Breast Cancer Prediction and Detection:
Comparison Of The Latest Machine Learning Techniques

Ornella Kelly Nya Yanga, Dr. Elaheh Homayounvala

Intelligent Systems Research Group, SCDM, London Metropolitan University, 166-220 Holloway Rd, London N7 8DB,UK

orn0023@my.londonmet.ac.uk, e.homayounvala@londonmet.ac.uk

November, 2023

Abstract

Breast cancer is a prevalent and potentially deadly disease affecting women in the UK, with 1 in 7 women and 1 in 1,000 men at risk Early diagnosis, effective treatment, awareness, lifestyle choices, genetic testing, and research efforts have helped reduce mortality and improve patient outcomes. Extensive research has enhanced our understanding of the disease and led to better patient survival rates and quality of life. However, breast cancer remains a significant global health challenge, requiring ongoing research and innovation. This paper discusses using machine learning and deep learning techniques, including Convolutional Neural Networks, Transfer Learning, and Ensemble Learning, to analyze a dataset primarily consisting of images. The main goal is to compare these methods based on performance, focusing on applying effective pre-processing techniques. using the Digital Database for Screening Mammography (DDMS) dataset. CNN exhibited favorable accuracy, and ResNet reached an impressive 93%.

Keywords: Machine learning (ML), Deep learning (DL), Convolutional Neural Network (CNN), Transfer Learning, ResNet, VGG, Inception.

1. Introduction

Breast cancer is a widespread and potentially life-threatening disease that impacts women in the UK, posing a significant risk for individuals [1]. It is a malignant growth that originates from the cells of the breast and is characterized by a cluster of cancerous cells with the potential to proliferate into adjacent tissues or metastasize to remote bodily regions.

Various classifications of breast cancer exist, including non-invasive **Ductal Carcinoma In Situ (DCIS)**, invasive **Ductal Carcinoma (IDC)**, and invasive **Lobular Carcinoma (ILC)**.

Other rarer breast cancer categories encompass inflammatory, triple-negative, and HER2-enhanced.

This paper categorizes breast conditions into three types:

- **Benign:** Non-cancerous conditions that do not spread and may show symptoms like lumps.
- **Malignant:** Cancerous conditions, which can invade nearby tissues and spread to other body parts.
- **Normal:** Describes breast tissue functioning without any abnormalities.

Reducing breast cancer mortality comes through early detection, effective treatment strategies, public health interventions, and research into more advanced and targeted treatments.

1.1. Contribution & Novelty

The primary objective is to generate fitting code for the methods introduced, facilitating a thorough comparison of performance, robustness, interpretability, and execution time derived from the final results.

We aim to elucidate the most suitable method for the Digital Database for Screening Mammography (DDSM) [2] dataset based on the outcomes.

This paper seeks to build upon existing research, specifically focusing on recent articles incorporating contemporary methodologies.

1.2. Limitations

Our discussion centers on the dataset's purpose and constraints. A key challenge is the need for a universally accepted data preparation method, requiring experimentation for optimal results. Managing unbalanced data poses another hurdle, often necessitating rigorous cleaning. While addressing imbalanced data, removing specific components reduces the dataset size. Feature selection is a challenge we should have extensively explored. Additionally, handling over 10,000 images requires extra storage space for efficient processing.

2. Literature Review

The research benefited from the article by **Varsha Nemade, Sunil Pathak, and Ashutosh Kumar Dubey** [3]. This article extensively explores breast cancer detection using machine learning and deep learning techniques. It aims to simplify the evolutionary journey of breast cancer diagnostic methods, focusing on histopathological images and mammography. The article encapsulates research from 2015-2021, highlighting major techniques and recent innovations in breast cancer detection. **Figure 1** illustrates all techniques used in a paper analyzed in that article.

The research's outstanding quality brought to the forefront several significant challenges, including unbalanced data, limited datasets, optimal preprocessing, feature extraction, and lesion localization. These challenges underscore the complexities of breast cancer diagnosis and the ongoing efforts to address them.

.

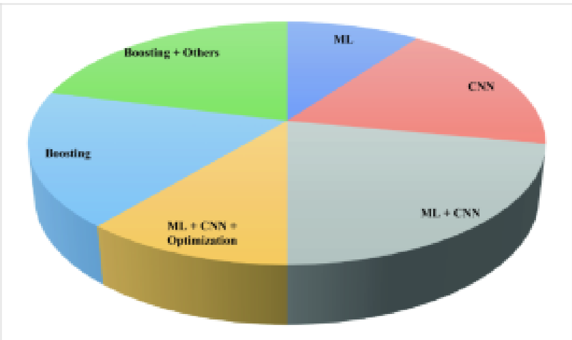

Figure 1: Partitioning of the ML algorithms analyzed in this article

They also evaluated the outcomes of both ML and DL techniques. In certain scenarios, traditional learning methods were examined, and transfer learning approaches were also considered.
Invaluably, the items are analyzed from multiple perspectives, considering the dataset type and the methodology employed. Accuracies are juxtaposed, and the benefits and limitations of each approach are emphasized. Additionally, there are numerous summary tables provided.
Numerous articles utilized for this research were sourced from the contributions of Varsha Nemade, Sunil Pathak, and Ashutosh Kumar Dubey.

Similar research was conducted by **Abdelhafz D, Yang C, Ammar R, and Nabavi S** [4]. They extensively evaluated the use of Convolutional Neural Networks (CNNs) in mammographic (MG) image analysis. The review covered 83 research papers that employed CNNs for various mammography tasks. Their work mainly focused on identifying effective strategies for improving diagnostic accuracy and exploring the structure of CNNs used in different studies.

