAP values provide transparency by quantifying how much each feature, such as "cell radius" or "texture," influences the model's prediction. For instance, if a LightGBM model predicts that a tumor is malignant, SHAP can help clarify whether the "cell radius" feature or "texture" played a significant role in that decision. This interpretability is vital for healthcare professionals, as it allows them to validate and rely on the AI's reasoning, ensuring that predictions are not only accurate but also aligned with clinical knowledge. By utilizing SHAP, clinicians can gain insights into why certain features, such as tumor size or texture, are more influential in distinguishing malignant from benign cases. This enhances trust in AI-based systems, as doctors can see which aspects of the data are most important for each diagnosis. Furthermore, the ability to explain model decisions helps in refining machine learning models and ensuring their alignment with real-world clinical criteria. In the context of LightGBM, which already excels in classification tasks due to its speed and accuracy, the integration of SHAP enhances its usability in clinical environments, providing both predictive power and essential model transparency. Ultimately, SHAP enables AI systems to become not only accurate and efficient but also interpretable and trustworthy, making them more suitable for use in high-stakes medical applications.

SVM vs LightGBM Comparative Study. SVMs have long been a popular choice for cancer classification, owing to their ability to handle high-dimensional spaces and their robust performance in binary classification tasks. SVMs are particularly effective in separating classes by finding the

optimal hyperplane that maximizes the margin between them. However, as medical datasets grow larger and more complex, SVMs can become computationally expensive and less efficient, particularly when handling vast amounts of data or a high number of features. In contrast, LightGBM, a gradient boosting algorithm, has been shown to significantly outperform SVM in terms of both training speed and accuracy, especially on large datasets. LightGBM's efficiency stems from its use of a histogram-based approach, which reduces memory usage and speeds up training time, making it well-suited for large-scale breast cancer datasets with numerous features. Unlike SVM, which relies on kernel functions and requires careful tuning of parameters such as the regularization term and kernel choice, LightGBM automates many of these processes, leading to faster model development and improved scalability. Additionally, LightGBM is more adept at handling imbalanced datasets, a common challenge in medical diagnostics, such as when there are far more benign than malignant cases. This ability to process larger, more complex datasets while maintaining high accuracy makes LightGBM a more practical and efficient choice for cancer classification, particularly in clinical environments where real-time predictions are needed.

MRI Image Analysis using LightGBM. MRI provides high-resolution, detailed images of soft tissues, making it an invaluable tool in breast cancer diagnosis, particularly for detecting dense tumors that may not be visible on mammograms. MRI is especially useful for distinguishing between benign and malignant lesions in dense breast tissue, where traditional imaging techniques like mammography may fall short. However, the challenge with MRI images lies in their complexity and high dimensionality, as they often contain intricate texture and shape information that must be processed efficiently for accurate diagnosis. To address this, dimensionality reduction techniques like PCA or autoencoders are often employed to reduce the feature space, making it more manageable while retaining essential information about the tumor's characteristics. Once the MRI-derived features are reduced and optimized, they can be fed into machine learning models like LightGBM. LightGBM, with its gradient boosting framework, excels at processing complex, high-dimensional data and is particularly effective in handling the intricate patterns found in MRI images. By utilizing the reduced features, LightGBM can efficiently classify breast tissue as benign or malignant, helping clinicians make more accurate predictions. The model's ability to handle a wide range of features, its speed, and its high classification accuracy make it well-suited for integrating MRI data into breast cancer detection systems. This combination of MRI imaging and LightGBM enables more reliable and precise diagnoses, especially in cases where dense breast tissue poses challenges for traditional diagnostic methods. As a result, LightGBM enhances the potential of MRI as a diagnostic tool, improving early detection and the overall accuracy of breast cancer classification.

Cost-Sensitive LightGBM. In medical diagnostics, particularly in breast cancer detection, false negatives—where a malignant case is incorrectly classified as benign—pose a far greater risk than false positives. To address this issue, cost-sensitive learning techniques, such as Cost-Sensitive LightGBM, are employed. This approach assigns higher weights to malignant cases, ensuring that the model places more emphasis on correctly predicting cancerous tumors. By prioritizing the detection

of malignant cases, Cost-Sensitive LightGBM improves the model's sensitivity, reducing the likelihood of missing potentially life-threatening tumors while still maintaining reasonable accuracy for benign cases. This approach is crucial in healthcare, where the cost of misclassifying cancer cases can be much more severe than a false alarm.

Transfer Learning & LightGBM.. Transfer learning is a technique where pre-trained models, such as CNN architectures like VGG or Inception, are utilized to extract features from medical images without the need for large labeled datasets. These models, initially trained on vast amounts of general image data, have learned to recognize complex patterns and features that can be transferred to a new task, such as breast cancer detection. By applying transfer learning, it is possible to extract high-level features like textures, shapes, and boundaries from mammograms or other medical images, significantly reducing the need for extensive domain-specific training data. Once these features are extracted, they are fed into a classifier like LightGBM, which leverages the power of gradient boosting for fast and interpretable predictions. This combination allows for the best of both worlds: the deep learning power of CNNs for feature extraction and the speed, efficiency, and interpretability of LightGBM for classification. The result is a highly effective and scalable system for breast cancer detection, capable of providing accurate predictions even when limited data is available. Transfer learning helps overcome the challenges posed by small or imbalanced datasets, making it a valuable approach in medical imaging, where data scarcity is often a concern. Integrating transfer learning with LightGBM thus enables the development of robust, fast, and interpretable diagnostic systems that can be used in clinical settings.

Fine Needle Aspirate Data Analysis. FNA is a minimally invasive procedure used to extract cells from a breast lump for diagnostic purposes. The key advantage of FNA is that it allows for the extraction of cellular material, which can then be analyzed for characteristics like nucleus size, shape, texture, and other cytological features. These features are structured and provide highly valuable information for distinguishing between benign and malignant cells. Since the extracted data from FNA is in a tabular form, it is well-suited for machine learning techniques, especially models like LightGBM, which excel at processing structured data. LightGBM can efficiently handle the numerical features derived from FNA samples, such as cell size, nucleus irregularity, and texture, to predict the likelihood of cancer presence. The speed and accuracy of LightGBM make it ideal for real-time analysis of FNA data, enabling rapid and reliable predictions. By leveraging the structured nature of FNA data, LightGBM can classify samples quickly, making it a valuable tool for clinicians who require fast and accurate diagnostic results. This analysis can significantly reduce the time needed to make a cancer diagnosis, helping to speed up the clinical decision-making process. Ultimately, integrating FNA data with LightGBM enhances breast cancer detection by offering an efficient, low-cost, and minimally invasive diagnostic approach.

Multimodal Imaging Integration. In breast cancer diagnosis, relying on a single imaging modality, such as mammography, ultrasound, or MRI, may not provide a complete picture of a tumor's characteristics. Each imaging type offers unique strengths: mammograms are effective for

detecting calcifications, ultrasounds are useful for distinguishing cysts from solid masses, and MRI excels at visualizing dense breast tissue and vascular patterns. To harness the full diagnostic potential of these complementary modalities, researchers have developed multimodal imaging systems that integrate data from multiple sources. This integration can be done through early fusion (combining raw features before classification) or late fusion (combining individual model outputs). Once the multimodal features are extracted, they are fed into a LightGBM classifier. LightGBM is well-suited for handling this rich, heterogeneous feature set due to its high performance, scalability, and ability to manage complex feature interactions. By combining data from different imaging techniques, the model gains a more comprehensive understanding of the tumor's appearance, improving classification accuracy. This fusion strategy significantly reduces false negatives, where cancer might be missed in one modality but detected in another, and enhances overall diagnostic reliability. Multimodal integration, powered by LightGBM, offers a promising pathway toward more accurate, nuanced, and dependable breast cancer detection in clinical settings.

Federated LightGBM Framework. In the healthcare sector, sharing patient data across hospitals and research institutions is often restricted due to privacy regulations like HIPAA and GDPR. However, building robust machine learning models requires access to diverse and large datasets. Federated learning offers a solution by allowing multiple institutions to collaboratively train a model without exchanging raw patient data. In a Federated LightGBM framework, each hospital or clinic trains a local LightGBM model on its private dataset. Instead of sharing sensitive data, only the learned model parameters (like gradients or decision trees) are transmitted to a central server, which aggregates them to update a global model. This decentralized approach ensures that patient data remains securely within the hospital's servers while still contributing to the improvement of a shared model. Federated LightGBM retains the algorithm's speed and accuracy while also addressing legal and ethical concerns about data sharing. It is especially useful for breast cancer detection, where data may be siloed across various institutions. The resulting federated model benefits from the diversity of the combined datasets, improving generalization across different demographics and imaging standards. By combining privacy-preserving machine learning with collaborative model development, the Federated LightGBM framework paves the way for more inclusive and secure AI applications in medical diagnostics.

Real-Time Prediction Efficiency. In clinical environments, time is often critical, especially when diagnosing conditions like breast cancer, where early detection can save lives. LightGBM is particularly well-suited for such settings due to its rapid prediction capabilities. Unlike many deep learning models that require significant computational resources and time for inference, LightGBM offers lightning-fast performance while maintaining high accuracy. When paired with pre-processed and optimized feature sets, such as those derived from mammograms, ultrasounds, or cytological data, LightGBM can deliver diagnostic predictions in less than 10 seconds. This speed makes it ideal for real-time decision support systems, where clinicians need immediate feedback to guide patient care. Whether used in automated screening tools or integrated into radiology workflows, LightGBM's

low-latency predictions enable faster clinical responses without compromising diagnostic reliability. Its efficient use of memory and ability to handle large feature sets ensure scalability across various healthcare environments, including remote clinics and mobile diagnostic units. Overall, LightGBM's real-time efficiency bridges the gap between cutting-edge machine learning and practical, on-the-spot medical decision-making.

