VIRGINIA TECH

UMUR KOSE,

DANNY SMYTH,

VANESSA EICHENSEHR,

KIMIYA MOHAMMADI-JOZAI

CS 5805

DECEMBER 11, 2024

BREAST CANCER DIAGNOSIS: A FEATURE EXTRACTION AND CLASSIFICATION APPROACH

INTRODUCTION

2.3m

new cases/year [1]

670k

deaths/year[2]

- Breast Cancer is the most common cancer in women worldwide
- Traditional detection methods (mammograms, biopsy) are subject to error and result in later detection
- Goal: improve cancer detection in images of cells using machine learning

RELATED WORKS

- There are several published studies that implement ML techniques on breast cancer data
- Wisconsin Breast Cancer Dataset (WBCD) is benchmark dataset
  - Images are pre-proccessed and segmented, and feature selection already performed
  - Includes 10 features for each cell nucleus such as radius, perimeter, texture, symmetry

ML Technique	Accuracy	Publish Year	Source
Decision Tree Forest	95.51%	2013	[3]
Single Decision Tree	95.75%	2013	[3]
Lagrangian Support Vector Machines	95.42%	2014	[4]
Tree Augmented Naïve Bayes	94.11%	2018	[5]
J48	93.41%	2018	[6]
Logistic Regression	94.16%	2023	[7]
Random Forest	95.62%	2023	[7]
K-Nearest Neighbor	94.16%	2023	[7]
Artificial Neural Network	96.35%	2023	[7]

METHODOLOGY

Algorithm Analysis

- New Data Collection

Data Exploration and Feature Extraction

- Algorithm Implementation

Random Forest and CNN Implementation

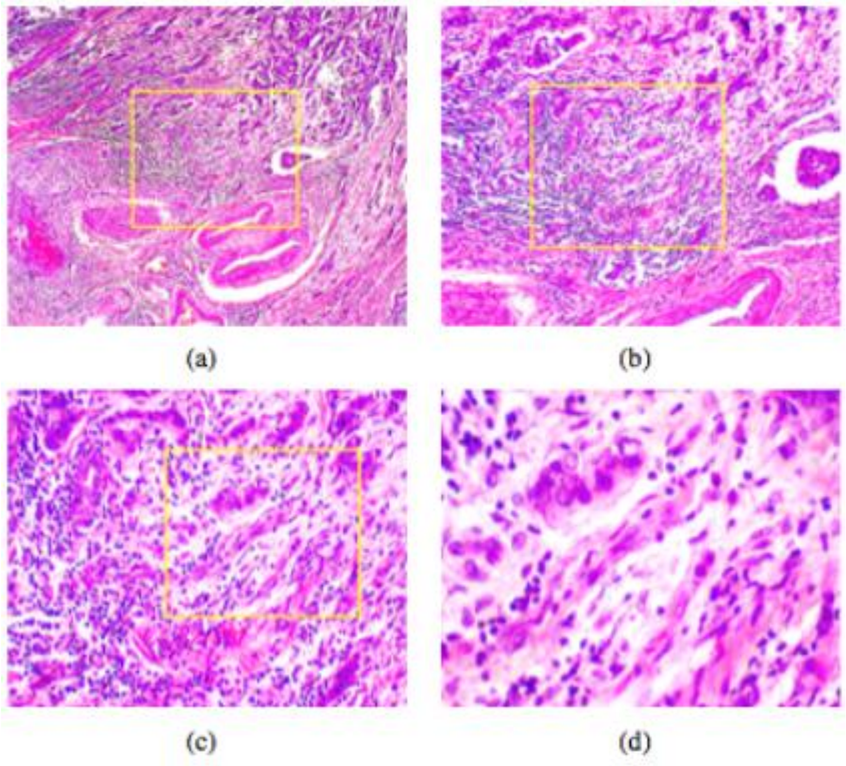
- Comparison

Results

- Analysis

DATA COLLECTION

- Original Wisconsin dataset released in 1995
- BreakHis Dataset 2016:
  - 9,109 images of tissue from 82 patients
  - 40x, 100x, 200x, 400x resolutions
  - Used for both binary and multiclass classification
- Goal: replicate results from the Wisconsin dataset on Breakhis using feature extraction



