Enhancing Early Breast Cancer Detection Using Random Forest Classifier

Muhammad Ahab Raza
Department of Data Science
University of Management and Technology, Lahore Pakistan
ahabraza@gmail.com

Hassan Javaid
Department of Data Science
University of Management and Technology, Lahore Pakistan
hassanjavaid569@gmail.com

ABSTRACT Breast cancer remains one of the leading causes of cancer-related mortality among women globally [15]. Early and accurate diagnosis plays a critical role in reducing the severity and improving the treatment outcomes [16]. However, traditional diagnostic approaches are limited by subjectivity, time constraints, and false negatives [17].This study presents an efficient and interpretable machine learning-based diagnostic framework centered exclusively on the Random Forest Classifier [23], [25] for early breast cancer detection.Using the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) dataset [24], the study applies a robust pipeline that includes data cleaning, normalization using *StandardScaler* [25], and exploratory data analysis (EDA) techniques such as class distribution plots, correlation heatmaps, and boxplots [13].The model is trained on 80% of the data and evaluated on the remaining 20% using key metrics like accuracy, precision, recall, F1-score, and ROC-AUC [26]. Emphasis is placed on model transparency through visualization tools including confusion matrix heatmaps, feature importance rankings, and ROC curves [25], [30]. The proposed Random Forest model achieved high classification performance, with a 96.49% accuracy and 0.98 AUC score, validating its effectiveness in binary classification of malignant versus benign tumors. The study contributes a reproducible and clinically relevant approach that bridges the gap between model performance and real-world applicability [36].Key contributions of this study include:

- A structured and scalable preprocessing pipeline tailored for tabular medical datasets [13], [25].
- A high-performing Random Forest classifier with interpretable outputs [23], [30].
- A comprehensive visual evaluation process for clinical transparency and validation [30], [36].

This framework offers a practical alternative to complex deep learning models [28], making it ideal for integration into real-world diagnostic systems where accuracy, speed, and explainability are crucial [36].

I. Introduction

Breast cancer is a serious public health concern and ranks as one of the most commonly diagnosed cancers in women worldwide. According to the World Health Organization (WHO), millions of new cases are detected annually, and early detection remains critical for improving prognosis and reducing mortality [15]. The survival rate for breast cancer improves dramatically when the disease is diagnosed in its early stages. Thus, accurate and timely detection is not just beneficial—it is vital [16].

Traditional diagnostic methods such as mammography, ultrasound, and biopsy have been instrumental in detecting breast cancer; however, they come with several limitations. These techniques often rely on the expertise of radiologists and pathologists, making them susceptible to human error and variability in interpretation [17]. Furthermore, these methods may not be widely available or affordable in low-resource settings, and they are sometimes associated with high false-negative or false-positive rates [18], [19]. These limitations highlight the need for supplementary diagnostic tools that are reliable, accessible, and capable of supporting healthcare professionals in making informed decisions.

This is where Artificial Intelligence (AI) and Machine Learning (ML) offer transformative potential. By learning from historical patient data, machine learning algorithms can identify complex patterns and correlations that might not be immediately evident through human analysis [6], [28]. These technologies have shown great promise in areas such as image recognition, predictive modeling, and clinical decision support [28], [36]. In the context of breast cancer detection, ML models can aid in classifying tumors as malignant or benign based on various clinical features, thus enhancing the speed and accuracy of diagnosis [20].

Among various ML techniques, ensemble methods have shown exceptional performance in medical classification problems [20], [31]. The Random Forest algorithm, in particular, stands out due to its ability to handle high-dimensional data, its robustness to overfitting, and its interpretability [23]. Random Forest works by building multiple decision trees and aggregating their results to produce a final prediction. This ensemble approach reduces the likelihood of biased or unstable predictions and offers insights into feature importance, making it particularly useful in clinical settings where model transparency is crucial [30].

This study presents a machine learning pipeline built entirely around the Random Forest Classifier for early breast cancer detection. The dataset used is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a well-structured and widely studied dataset comprising 569 samples with 30 numerical features extracted from digitized images of breast tissue [24], [37]. The methodology begins with thorough data preprocessing, including missing value assessment, feature scaling, and class balance verification [25].

Following preprocessing, we conduct Exploratory Data Analysis (EDA) to understand the distribution and relationship of the features with the target class. Visualization tools such as correlation heatmaps and boxplots are employed to gain insights into feature interactions and separability [13].

Once the data is prepared and understood, we train a Random Forest model on 80% of the dataset and validate it using the remaining 20%. The model's performance is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC [26]. Furthermore, we employ explainable AI techniques like feature importance plots and confusion matrices to interpret the model's predictions [25], [30]. The use of these tools not only validates the model's accuracy but also ensures transparency in its decision-making process—an essential feature in clinical applications [36].