2.1. Embryonic Stages - Texture Analysis:
da Rocha et al. [5] pioneered the integration of Local Binary Patterns (LBPs) with ecological indices, laying the groundwork for texture-based mammogram analysis. Promising applications beyond breast cancer detection, such as detecting anomalies in lung nodules or prostate anomalies, were explored.

2.2. Deep Learning - A New Dawn:
In 2015, **Dhungel, Carneiro, and Bradley** [6] introduced an amalgamation of structured support vector machines and deep learning, setting a new benchmark in mammogram analysis. Deep learning emerged as a powerful tool, showcasing improved accuracy and efficiency.

2.3. Transfer Learning - Bridging Past and Present:
In the latter half of the 2010s, **Suzuki et al.** [7] introduced transfer learning by infusing Deep Convolutional Neural Networks (DCNNs) with mammographic data. **Samala et al.** [8] demonstrated the benefits of multi-stage transfer learning, especially in digital breast tomosynthesis.

2.4. Computer-Aided Diagnosis (CAD) - The Modern Frontier:
In 2020, **Mohanty et al.** [9] explored new territories with their advanced CAD system, redefining mammogram classification standards. Focus shifted to creating sophisticated CAD systems for highly accurate mammogram classification.

2.5. A Departure - Embracing Artificial Intelligence in a Novel Manner:
The Adaptive Artificial Immune Networks (A2INET) ensemble offered a fresh perspective, showcasing the possibilities of automated diagnostic tools. Integration of AI into mammographic analysis took a novel turn. Additionally, specific methodologies in mammogram classification were discussed:

2.6. Support Vector Machines (SVMs):
Vijayarajeswari et al. [10] demonstrated a high classification accuracy of 94% using the Hough transform for feature extraction in conjunction with SVM classification. Innovative Integration:

Mohanty et al. [11] in 2020 fused CAD, PCA, WC-SSA, and KELM, achieving an outstanding accuracy rate of 99.92% in breast cancer detection.

2.7. Deep Learning Advancements:
Chougrad et al. [12] introduced a multi-label image categorization method in 2020. **Rahman et al.** [13] and **Alkhaleefah et al.** [14] focused on CNN architectures and transfer learning, demonstrating significant improvements in accuracy and precision.

These studies collectively illustrate the evolving landscape of breast cancer detection, showcasing a spectrum of methodologies with notable accuracy rates, enhancing our ability to detect breast cancer precisely.

3. Research Design & Methodology

Various algorithms are introduced and implemented, and consistent preprocessing methods are applied for a fair comparison of their performance.

The study emphasizes evaluating these algorithms to achieve the highest possible accuracy in cancer detection, primarily using detection accuracy as the main evaluation metric. It acknowledges the significance of preprocessing in machine learning but maintains consistent preprocessing techniques without extensive optimization, recognizing that preprocessing intricacies in medical datasets are separate research topics beyond the study's primary focus.

3.1. Dataset

For this research, We have chosen to utilize the DDSM dataset, Digital Database for Screening Mammography, a substantial collection of mammographic images that plays a pivotal role in breast cancer research, particularly in assessing computer-aided diagnosis systems. This dataset encompasses a wide array of cases from diverse patient populations, spanning various age groups, breast densities, and pathologies, including both normal and abnormal cases featuring benign and malignant calcifications and masses. Despite certain limitations, the DDSM dataset has significantly contributed to advancing mammography interpretation algorithms and serves as a standard for comparing CAD systems in academic research.

3.1.1. Data Collection & Data Size

We employed "The Complete Mini-DDSM" dataset [2], a condensed version of DDSM updated by ABBAS CHEDDAD in 2021, available on Kaggle.com. Stored in a 53 GB zipped file, it comprises three subdirectories, with "MINI-DDSM-Complete.jpeg" being the main focus for our analysis.

Formerly "MINI-DDSM-Complete.jpeg," the directory holds 17,608 files in three folders: "Benign," "Cancer," and "Normal," accompanied by an Excel file consolidating image categorization data.

3.1.2. Type of Data

In mammography, two standard views are commonly used for each breast: the Craniocaudal (CC) view and the Mediolateral Oblique (MLO) view(as shown in **figure 2**).

Craniocaudal (CC) View:

- Captured from above the breast.
- Involves vertical breast compression.
- Offers a good view of central and superficial breast tissue.

Mediolateral Oblique (MLO) View:

- Taken at a 45-degree angle.
- Uses oblique breast compression.
- Provides a view of lateral and deeper breast tissue, including the breast tail.

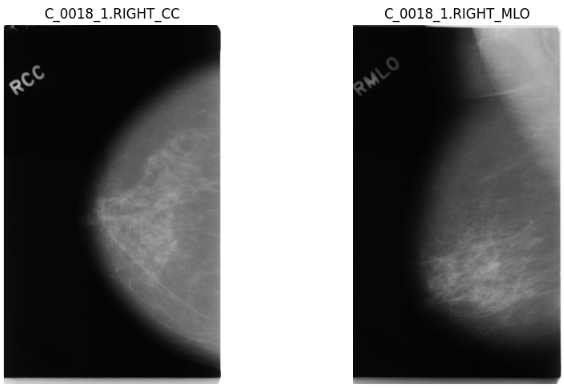

Figure 2: Example CC and MLO images in the dataset

Radiologists typically review both CC and MLO views for a comprehensive assessment.

