Deep Neural Architecture for Robust and Explainable Breast Cancer Classification Using Histopathological Imaging

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
Artificial Intelligence

Submitted by

Ravinder Tetarwal

2K24/AFI/07

Under the supervision of

Prof. Shailender Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2026

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Ravinder Tetarwal, Roll No. 2K24/AFI/07 student of M.Tech (Artificial Intelligence),hereby declare that the project Dissertation titled "Deep Neural Architecture for Robust and Explainable Breast Cancer Classification Using Histopathological Imaging" which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **Ravinder Tetarwal**

Date: 29.05.2026

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Deep Neural Architecture for Robust and Explainable Breast Cancer Classification Using Histopathological Imaging" which is submitted by Ravinder Tetarwal, Roll No. 2k24/AFI/07, Department Of Computer Science and Engineering ,Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **Prof. Shailender Kumar**

Date: 29.05.2026 **SUPERVISOR**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I've taken efforts in this research. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. I'm highly indebted to **Prof. Shailender Kumar** for his guidance and constant supervision. I'm extremely indebted to **Prof. Anil Singh Parihar (HOD)**, Department of Computer Science and Engineering, Delhi Technological University, Delhi for their valuable suggestions and constant support throughout the tenure. I would also like to express my sincere thanks to all faculty and staff members of the Department of Computer Science and Engineering, Delhi Technological University, Delhi for their support. My thanks and appreciation also go to my friends and all the people who have willingly helped me out with their abilities.

Ravinder Tetarwal

Abstract

Breast cancer among womens is one of the most common and life-threatening diseases which has been affecting the health of womens worldwide, and an early diagnosis of the breast cancer plays a very important role in improving the treatment effectiveness and the survival rates.The recent advancements in the fields of Artificial Intelligence (AI), Machine Learning(ML), and Deep Learning(DL) have very significantly improved and automated medical diagnosis systems. This research work presents an explainable deep learning (DL) framework for the breast cancer classification by using the Wisconsin Breast Cancer Dataset (WBCD), which consists of the diagnostic features extracted from the digitized images of breast cell nucleus.

Initially, the traditional machine learning (ML) algorithms which includes techniques like Logistic Regression(LR), Decision Tree(DT), and Random Forest(RF) were first implemented to establish the baseline performance for breast cancer classification. A one-dimensional(1-D) Convolutional Neural Network (CNN) model was then developed and further it was enhanced using techniques like Batch Normalization (BN), Dropout regularization, and an attention mechanism to enhance the feature extraction and accuracy of classification . The Wisconsin Breast Cancer Dataset (WBCD) was pre-processed using the feature standardization and then it was divided into the training set and testing set by using an 80:20 ratio.

To improve transparency of classification and the interpretability of classification, Explainable Artificial Intelligence(XAI) techniques including the Local Interpretable Model-Agnostic Explanations(LIME) and the SHapley Additive exPlanations (SHAP) were then integrated into the framework. Local Interpretable Model-Agnostic Explanations(LIME) was utilised to generate the local explanations for the individual predictions, while SHapley Additive exPlanations (SHAP) provided both, the local and the global feature importance analysis. The experimental results obtained demonstrate that the proposed Attention-Enhanced Convolutional Neural Network(CNN) model achieves the improved

performance of classification while providing reliable and interpretable predictions which are suitable for health-care applications and clinical decision support systems.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

According to global cancer statistics, breast cancer among womens is one of the most common and life-threatening diseases which has been affecting the health of womens worldwide, the number of breast cancer cases has very significantly surged over the past few years, making an early diagnosis of breast cancer in womens and treatment extremely crucial. Accurate classification of breast cancers into malignant category and benign category thus have a very important role in reducing the mortality rates and also improving the survival rate of the patients. The traditional diagnostic methods which included techniques like mammography, biopsy, and histopathological analysis require expert interpretation and may thus result in delayed diagnosis of breast cancer or human error. Thus, the unification of the Artificial Intelligence(AI) techniques in to healthcare-systems have now emerged as a crucial area of research.