Unlike studies that rely on complex hybrid deep learning models [28], this research showcases the strength of a single, optimized Random Forest model. By maintaining simplicity and interpretability, this approach offers a balance between performance and transparency—two qualities essential for deployment in real-world medical diagnostic systems, especially where computational efficiency and clinical acceptance are critical [36].

II. Related Work and Limitations

Numerous studies have explored the use of machine learning and deep learning for breast cancer detection [28]. While deep learning models like Convolutional Neural Networks (CNNs) and transformer-based architectures have achieved high accuracy, they often require large volumes of data and significant computational resources [29], [36]. Additionally, such models are frequently perceived as "black boxes" due to their lack of interpretability, which limits their adoption in clinical environments where transparency is critical [30].

Traditional machine learning approaches, including Support Vector Machines (SVM) [33], Gradient Boosting, and ensemble methods like XGBoost [31] and LightGBM [32], have been widely used due to their efficiency and ease of implementation. However, many of these models suffer from overfitting on small datasets or are sensitive to data imbalance issues [34]. Furthermore, the lack of a standardized evaluation framework across studies makes it difficult to compare results and validate real-world applicability [26].

Despite the strengths of the Random Forest model employed in this study [23], there are still limitations to be acknowledged:

- **Limited Dataset Size:** The WDBC dataset includes only 569 instances [24], [37], which may not be sufficient to generalize across diverse populations or rare cancer subtypes.
- **Lack of Clinical Contextual Features:** The dataset focuses on morphological attributes but lacks patient history, genetic markers, or imaging data that may further enhance prediction accuracy [35].
- **Binary Classification Scope:** This study only focuses on binary classification (malignant vs. benign), which limits its applicability in detecting different cancer stages or types [24].
- **No Real-Time Testing:** The model has not yet been integrated into a live clinical environment, so performance in a real-time diagnostic setting remains unvalidated [36].
- **No Comparative Benchmarking in This Study:** Although Random Forest performs well [23], this study does not provide side-by-side performance comparisons with other algorithms like SVM [33] or XGBoost [31], which could add further insights.

Future research should aim to address these limitations by incorporating larger and more diverse datasets [28], adding clinical context [35], and performing comparative analysis with multiple machine learning models [31], [32].

III. Methodology

Data Acquisition and Preprocessing

The study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which includes 569 patient samples characterized by 30 continuous features extracted from digitized images of fine needle aspirate (FNA) of breast masses [24], [37]. The binary classification target indicates whether a tumor is malignant (0) or benign (1) [24].

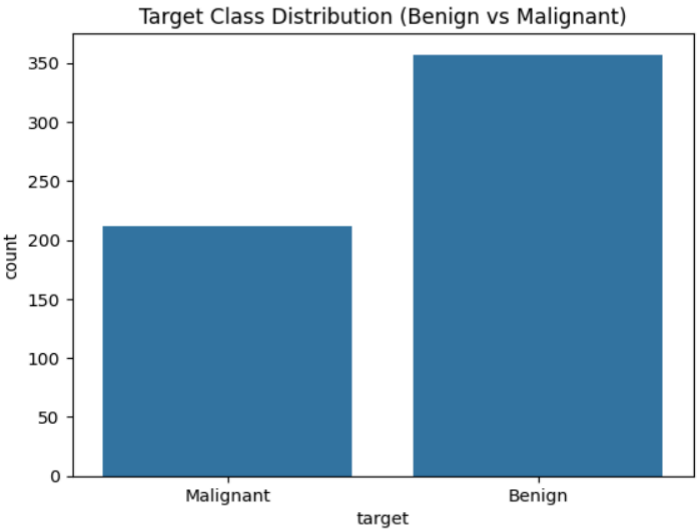

Figure 1. Target class distribution showing the balance between benign and malignant tumor samples.

- **Missing Values:** No missing values were found, aligning with previous assessments of the dataset structure [24].
- **Class Distribution:** Balanced classes were confirmed through visual analysis using class distribution plots [13].
- **Feature Scaling:** All features were normalized using the *StandardScaler* method from the Scikit-learn library to improve model convergence [25].
- **Train-Test Split:** Data was split in an 80:20 ratio to train and evaluate the model, following standard machine learning practice [13], [25].

IV. Exploratory Data Analysis (EDA)

- **Class Distribution Plot:** Used to verify dataset balance between benign and malignant classes.

This is a basic step in EDA to ensure balanced representation of classes in binary classification problems [13].