Machine learning algorithms consider both views to detect potential anomalies effectively in mammograms.

3.1.3. Data Cleaning & Data Split

We use "masks", images outlining specific regions of interest (ROI) in original images, crucial for tasks like lesion segmentation in mammograms. They accurately represent abnormalities and are employed in machine learning for training models to improve segmentation and in data augmentation for neural networks. In this analysis, we removed "mask" images, leaving 7898 images. Each folder contains mammograms for both breasts, but not all individuals have tumors in both. We created a separate "no_mask" folder for mammograms without tumors to prevent skewing the results. **Figure 3** shows what we intend for "mask" and classical mammography images.

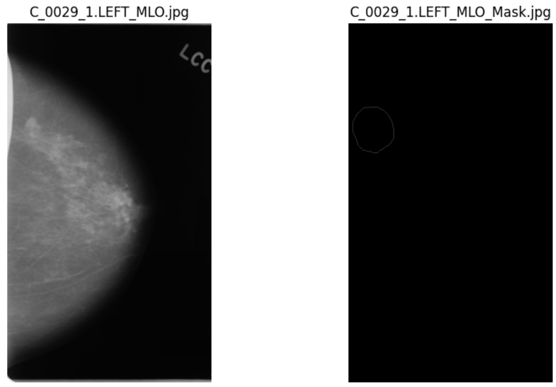

Figure 3: Example of classical mammography and mask image

We partitioned the dataset pre-analysis for an organized approach, creating three subfolders (Train, Test, and Validation) within each category (Benign, Cancer, and Normal). The data distribution for each subfolder was 70% for training, 15% for testing, and 15% for validation.

3.2. Algorithm Selection and Model Design

We will offer a brief summary of the chosen algorithm and delve into its practical implementation.

3.2.1. Traditional Machine Learning Algorithms

Five distinct machine learning classifiers were employed to achieve accurate classification. These algorithms were selected due to their varying mechanisms and theoretical foundations, ensuring a comprehensive examination of the dataset:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVC)
- Random Forest
- AdaBoost
- XGBoost

Data Preprocessing. Initially, we structured the dataset into three folders: Train, Test, and Validation. We read each image, resized it to 128x128 pixels, and normalized it to a scale between 0 and 1 for computational efficiency. We encoded the labels (Benign, Normal, Cancer) into a numerical format using LabelEncoder for compatibility with machine learning models.

Training and Evaluation. We trained each of the five machine-learning algorithms on the reshaped training data. Following training, we assessed the algorithms on the Train, Test, and Validation datasets. We will calculate accuracy, precision, recall, F1 score, Cohen's Kappa, and present a confusion matrix for a comprehensive performance overview. Through systematic evaluation, we thoroughly analyzed each model's effectiveness across diverse data segments.

3.2.2. CNNs:

Commencing with an introduction to Convolutional Neural Networks (CNNs) and their pivotal role in breast cancer detection through medical images like mammograms, we present a concise, step-by-step implementation of the CNN model. Our detailed approach covers various aspects:

- **Data Preprocessing:** We organize the dataset, apply image augmentation, and utilize DataLoader for efficient batching, ensuring data integrity for training, testing, and validation.
- **Architectural Decisions:** Our choice is ResNet-18, valued for its residual connections tailored for the binary classification task.
- **Training Dynamics:** Explaining the selection of cross-entropy loss and the Stochastic Gradient Descent (SGD) optimizer, we introduce a dynamic learning rate scheduler to adapt during training.
- **Training Mechanism:** Describing the "training_model" function as our primary orchestrator, managing both training and validation datasets, we delve into the training loop, encompassing forward and backward passes. The validation phase is elucidated, focusing on optimal performance-based model selection.
- **Visualization and Interpretation:** We introduce the "visualize_model" function, enhancing clarity by displaying images alongside predicted classes.
- **Extended Training Exploration:** We emphasize developing an advanced ResNet-18 model with optimized parameters, underscored by rigorous training and knowledge transfer.

Our explication provides a comprehensive understanding of the methodology and rationale governing our formal and structured CNN implementation for breast cancer detection.

3.2.3. ResNet

We chose the ResNet34 architecture for its ability to train deeper neural networks with residual connections, mitigating vanishing gradient issues while maintaining performance. Key aspects:

- **Model Insights:** ResNet34 captures spatial hierarchies using 3x3 and 1x1 convolutional layers. It employs BasicBlock and Bottleneck modules for residual learning, addressing vanishing gradient challenges. Batch Normalization and strategic strides for downsampling enhance the architecture.
- **Training Strategy:** We meticulously configured the ResNet-34 model for training, using Stochastic Gradient Descent (SGD) with momentum. An adaptive learning rate scheduler adjusts rates based on validation accuracy.
- **Training Epochs & Observations:** The model undergoes rigorous training, monitoring metrics like loss and accuracy for both datasets to provide insights into its progression and convergence.
- **Visual Analysis:** Visualizations include a loss curve, accuracy plot, example predictions, and a confusion matrix, offering a comprehensive view of the model's performance.
- **Evaluation & Concluding Remarks:** The trained ResNet34 model undergoes a rigorous evaluation using a dedicated test set. A confusion matrix quantitatively assesses precision and recall for each class (Benign, Cancer, and Normal).

Our choice of the ResNet34 architecture addresses deep learning challenges in gradient descent. The training and evaluation adhere to rigorous standards, crucial in medical applications where accuracy and reliability are paramount.

