Breast Cancer Detection Using Machine Learning

Vasi Hussain Sayed[1], Atharva Raut[2], Hemlata Patil[3], Niharika Ranjerla[4], Prof. Chandrakant Rane[5]

[1,2,3,4,5] *COMPUTER ENGINEERING, INDALA COLLEGE OF ENGINEERING, KALYAN*

---***---

Abstract

Breast cancer continues to be a major health challenge for women around the world, with countless new cases emerging each year. Catching the disease early makes a significant difference in treatment success and survival rates. However, traditional diagnostic tools—though helpful—can fall short due to time delays, human error, or limited precision. In this study, we looked into how machine learning (ML) techniques can support tumor classification by drawing insights from both clinical records and medical images. We experimented with algorithms like Support Vector Machines (SVM), Random Forest to access their performance in classifying tumors. Using the CBIS-DDSM: Breast Cancer Image Dataset as a training base, we developed a system that could potentially support real-time clinical decisions. The dataset was split using train_test_split to ensure unbiased evaluation. The SVM was fine-tuned using cross-validation to find the best parameters, significantly enhancing accuracy. Results indicated strong classification performance in terms of precision, recall, and F1-score. This highlights SVM's potential in medical image analysis. Our goal wasn't just to showcase the technology but to demonstrate how ML could serve as a practical tool in modern diagnostics, offering faster, more consistent, and scalable solutions in cancer detection.

1.Introduction

Breast cancer has become an increasingly urgent concern, particularly for women over 40. Even with improvements in treatments and therapies, delays in diagnosis still contribute to high mortality rates. While current diagnostic approaches like mammograms, biopsies, and clinical exams are essential, they are not always reliable on their own. Often, these methods are affected by human judgment or limited by the available technology.

That's where machine learning (ML) comes in. As a part of artificial intelligence (AI), ML can be trained to recognize patterns from vast amounts of past data and make predictions based on that learning. In this project, we applied different ML models to streamline breast cancer detection and make the classification process more accurate. Specifically, we used the CBIS-DDSM: Breast Cancer Image Dataset, a widely used referenced dataset in this field, to train and test our models By doing so, we aimed to evaluate how effectively ML can distinguish between benign and malignant tumors. Beyond just saving time, these intelligent systems could provide a valuable second opinion to oncologists and help improve early diagnosis, potentially leading to better outcomes for patients. Our system design emphasizes simplicity, reproducibility, and effectiveness, ensuring it can be adapted in real-world clinical environments.

2.Literature Review

The cause of Breast Cancer includes changes and mutations in DNA. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common. Wang, D.; Zhang and Y.-H Huang (2018) et al. [1]

Dongdong Sun et al. have proposed a deep learning (DL) method named D-SVM for the prediction of human breast cancer prognosis. The algorithm effectively learned hierarchical and abstract representation from raw input data and successfully integrated traditional classification method [2].

Ch. Shravya et al. focuses a relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN) on a particular dataset. [3].

Mariam et. al. [4] uses two different classifiers namely Naive Bayes and K Nearest Neighbors for breast cancer classification on comparing accuracy using cross-validation and KNN achieved that 97.51% accuracy with lowest error rate then Naive Bayes Classifier 96.19% accuracy.

Aruna et al. [5] uses three different classifiers namely Naive Bayes, Support Vector Machine, and Decision Tree to classify a Wisconsin breast cancer dataset and got the best outcome by utilizing a support vector machine with an accuracy score of 96.99%.

3.Methodologies Used

For this project, we used the CBIS-DDSM breast cancer image dataset, which contains curated mammography images labelled with important diagnostic information. The dataset includes various types of breast lesions, such as calcifications and masses, along with labels indicating whether they are benign or malignant. Since this is an image dataset, our workflow started with image processing rather than working directly with numerical features. We began by preprocessing the images to prepare them for modelling. This involved resizing all the images to a consistent dimension to ensure uniformity and reduce computational load. We also converted the grayscale images to a standardized format, normalized the pixel values to fall between 0 and 1, and applied basic augmentation techniques like flipping and rotation to expand the dataset and improve generalization. Next, we extracted relevant features from the images using traditional image processing techniques. We focused on extracting texture, shape, and intensity-based features using methods such as histogram analysis, edge detection (e.g., Sobel or Canny), and possibly Gray-level co-occurrence matrix (GLCM) features. These features were then flattened and transformed into a format suitable for a machine learning model. We chose a Support Vector Machine (SVM) for classification because it works well with smaller datasets and can effectively handle high-dimensional feature spaces. We tested different kernels—linear, polynomial, and RBF (Radial Basis Function)—and found that the RBF kernel offered the best performance. To optimize the model, we used grid search with cross-validation to fine-tune parameters like C and gamma. Finally, we evaluated the model using common metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results showed that the SVM performed well, especially in distinguishing between benign and malignant cases. By combining image preprocessing, feature extraction, and a carefully tuned SVM, we were able to build a model capable of assisting in early breast cancer detection.