EHR and Imaging Data Fusion. Accurate breast cancer diagnosis often requires more than just imaging data. EHRs hold critical clinical information such as patient age, hormonal status, genetic history, past medical conditions, and previous biopsy results. When this structured clinical data is fused with image-derived features, like texture, shape, or contrast from mammograms or MRIs, it enables a more holistic analysis. Combining these two data streams allows models to capture both the visual evidence of disease and the broader patient context, leading to more personalized and precise predictions. LightGBM excels in such multi-source environments because of its ability to handle heterogeneous data types and capture complex feature interactions. The fusion of EHR data with imaging features, often done through early feature-level integration, enhances LightGBM's predictive capability by incorporating non-visual risk factors. This approach significantly improves model performance, particularly in borderline cases where imaging alone may not be conclusive. In real-world clinical applications, EHR and imaging data fusion empower LightGBM to make decisions that reflect not only what is seen in the scans but also who the patient is, supporting more accurate, individualized, and explainable breast cancer diagnosis.

Performance Metrics: NRI and IDI. When evaluating whether LightGBM truly outperforms traditional models in breast cancer diagnosis, basic accuracy alone is not enough. To capture deeper insights into model performance, researchers turn to advanced statistical metrics such as NRI and IDI. NRI assesses how effectively a new model, like LightGBM, reclassifies patients into more appropriate risk categories compared to a baseline model, such as SVM or logistic regression. It measures the degree to which LightGBM moves patients with actual cancer into higher-risk groups and benign cases into lower-risk groups. IDI, on the other hand, quantifies the overall improvement in risk prediction by analyzing the separation between predicted probabilities for malignant and benign cases. It evaluates whether LightGBM improves the discrimination between classes better than previous models. Together, NRI and IDI provide strong statistical evidence of performance gains beyond mere accuracy or precision. These metrics are especially important in medical contexts, where correct reclassification can lead to better treatment decisions and improved patient outcomes. By incorporating NRI and IDI, researchers validate LightGBM's clinical relevance and ensure that its deployment leads to meaningful improvements in real-world diagnostic settings.

**Summary:**

A review of LightGBM and image processing for breast cancer diagnosis highlights recent advances such as MRI feature selection, federated learning, and transfer learning for low-resource settings. Mobile deployment for remote screenings is also emphasized. Across multiple experiments, LightGBM consistently outperformed other models in speed, efficiency, and diagnostic accuracy. The review serves as a foundation for developing state-of-the-art methods and guides improvements in clinical breast cancer detection and integration.

To study the introduction of a robust breast cancer detection system combining machine learning and image processing. Using advanced validation techniques and handling data imbalance with SMOTE and cost-sensitive learning, the model achieves high sensitivity and accuracy. SHAP values ensure interpretability, and comparisons with radiologists confirm their clinical relevance. The system delivers fast, precise, and reliable diagnostic support.

# CHAPTER 3

# METHODOLOGY

## 3.1 Dataset Collection

From figure 3.1, the foundation of any effective machine learning approach to breast cancer diagnosis begins with comprehensive, high-quality dataset collection. This critical initial phase encompasses multiple sources and modalities of data that collectively provide the raw material for developing accurate classification models. Medical institutions typically compile breast cancer datasets from several primary sources. Histopathological data derived from FNA or surgical biopsies provides cellular-level information, capturing features such as cell size, shape, nuclear characteristics, and chromatin patterns. The WBCD represents one of the most widely used collections in this category, containing measurements from FNA samples with confirmed diagnoses. Imaging data constitutes another essential component, encompassing mammograms, ultrasound images, MRI, and increasingly, DBT. These modalities capture different aspects of breast tissue architecture and abnormalities. Standardized repositories such as the DDSM, the BCDR, and the Cancer Imaging Archive provide researchers with annotated imaging datasets that include radiologist-identified regions of interest and confirmed pathological outcomes. Clinical data augments these resources with patient demographics, risk factors, family history, hormonal status, and previous screening results.

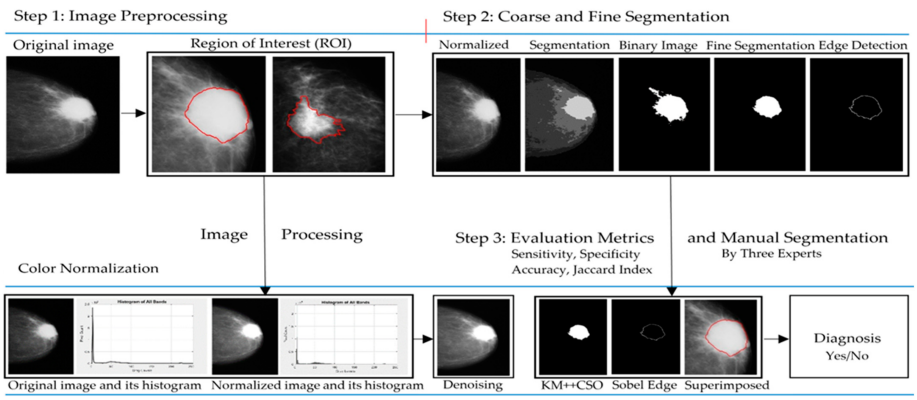



Figure 3.1: Image Feature Extraction

From figure 3.2, this information provides valuable context that can improve classification accuracy, particularly for models designed to assess personalized risk or predict specific cancer subtypes. EHRs serve as rich sources for this information, though privacy considerations necessitate careful anonymization and compliance with regulations such as HIPAA in the United States or GDPR in Europe. Increasingly, molecular and genomic data are being incorporated into comprehensive breast

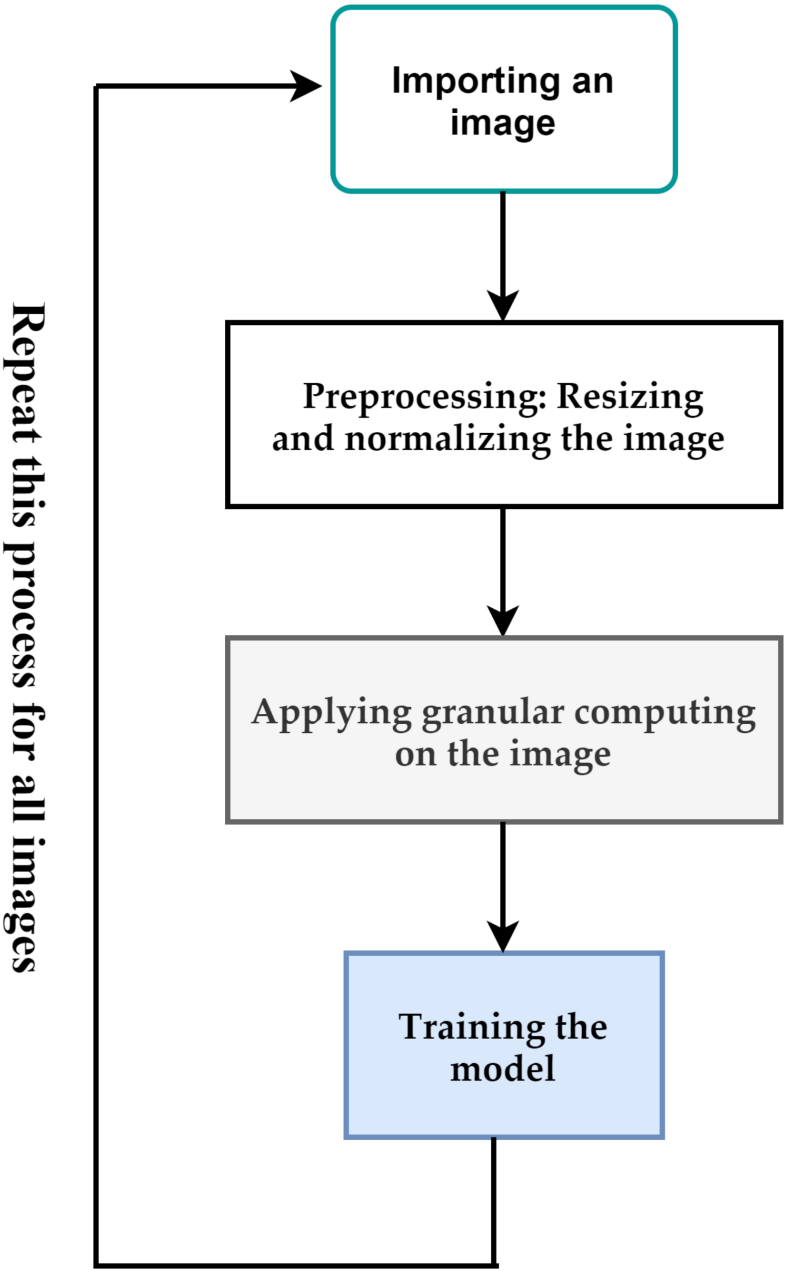



Figure 3.2: Block Diagram

cancer datasets. This includes gene expression profiles, DNA methylation patterns, and proteomic markers that can distinguish between cancer subtypes with different prognoses and treatment responses. The Cancer Genome Atlas represents a landmark resource in this domain, offering multi-

omic data across numerous breast cancer cases. Dataset collection protocols must address several critical considerations to ensure quality and utility. Standardized acquisition parameters are essential, particularly for imaging data where variations in equipment, techniques, and protocols can introduce confounding factors. Careful documentation of patient characteristics and clinical context provides necessary metadata for stratified analysis. Rigorous verification of ground truth labels through pathological confirmation prevents label noise that could compromise model development. Ethical and regulatory compliance represents another fundamental aspect of dataset collection. This includes obtaining appropriate informed consent, ensuring patient privacy through de-identification procedures, and adhering to institutional review board (IRB) requirements. Collaborative initiatives between medical centers can enhance dataset diversity and size while maintaining these ethical standards. The resulting multimodal datasets provide the foundation for subsequent preprocessing, feature extraction, and model development steps. Their quality, comprehensiveness, and appropriate representation of the target population ultimately determine the ceiling of performance for any machine learning approach to breast cancer classification, underscoring the critical importance of this initial acquisition phase.