3.2.4. VGGNet

The VGGNet, and specifically the VGG16 architecture, was selected due to its well-known deep convolutional structure. This architecture offers a comprehensive stack of convolutional blocks, making it suitable for intricate feature extraction from images. The model was adapted to handle a trichotomous classification task, aligning with the three medical image categories: 'Benign,' 'Cancer,' and 'Normal.'

The architecture comprises multiple convolutional blocks, each building on the previous one to capture increasingly complex features. The final layers consist of a classifier that maps the extracted features to the three medical classes. Dropout layers were introduced to mitigate overfitting.

The weight matrices were initialized using He initialization for effective optimization. Bias correction was implemented to ensure a neutral starting point for training.

The model was trained using the Stochastic Gradient Descent (SGD) optimizer with momentum. A learning rate scheduler was employed to adapt the learning rate based on validation performance.

The training process was documented over a series of epochs, and loss and accuracy metrics were visualized to understand the model's training dynamics. The model's performance was evaluated using a confusion matrix, providing insight into its classification accuracy across the medical classes.

This architectural exploration showcases the adaptability of VGGNet for medical image classification and underscores the rigorous training and evaluation standards employed in this paper.

3.3. Machine Learning Framework and Tools

We delve into the machine learning framework and tools that were integral to our research. These components played a crucial role in the success of our paper.

3.3.1. Environment:

Python, chosen for its simplicity and extensive library ecosystem, was our primary programming language in this research. Notable libraries like NumPy, PIL, scikit-learn, and XGBoost enriched our Python environment, streamlining complex tasks and enabling focused problem-solving. Meticulous organization of the dataset into distinct paths, such as 'Final/train,' 'Final/test,' and 'Final/valid,' ensured systematic data separation for various research phases and maintained data integrity.

3.3.2. Frameworks & Libraries:

- Scikit-learn (sklearn) is a versatile Python library that offers accessible and reusable design principles. We utilized it for preprocessing, employing the LabelEncoder utility and various metrics for comprehensive classifier performance evaluation.
- XGBoost, known for its speed and performance, was crucial for classification tasks, particularly in handling large datasets.
- Numpy, a fundamental scientific computing package, was essential for working with large multi-dimensional arrays and matrices.
- Pillow (PIL) facilitated image processing and dataset handling by opening and standardizing image files.
- The OS library enabled directory navigation and file operations, ensuring efficient preprocessing and modeling.

3.3.3. Tools:

- Jupyter Notebook played a central role, providing an interactive platform that seamlessly combined code, text, and visuals.
- Our research involved image processing, including image resizing and normalization to maintain dataset consistency.

- Machine learning classifiers, such as KNeighborsClassifier, SVC, RandomForestClassifier, AdaBoostClassifier, and XGBClassifier, were instrumental in performance evaluation on train, test, and validation datasets.
- Torchvision, an extension of PyTorch [16], offered valuable datasets and model configurations for deep learning tasks.
- DataLoader, a PyTorch component, managed data processing, including batching, randomization, and parallelization.

3.3.4. Hardware:

Emphasizing the computational demands of image processing and machine learning, we prioritized hardware with robust RAM. Checks for CUDA-compatible GPUs were included for speed optimization, gracefully falling back to CPU usage when GPUs were unavailable.

For reproducibility, we set a random seed (RANDOM_SEED = 123) to achieve consistent results despite inherent training process randomness. Well-defined hyperparameters, including BATCH_SIZE and NUM_EPOCHS, balanced computational efficiency and model performance.

3.4. Infrastructure Limitations

The research faced several significant limitations and made certain assumptions throughout the study:

- **Infrastructure Limitations:** Due to limitations in the researcher's personal computing device, the university's laboratory infrastructure was used. While this provided better computational capacity, it introduced spatial constraints inherent to the laboratory setting, limiting the study's full potential.
- **Extended Runtime:** The complexity of some algorithms resulted in extended runtime, sometimes exceeding 10 hours. While indicative of exhaustive analysis, this runtime might not be practical in real-time clinical settings, potentially limiting the algorithm's applicability.
- **Dataset Size:** The large DSMM dataset (over 53 GB) posed challenges in terms of downloading and handling. While it was rich for algorithmic training and validation, the data's volume introduced potential overfitting risks and demanded significant computational and memory resources.
- **Assumption of Uniformity:** The study assumed that the laboratory setting and the selected dataset uniformly represented real-world scenarios. However, this may limit the generalizability of findings, as the dataset might not capture all nuances of global mammogram data.
- **Comparative Nature:** The study compared various machine learning algorithms based on performance and applicability. This assumes consistency in metrics used for comparison, potentially overlooking unique strengths or weaknesses of individual algorithms.

3.5. Ethical Considerations

Incorporating mammograms into machine learning for medical diagnostics offers significant promise, especially with datasets like DDSM. However, this intersection raises important ethical considerations:

Patient Privacy: The foremost concern is safeguarding patient privacy. Mammograms contain highly personal medical information, and while anonymization is practiced, the evolving landscape of data security necessitates ongoing efforts to protect patient anonymity.

Informed Consent: Obtaining informed patient consent is crucial. Mammographic data is primarily collected for clinical purposes, and any secondary use for research, like contributing to machine learning datasets, should involve explicit and informed consent to respect patient autonomy.

Algorithmic Bias and Fairness: Ensuring algorithmic fairness is vital. Biases in the dataset can lead to biased algorithms, potentially resulting in unequal diagnostic outcomes for different demographic groups, and contributing to health disparities.