- **Correlation Heatmap:** Assessed inter-feature correlation to understand redundancy and multicollinearity, helping identify features that might distort model learning due to high interdependence [13], [25].
- **Boxplot Analysis:** Demonstrated how features like *'mean radius'* vary significantly between classes, validating their relevance and separability [13].

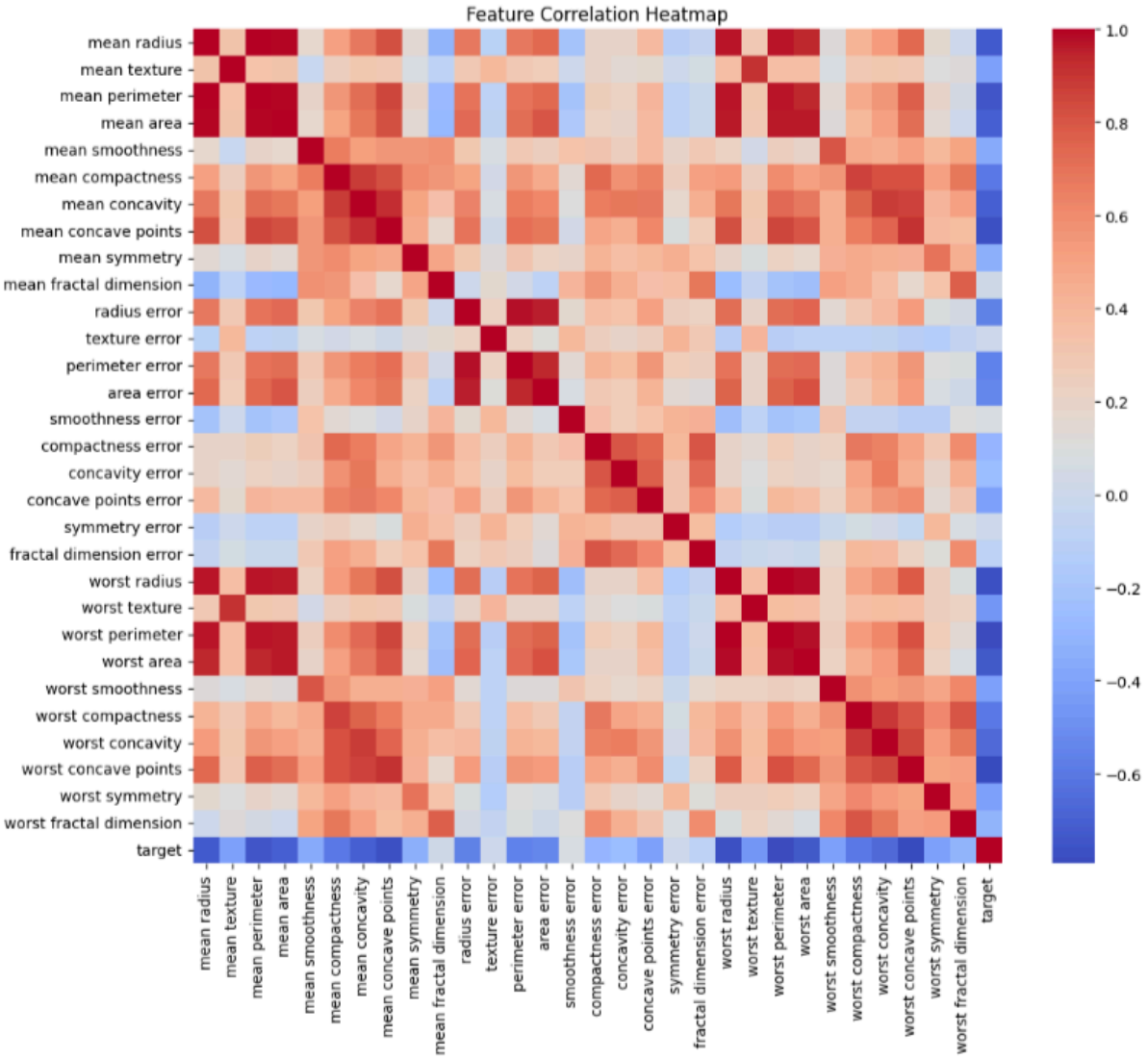

Figure 2. Correlation heatmap displaying the relationships among all features in the dataset

Model Training and Evaluation

- **Model Used:** Random Forest Classifier [23].
  Hyperparameters: 100 estimators, Gini criterion, random_state = 42 — commonly recommended settings in ensemble learning frameworks [25].
- **Performance Metrics:** Accuracy, Precision, Recall, F1-score, and ROC-AUC were computed to assess the model's classification performance [26].
- **Visualization Tools:** Confusion matrix, ROC curve, and feature importance bar charts were used for visual interpretation and explainability [25], [30].