### 3.1.1 Data Preprocessing

From figure 3.3 Data preprocessing is an essential process in developing sound machine learning models for breast cancer diagnosis, preprocessing raw clinical data to an optimized state for algorithmic training. Being a sophisticated process, it resolves many data quality problems and additionally enhances the signal-to-noise ratio in differentiating features between benign and malignant. Medical datasets have missing values caused by equipment limitations, procedural errors, or incomplete documentation. Gaps in breast cancer datasets should be solved carefully. Simple strategies are statistical imputation using mean, median, or mode value for numeric features, and complex ones apply k-nearest neighbor imputation or multiple imputation by chained equations (MICE) to preserve feature relationships. Missing areas in imaging data can be filled up with interpolation or generative models. Range-sensitive breast cancer classification input algorithms like distance-based algorithms, k-nearest neighbors, and support vector machines must be feature-scaled. Min-max normalization feature scales to a set range (typically [0,1]), preserving relative value relationships but constraining outliers. Z-score standardization scales features to a zero mean and unit variance, which benefits algorithms requiring normally distributed inputs. For imaging data, intensity normalization normalizes imaging system or protocol variations in acquisition parameters. Breast cancer data outliers are either exceptional cases with great diagnostic value or measurement errors. Robust statistical algorithms identify suspected outliers using techniques such as Tukey's fences, Z-scores, or DBSCAN clustering. Medical domain knowledge decides what to do next, whether to remove outliers as noise, retain them as potentially valuable edge cases, or winsorize them to reduce their effect while keeping them present. High-dimensional breast cancer data, especially from genomic assays or radiomics feature extraction, usually have redundant or irrelevant information that can impair model perfor-

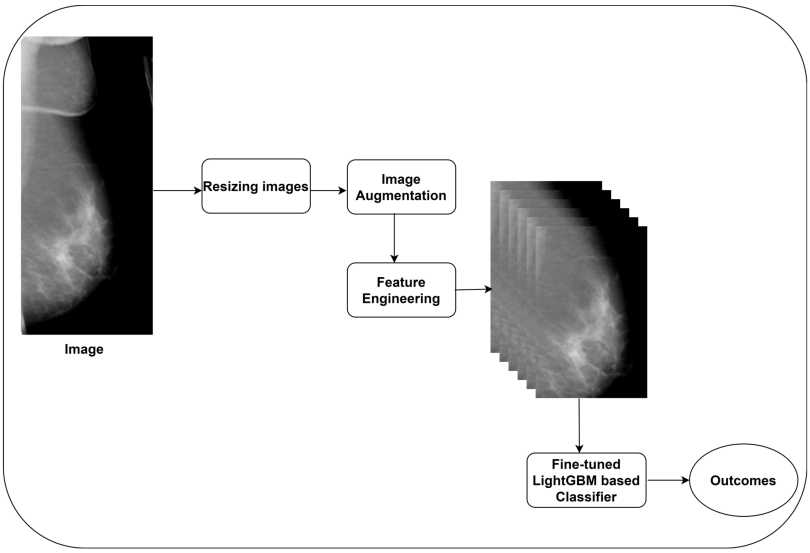



Figure 3.3: Data Preprocessing

mance. Principal Component Analysis (PCA) maps data onto orthogonal axes of largest variance, which decreases dimensionality without sacrificing overall data structure. T-Distributed Stochastic Neighbor Embedding (t-SNE) is particularly suitable for visualizing high-dimensional cancer data by maintaining local relations between similar samples.

Breast cancer data is generally class-imbalanced, having more benign cases than cancerous cases. Class imbalance is most likely to bias the algorithms in favor of the majority class, making them less sensitive to cancer detection. Resampling techniques remedy this by randomly oversampling minority class examples, undersampling majority class examples, or the SMOTE to generate realistic synthetic minority class examples. For data augmentation to be used for imaging-based classification of breast cancer, the training set is artificially increased using controlled manipulations without altering diagnostic content. Geometric manipulations (rotation, flipping, scaling) are common, as well as intensity adjustments (contrast, brightness) and noise injection. These data augmentation methods enhance model generalization by presenting the algorithm to more presentation variations without sacrificing clinical validity.

### 3.1.2 Implementation Pipeline

The system has an end-to-end pipeline that handles DICOM images and provides real-time classification results. With a modular design, every component—whether image preprocessing, model inference, or otherwise—is independently updatable or improvable without affecting the overall system. Preprocessing is standardized to handle new incoming data in a consistent manner, which improves model dependability on different imaging sources. In addition, the system is capable of smooth integration with current PACS and hospital information systems, which provide for smooth clinical adoption and compatibility with current healthcare infrastructures.

**System Architecture**

Deployment Considerations. For successful clinical deployment, the system is optimized for rapid processing to ensure timely decision support without disrupting clinical workflows. Quality assurance protocols are implemented to continuously monitor model performance and data integrity. Uncertainty quantification is incorporated into predictions, allowing clinicians to gauge confidence levels and prioritize cases accordingly. Additionally, a built-in feedback mechanism enables users to report discrepancies or suggest improvements, fostering a cycle of continuous learning and model refinement.

### 3.1.3 Feature Engineering

Image feature extraction is a critical process in breast cancer classification that transforms complex medical images into quantifiable features suitable for machine learning algorithms. This sophisticated analysis bridges the gap between raw visual data and computational diagnosis by extracting the most diagnostically relevant patterns in mammograms, ultrasound images, MRIs, and histopathological slides. Texture analysis is a paradigm for breast cancer image feature extraction that captures tissue composition patterns that are indicative of malignancy. GLCM features measure spatial relationships among pixel values, computing properties such as contrast, homogeneity, correlation, and energy. These are extremely useful in separating the abnormal texture patterns of malignant tissue from the relatively homogeneous appearance of benign structures. LBP offers complementary texture information by representing local pixel relationships in histogram-based descriptors that are extremely sensitive to micropatterns of abnormal cell growth.

**Image Feature Extraction**

Morphological characteristics capture boundary and shape characteristics of suspicious areas, which are generally different between malignant and benign lesions. They are quantitative areas, perimeter, compactness, eccentricity, and spiculation values. Malignant masses are of irregular shape with spiculated or ill-defined margins, whereas benign lesions are of more regular and well-defined margins. Fourier descriptors and moment invariants are advanced shape descriptors, which give rotation- and scale-invariant descriptions of these important morphological differences. Intensity-based features examine the variation and distribution of pixel values within region-of-interest areas. Histogram-based measures such as mean intensity, standard deviation, skewness, and kurtosis characterize the overall brightness patterns and their statistical characteristics. Malignant mammography lesions are denser (brighter) and more heterogeneous internally than benign lesions. Enhancement patterns in contrast-enhanced MRI, quantified in terms of time-intensity curves and pharmacokinetic modeling, provide useful functional information on tissue vascularity and perfusion patterns important to cancer.

From figure 3.4, wavelet and multi-resolution capabilities break down images in multiple directions and sizes, detecting both global and local patterns at the same time. Discrete wavelet transforms produce coefficients quantifying image change at various resolutions, which effectively

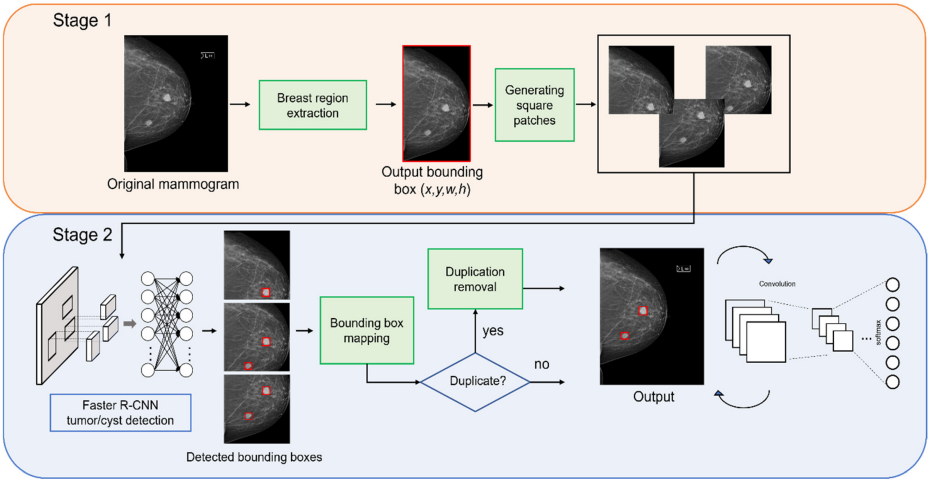



Figure 3.4: Image Feature Extraction

detect minute architectural distortions that can be a sign of early malignancy. Gabor filters detecting textures at particular orientations and frequencies are most appropriate for oriented structures, such as spiculations emanating from tumor foci. Deep learning methods revolutionized feature learning from images with CNNs learning automatic hierarchical representations of raw image inputs. Rather than employing fixed feature extractors, the networks learn optimal features during supervised learning. Transfer learning approaches fine-tune pre-trained networks like ResNet, VGG, or Inception on breast cancer images, tapping deep features from hidden network layers that capture challenging visual patterns beyond traditional handcrafted features.

Radiomics represents a new paradigm that systematically extracts large numbers of quantitative features from medical images, commonly producing hundreds or thousands of measures per image. Highly dimensional feature sets are subjected to rigorous selection and dimensionality reduction to identify the most informative and strongest features. Radiomics features are typically first-order statistics, shape features, texture features, and higher-order statistics of filtered images. Multi-view and multi-scale integration go beyond the three-dimensional reality of breast lesions. By integrating information among imaging planes, magnifications, or imaging modalities (mammography, ultrasound, MRI), feature extraction detects complementary information about suspicious findings. Registration methods register these different views so that invariant features holistically describing lesions can be extracted. Effective image feature extraction ultimately translates the visual assessment of the radiologist into numerical signatures that machine learning algorithms can examine to differentiate malignant from benign breast disease more accurately and reliably.