Interpretability of Algorithms: The interpretability of algorithms is critical in medical diagnostics. In this context, understanding the rationale behind algorithmic recommendations is essential, as black-box models may erode trust and collaboration between clinicians and algorithms.

Human-Machine Collaboration: As these algorithms are integrated into clinical settings, it's essential to define the roles of humans and machines. Algorithms should complement rather than replace human expertise, fostering a collaborative and synergistic diagnostic approach that leverages the strengths of both.

4. Results & Discussion

We aim to present and evaluate outcomes from our extensive investigation into mammogram classification. Our work aligns with the commitment to enhance diagnostic accuracy within the broader academic research. Beyond conveying results, this section sets the stage for an in-depth analysis, discussing the wider implications of our discoveries, their interaction with existing knowledge, and proposing novel research directions for future academic endeavors.

4.1. Machine Learning Algorithm

We investigated five algorithms: K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest Classifier, AdaBoost Classifier, and XGBoost Classifier. Each algorithm's performance was assessed using various metrics on training, testing, and validation datasets.

1. K-Nearest Neighbors (KNN):
   - Training Dataset: Accuracy of 0.75.
   - Test Dataset: Accuracy of 0.65.
   - Validation Dataset: Accuracy of 0.64.

   KNN model is performing reasonably well on the training data, but there might be some overfitting as it generalizes less effectively to new, unseen data.

```
Accuracy: 0.75
Precision: 0.76
Recall: 0.75
F1 Score: 0.76
Cohen's Kappa: 0.62
Confusion Matrix:
[[ 734  145  102]
 [ 310  607   82]
 [ 156  104 1425]]
```

```
Accuracy: 0.65
Precision: 0.67
Recall: 0.65
F1 Score: 0.65
Cohen's Kappa: 0.46
Confusion Matrix:
[[138  47  26]
 [ 92 100  23]
 [ 60  31 271]]
```

```
Accuracy: 0.64
Precision: 0.65
Recall: 0.64
F1 Score: 0.64
Cohen's Kappa: 0.45
Confusion Matrix:
[[124  58  28]
 [ 89  98  27]
 [ 47  33 281]]
```

2. Support Vector Classifier (SVC):
   - Training Dataset: Accuracy of 0.74.
   - Test and Validation Datasets: Accuracies of 0.67.

```
Accuracy: 0.74
Precision: 0.74
Recall: 0.74
F1 Score: 0.74
Cohen's Kappa: 0.60
Confusion Matrix:
[[ 586  242  153]
 [ 208  651  140]
 [ 113   88 1484]]
```

```
Accuracy: 0.67
Precision: 0.67
Recall: 0.67
F1 Score: 0.67
Cohen's Kappa: 0.49
Confusion Matrix:
[[102  71  38]
 [ 49 127  39]
 [ 38  23 301]]
```

```
Accuracy: 0.67
Precision: 0.67
Recall: 0.67
F1 Score: 0.67
Cohen's Kappa: 0.48
Confusion Matrix:
[[114  51  45]
 [ 58 109  47]
 [ 40  16 305]]
```

   SVC exhibited stable performance across all datasets.

3. Random Forest Classifier:
   - Training Dataset: Perfect accuracy of 1.00.
   - Test and Validation Datasets: Accuracies of 0.73 and 0.72, respectively.

```
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
Cohen's Kappa: 1.00
Confusion Matrix:
[[ 981    0    0]
 [   0  999    0]
 [   0    0 1685]]
```

```
Accuracy: 0.73
Precision: 0.72
Recall: 0.73
F1 Score: 0.72
Cohen's Kappa: 0.57
Confusion Matrix:
[[117  56  38]
 [ 39 140  36]
 [ 24  22 316]]
```

```
Accuracy: 0.72
Precision: 0.72
Recall: 0.72
F1 Score: 0.72
Cohen's Kappa: 0.57
Confusion Matrix:
[[125  51  34]
 [ 47 127  40]
 [ 23  21 317]]
```

The perfect training accuracy suggests overfitting, while the lower test and validation accuracies indicate a performance drop.

4. AdaBoost Classifier:
   - Training Dataset: Accuracy of 0.72.
   - Test and Validation Datasets: Accuracies of 0.66.

```
Accuracy: 0.72
Precision: 0.71
Recall: 0.72
F1 Score: 0.72
Cohen's Kappa: 0.56
Confusion Matrix:
[[ 566  281  134]
 [ 305  575  119]
 [ 108   87 1490]]
```

```
Accuracy: 0.66
Precision: 0.66
Recall: 0.66
F1 Score: 0.66
Cohen's Kappa: 0.47
Confusion Matrix:
[[ 99  68  44]
 [ 56 123  36]
 [ 30  32 300]]
```

```
Accuracy: 0.66
Precision: 0.65
Recall: 0.66
F1 Score: 0.65
Cohen's Kappa: 0.46
Confusion Matrix:
[[109  59  42]
 [ 61 110  43]
 [ 37  28 296]]
```

   AdaBoost Classifier showed consistent performance but struggled with certain classes in the test set.

5. XGBoost Classifier:
   - Training Dataset: Perfect accuracy of 1.00.
   - Test and Validation Datasets: Accuracies of 0.75.

```
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
Cohen's Kappa: 1.00
Confusion Matrix:
[[ 981    0    0]
 [   0  999    0]
 [   0    0 1685]]
```

```
Accuracy: 0.75
Precision: 0.75
Recall: 0.75
F1 Score: 0.75
Cohen's Kappa: 0.61
Confusion Matrix:
[[126  58  27]
 [ 45 143  27]
 [ 21  18 323]]
```

```
Accuracy: 0.75
Precision: 0.75
Recall: 0.75
F1 Score: 0.75
Cohen's Kappa: 0.62
Confusion Matrix:
[[143  37  30]
 [ 61 129  24]
 [ 20  21 320]]
```

   Similar to Random Forest, XGBoost displayed potential overfitting due to the perfect training accuracy.