INITIAL ALGORITHM ANALYSIS

Neural Network

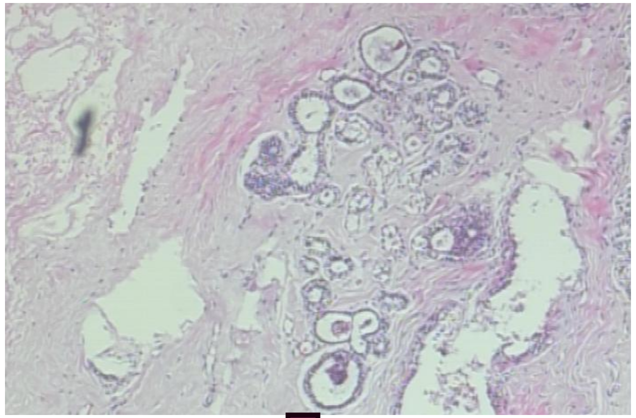
- Works for complex, non-linear data
- Higher computation costs
- Requires more input data
- Hich accuracy, low interpretability
- Better for image data
- 96.35% accuracy on Wisconsin Dataset

Decision Tree

- Better for structured data
- Works well on small datasets
- Bagging to reduce dataset variance (Random Forest)
- 95.75% accuracy on Wisconsin Dataset

EXPLORATORY DATA ANALYSIS

- Feature Extraction
  - We aim to extract features of Wisconsin Dataset from BreakHis Dataset
- Otsu Thresholding [11]
  - A widely used non-parametric and unsupervised technique in image processing for automatic thresholding
  - separate an image into foreground and background by finding a threshold that minimizes intra-class variance
- Define Different Features as Pixels
- Extract the Diagnosis from the Name of Files



resolution	patient_Diag	tumor_type	patient_ID	radius_mean
100	B	A	14-22549AB	26.196872
100	B	A	14-22549AB	54.373359
100	B	A	14-22549AB	15.594510
100	B	A	14-22549AB	11.084567
100	B	A	14-22549AB	48.082062
100	B	A	14-22549AB	33.728912
100	B	A	14-22549AB	2.459245
100	B	A	14-22549AB	1.492705
100	B	A	14-22549AB	33.530143
100	B	A	14-22549AB	53.988551
100	B	A	14-22549AB	0.797885

RESULTS OF DECISION TREE

- Dataset Resolution (200)
- Test-Train Split (20)
- CV Folds (10)
- Max Depth of Tree (20)
- Number of Trees (500)
- Binary vs. Multiclass
- Feature Removal

	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FN}$	$\frac{T}{T+F}$
Class	**Precision**	**Recall**	**F-1Score**	**Accuracy**
B	0.81	0.78	0.79	0.80
M	0.78	0.82	0.80	

RESULTS OF DECISION TREE

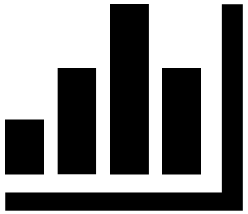
- Dataset Resolution (200)
- Test-Train Split (20)
- CV Folds (10)
- Max Depth of Tree (20)
- Number of Trees (500)
- Binary vs. Multiclass
- Feature Removal

	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FN}$	$\frac{T}{T+F}$
Class	**Precision**	**Recall**	**F-1Score**	**Accuracy**
A	0.83	0.94	0.88	0.79
F	0.73	0.39	0.47	
PT	0.84	0.84	0.84	
TA	0.81	0.90	0.85	
DC	0.61	0.39	0.47	
LC	0.81	0.83	0.82	
MC	0.82	0.79	0.81	
PC	0.79	0.82	0.80	