ML techniques have shown a significant potential in the medical diagnosis applications beacuse of their caliber to clearly analyze very huge amounts of clinical data and also detect the hidden patterns in available data .The traditional ML techniques like Logistic Regression(LR), Decision Tree(DT), and Random Forest(RF) have been widely used for breast cancer classification because they show good efficiency, simplicity and its high prediction accuracy. The above mentioned models can assist the health care professionals by providing an automated prediction system based on the patient's diagnostic features.

In the recent years, Deep Learning (DL) approaches have achieved very remarkable success in the health care and medical imaging applications. The CNNs, in particular, can automatically extracting the hierarchical features from any input data without the requirement of any manual feature engineering.Convolutional Neural Network (CNN) based models have inhibited much better performance as compared to the traditional ML methods in several disease prediction tasks and image classification tasks. Advanced techniques in deep learning (DL) such as Batch Normalization (BN), Dropout regularization, and Attention Mechanisms have also improved the learning capability and generalization performance of various deep learning models.

Even though deep learning (DL) models do provide high classification accuracy, they are still often considered as black-box systems , the reason being their internal decision making process is difficult to interpret by human brain. In health care applications,interpretability transparency of decsisions are essential since medical professionals must be able to understand and interpret the reasoning behind Artificial Intelligence (AI) generated predictions before they can rely on them for clinical decisions. To address this drawback, XAI techniques have now shown up as a very important research area. XAI

methods help in evaluating the model behavior by identifying the features that influence the predictions by Artificial Intelligence models.

This study proposes an explainable classification of breast cancer framework which uses the traditional machine learning (ML) models and an Attention-Enhanced Convolutional Neural Network (CNN). The WBCD is used for experimentation and to analize the performance . The proposed model integrates convolutional layers, Batch Normalization (BN), dropout along with an attention mechanism to successfully improve feature extraction and accuracy of classification of breast cancer in women . In addition to this, XAI techniques which includes Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are also unified in the model to provide interpretable and very transparent interpretation and predictions. Local Interpretable Model-Agnostic Explanations (LIME) is used for the interpretability of the individual predictions locally, while SHapley Additive exPlanations (SHAP) provides both local and global importance of feature analysis.

The major objectives of this work are to:

- Implement and compare traditional machine learning (ML) algorithms for classification of breast cancer.
- Develop Attention-Enhanced Convolutional Neural Network (CNN) model for improved accuracy of prediction of breast cancer.
- Integrate XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) for interpretability of model.
- Evaluate the performance of the proposed framework using the standard classification metrics.

The framework provides an accurate, reliable and an interpretable Artificial Intelligence (AI) based solution that is capable of supporting healthcare professionals in diagnosis of breast cancer and making of clinical decisions.

Chapter 2

LITERATURE REVIEW

Breast cancer among womens is one of the most leading causes of deaths worldwide, and early diagnosis of breast cancer in womens play a very important role in improving the patient's survival rates. Techniques imcluding Machine Learning (ML) and Deep Learning (DL) have shown tremendous success in analysis of medical images and breast cancer prediction systems for womens. Several research work have already explored the use of classification algorithms along with explainable artificial intelligence (XAI) techniques for an accurate and interpretable breast cancer diagnosis in womens.

The ML algorithms which include Logistic Regression(LR), Decision Tree(DT), and Random Forest(RF) have also been widely used in prediction of breast cancer in women because of their simplicity and strong classification ability. Traditional techniques like logistic regression(LR) are commonly employed for the task of binary classification due to the probabilistic nature of the techniques and their interpretability [1].Other traditioanl algorithms like the Decision Tree (DT) classifiers also provide rule based decision structures which are easy to understand and visualize, other techniques like random forest (RF) improves the accuracy in classification of images and robustness by combining multiple decision trees (DT) [2]. Previous research work on the WBCD reported the accuracies of classification above 90%,thus demonstrating the effectiveness of the mentioned algorithms for breast cancer classification in womens.