Key considerations include addressing overfitting, exploring different hyperparameters, and examining feature importance. We highlighted the significance of analyzing confusion matrices to gain insights into model performance.

Overall, the study underscores the need for iterative refinement, validation, and the incorporation of domain-specific knowledge to establish a robust mammogram classification system, given the complexities and real-world implications of this domain.

4.2. CNNs

The CNN achieved an 82% accuracy in mammogram interpretation, a notable result but falls below benchmarks surpassing 90% in the field. Acknowledging the challenges in this critical task, improvements are crucial. Despite the achievement in 129 minutes, areas for enhancement include advanced pre-processing, exploring advanced model architectures, addressing imbalances, and utilizing transfer learning.

Given the critical nature of 'Cancer' diagnosis, achieving a higher accuracy rate is imperative. The model's training efficiency in 128 minutes and 57 seconds is commendable, but exploring longer training or optimizations may yield better results. Model interpretability, essential in clinical contexts, can be enhanced using techniques like Class Activation Mapping.

In conclusion, the 82% accuracy lays a promising foundation, emphasizing the need for refinement and improvement to align with state-of-the-art benchmarks in mammogram interpretation.

4.3. RestNet

The performance evaluation of the ResNet model encompassed various key metrics:

- **Epochs and Batches:** The model was trained over 50 epochs, allowing it to revisit the dataset 50 times. Each epoch was divided into 20 batches, enabling the model to update its weights more frequently for faster convergence (refer to **Figure 4**).
- **Loss:** The training process exhibited a promising reduction in loss, starting at 0.9886 and decreasing to 0.1106. This reduction indicates that the model improved its predictions, minimizing the disparity between them and the actual results (refer to **Figure 4**).
- **Training and Validation Accuracy:** The model demonstrated substantial progress in both training and validation accuracy. Training accuracy increased from 27.25% to 92.78%, while validation accuracy improved from 27.04% to 93.88% (as illustrated in **Figure 5**). The consistency between training and validation accuracies suggested effective model generalization without overfitting.
- **Time:** An anomaly observed during training was increased training time with each epoch, with a particularly perplexing jump between the 5th and 6th epochs.
- **Test Accuracy:** The model's testing phase resulted in an impressive accuracy of 93.94%, closely aligning with the validation accuracy, affirming the model's reliability.

The metrics indicated a successful training process, especially the close match between training and validation accuracies, signifying the model's ability to generalize effectively. Nonetheless, the time-related anomaly requires further investigation before deploying the model extensively.

In visualization, monitoring the loss is crucial during deep learning training. To mitigate fluctuations in loss due to mini-batch training, a running average is employed, providing a smoothed trend in the loss. This technique filters out random fluctuations and facilitates better interpretation of the overall loss trend.

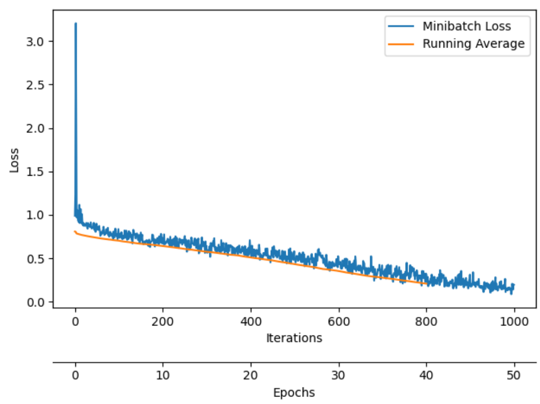

Figure 4: Training loss ResNet

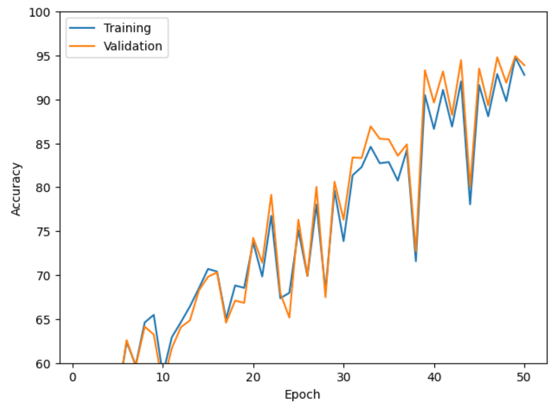

Figure 5: Accuracy Plot RestNet

The confusion matrix comprehensively analyzes the model's performance for each class. It offers insights into accuracy, precision, recall, F1-score, class-specific performance, and potential data imbalances, as highlighted in **Figure 6**. This information aids in understanding the model's strengths and areas for improvement, particularly regarding specific classes where precision or recall may need enhancement.

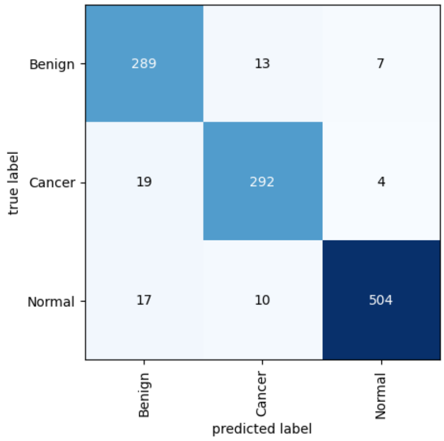

Figure 6: Confusion Matrix RestNet

Overall, these evaluations provide valuable insights into the model's performance, its ability to generalize, and areas where improvements can be made, all of which are essential for informed decision-making in applying this machine learning model.