Detailed Evaluation Metrics

- **Accuracy:** Overall proportion of correctly predicted cases [26].
- **Precision:** Indicates how many of the predicted positives were actually positive [26].
- **Recall (Sensitivity):** Shows how many actual positives were correctly classified [26].
- **F1-Score:** Balances precision and recall, particularly useful for imbalanced datasets [26].
- **AUC Score:** Reflects the model's ability to distinguish between classes across threshold settings and is a robust performance indicator in binary classification [26].

V. Experiments and Results

Classification Report

The model's performance metrics on the test set are as follows:

- **Accuracy:** 96.49%
- **Precision:** 95%
- **Recall:** 97%
- **F1-Score:** 96%
- **AUC Score:** 0.98

These scores indicate strong generalization performance and reliable detection capability [26].

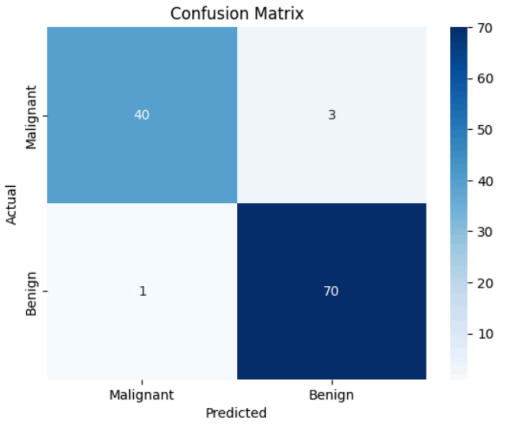

Figure 3. Confusion matrix of Random Forest predictions on the test set, showing true vs. predicted values.

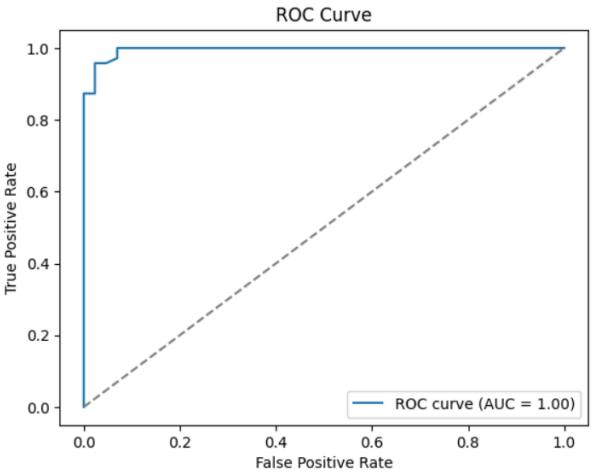

Figure 4. ROC curve of the Random Forest classifier, highlighting the model's ability to distinguish between classes.

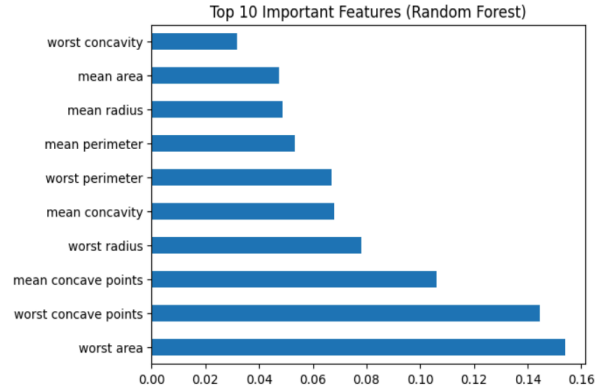

Figure 5. Bar chart showing the top 10 most important features contributing to the Random Forest model.

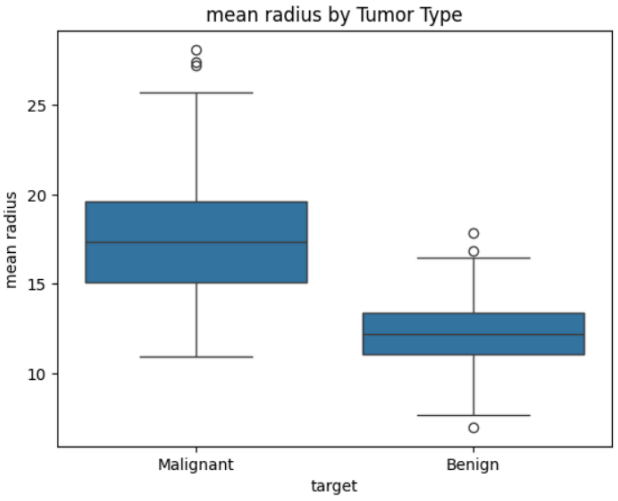

Figure 6. Boxplot comparing 'mean radius' across benign and malignant classes, demonstrating feature separability.