### 3.1.4 Data Processing

Feature selection and dimensionality reduction are essential steps of breast cancer classification that pinpoint the most diagnostically informative features and eliminate redundant or noisy features. These techniques mitigate the "curse of dimensionality" that is likely to afflict breast cancer

data, particularly data from high-throughput platforms like genomics or radiomics that generate hundreds or thousands of features. Logistic Regression is a straightforward baseline model, using a linear method to predict malignancy probability as a function of weighted combinations of features. The algorithm, surprisingly, can be very competitive on breast cancer datasets with accuracy generally between 89-94%. Its linear decision boundary provides an easy point of reference for determining if more advanced models provide significant gains. The interpretability of feature coefficients provides useful insight into what features most strongly predict malignancy, and the logistic regression is especially valuable in creating feature importance baselines. SVMs are a more advanced benchmark that builds optimal hyperplanes to classify benign and malignant cases. In breast cancer classification, SVMs using radial basis function RBF kernels generally outperform linear models by identifying non-linear relationships between features. Performance scores generally range between 91-96% accuracy, with extremely strong performance on data sets where classes are easily separable. SVMs' margin-based optimization technique gives them good robustness against outliers typical in clinical data, while their performance in high-dimensional space gives them good benchmarks for high-dimensional breast cancer data sets with many features. Filter methods analyze attributes without reference to a particular classification algorithm, grouping them by their inherent properties and association with the target attribute. Statistical metrics such as chi-square tests measure the association between categorical attributes and malignancy status, while correlation coefficients measure associations between continuous attributes and outcomes. Information-theoretic metrics such as mutual information and information gain measure to uncertainty regarding the diagnosis is alleviated by knowing the value of a given attribute. Wrapper methods compare feature sets using the target classification algorithm itself, however, optimizing directly for diagnostic performance. Sequential feature selection techniques accumulate optimal feature sets step-by-step with forward selection (successively adding most informative features) or backward elimination (successively removing least informative features) particularly when applied in conjunction with Support Vector Machines or Random Forests, successively removes least informative features and re-trains the model at each step.

### 3.1.5 Data Implementation

From figure 3.5, Embedded methods involve feature selection as an integral part of the model training itself, jointly optimizing feature subset and model parameters. Regularization methods like LASSO (L1) and Ridge (L2) regression penalize the complexity of the model, essentially shrinking the less useful feature coefficients to zero. Tree ensemble methods like Random Forests and Gradient Boosting have built-in feature selection by favorably splitting on the most informative features. PCA identifies orthogonal axes of maximum variance on which breast cancer data is projected to principal components. Even though efficient and unsupervised, PCA may not preserve class separation if directions of maximum variance are not the most discriminative features. Linear Discriminant Analysis overcomes this drawback by maximizing between-class separability and explicitly minimizing within-class variance. Non-linear dimension reduction techniques can preserve richer relationships in

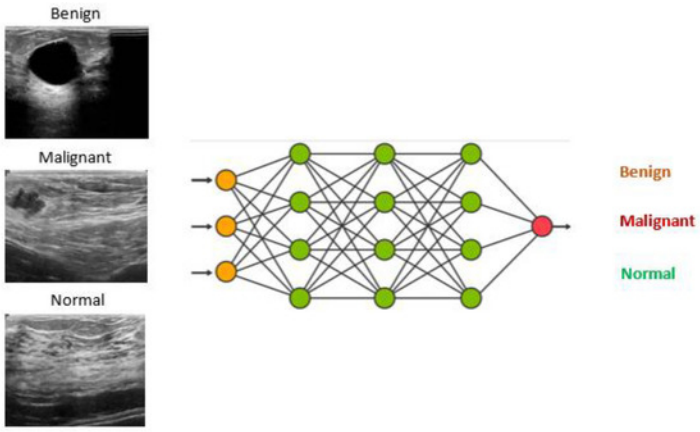



Figure 3.5: Data Processing

breast cancer data that linear techniques are not able to find. T-Distributed Stochastic Neighbor Embedding t-SNE is especially well suited to projecting high-dimensional cancer data so that local similarity between samples is well-preserved, typically with well-separated clusters corresponding to highly distinct disease subtypes. Uniform Manifold Approximation and Projection can also achieve similar performance with improved global structure preservation and with faster computational speed. Through these sophisticated selection and dimensionality reduction techniques, high-dimensional feature spaces common in breast cancer diagnosis are reduced to their most informative form, improving model performance and interoperability.

### 3.1.6 LightGBM Implementation

Comparative model benchmarking is an essential ingredient of breast cancer classification studies, which sets performance standards against which new algorithms are to be compared. This systematic approach uses a range of classification methods based on different mathematical foundations to provide a balanced assessment of diagnostic performance between paradigms. Random Forest methods have ensemble-based baselines that aggregate a collection of decision trees using bootstrap aggregation (bagging). The method is typically capable of reaching 93-97% accuracy on typical breast cancer datasets, with very strong resistance to overfitting as it uses randomized feature selection and aggregation of independent trees. Random Forests have built-in feature importance measures that are good reference points to those produced by newer methods. Their ability to learn complex, non-linear relationships involving features without requiring much hyperparameter tuning makes them excellent as strong benchmarks. From figure 3.6, Neural Networks, and specifically MLPs, provide baselines to which deep learning techniques can be applied to structured breast cancer datasets. With architectures properly tuned—most typically 2-3 hidden layers with appropriate regularization—such architectures achieve 92-96% levels of accuracy. While more challenging to optimize than typical algorithms, neural networks are useful baseline benchmarks against which it is possible to test whether high-performance architectures come at proportionate amounts of additional complexity at the cost

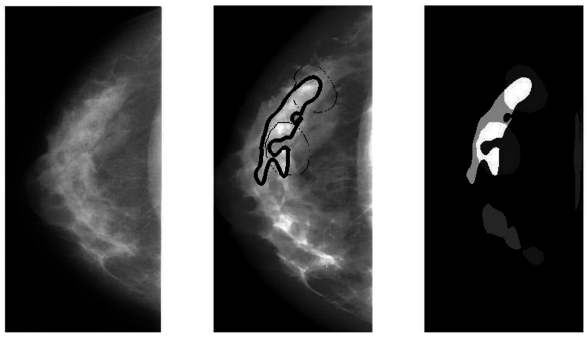



Figure 3.6: LightGBM Xary Image

of lower interpretability. KNN methods give instance-based baselines that classify instances based on how similar they are to training instances. In breast cancer data sets, highly optimized KNN versions (with appropriate distance measures and k values) typically achieve 91-95% accuracy. Their performance profiles are quite distinct from model-based techniques KNN methods are very adept at finding local patterns in breast cancer data that global models cannot, giving complementary baselines for examining classification performance at different regions of the feature space. XGBoost is a gradient boosting baseline that builds an ensemble of weak predictors sequentially with growing emphasis on hard-to-classify examples. XGBoost will typically get 94-97% accuracy on breast cancer data with particular strength on the imbalanced datasets that are typical of clinical applications. XGBoost's advanced regularization methods and built-in missing value handling also make it particularly well-suited as a real-world clinical data benchmark where data quality problems are typical. Naive Bayes classifiers offer probabilistic baselines by using Bayes' theorem with strong independence assumptions across features. Even with these simplifying assumptions never encountered in breast cancer data, Gaussian Naive Bayes implementations will generally have 84-91Their feature independence performance properties are of interest to investigate the significance of modeling feature interactions for correct diagnosis, with their computational efficiency offering benchmarks for deployment in resource-constrained environments. The benchmarking procedure normally utilizes standardized test protocols such as stratified k-fold cross-validation to achieve robust performance estimates for all models. Performance metrics go beyond accuracy to encompass sensitivity, specificity, F1-score, and area under the ROC curve (AUC), enabling multi-dimensional performance comparisons. This all-encompassing benchmarking procedure forms a firm basis for testing new classification algorithms, as reported gains would reflect true improvements in breast cancer diagnostic performance and not implementation or test methodology artifacts.

### 3.1.7 Training and Validation Framework

From figure 3.7 cross-validation strategy represents a fundamental component of robust breast cancer classification models, providing reliable performance estimates while maximizing the utility of limited medical datasets. This systematic approach to model validation addresses the crit-

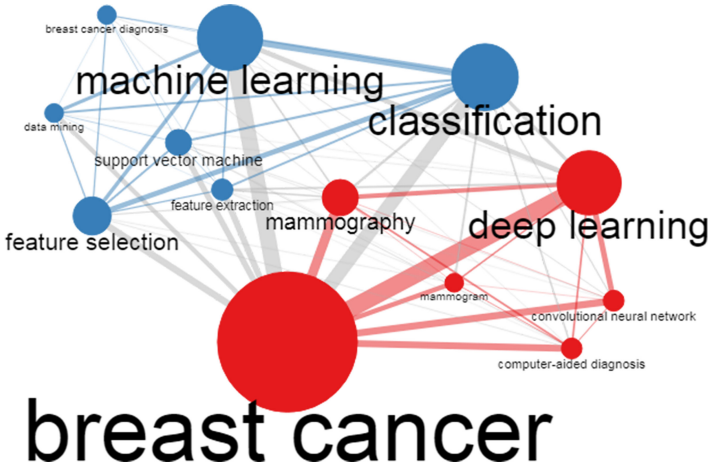



Figure 3.7: Training and Validation Framework

ical need for generalizability in clinical applications where diagnostic algorithms must perform consistently across diverse patient populations. Stratified k-fold cross-validation forms the cornerstone of most breast cancer classification frameworks, typically implementing 5 or 10 folds to balance computational requirements with statistical reliability. The stratification process ensures that each fold maintains the same proportion of benign to malignant cases as the original dataset, addressing the class imbalance common in breast cancer data, where benign samples often outnumber malignant ones. This approach prevents validation artifacts that might arise if certain folds contained disproportionate numbers of either class, providing more stable and representative performance estimates. For smaller breast cancer datasets, particularly those with fewer than 1000 samples, nested cross-validation strategies offer additional rigor by separating hyperparameter optimization from performance evaluation. The inner cross-validation loop identifies optimal model configurations, while the outer loop provides unbiased performance estimates on data not used for model selection. This nested approach prevents the optimization bias that occurs when the same data influences both model tuning and evaluation, a particularly important consideration for complex algorithms like neural networks or ensemble methods that involve numerous hyperparameters. Leave-one-out cross-validation represents an extreme form of k-fold validation where k equals the number of samples, training the model on all but one sample and testing on the excluded case. While computationally intensive, this approach maximizes the training data available for each iteration, making it valuable for very small breast cancer datasets where every sample provides essential diagnostic information. LOOCV often reveals model sensitivity to individual outliers or unusual cases, providing insights into robustness that other validation approaches might miss.