4.4. VGGNet

We thoroughly evaluated a machine learning model, focusing on various critical aspects of its performance. Key findings and observations from this evaluation include:

- **Training Loss and Time:** The model exhibited a consistent decrease in training loss from an initial high of 2.0450 to a final value of 0.6460 after 50 epochs. However, there was an inconsistency in training time between epochs, particularly during the transition from the 7th to the 8th epoch, which may be attributed to various factors, including background processes and model checkpoint saving.
- **Overfitting and Accuracy Plateau:** The alignment between training and validation accuracy indicated that overfitting was not a significant concern. Notably, the model's accuracy exhibited a growth plateau around the 48th epoch, suggesting that further training epochs might yield diminishing returns regarding accuracy improvement.
- **Performance Evaluation:** The model achieved a test accuracy of 71.26%, and the significance of this accuracy depended on the complexity of the task and dataset. For intricate classifications in a heterogeneous dataset, this accuracy was considered commendable.
- **Suggestions for Improvement:** The researcher provided valuable recommendations for enhancing the model's performance. These suggestions covered diverse aspects of model development, including data augmentation, regularization methods, learning rate adjustments, model architecture modifications, early stopping, and the potential benefits of transfer learning.
- **Analysis of Minibatch Loss and Accuracy Trends:** The minibatch loss consistently decreased over training epochs (illustrated in **Figure 7**), and it was suggested that visualizing the running average of the loss could provide a more stable insight into training progress. Training and validation accuracy demonstrated steady improvements, as depicted in **Figure 8**, indicating effective

generalization.

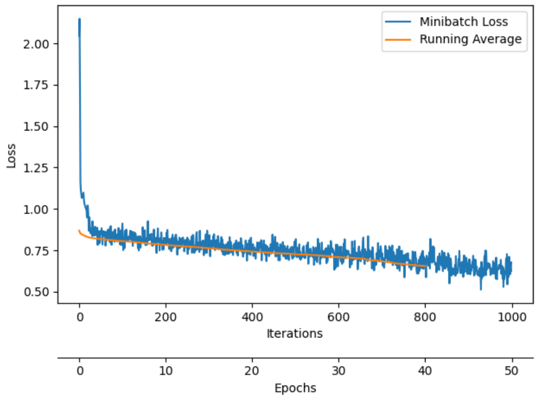

Figure 7: Training Loss VGGNet

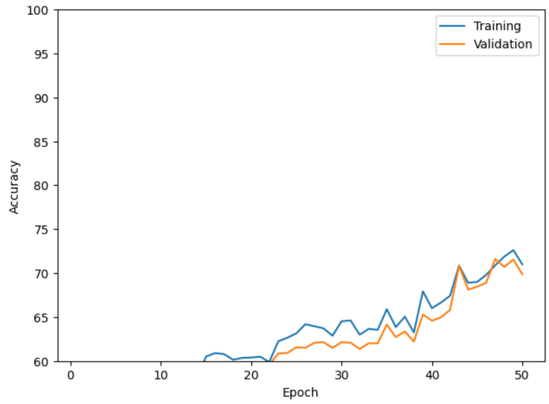

Figure 8: Accuracy Plot VGGNet

- **Class-Specific Performance and Confusion Matrix:** Class-specific performance was assessed using a confusion matrix (as shown in **Figure 9**), revealing class misclassifications and potential feature similarities. Further analysis of these misclassifications was recommended to understand the underlying patterns.

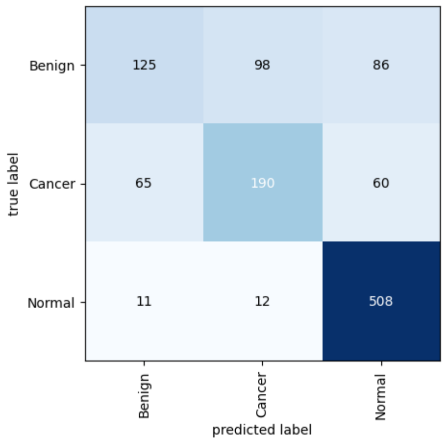

Figure 9: Confusion Matrix VGGNet

5. Conclusion & Future Works

In this research, we analyzed mammographic images from the DDSM dataset, exploring traditional algorithms like KNN, SVC, Random Forest, AdaBoost, and XGBoost. Each had strengths and limitations; KNN needed optimization, SVC achieved a balanced trade-off, Random Forest struggled with unseen data, and AdaBoost/XGBoost provided competitive results.

Transitioning to deep learning, we implemented CNN and pre-trained models like AlexNet, VGGNet, ResNet, and VGGNet. The CNN showed good accuracy, and ResNet achieved an impressive 93% accuracy. Not as high as the 98.8% achieved by de Oliveira et al. [15] using SVM on the same dataset, but superior to the 80% reported by Swiderski et al. [17] employing ensemble learning methods like SVM, DT, RF, KNN, nonlinear autoencoder, and logistic classifiers. Figure 10 recaptures the entirety of the accuracy performance. **Figure 10** recaptures the entirety of the accuracy performance applied in this paper.