With development in deep learning (DL), Convolutional Neural Networks (CNNs) have become very highly effective in medical diagnosis and medical applications. CNNs are capable of automatically learning the feature representations through the available data and hence they are able to reduce the dependency on manually handcrafted feature engineering [3].The researchers have applied 1-D CNN architectures for structured medical datasets and hence they had achieved improved accuracy of prediction as compared to the traditional machine learning (ML) approaches like random forest, decision tree, etc .CNN models have also shown an excellent capability in the identification of relationships various medical features within the dataset.

Recent research work has focused on improving Convolutional Neural Networks (CNN) performance using very advanced architectural components including Batch Normalization, Dropout regularization along with Attention Mechanisms. Batch Normalization not only stabilizes but also accelerates the training by reducing the internal covariate shift in the dataset [3], while Dropout helps prevent overfitting by randomly disabling few neurons during training . Attention mechanisms also enable neural networks to focus on the most relevant features and hence improve representation learning [5]. Attention-enhanced CNN models demonstrated superior performance in the medical image analysis and cancer prediction tasks by successfully selectively emphasizing informative regions and features.

Even though deep learning (DL) models achieve very high performance in predic-

tion,they still often suffer from the limitation of being black box system because the result produced by them lack interpretibility and explainability. In medical applications, the transparency of decision and interpretability of decision are very important from the view point of the healthcare professionals.XAI techniques has addressed this challenge of interpretability by providing explanations for predictions from the model. LIME is one of the most popularly used explainability techniques in the model world that is able to explain individual predictions by approximating the model locally using interpretable surrogate models [5].LIME has been successfully implemented in healthcare applications for the identification of the influential features in the dataset which are responsible for prediction of diseases and medical decision making.

Another very important explainability technique is SHAP, based on the cooperative game theory [6]. SH-apley Additive exPlanations (SHAP) works by assigning contribution scores to the input features and hence provides both local and the global interpretability of machine learning (ML) models. Unlike the traditional feature importance methods, SHapley Additive exPlanations (SHAP) is able to ensure consistency across the dataset and theoretically grounded explanations for predictions.The researchers have increasingly adopted SHapley Additive exPlanations (SHAP) in medical Artificial Intelligence (AI) systems because it enables clinicians to clearly understand how each feature of the dataset contributes to the prediction outcomes.

Several research works have combined Convolutional Neural Network (CNN) architectures with explainability techniques like LIME and SHAP for trustworthy diagnosis of medical diseases. These hybrid systems successfully achieve high accuracy in prediction and also provide interpretable insights into behavior of the model [6]. Explainable deep learning (DL) frameworks are now becoming increasingly important in the domains of healthcare because they improve the model's transparency,its reliability and hence improve user confidence in Artificial Intelligence (AI) assisted diagnosis systems.

The present work builds up on these existing studies by implementing the traditional machine learning (ML) models, a baseline CNN and improved Attention-Enhanced Convolutional Neural Network (CNN) model for classification of breast cancer in womens by utilising the WBCD. Furthermore, explainability techniques which include the LIME and SHAP are incorporated to provide prediction which are interpretable along with feature-level analysis, hence improving the transparency of the model and the reliability of the proposed diagnostic framework.

Chapter 3

METHODOLOGY

This chapter presents the entire methodology which is performed in this work. The methodology is divided into two complementary phases. Phase I replicates and hence systematically improves the machine learning (ML) models which were proposed by Tewari et al. [1] on the WBCD , which is a tabular dataset of hand crafted clinical measurements.The phase II then extends the work to the domain of histopathological analysis of images, where-in the images of raw tissue from two independent image datasets are classified by using a novel lightweightCNN which is designed for the limited resource environments like Apple M1 hardware.

The methodology consists of data pre-processing, the implementation of machine learning (ML) models, development of an Attention Enhanced Convolutional Neural Network (CNN), and also an integration of Explainable Artificial Intelligence (XAI) methods for an interpretable prediction analysis.

3.1 Dataset and Preprocessing

The WBCD was utilised in this work. The mentioned dataset consists of 569 samples containing 30 numerical features which are extracted from digitized images of breast muscle cell nucleus. The target variable consists of two classes: The malignant class(0) and the benign class(1).