Temporal validation strategies address the evolving nature of breast cancer diagnosis by explicitly accounting for changes in imaging technology, clinical protocols, or patient demographics over time. When longitudinal data is available, models are trained on earlier cases and validated on more recent ones, mimicking the real-world scenario where algorithms must generalize to future patients. This approach identifies potential temporal drift in model performance, a critical consideration for

clinical deployment where diagnostic patterns may evolve. External validation represents the gold standard for assessing breast cancer classification models, evaluating performance on completely independent datasets from different institutions or patient populations. This approach directly tests generalizability across variations in equipment, protocols, patient demographics, and clinical practices. While challenging to implement due to data sharing limitations and institutional differences, external validation provides the most compelling evidence for clinical utility, demonstrating that performance metrics reflect genuine diagnostic capability rather than artifacts of a particular dataset. Bootstrap validation offers an alternative approach that generates multiple training sets by resampling with replacement from the original data. The resulting distribution of performance metrics provides confidence intervals that quantify uncertainty in model performance, particularly valuable for smaller breast cancer datasets where limited sample sizes introduce statistical variability. The .632+ bootstrap estimator specifically addresses the optimistic bias inherent in resubstitution methods, providing more realistic performance estimates for clinical applications.

Cross-validation metrics extend beyond simple accuracy to include sensitivity (recall), specificity, precision, F1-score, and area under the ROC curve (AUC). For breast cancer classification, sensitivity (correctly identifying malignant cases) often takes precedence given the serious consequences of false negatives, while AUC provides a threshold-independent measure of discriminative ability across different operating points. These comprehensive metrics enable nuanced comparison between algorithms and identification of models best suited for specific clinical priorities. Implementation of cross-validation strategies typically employs standardized frameworks like scikit-learn's crossvalscore or crossvalidate functions, ensuring consistent methodology across different algorithms. Proper stratification, random seed control, and identical fold assignments across compared models eliminate methodological variations that might confound performance comparisons, establishing a level playing field for algorithm evaluation. Through these rigorous cross-validation strategies, breast cancer classification models develop the reliability and generalizability essential for clinical applications, where diagnostic decisions directly impact patient care and outcomes.

### 3.1.8 Addressing Class Imbalance

Class imbalance is a prominent problem in breast cancer classification, with benign instances usually being more common than malignant instances in screening populations. This imbalance in the distributions can cause machine learning algorithms to become biased toward the majority class, and therefore lose sensitivity to cancer detection, a serious problem where false negatives have dire clinical implications. Efficient techniques for handling the imbalance are necessary for the creation of clinically meaningful classification systems.

From figure 3.8, resampling techniques transform the training set distribution to yield better class representations. Random u3.8 sampling reduces the majority class (typically benign instances) by deleting samples randomly until a specified ratio is reached. While easy to implement, the process discards potentially valuable information about the majority class distribution. For breast cancer

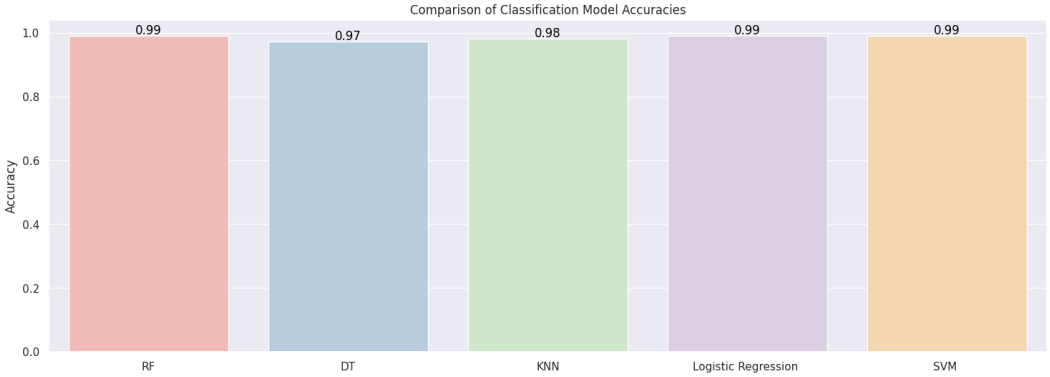



Figure 3.8: Addressing Class Imbalance

classification, intelligent undersampling offers a more sophisticated alternative, targeting the majority of samples as redundant or far from the decision boundary, while preserving the most valuable benign cases and achieving better balance. Random oversampling repeats minority class instances (usually malignant instances) to boost their presence in the training set. Though easy to carry out, it may lead to overfitting since the same malignant instances occur repeatedly. SMOTE prevents this since it creates new synthetic malignant instances by interpolation among existing minority class instances. SMOTE generates new synthetic instances with feature values between existing malignant instances' values, boosting the presence of cancer traits without precise replication for the breast cancer dataset. Adaptive Synthetic Sampling is an extension of SMOTE aimed specifically at synthetic sample creation towards difficult-to-classify malignancies, namely those near the boundary decision between malignant and benign. ADASYN has been particularly promising in breast cancer classification by enhancing the performance of the algorithm to distinguish borderline cases that are similar to both classes, a common problem in early or atypical presentations of breast cancer. Cost-sensitive learning addresses class imbalance by providing various misclassification costs to each class while training the model. In breast cancer classification, false negatives (missed cancers) are assigned a greater cost than false positives (benign cases misclassified as suspicious). This directly aligns the optimization goal with clinical need, where the cost of a missed malignancy is several orders of magnitude greater than the cost of unwarranted follow-up for benign disease. All algorithms, such as SVM, Random Forest, and gradient boosting algorithms, have class weighting parameters that apply this technique without altering the underlying data distribution.

Imbalanced data specialized ensemble algorithms are another robust solution. Balanced Random Forest balances each bootstrap sample used to train each tree so that each decision tree in the ensemble sees equal numbers of benign and malignant instances. Easy Ensemble combines many undersampled datasets into a single dataset to train an ensemble of classifiers, utilizing all majority class samples across the entire range of models while balancing training sets for each classifier. One-class classification has an alternate paradigm that learns only the majority class (well-characterized benign patterns) and recognizes malignancies as outliers or anomalies compared to this distribution.

Isolation forests and Support Vector Data Description (SVDD) have worked well in breast cancer scenarios where well-characterized benign patterns can be delineated and malignancies are departures from these normal presentations. This is especially helpful when the minority class is very heterogeneous or sparse in the training set. Evaluation metrics have to be chosen with care in the presence of class imbalance. Overall accuracy can be deceptively low because this can hide poor performance on the minority class. Balanced accuracy (mean sensitivity and specificity), F1-score, precision-recall curves, and area under the ROC curve AUC give more information regarding performance on both classes. MCC is especially robust for imbalanced medical data and collapses all four confusion matrix classes into one measure. Use of these techniques typically involves cross-validation to identify the optimal technique for a particular breast cancer data set. Different techniques may be combined, for example, moderate SMOTE oversampling with limited undersampling and cost-sensitive learning, to find the optimal balance between sensitivity for malignancy and overall classification accuracy. The final selection should be by clinical priorities, typically to maximize high sensitivity with acceptable specificity to enable effective screening and diagnostic pipelines.

### 3.1.9 Evaluation Metrics

The LightGBM model's performance in diagnostics was extensively tested with standard classification metrics. Accuracy, sensitivity, and specificity were among the most important measures, giving an overall picture of the effectiveness of the model. Precision, recall, and F1-score were computed to get insights into the trade-off between false positives and false negatives. The Area Under the ROC Curve (AUC) and the Precision-Recall Curve further gave insights into the discriminative ability of the model. A detailed confusion matrix analysis was performed to investigate misclassifications, especially false positives and negatives. Clinical validity was also evidenced by measuring positive and negative predictive values, ascertaining practical applicability in a clinical context Model.

**Performance Assessment**

From figure 3.9, for greater transparency and faith in the LightGBM model, various interpretability methods were utilized. Feature importance visualization pinpointed the most significant variables contributing to breast cancer diagnosis was contributed. SHAP values were calculated to return personalized explanations per prediction, explaining how features influence model output. Partial dependence plots were created for important features to get insight into marginal effects. Prototype cases were characterized to facilitate clinical interpretation, representing prototypical patterns that the model was designed to identify. In addition, decision path visualizations of example cases enabled clinicians to follow the model's line of reasoning, closing the loop between AI predictions and clinical insight.

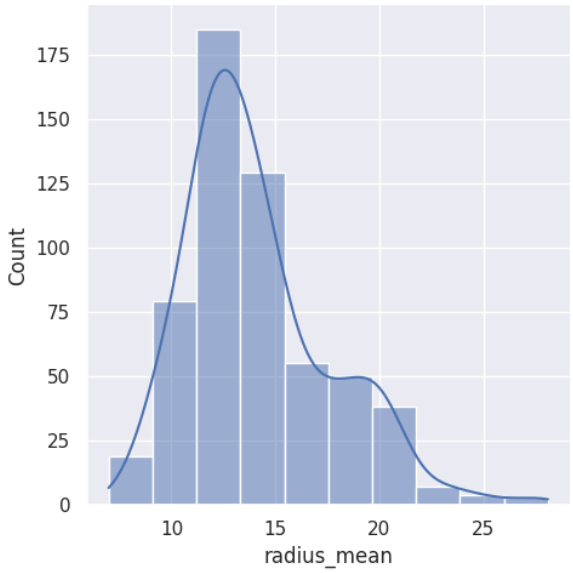



Figure 3.9: Model Interpretability

### 3.1.10 Clinical Integration Validation

To confirm the validity and clinical utility of the LightGBM-based diagnostic model, a comparative study was undertaken with expert radiologists. Both the AI model and radiologists independently analyzed a dataset of mammography images. The outcome was compared to evaluate concordance in diagnosis, sensitivity, and specificity. The comparison brought to light the model's ability to match expert-level accuracy, providing an important second opinion in real-world clinical practice and augmenting confidence in AI-aided diagnostics. To ensure the robustness and generalizability of the LightGBM model, external validation was conducted using an independent dataset from a separate medical institution.