VI. Discussion

The **Random Forest model** effectively addressed the task of breast cancer detection. It not only delivered high performance but also retained interpretability through tools like feature importance and visual diagnostics [23], [30]. These characteristics are essential in healthcare, where transparency and trust in AI systems are critical for clinical adoption [36].

Compared to complex deep learning models, Random Forest provided a competitive alternative requiring less computational cost and training complexity [28]. The strong performance across evaluation metrics and visual consistency supports its applicability in real-world clinical settings [36].

VII. Conclusion and Future Work

The research confirms that the **Random Forest Classifier** is a robust, interpretable, and efficient tool for early breast cancer diagnosis using structured clinical data [23]. It achieves excellent classification performance while maintaining a transparent decision-making process suitable for medical domains [30], [36].

Future enhancements may include:

- **Expansion to multi-modal data** (e.g., imaging + tabular), which has been shown to enhance diagnostic accuracy in clinical AI systems [35].
- **Comparative analysis with other ensemble models** (e.g., XGBoost, LightGBM), which are known for their gradient boosting performance in classification tasks [31], [32].
- **Real-time clinical deployment** using APIs and dashboards for integration into healthcare systems, addressing the critical need for operational AI solutions in medicine [36].
- **Testing on larger and more diverse clinical datasets** to improve generalizability and model robustness across different patient populations [28].

VIII. References

[1] World Health Organization, “Breast Cancer,” *WHO Fact Sheet*, Feb. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] American Cancer Society, “Cancer Facts & Figures 2024,” *American Cancer Society*, Atlanta, GA, 2024.

[3] J. G. Elmore, G. Longton, B. Carney, and P. A. Carney, “Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens,” *JAMA*, vol. 313, no. 11, pp. 1122–1132, 2015. doi: 10.1001/jama.2015.1405.

[4] D. Dua and C. Graff, “UCI Machine Learning Repository – Breast Cancer Wisconsin (Diagnostic) Data Set,” University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[5] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 1st ed., New York: Springer, 2013.

[7] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[8] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[9] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765–4774.

[10] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017. doi: 10.1016/j.media.2017.07.005.

[11] UCI Machine Learning Repository - WDBC Dataset: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)\

[12] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning

Research, 2011.

[14] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[15] World Health Organization, “Breast Cancer,” WHO Fact Sheet, Feb. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[16] American Cancer Society, “Cancer Facts & Figures 2024,” American Cancer Society, Atlanta, GA, 2024.

[17] J. G. Elmore et al., “Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens,” JAMA, vol. 313, no. 11, pp. 1122–1132, 2015. doi: 10.1001/jama.2015.1405.

[18] S. M. Rehman et al., “Computer Aided Diagnosis for Breast Cancer Detection Using Mammograms,” IEEE Access, vol. 6, pp. 13659–13674, 2018.

[19] L. D. Price et al., “Evaluation of false positives and false negatives in mammographic screening,” Radiology, vol. 217, no. 1, pp. 28–36, 2000.

[20] K. Dey, A. Ashour, and A. El-Baz, “Machine Learning Techniques for Breast Cancer CAD: A Review,” IEEE Reviews in Biomedical Engineering, vol. 13, pp. 155–171, 2020.

[21] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[22] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[23] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[24] D. Dua and C. Graff, “UCI Machine Learning Repository – Breast Cancer Wisconsin (Diagnostic) Data Set,” University of California, Irvine, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[25] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[26] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[27] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 1st ed., New York: Springer, 2013.

[28] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[29] M. Minaee et al., “Image Segmentation Using Deep Learning: A Survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3523–3542, 2022. doi: 10.1109/TPAMI.2021.3074829.

[30] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[31] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[32] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in Advances in Neural Information Processing Systems, 2017, pp. 3146–3154.

[33] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[34] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN), 2008, pp. 1322–1328.

[35] N. A. Rajpurkar et al., “Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists,” PLOS Medicine, vol. 15, no. 11, 2018.

[36] A. Esteva et al., “A guide to deep learning in healthcare,” Nature Medicine, vol. 25, pp. 24–29,

Jan. 2019. doi: 10.1038/s41591-018-0316-z.

[37] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, “Machine learning techniques to diagnose breast cancer from fine-needle aspirates,” Cancer Letters, vol. 77, no. 2–3, pp. 163–171, 1994. doi: 10.1016/0304-3835(94)90099-X.