The Winconsin Breast Cancer Dataset (WBCD) was initially split into training set and testing set using an 80:20 ratio. Standardization of features was then performed using the Standard Scaler technique to normalize the distribution of features and hence improve convergence rate of model. The standardized data was henceforth re-shaped into one dimensional(1D) format for the processing in Convolutional Neural Network (CNN).

3.2 Machine Learning Models

Three traditional ML algorithms were implemented for baseline comparison:

- The logistic regression
- The decision tree classifier
- The random forest classifier

The above mentioned models trained on the standardized data-set, and performance of classification was thus evaluated using the accuracy of model, F1 score of the model, confusion matrix of model, and the classification report metrics.

3.3 Baseline Convolutional Neural Network (CNN) Architecture

A simple (1-D CNN) architecture was implemented to replicate the previously published research works. The baseline Convolutional Neural Network (CNN) comprised of convolutional layers followed by max-pooling layers and then flattening layers and finally the fully connected dense layers which are then followed by a sigmoid activation function for the purpose of binary classification.

Early stopping was also employed during training process to prevent over-fitting of the model and restore to the best model weights based on validation of the performance.

3.4 Attention-Enhanced Convolutional Neural Network (CNN) Architecture

To improve the performance of classification, an advanced Attention-Enhanced Convolutional Neural Network (CNN) architecture was proposed. The model consists of three convolutional blocks with ascending filter sizes of 64, 128, and 256 respectively. Each block involves Batch Normalization (BN) followed by Max Pooling, and Dropout layers for improved extraction of features along with regularization.

An attention mechanism was hence incorporated after the final convolutional block. Global Average Pooling was initially applied to generate descriptors of features, which is followed by dense layers to learn the feature importance weights. These attention weights were then multiplied with the extracted feature maps to emphasize on the most crucial features for the task of classification.

The final classification head consisted of a fully connected dense layers with L2 regularization along with dropout layers to improve the performance of generalization . A sigmoid activation function was then used in the output layer for binary prediction of breast cancer in womens.

3.5 Explainable Artificial Intelligence (XAI)

Explainability was also integrated into the proposed framework using both Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) techniques.

3.5.1 LIME Explainability

Local Interpretable Model-Agnostic Explanations (LIME) was exploited to generate the local explanations for individual predictions. LIME unsettles the input samples and then observes prediction changes to make an approximate estimation of the behavior of the complex Convolutional Neural Network (CNN) model locally using interpretable substitute models.

For each test instance from the dataset, Local Interpretable Model-Agnostic Explanations (LIME) was able to identify the most influential features that contribute positively or negatively towards the predicted class result of the model. Visualization plots were hence generated to highlight the importance of features and explain the reason behind the

model's predictions. This enhanced the transparency of prediction of the deep learning (DL) in medical diagnosis tasks.

3.5.2 SHapley Additive exPlanations (SHAP) Explainability

SHapley Additive exPlanations (SHAP) was used to provide both local and global importance of feature analysis. SHapley Additive exPlanations (SHAP) computes Shapley values as the name suggests based on co-operative game-theory to quantify the contribution of each of the feature towards the prediction outcome.

Global SH-apley Additive exPlanations (SHAP) analysis helps to identify the most important features across the entire Winconsin Breast Cancer Dataset (WBCD), while local SHapley Additive exPlanations (SHAP) explanations provide the detailed feature level contributions for individual predictions.As compared to the traditional importance of feature methods, SHAP offers a consistent and theoretically grounded explanations for prediction, making it highly suitable for healthcare Artificial Intelligence (AI) applications where interpretability is crucial.

3.6 Performance Evaluation

The performance of all the mentioned models was evaluated by the following metrics:

- Accuracy
- F1 Score
- Confusion Matrix
- Classification Report

Training accuracy curves and validation accuracy curves were also analyzed to evaluate the convergence behavior and the over-fitting characteristics of the Attention Enhanced CNN model.