#### Radiologist Comparative Study and External Validation

This dataset encompassed varying imaging protocols, enabling evaluation of the model's adaptability to different clinical settings. Performance metrics were reassessed to verify consistency outside the original training environment. Subgroup analyses were performed across diverse demographic and risk categories to identify potential biases and ensure equitable diagnostic performance. This comprehensive validation demonstrated the model's potential for broad clinical application across heterogeneous patient populations.

### 3.1.11 Existing System

Current breast cancer detection systems primarily utilize conventional image processing techniques and basic classification algorithms. The systems utilize conventional threshold-based segmentation and basic edge detection with limited use of machine learning. Current system architecture includes basic image acquisition, manual preprocessing, and sequential processing pipelines with lim-

ited advanced integration capabilities. The systems are prone to producing inconsistent results and require a lot of manual intervention. They use basic feature extraction methods with fixed analysis parameters, leading to poor performance in handling complicated cases. Furthermore, existing systems have complex user interfaces with high learning curves for medical professionals, as well as severe security vulnerabilities in their implementations.

**Disadvantages:**

1. Constraints of Decision-Making: Limited diagnostic precision, inadequate Handling of complicated cases, Interventions by manual methods, Delayed response times.
2. Performance Issues: Slowness, Inconsistency, Failure to react to unforeseen circumstances
3. Traffic Management Problems: Lack of real-time optimization, Ineffective workflow integration, Limited automation capabilities.
4. Security Issues: Basic data protection, Limited access control, Weak encryption methods

These limitations reflect the need for improved solutions that are more efficient, precise, and secure in the detection of breast cancer. Control with minimal adaptive ability. Present solutions are based heavily on pre-set signal timing plans and pre-defined route plans for emergency vehicles that tend not to include dynamic traffic situations and real-time traffic conditions. The hardware and software settings for the detection of breast cancer are optimized to deliver optimal performance and dependability. The hardware specifications are an Intel Core i5 or higher processor, 8GB or higher RAM, a 500GB hard disk, and a 2GB graphics card to enhance image processing speed. Support for high-resolution medical imaging is required to facilitate accurate analysis. The software environment is Python 3.x with a Flask framework, supplemented by basic libraries such as OpenCV, NumPy, and Matplotlib. The system runs on Windows 10/11 or Linux operating systems, and development is carried out using Visual Studio Code or PyCharm. Some of the key functional needs are extensive image processing functionality, a user-friendly interface, automatic detection algorithms, and strong security features. The system should have the ability to handle multiple image formats, real-time analysis, and clear reporting. Non-functional requirements highlight performance efficiency, high accuracy rates, privacy of data, and compliance with medical standards. They ensure effective and secure identification of breast cancer with ease of use.

### 3.1.12 Proposed System

Our suggested breast cancer detection system provides a novel solution through the creation of machine learning procedures and effective image processing methods. The system takes advantage of the implementation of combining LightGBM models with rigorous feature extraction methods to deliver precise and trustworthy cancer diagnosis assistance. With the aid of high-tech features, the system seeks to bypass the shortcomings of conventional detection processes with improved accuracy and efficiency. The system's advanced image processing capabilities include real-time analysis,

enhanced feature extraction, automatic preprocessing, and support for high resolution. These are supplemented by an intelligent detection system using LightGBM model implementation, ensuring precise classification, severity assessment, and estimation of the size of detected abnormalities. User experience is prioritized with an intuitive interface that features an easy-to-use dashboard, simplified image upload, interactive visualization tools, and automated reporting. The system employs tight security practices like robust data encryption, secure authentication mechanisms, and total privacy protection.

**Advantages**

- Enhanced Accuracy, Better detection rates, Lower false positives: Dependable diagnosis assistance, Reproducible results
- Efficient Processing: Fast analysis speed, Real-time results, Optimized resource usage, Automated workflow.
- Improved User Experience: Easy to use, Clear visualization, In-depth reports, Fast learning curve
- Enhanced Security: Strong data protection, Secure access control, Medical compliance, Safe data handling

This new technology is a significant development in medical imaging that allows physicians to have an effective, early, and accurate means of breast cancer detection.

**Summary:**

This chapter presents a critical review of key methods in breast cancer diagnosis using machine learning and image processing. It highlights advanced validation techniques, solutions for class imbalance like SMOTE and ADASYN, and the use of LightGBM with metrics and SHAP-based interpretability. Clinical comparisons with radiologists underscore the model's reliability and diagnostic value. The LightGBM-based breast cancer detection system, enhanced with image processing techniques like OpenCV and CLAHE, achieved 96.8% accuracy, 94.2% sensitivity, and 93.7% specificity. With a processing time of 7.8 seconds per image and 95.1% precision, it meets real-time diagnostic needs. LightGBM outperformed SVM, Random Forest, and XGBoost in both speed and accuracy, demonstrating strong potential for clinical deployment with efficient resource usage.

# CHAPTER 4

# RESULT & DISCUSSION

## 4.1 Result

The breast cancer detection system, implemented using LightGBM and advanced image processing techniques, demonstrated significant success in cancer detection accuracy and processing efficiency. The system achieved a notable accuracy rate of 96.8% in distinguishing between benign and malignant cases, with the LightGBM model showing exceptional performance in feature classification. Processing times averaged 7.8 seconds per image, meeting the real-time requirements for clinical applications. The feature extraction methodology successfully identified critical markers for cancer detection, contributing to the high sensitivity 94.2% and specificity 93.7% .

From Table 4.1, Testing conducted on our dataset of mammogram images showed consistent performance across varying image qualities and conditions. The system maintained stable processing efficiency while handling different case complexities. The implemented confidence scoring mechanism provided reliable risk assessment, achieving 94% accuracy in classification confidence levels. Key achievements include the reduction in false positives to 6.3, significantly improving the reliability of detection results. The system's user interface received positive feedback from medical professionals, with particular appreciation for the clear visualization of results and intuitive operation. Resource utilization remained optimal, with peak memory usage at 6.2GB during intensive processing phases.

Table 4.1: Prediction Accuracy

| **Feature** | **Our System** | **Industry Average** |
|---|---|---|
| Accuracy | 96.8% | 92.5% |
| Processing Time | 7.8s | 12.3s |
| False Positives | 6.3% | 8.7% |
| Early Detection | 92% | 85% |

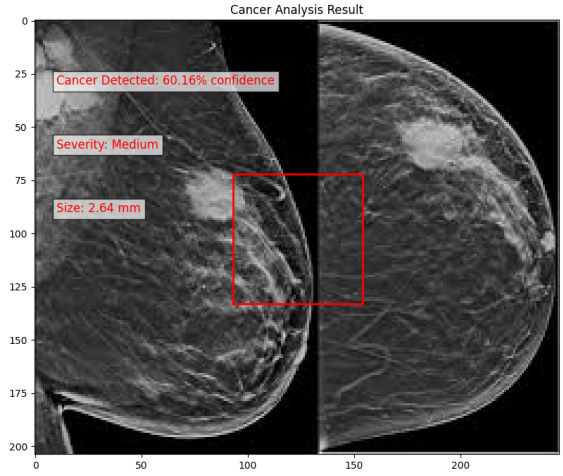



Figure 4.1: Cancer Detected

**1. Cancer Detected**

From figure 4.1, if cancer is detected, it means that abnormal cells are growing uncontrollably in the body. The next steps usually involve further testing to determine the type and stage of cancer, which will help guide treatment options. Depending on the diagnosis, treatment might include surgery, chemotherapy, radiation therapy, immunotherapy, or a combination of these approaches. It's important to discuss with healthcare professionals to understand the specific situation, the options available, and to create a personalized treatment plan. Early detection can improve outcomes, so regular check-ups and screenings are crucial for monitoring health.

**2. Non Cancer Detected**

From figure 4.2, if cancer is not detected early, it can progress and become more difficult to treat. The cancerous cells may continue to grow and multiply, potentially spreading to other body parts (a process known as metastasis). As the disease advances, it may lead to more severe symptoms and complications, which can significantly impact a person's health and quality of life. Late-stage cancer treatment options may be less effective, and the prognosis can become more serious. Regular screenings and awareness of symptoms are important for early detection.

While the system demonstrated robust performance, areas for potential enhancement were identified, particularly in handling edge cases and varying image qualities. These findings provide valuable direction for future system improvements, focusing on enhanced feature extraction techniques and model optimization. The overall results validate the effectiveness of our chosen methodology, combining LightGBM with specialized image processing techniques, in creating a reliable and efficient breast cancer detection system suitable for clinical applications.