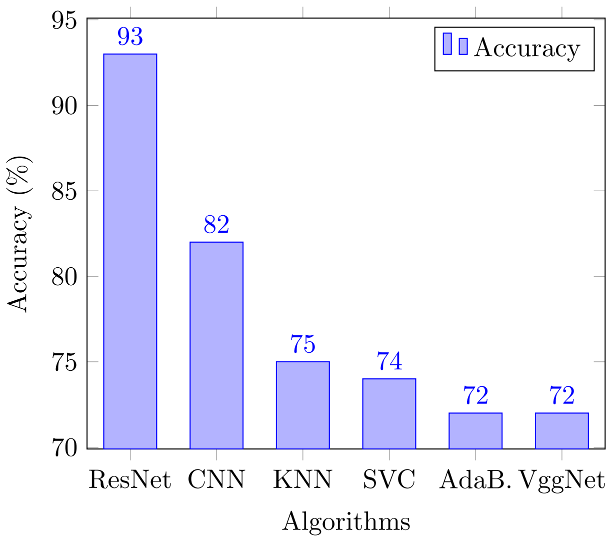

Figure 10: Algorithm Comparison

Our main contribution is a comprehensive analysis of existing methodologies, acknowledging prior research accomplishments and building upon this foundation. In light of the critical need for more efficient breast cancer diagnostic tools, our research paves the way for the potential development of an innovative software solution. This envisioned software aims to support healthcare professionals in identifying breast cancer, presenting a promising avenue for enhancing diagnostic capabilities in the field.

5.1. Future Works

Potential future endeavors involve the establishment of universally accepted data preparation methods. Exploring strategies for managing unbalanced data without substantial size reduction is crucial. Further research might investigate techniques for feature selection to improve overall data processing effectiveness. Additionally, efficient management of extensive datasets, especially those with over 10,000 images, could be a central focus in future investigations, encompassing optimization of device storage space allocation.

References

[1] Cancer Research UK https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer

[2] The Complete Mini-DDSM https://www.kaggle.com/datasets/cheddad/miniddsm2(Accessed: September 2023).

[3] Varsha Nemade, Sunil Pathak, Ashutosh Kumar Dubey (2022) *A Systematic Literature Review of Breast Cancer Diagnosis Using Machine Intelligence Techniques* 29, pages4401–4430 https://doi.org/10.1007/s11831-022-09738-3 (Accessed: September 2023)

[4] Dina Abdelhafiz, Clifford Yang, Reda Ammar & Sheida Nabavi(2019) *Deep convolutional neural networks for mammography: advances, challenges and applications*

[5] da Rocha SV, Junior GB, Silva AC, de Paiva AC, Gattass M (2016) *Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution.* Expert Syst Appl 66:7–19

[6] Dhungel N, Carneiro G, Bradley AP (2017) *A deep learning approach for the analysis of masses in mammograms with minimal user intervention.* Med Image Anal 37:114–128

[7] Suzuki S, Zhang X, Homma N, Ichiji K, Sugita N, Kawasumi Y, Ishibashi T, Yoshizawa M (2016) *Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis.* In: 2016 55th annual conference of the Society of Instrument and control engineers of Japan (SICE). IEEE, pp 1382–1386

[8] Samala RK, Chan HP, Hadjiiski L, Helvie MA, Richter CD, Cha KH (2018) *Breast cancer diagnosis in digital breast tomos ynthesis: efects of training sample size on multi-stage trans fer learning using deep neural nets.*IEEE Trans Med Imaging 38(3):686â€“696

[9] Mohanty F, Rup S, Dash B, Majhi B, Swamy MN (2020) *An improved scheme for digital mammogram classifcation using weighted chaotic salp swarm algorithm-based kernel extreme learning machine.* Appl Soft Comput 91:106266

[10] Vijayarajeswari R, Parthasarathy P, Vivekanandan S, Basha AA (2019) *Classification of mammogram for early detection of breastcancer using SVM classifier and Hough transform.* Measurement 146:800–805

[11] Mohanty F, Rup S, Dash B, Majhi B, Swamy MN (2020) *improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine.* Appl Soft Comput 91:106266

[12] Chougrad H, Zouaki H, Alheyane O (2020) *Multi-label transfer learning for the early diagnosis of breast cancer.* Neurocomputing 392:168–180

[13] Rahman AS, Belhaouari SB, Bouzerdoum A, Baali H, Alam T, El-daraa AM (2020) *Breast mass tumor classifcation using deep learning.* In: 2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT). IEEE, pp 271–276

[14] Alkhaleefah M, Chittem PK, Achhannagari VP, Ma SC, Chang YL (2020) *The infuence of image augmentation on breast lesion classification using transfer learning. In:* 2020 International conference on artifcial intelligence and signal processing (AISP). IEEE, pp 1–5

[15] de Oliveira FS, de Carvalho Filho AO, Silva AC, de Paiva AC, Gattass M (2015) *Classifcation of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM.* Comput Biol Med 57:42–53

https://github.com/rasbt/stat453-deep-learning-ss21/tree/main/L14/2-resnet-ex

[16] PyTorch (2017) *ImageFolder — Torchvision 0.14 documentation, pytorch.org.* Available at: https://pytorch.org/vision/stable/generated/torchvision.datasets.ImageFolder.html (Accessed: September 2023).

[17] Swiderski B, Osowski S, Kurek J, Kruk M, Lugowska I, Rutkowski P, Barhoumi W (2017) *Novel methods of image description and ensemble of classifiers in application to mammogram analysis.* Expert Syst Appl 81:67–78