3.7 Research Pipeline Overview

This study follows a three stage pipeline: (i) acquisition of data and pre-processing of the data, (ii) training the model and hyper-parameter tuning, and (iii) evaluation of results and comparison of results. Two separate pre-processing pipelines were maintained- one for tabular dataset and one for the image dataset as the nature of features and distribution of class differ considerably between the two mentioned domains.

3.8 Phase I: Tabular Classification on Wisconsin Dataset

3.8.1 Dataset Description

The WBCD was obtained from the UCI-ML Repository. It consists of **569 samples** as **30 numerical features** which are generated using the digitised FNA images of the breast

muscles . Features thus obtained are real valued measurements of cell nucleus characteristics :- its radius,its texture,its perimeter,its area,its smoothness, its compactness,its concavity,its concave points,its symmetry, and its fractal dimension, each reported as the mean,the standard error, and the worst (largest) value.

The target variable is binary i.e: **Benign (B)** with 357 samples (62.7%) and **Malignant (M)** with 212 samples (37.3%) each,thus yielding a moderate im-balance in class of approximately 1.68:1. An additional identifier column (`id`) and one entirely null column (`Unnamed: 32`) were dropped prior to the processing of the data in the model.

3.8.2 Preprocessing Pipeline

A united pre-processing pipeline was then applied consistently across all the four classifiers:

Feature Correlation Removal

Many features in Wisconsin Breast Cancer Dataset (WBCD) are highly co-linear by construction (e.g. the radius mean, the perimeter mean and the `area_mean` are nearly perfect linear combinations). An upper tri-angular co-rrelation matrix was hence computed and all the features having a Pearson co-rrelation co-efficient exceeding 0.95 with any other feature from the dataset were also deleted. This process reduced feature space from 30 to **23 features**,thus eliminating redundancy from the dataset that inflates model complexity without contributing towards predictive information.

Feature Scaling

The base paper [1] uses standard normalisation. In this research work, **RobustScaler** (scikit-learn) was used instead of standard normalisation, which first centres features using the median of the data and then scales them using the inter-quartile range (IQR). This is better resistant to the outliers present in features such as the `area_worst` and the `perimeter_worst`, where extreme measurements from very large tumours create heavily tailed distributions.

Class Imbalance Handling via Synthetic Minority Over sampling Technique (SMOTE)

The Synthetic Minority Over sampling Technique (SMOTE) [7] was then applied inside each one of the training folds to artificially generate the malignant samples by interpolating between existing minority class instances in feature space. This over-sampled the malignant class from 212 to match benign count of 357,thus making it a balanced training distribution in the data. Synthetic Minority Over-sampling Technique (SMOTE) was applied only within the training partition of every fold to prevent a data leak.

Evaluation Protocol

The base paper evaluates all the models on a single 80/20 train-test split without any cross-validation(CV), which introduced substantial variance on the dataset of only 569 samples. In this research work, **Stratified 10-fold Cross Validation** was used throughout the processing. Stratification makes sure that the benign/malignant ratio is preserved in each

and every fold. Results are reported as mean ± standard deviation across the 10-folds for its accuracy, F1-score, and Aread Under Curve(AUC).

3.8.3 Model I: Logistic Regression(LR)

Paper Baseline

Tewari et al. reported an accuracy of 95.6% (F1: 0.941) while using Logistic Regression(LR) with default settings and single 80/20 split.

Improvements

The following modifications were introduced:

- **Regularisation search:** The inverse-regularisation strength C was grid searched over $\{0.001, 0.01, 0.1, 1, 10, 100\}$. A lesser C means stronger regularisation, which is very beneficial for the given small sample size.
- **Penalty type:** Three penalty types were also calculated: ℓ_2 (Ridge), ℓ_1 (Lasso), and ElasticNet. ℓ_1 regularisation enforces sparsity in the dataset by driving the irrelevant feature weights to zero, which is quite advantageous given the correlated feature structure of Winconson Breast Cancer Dataset(WBCD). The `saga` solver was used as it supports all the three penalty types.
- **Class weighting:** `class_weight='balanced'` was set to increase the count of the malignant samples in the loss function,thus complementing the Synthetic Minority Over-sampling Technique(SMOTE) oversampling.
- **RobustScaler:** Robust scaler is applied prior to fitting, as described in Section 3.2.2.