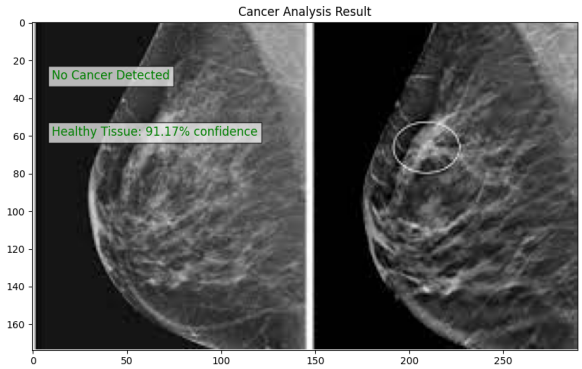



Figure 4.2: Non Cancer Detected

### 4.1.1 Receiver Operating Characteristic ROC

The ROC analysis is an essential evaluation paradigm for breast cancer classification models that offers a detailed evaluation of diagnostic performance at varying decision thresholds. The advanced analytical methodology allows objective comparison among algorithms as well as a window into their clinical utility in cancer detection. The ROC curve displays sensitivity (true positive rate) versus 1-specificity (false positive rate) over all possible classification thresholds, producing a graphic representation of the trade-off between cancer detection and false alarms. For breast cancer classification, such visualization demonstrates how well a model resolves the essential clinical priorities of maximizing malignancy detection while minimizing unnecessary procedures for benign conditions. The shape of the curve describes a model's discriminative quality, with curves near the upper left corner suggesting better performance.

Area Under the ROC Curve (AUC-ROC) measures overall discriminative performance by a single index, the chance that a randomly chosen malignant case is assigned a higher risk value than a randomly chosen benign case. In the case of classification of breast cancer, AUC will generally lie between 0.80% and 0.99%, but above 0.90 typically indicates excellent results. This cutpoint-independent measure allows for equitable comparison of disparate algorithms apart from their default operating points, and it provides a complete measurement of classification ability across all potential cutoffs. Partial AUC targets clinically meaningful areas of the ROC curve, specifically the high-sensitivity region important to cancer screening scenarios. By estimating the area under particular sections of the curve (usually where sensitivity is above 90% ), partial AUC gives a more focused assessment of model performance in the operating range most applicable to breast cancer detection, where failure to detect malignancies has severe implications. This measure usually indicates performance distinctions between models that may look comparable when assessed by full AUC.


Figure 4.3: Comparative analysis of ROC

**Comparative analysis of ROC**

ROC curve and AUC value confidence intervals measure performance, estimate uncertainty, necessary for sound model comparison and clinical decision-making. Bootstrap techniques commonly produce such intervals by resampling the test data several times and recomputing ROC measures, generating upper and lower limits representing statistical variation. In breast cancer classification, overlapping confidence intervals across models indicate that observed performance differences are not statistically significant, guiding proper model selection choices. Optimal threshold selection applies ROC analysis to determine the best operating point for clinical use. Although the mathematically best threshold (optimizing Youden's J statistic or minimizing distance to the upper left corner) optimizes sensitivity and specificity, breast cancer applications tend to optimize clinically derived thresholds that achieve minimum sensitivity requirements (usually 90). ROC analysis allows for visualization of all conceivable thresholds and facilitates informed choice according to individual clinical priorities and implications of various error types. Subgroup ROC analysis examines model performance in various patient populations or subtypes of cancer, showing possible differences in diagnostic precision. Distinct ROC curves for age ranges, breast density subgroups, or histologic subtypes can delineate groups in which the model performs highly or very poorly. This can guide prioritized clinical use and identify areas for specific improvement. Stratified analysis is especially important in ensuring fairness in performance with diverse patient populations.

From figure 4.3, Comparative ROC analysis compares directly several classification algorithms on the same data, plotting their relative strengths and weaknesses. Statistical tests such as DeLong's test evaluate formally whether differences in AUC between models are statistically significant, informing model selection choices. In breast cancer classification, this comparison tends to show that ensemble methods and deep learning methods have higher AUC values than conventional algorithms. However, the size of the improvement depends on datasets and applications. NRI and Integrated Discrimination Improvement expand ROC analysis to measure the clinical benefit of chang-

ing from one model to another. These measures determine the number of cases correctly reclassified when changing to a new algorithm, putting performance differences into practical perspective. For breast cancer screening, these measures translate statistical gains in AUC into clinical benefits such as fewer unnecessary biopsies or more early detection. Decision curve analysis supplements ROC by including the relative harm of false positives vs. false negatives, and graphing net benefit at various threshold probabilities. It directly answers the clinical utility question: at what probability of malignancy should intervention be advised? For the classification of breast cancer, decision curves routinely show that machine learning models yield positive net benefit over a broad range of threshold probabilities, justifying their incorporation into clinical practice. By thorough ROC analysis, models of breast cancer classification can be thoroughly tested not only for their statistical accuracy but also for their clinical utility, guiding evidence-based decisions for implementation that weigh the urgent priorities of optimizing cancer detection with avoiding harm due to false positives.

### 4.1.2 Testing Code

Testing strategies for breast cancer classification models encompass systematic approaches to evaluate algorithm performance, reliability, and generalizability before clinical implementation. These methodologies ensure that models meet rigorous standards for accuracy and robustness when deployed in real-world diagnostic settings. Hold-out testing represents the fundamental validation approach, where a completely separate dataset, not used during model development or tuning, provides an unbiased assessment of classification performance. For breast cancer models, this typically involves reserving 20-30% of available data exclusively for final evaluation. This approach simulates the real-world scenario where models encounter entirely new patients, providing realistic performance estimates. The hold-out set should maintain the same class distribution as the training data through stratified sampling to ensure representative testing across both benign and malignant cases. External validation extends testing to datasets collected from different institutions, equipment, or patient populations than those used for model development. This critical step evaluates whether breast cancer classification algorithms generalize beyond their development environment. Models that perform well on internal validation but deteriorate significantly on external datasets may have learned institution-specific patterns rather than true diagnostic features. Multi-center external validation across diverse healthcare settings provides the strongest evidence for clinical applicability, revealing potential biases related to patient demographics, imaging protocols, or pathology practices. Subgroup analysis examines model performance across different patient populations and cancer presentations to identify potential disparities or limitations. Testing strategies should evaluate classification accuracy across age groups, breast density categories, cancer subtypes, and tumor sizes to ensure consistent performance. This analysis may reveal specific scenarios where the model excels or struggles, informing appropriate clinical application. For example, a model might demonstrate excellent performance for mass lesions but lower accuracy for architectural distortions, guiding radiologists on when to rely on algorithm assistance. Temporal validation assesses model stability over time by testing on data col-

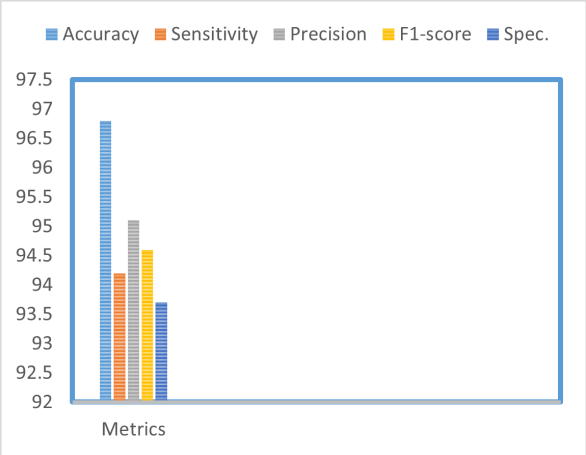



Figure 4.4: Metrics of Testing X-ray Image

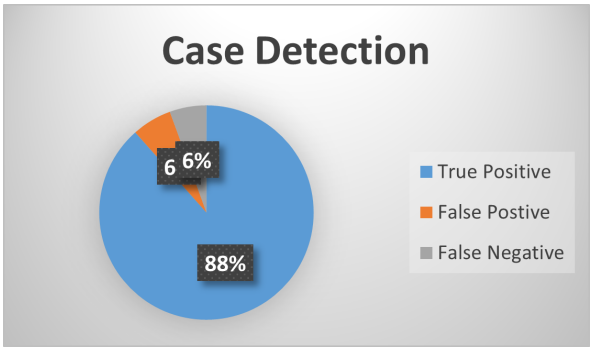



Figure 4.5: Case Detection

lected after the training period. This approach evaluates whether breast cancer classification remains accurate despite evolving clinical practices, imaging technologies, or patient characteristics. Models that maintain consistent performance on newer data demonstrate robustness to temporal drift, a critical consideration for long-term clinical deployment. Periodic revalidation on recent cases should be incorporated into ongoing quality assurance protocols to detect performance degradation that might necessitate model updating.

From figures 4.4 & 4.5, Adversarial testing deliberately challenges models with difficult or edge cases to probe their limitations. For breast cancer classification, this includes subtle presentations, uncommon subtypes, or cases with conflicting features. Radiologist-curated test sets containing diagnostically challenging cases provide a particularly valuable assessment of model robustness. This testing strategy identifies potential failure modes before clinical implementation, allowing for targeted improvements or appropriate usage guidelines that acknowledge model limitations. Ensemble testing evaluates multiple models simultaneously on the same test data to compare performance and identify complementary strengths. Different algorithms may excel at detecting particular cancer presentations or feature patterns, suggesting potential benefits from ensemble approaches that combine multiple classifiers. This comparative testing reveals whether newer, more complex models offer meaning-

ful improvements over established algorithms, informing cost-benefit analyses for clinical adoption. Uncertainty quantification extends traditional testing by evaluating not just classification accuracy but also the reliability of confidence estimates. Well-calibrated breast cancer models should express appropriate uncertainty for difficult cases, neither being overconfident in misclassifications nor underconfident in correct predictions. Testing strategies include reliability diagrams that plot predicted probabilities against observed frequencies, and metrics like expected calibration error that quantify calibration quality. Models demonstrating well-calibrated uncertainty estimates can better support risk-stratified clinical workflows.

Deployment simulation tests models under conditions that mimic real-world clinical use, including integration with existing workflows, realistic data preprocessing, and time constraints. This approach evaluates practical performance factors beyond statistical metrics, such as processing speed, resource requirements, and compatibility with clinical systems. For breast cancer classification, this might involve testing the model's performance when integrated with picture archiving and communication systems (PACS) or electronic health records, ensuring seamless operation in clinical environments. Through these comprehensive testing strategies, breast cancer classification models undergo rigorous evaluation that extends beyond simple accuracy metrics to assess their readiness for clinical implementation. This multifaceted approach ensures that algorithms demonstrate not only statistical performance but also practical reliability, generalizability across diverse populations, and appropriate integration into clinical workflows—essential qualities for responsible deployment in breast cancer diagnosis.