Applying SVM for breast cancer detection lies in its ability to:

- **Increase Diagnostic Accuracy**: SVM has been shown to consistently perform well in terms of classification accuracy, reducing the chances of misdiagnosis and ensuring more reliable tumor detection.

- **Minimize Human Error**: By automating the classification process, SVM reduces the potential for human error, making it an effective decision support tool for doctors.

- **Faster Diagnosis**: Machine learning models, once trained, can rapidly process large amounts of data, leading to quicker decision-making and faster diagnoses, which is crucial for early-stage cancer detection.

- **Non-invasive Detection**: SVM-based systems can classify tumors from images or medical reports without the need for invasive procedures, thus improving patient comfort and reducing healthcare costs.

- **Enhance Screening Programs**: SVM can be used in large-scale screening programs, helping detect breast cancer early in populations, leading to better public health outcomes and reduced mortality rates.

- **Assist in Personalized Medicine**: By identifying specific tumor characteristics, SVM can potentially assist in customizing treatment plans, contributing to the advancement of precision medicine.

- **Scalability and Cost-Effectiveness**: Once developed, machine learning models can be deployed across various healthcare settings, making breast cancer detection more accessible and affordable, particularly in resource-constrained areas.

4.System Architecture

The system we've designed is built in layers, each playing a distinct role in transforming raw data into actionable clinical insights:

- **Data Input Layer:** This component gathers information from various sources, including patient records and imaging data.

- **Preprocessing Module:** Here, the system cleans and prepares the data by normalizing values, filtering out noise, and extracting meaningful features.
- **ML/DL Engine:** At the core, different model such as SVM process the data to detect cancer patterns.
- **Prediction Layer:** After analysis, the system predicts whether a tumour is benign or malignant and assigns a confidence level to that prediction.
- **Explainability Tools:** Modules like SHAP and LIME are integrated to help explain why the model made a particular decision, which is crucial in a medical context.

On the technical side, the system runs in a Python environment using frameworks like TensorFlow and Scikit-learn. We've also built a web interface that allows medical professionals to access predictions easily, with a backend support. This modular design helps the system run in real-time and makes it easier to expand for use in larger hospitals or clinics.

5.Proposed System

Our approach blends machine learning with thoughtful interface design to create a practical breast cancer detection tool. Here's how it's structured:

- Data is collected from multiple inputs—such as mammograms and patient histories.
- Several machine learning models (including SVMs, and Random Forests) are trained to interpret this data.
- We use ensemble methods to combine outputs from these models, boosting accuracy and reducing the risk of false predictions.
- A clinician-facing web dashboard displays the results in an intuitive way.
- Explainable AI components are included so doctors can understand how the system arrived at its conclusions.

This study set out with the following key goals:

- To develop a reliable machine learning system that can help detect breast cancer early.
- To reduce diagnostic errors by making the process more automated and explainable.
- To create a solution that's scalable and interpretable, so it can be trusted by clinicians.
- To ensure fast processing times for real-time use in clinical environments.

Getting clean, reliable input data is crucial for any machine learning model. Here's how we tackled preprocessing:

- **Missing Data:** We either filled in the gaps using statistical methods (mean/median) or removed incomplete records, depending on the scenario.
- **Normalization:** All features were scaled to ensure fair treatment across the dataset.
- **Encoding Categorical Variables:** When needed, we transformed non-numeric data into numeric form using encoding techniques like one-hot encoding.
- **Outlier Detection:** We flagged and removed values that didn't fit expected patterns, which could distort the model.
- **Dimensionality Reduction:** PCA helped us condense the dataset while retaining its essential characteristics.