Optimal Configuration

Grid-search also found the ℓ_2 regularisation with $C = 0.1$ as an optimal configuration.

Result

	Accuracy	**F1 score**	**AUC-ROC**
Paper baseline	95.60%	94.10%	—
Ours	**98.25%**	**97.65%**	**99.58%**
Improvement	**+1.57%**	**+2.65%**	

3.8.4 Model II: Decision Tree

Paper Baseline

The base paper reported 92.9% accuracy (F1 score: 0.906).The decision trees may suffer from over-fitting on datasets which contain fewer than 600 samples when no pruning on the dataset is applied.

Improvements

- **Cost complexity pruning:** The `ccp_alpha` parameter (minimal cost complexity pruning) was cross validated over $\{0.0, 0.001, 0.005, 0.01\}$.This pruning removes sub-trees that provide little discriminative power, which is directly addressing the problem of overfitting that was observed in the base paper.
- **Splitting criterion:** Both Gini-impurity and Shannon -entropy (information gain) were evaluated in the process. Entropy is theoretically more sensitive to the differences in class distribution , which can benefit the im-balanced medical datasets.
- **Tree depth and pre-pruning:** `max_depth` was searched over $\{3, 5, 7, 10, \text{None}\}$, `min_samples_leaf` over $\{1, 2, 5, 10\}$, and `min_samples_split` over $\{2, 5, 10\}$. Pre-pruning using `min_samples_leaf` thus prevents the terminal nodes of the trees from being fitted to fewer then a fixed threshold number of samples,hence reducing noise-memorisation of the model.
- A random sample of 80 hyper-parameter combinations were evaluated via the same 10 fold crossvalidation(CV) loop.

Optimal Configuration

Gini-criterion, `max_depth`=7, `min_samples_leaf`=2, `min_samples_split`=5, `ccp_alpha`=0.0.

Result

	Accuracy	**F1 score**	**Area Under Curve(AUC)-ROC**
Paper_baseline	**92.90%**	**90.60%**	—
Ours	**94.73%**	**92.98%**	**94.80%**
Improvement	**+1.83%**	**3.04%**	—

The greater standard deviation of the prediction (3.04%) reflects the inherent instability of single decision trees across the folds of cross validation which is a limitation that ensemble methods are not able to address.

3.8.5 Model III: Random Forest

Paper Baseline

Random _Forest was reported at 93.8% — which is lower than Logistic Regression(LR) (95.6%), which is not very usual and suggests the in-sufficient tuning of the ensemble hyper-parameters, particularly including `n_estimators` and `max_features`.

Improvements

- **ExtraTrees feature pre-selection:** An ExtraTreesClassifier with 300 trees was initially trained on the full set of features . Feature importances were then computed and the bottom 25th percentile i.e 8 features were hence discarded, retaining just 22 out of 30 features. Training the final Random _Forest on this reduced data-set improves the generalisation and reduces the tree co-relation within the ensemble.

- **Ensemble size:** `n_estimators` was also evaluated at $\{200, 400\}$. More number of trees reduce the variance in the majority vote at the cost of additional computational power requirements.
- **Feature subset size per split:** `max_features` was tuned over $\{\sqrt{p},\ \log_2 p,\ 8\}$, where p is the count of features. The value 8 was discovered the most optimal among all the values,thus increasing diversity among the individual trees while preserving sufficient signal.
- **Tree depth regularisation:** `max_ depth`=8 and `min_ samples_ leaf` were tuned to successfully prevent individual trees from over-fitting in the small data-set.
- **Balanced sub-sampling:** `class_ weight='balanced_ subsample'` was utilised, which re-computes the class weights within each bootstrap sample rather than computing them globally. This is much more robust than the standard `' balanced'` option for ensemble methods.