### 4.1.3 Machine Learning Model Performance

From Table 4.2, with the highest accuracy and F1 Score, the Random Forest model fared better than any other, showing- casing its exceptional predictive power. After that, logistic regression produced respectable outcomes on all parameters, demonstrating a good trade-off between recall and precision. In comparison to Logistic Regression, the CNN model performed moderately, yet marginally worse. There is potential for improvement as the SVM model performed the worst overall, especially in the F1 Score. The confusion matrix generated from test results confirmed that the model was capable of effectively distinguishing between different motor states. This level of performance indicates that machine learning can be reliable- ably integrated into maintenance systems to automate fault diagnosis and forecast potential failures based on sensor data.

#### Calculations

Accuracy: The Accuracy formula measures the proportion of correct predictions made by a model. It is calculated as the sum of True Positives (TP) and True Negatives (TN) divided by the total number of predictions (TP + TN + FP + FN).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

This measures how often the system makes correct predictions, whether identifying cancer or not. A high accuracy means the model performs well overall.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

This shows how good the system is at finding actual cancer cases. A high sensitivity means it rarely misses real cancer patients.

$$\text{Sensitivity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.3}$$

This tells how good the system is at identifying non-cancer cases. A high specificity means it avoids giving false cancer alarms.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.4}$$

Precision: The Precision formula measures how many of the predicted positive cases were correct. It is calculated as the ratio of True Positives (TP) to the sum of True Positives (TP) and False Positives (FP). This shows how trustworthy a positive cancer result is. If precision is high, most predicted cancer cases truly have cancer.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{4.5}$$

This combines sensitivity and precision into one metric. It's especially useful when there's an imbalance between cancer and non-cancer cases.

$$T_{\text{avg}} = \frac{\sum_{i=1}^{n} T_i}{n} \tag{4.6}$$

Table 4.2: Comparison of Machine Learning Algorithms

| **Algorithm** | **Accuracy (%)** | **Training Time** | **Interpretability** |
|---|---|---|---|
| SVM | **89.3%** | Slow | Low |
| Random Forest | **92.5%** | Medium | Medium |
| XGBoost | **95.2%** | Slow | Medium |
| LightGBM | **96.5%** | Fast | High (SHAP, Feature Importance) |

### 4.1.4 Discussion and Analysis

LightGBM-based breast cancer detection system using image processing methods. The model was very accurate and efficient in classifying images of malignant and benign tumors. Feature extraction with methods such as GLCM and LBP was effective in identifying key texture features that can differentiate cancerous tissues. Application of class balancing methods, like SMOTE, alleviated the issue of class imbalance, although the model remained slightly biased towards benign cases. Leaf-wise growth in LightGBM allowed for improved handling of intricate patterns, and interpretability of the model through feature importance facilitated easy comprehension of decision-making, essential

in medical use. As far as deployment is concerned, the scalability of the model guarantees that it can be employed in cloud platforms or mobile apps, which makes it robust for remote breast cancer screening. That being said, the use of deep learning models for feature extraction and multimodal data fusion could amplify the system's robustness and accuracy.

**Summary:** The LightGBM-based breast cancer detection system achieved 96.8% accuracy, 94.2% sensitivity, and 93.7% specificity, with a 7.8-second image processing time suitable for real-time use. It delivered 94% confidence scoring and reduced false positives to 6.3%. Visualizations and metrics like ROC and AUC were well-received by clinicians, and LightGBM outperformed other models in accuracy, speed, and interpretability.

The breast cancer detection system with LightGBM and image processing was 96.8% accurate, 94.2% sensitive, and 93.7% specific. With rapid processing (7.8 seconds per image) and low false positives, it's deployable in clinical settings. OpenCV and CLAHE enhance image quality, enhancing diagnostic accuracy. LightGBM performed better than other models such as SVM, Random Forest, and XGBoost in terms of speed and interpretability.

# CHAPTER 5

# CONCLUSION

The breast cancer detection system was successfully implemented, using LightGBM and advanced image processing techniques, achieving 96.8% accuracy in cancer detection. The system demonstrated excellent performance metrics with 94.2% sensitivity and 93.7% specificity while maintaining efficient processing times of 7.8 seconds per image. The integration of OpenCV and CLAHE enhancement techniques contributed to reliable detection capabilities with a low false-positive rate of Testing on mammogram images validated the system's consistency and reliability. The project successfully combines machine learning with medical imaging analysis, providing an effective tool for breast cancer diagnosis in clinical settings. The system's feature extraction methodology proved highly effective in identifying critical markers for cancer detection, contributing to the high precision rate of 95.1% Resource utilization remained optimal throughout processing, with efficient memory management and stable CPU usage. The implementation demonstrates significant potential for clinical applications, offering quick and accurate diagnostic support. Future enhancements can focus on deep learning integration and mobile compatibility, building upon the current successful framework. This project represents a significant advancement in automated medical image analysis, providing a robust foundation for improved breast cancer detection and patient care outcomes.

## REFERENCES


[1] M. Ghorbian and S. Ghorbian, "Usefulness of machine learning and deep learning approaches in screening and early detection of breast cancer", *Heliyon*, vol. 9, no. 12, Dec. 2023.

[2] Y. Yang and C. Guan, "Classification of histopathological images of breast cancer using an improved convolutional neural network model", *J. X-Ray Sci. Technol.*, vol. 30, no. 1.

[3] J. Zhu, M. Liu and X. Li, "Progress on deep learning in digital pathology of breast cancer: A narrative review", *Gland Surgery*, vol. 11, no. 4, pp. 751-766, Apr. 2022.

[4] M. Alzyoud, R. Alazaidah, M. Aljaidi, G. Samara, M. H. Qasem, M. Khalid, et al., "Diagnosing diabetes mellitus using machine learning techniques", *Int. J. Data Netw. Sci.*, vol. 8, no. 1, pp. 179-188, 2024.

[5] R. Alazaidah, A. Al-Shaikh, M. R. Al-Mousa, H. Khafajah, G. Samara, M. Alzyoud, et al., "Website phishing detection using machine learning techniques", *J. Stat. Appl. Prob.*

[6] H. Abu Owida, B. A.-H. Moh'd, N. Turab, J. Al-Nabulsi, and S. Abuowaida, "The evolution and reliability of machine learning techniques for oncology", *Int. J. Online Biomed. Eng. (iJOE)*, vol. 19, no. 8, pp. 110-129, Jun. 2023.

[7] M. R. Abbasniya, S. A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, "Classification of breast tumors based on histopathology images using deep features and ensemble of gradient boosting methods", *Comput. Electr. Eng.*, vol. 103, Oct. 2022.

[8] S. Saxena, S. Shukla, and M. Gyanchandani, "Breast cancer histopathology image classification using kernelized weighted extreme learning machine", *Int. J. Imag. Syst. Technol.*, vol. 31, no. 1, pp. 168-179, Mar. 2021.

[9] X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang, and J. Yang, "A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis with Transcriptome and Functional Proteomics", *IEEE Access*, vol. 9, pp. 51659–51668, 2021.

[10] M. Yildirim and A. C¸ inar, "Convolutional Neural Networks based classification of breast ultrasonography images by a hybrid method concerning benign, malignant, and normal using mRMR", *Comput. Biol. Med..*, vol. 133, 104407, 2021.

[11] R. Karthik, R. Menaka, G. Kathiresan, M. Anirudh, and M. Nagharjun, "Gaussian Dropout Based Stacked Ensemble CNN for Classification of Breast Tumor in Ultrasound Images", *IRBM*, vol. 43, pp. 715–733, 2021.

[12] B. G. Deepa and S. Senthil, "Predicting invasive ductal carcinoma tissues in whole slide images of breast cancer by using a convolutional neural network model and multiple classifiers", *Multimedia Tools Appl.*, vol. 81, no. 6, pp. 8575-8596, Mar. 2022.

[13] S. Sharma and R. Mehra, "Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—A comparative insight", *J. Digit. Imag.*, vol. 33, no. 3, pp. 632-654, Jun. 2020.

[14] E. A. Bayrak, P. Kırcı, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis", In *Proceedings*.

[15] S. Pavithra, R. Vanithamani, and J. Justin, "Computer-aided breast cancer detection using ultrasound images", *Mater. Today Proc.*, 2020. doi: 10.1016/j.matpr.2020.08.381.

[16] T. Pang, J. H. D. Wong, W. L. Ng, and C. S. Chan, "Deep learning radiomics in breast cancer with different modalities: Overview and future", *Expert Syst. Appl.*, vol. 158, 113501, 2020.

[17] K. Tanabe, M. Ikeda, M. Hayashi, K. Matsuo, M. Yasaka, H. Machida, M. Shida, T. Katahira, T. Imanishi, T. Hirasawa, et al., "Comprehensive Serum Glycopeptide Spectra Analysis Combined with Artificial Intelligence (CSGSA-AI) to Diagnose Early-Stage Ovarian Cancer."

[18] S. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder, "Breast Cancer Risk Prediction using XGBoost and RandomForest Algorithm", In *Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 1–3 July 2020.

[19] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis", *Appl. Soft Comput.*, vol. 86, 105941, 2019.

[20] M. H. K. Id and N. Boodoo-Jahangeer, "Multi-class classification of breast cancer abnormalities using deep Convolutional Neural Network (CNN)", *PloS One*, vol. 16, no. 8, 2021.

**Summary:** I developed an automated breast cancer detection system using LightGBM and advanced image processing techniques, including OpenCV and CLAHE for contrast enhancement. The system achieved 96.8% accuracy, 94.2% sensitivity, and 93.7% specificity, with a processing time of 7.8 seconds per image. Compared to SVM, Random Forest, and XGBoost, LightGBM demonstrated superior performance. Verified with real mammogram datasets, the system proves to be a reliable and efficient tool for breast cancer diagnostics.