These steps helped ensure the model had clean, relevant data to work with, which is key to strong performance.

We've also planned for hospital integration, ensuring the system works with standard EHR formats and can be scaled via cloud platforms as needed.

We used a variety of tools and libraries to bring this system to life:

- **Python:** Our primary language due to its readability and strong ecosystem.
- **TensorFlow & Kera's:** For building and fine-tuning neural network models.
- **Scikit-learn:** Useful for implementing traditional ML algorithms and evaluation tools.
- **OpenCV:** Assisted in image preprocessing tasks.
- **Pandas & NumPy:** Handled data transformation and numerical operations.
- **Matplotlib & Seaborn:** Helped us visualize model performance and dataset trends.
- **Jupyter Notebook / VS Code:** Provided flexible environments for code development and experimentation.

These technologies work together to create a detection system that's not just accurate but also maintainable and ready for deployment.

6.Limitations & Research Gaps

Despite the progress in AI-based diagnostics, a few critical challenges remain:

Limitation-

- **Hand-Crafted Feature Ceiling**
  Relying on predefined texture (e.g. GLCM, LBP), shape (e.g. circularity, compactness) and intensity histograms means the model can't learn novel imaging patterns. Once these engineered descriptors reach their expressive limit, adding more won't improve separability.
- **Kernel Complexity & Scalability**
  Non-linear kernels like RBF offer flexibility but incur **O(n³)** training time and **O(n²)** memory for n samples. As the number of mammograms grows, both tuning and deployment become prohibitively expensive.
- **Hyperparameter Sensitivity**
  The regularization parameter C and kernel coefficient γ must be meticulously searched—often with grid or random search across dozens of candidate values. In high-dimensional feature spaces, this factorial explosion of combinations can render even cross-validation infeasible.
- **Imbalanced Lesion Representation**
  Rare but clinically critical cases (e.g. tiny micro-calcifications) make up a small fraction of the dataset. Standard SVM treats all misclassifications equally, so without tailored cost weights or specialized sampling (SMOTE, ADASYN), sensitivity on these minority classes remains low.
- **Support Vector Explosion**
  In complex decision boundaries, the number of support vectors can approach the size of the training set, bloating model size and slowing down inference—undermining real-time or edge-device diagnostics.
- **Noise & Outlier Vulnerability**
  SVM's margin maximization treats mislabeled or noisy points as hard constraints. Outliers lying near the boundary can drastically shift the hyperplane unless soft-margin parameters are heavily tuned.
- **Opaque Non-Linear Decisions**
  While linear SVMs let you inspect weight vectors, non-linear kernels map to infinite-dimensional spaces. Clinicians can't trace back which specific image characteristics yielded a "malignant" decision, hampering trust and adoption.

Research Gap-

- **Improved Feature Selection Strategies**
  Most current SVM pipelines rely on basic image features (e.g. texture, shape), but there's a lack of systematic comparison between feature selection techniques (e.g., Recursive Feature Elimination, L1-based selection, mutual information). Exploring which features are most predictive for different lesion types could significantly boost model performance.
- **Multi-Kernel SVM Approaches**
  The use of a single kernel (e.g., RBF) limits the model's flexibility. There's room to explore **multi-kernel learning**, where different kernels capture different image aspects (e.g., one for shape, another for texture). Research is needed to effectively combine them and assign weights dynamically.
- **Cost-Sensitive and Imbalance-Aware SVM**
  CBIS-DDSM has a class imbalance, especially between benign and malignant cases or between calcifications and masses. Traditional SVM does not inherently account for this. A gap exists in developing or adapting **cost-sensitive SVMs** or integrating **class-weighted loss functions** that better handle medical data imbalance.
- **Incremental and Online SVM for Real-Time Systems**
  Current SVM models are batch-trained, which means the model must be retrained entirely when new data is added. This is inefficient for real-world clinical systems. There is a gap in exploring **incremental or online SVM algorithms** that can update themselves as new mammogram images become available.
- **Generalization Across Institutions (Domain Adaptation)**
  SVM models trained on CBIS-DDSM may not perform well on mammograms from other hospitals due to imaging protocol differences. **Domain adaptation techniques for SVM** (like transfer

learning in the kernel space or reweighting strategies) are underutilized in this context.