COMPARISON TO CNN USING IMAGE PROCESSING

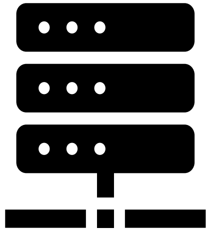
- CNN with 20/80 Test-Train Split
- Activation: Relu for Binary, Softmax for Multiclass
- Batch Size 32 and10 Epochs (one full cycle through all the batches in the entire training dataset)
- Binary Classification: 0.87 Accuracy
- Multiclass Classification: 0.99 Accuracy

CONCLUSION



<u>Model Performance</u>

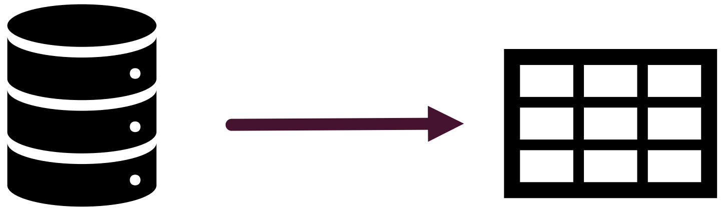
- CNN performs better in all metrics
- Random Forest accuracy is still promising



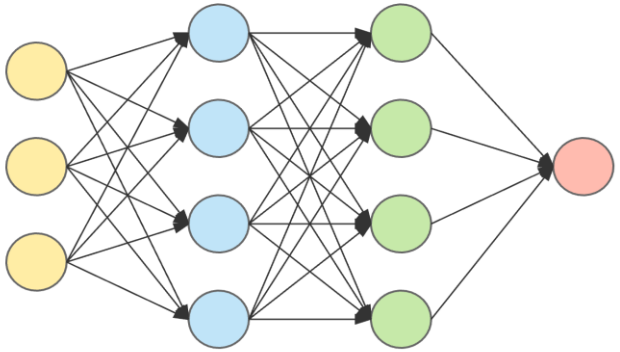
<u>Computational Efficiency</u>

- CNN is computationally expensive
- Feature extraction step is necessary for Random Forest

FUTURE WORK



Improving the Feature Extraction



Deep learning with limited labeled data points + Data Augmentation (GNN)

REFERENCES

[1] World Health Organization. Breast cancer-WHO Fact Sheet. https://www.who.int/

news-room/fact-sheets/detail/breast-cancer. 2024.

[2] National Breast Cancer Foundation (NBCF). Breast Cancer Facts & Stats. https://www.nationalbreastcancer.org/breast-cancer-facts/. 2024.

[3] Azar, A.T., & El-Metwally, S.M. (2013). Decision tree classifiers for automated medical diagnosis. Neural Computing and Applications, 23, 2387-2403.

[4] Azar, A.T., El-Said, S.A. (2014) Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Applic 24, 1163–1177. [14] Banu A, B., & Thirumalaikolundu

[5] Banu A, B., & Thirumalaikolundusubramanian, P. (2018). Comparison of Bayes Classifiers for Breast Cancer Classification. Asian Pacific journal of cancer prevention : APJCP, 19(10), 2917–2920. https://doi.org/10.22034/APJCP.2018.19.10.2917

[6] Chaurasia, V., Pal, S., Tiwari, B.B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology, 12, 119 - 126.

[7] Srivastava, U. P., Vaidehi, V., Koirala, T. K., & Ghosal, P. (2023, February). Performance Analysis of an ANN-based model for Breast Cancer Classification using Wisconsin Dataset. In 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) (pp. 1-5). IEEE.

REFERENCES

[8] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[9] Kaggle. (2019). BreakHis dataset. Retrieved from https://www.kaggle.com/datasets/ambarish/breakhis

[10] Spanhol, F., Oliveira, L. S., Petitjean, C., Heutte, L., A Dataset for Breast Cancer Histopathological Image Classification, IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016.

[11] Xu, X., Xu, S., Jin, L., & Song, E. (2011). Characteristic analysis of Otsu threshold and its applications. Pattern recognition letters, 32(7), 956-961.