- **Integration with Radiomics Pipelines**
  There is an underexplored opportunity to integrate **advanced radiomic feature sets** (e.g., using PyRadiomics) with SVM classification. Most current pipelines use a small set of handcrafted features; expanding and optimizing these systematically could improve prediction significantly.

Addressing these gaps will require a more inclusive approach to data, as well as systems that are both transparent and adaptable to real-world settings.

7.Results & Evaluation

Using the CBIS-DDSM: Breast Cancer Image Dataset, we tested multiple models and assessed their strengths:

- **SVMs** achieved over 98% accuracy when analysing image-based data, showing strong potential for tasks involving medical scans.
- **Effective Hyperparameter Tuning**: Research improved model performance by optimizing the decision boundary.
- **Real-Time Application Ready**: Its lightweight architecture and minimal training time make it suitable for real-time and resource-constrained environments.
- **Outperformed Baseline Models**: The SVM model showed superior diagnostic accuracy and speed compared to other baseline approaches.
- **ROC-AUC scores**: Indicated that our models consistently performed well at distinguishing between classes..
- **Precision:** Tells us how many of our positive predictions were actually correct.
- **Recall (Sensitivity):** Measures how well the model identified all actual positive cases.
- **F1-Score:** Combines precision and recall into a single number, especially useful when we care about both false positives and negatives.
- **Confusion Matrix:** Gives a detailed breakdown of where the model got things right—and where it didn't.

Overall, ensemble models stood out for their ability to combine speed, precision, and scalability—making them a practical choice for future deployment.

8.Conclusion & Future Scope

Machine learning clearly has a role to play in improving breast cancer diagnostics. Our work demonstrates that automated tools can support clinicians by providing accurate, timely, and explainable predictions. But there's more to come.

In conclusion, this project demonstrates the effectiveness of Support Vector Machine (SVM) in classifying breast cancer using the CBIS-DDSM mammogram image dataset. By applying proper preprocessing, feature extraction, and model tuning techniques, the SVM was able to distinguish between benign and malignant cases with promising accuracy. The method's strength lies in its ability to perform well even with a relatively small and imbalanced dataset, making it a suitable choice when deep learning models are not feasible. Despite its limitations in interpretability and scalability, the SVM proved to be a robust and reliable model for image-based medical diagnosis in this context.

Looking ahead, there is significant scope to enhance this work. One major area is improving the feature extraction pipeline—using more diverse radiomic features or dimensionality reduction techniques could help the SVM better capture complex lesion patterns. Additionally, implementing multi-kernel SVMs could allow the model to leverage different aspects of mammogram features simultaneously. Addressing class imbalance through cost-sensitive learning or oversampling methods could also improve the model's sensitivity to malignant cases. Furthermore, exploring domain adaptation techniques would help generalize the model to mammograms from different institutions or imaging setups. Finally, enhancing model transparency by incorporating explainability methods into the SVM framework could increase clinical trust and usability. These directions open up promising opportunities to further develop accurate, efficient, and interpretable breast cancer prediction systems using traditional machine learning approaches like SVM. In Future systems could adopt **federated learning** to train on decentralized data without compromising patient privacy. There's also great potential in combining imaging with **genetic profiles** for more personalized treatment plans. Real-time updates based on clinical feedback and **cloud-based platforms** could make these tools globally accessible.

As AI becomes more integrated into healthcare, such advancements could lead to a smarter, more responsive system of care.

9.References

[1]Wang, D. Zhang and Y. H. Huang “Breast Cancer Prediction Using Machine Learning” (2018), Vol. 66, NO. 7.

[2] Dongdong Sun, M.Wang, H. Feng and Ao Li , “Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction”, 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)

[3] Ch. Shravya, K. Pravalika, Sk. Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques”, 2019 International Journal of Innovative Technology and Exploring Engineering.

[4] Mariam Amrane, Saliha Oukid, Ikram Gagaoua and Tolga Ensari, "Breast cancer classification using machine learning"2018 Electric Electronics, Computer Science, Biomedical Engineerings, Meeting (EBBT).

[5] Aruna S, Rajagopalan S and Nandakishore L, "Knowledge based analysis of various statistical tools in detecting breast cancer"2011 Computer Science Information Technology.