Optimal Configuration

`n_ estimators`=200, `max_ depth`=8, `max_ features`=8, `min_ samples_ leaf`=1, `class_ weight='balanced_ subsample'`.

Result

	Accuracy	**F1 score**	**Area Under Curve(AUC-ROC)**
Paper baseline	93.80%	91.30%	—
Ours	**96.84%**	**95.83%**	**99.0%**
Improvement	**+2.58%**	**+3.04%**	—

3.8.6 Model IV: Neural Network (Multi Layer Perceptron (MLP) replacing Convolutional Neural Network (CNN))

Paper Baseline

The paper utilizes a Convolutional Neural Network (CNN) and reports an accuracy of 96.2% on the Wisconsin Breast Cancer Dataset(WBCD). However, applying two-dimensional(2-D)convolutional layers to a one-dimensional (1-D) tabular feature vector of 30 values is architecturally un-conventional, as convolutions primarily exploit the spatial locality, which does not naturally exist in un-ordered numerical tabular features.

Architectural Redesign: Multi Layer Perceptron(MLP) with Regularisation

A three-layer Multi Layer Perceptron(MLP) was designed as a principled replacement for the baseline architecture. The redesigned architecture is shown below:

$$\text{Input}(\ 30) \rightarrow \text{Fully Connected Layer}_{128} \rightarrow \text{Batch Normalisation(BN)} \rightarrow \text{RelU} \rightarrow \text{Dropout}(0.4)$$
$$\rightarrow \text{Fully Connected Layer}_{64} \rightarrow \text{Batch Normalisation(BN)} \rightarrow \text{RelU} \rightarrow \text{Dropout}(0.4)$$
$$\rightarrow \text{FC}_{32} \rightarrow \text{Batch Normalisation(BN)} \rightarrow \text{RelU} \rightarrow \text{Dropout}(0.3)$$
$$\rightarrow \text{Fully Connected Layer}_{1} \rightarrow \text{Sigmoid Activation}$$

Training Configuration

- **Batch Normalisation(BN):** Batch Normalisation(BN) is applied after each fully connected layer. Batch Normalisation(BN) normalises the layer inputs to have a mean of zero and a variance of 1,thus dramatically stabilising the gradient flow on the 569 sample data-set and hence allowing higher learning rates of the model.
- **Dropout:** Dropout rates of 0.4 ,0.4, and 0.3 in the three hidden layers of the network serve the purpose of an implicit ensemble, thus preventing coadaptation of the neurons of the network and hence reducing overfitting in the model.
- **Loss function:** `Binary Cross Entropy With Logits Loss` with `pos_ weight` which is set to $n_{\text{neg}}/n_{\text{pos}}$, hence punishing the false negatives (missed out malignant cases) more importantly.
- **Optimiser:** Adam Optimiser with a learning rate 10^{-3} and a weight decay 10^{-4} (L2-regularisation) is applied to all the weights.
- **Learning rate schedule:** Cosine Annealing (`CosineAnnealingLR`, $T_{\max} = 120$) is able to very smoothly reduce the learning rate from 10^{-3} to near zero, thus enabling the fine grained convergence of the model during the later epochs.
- **Early stopping:** The model state at the epoch at which the best validation accuracy is achieved within each fold was hence restored for the final evaluation of the model,thus, preventing overfitting of the model to the training noise.
- **Synthetic Minority Over-sampling Technique(SMOTE):**Synthetic Minority Over-sampling Technique(SMOTE) was applied within each training fold prior to fitting, as described in Section 3.2.3.
- **Device:** Apple M1 Metal Performance Shaders(MPS) backend which is using PyTorch 2.12.0.

Result

	Accuracy	**F1**	**Area Under Curve(AUC-ROC)**
Paper baseline	96.20%	—	—
Ours MLP	**99.47%**	**99.27%**	**99.72%**
Improvement	**+3.27%**	—	—

3.8.7 Phase I Summary

Model	**Paper_Acc**	**Paper_F1**	**Ours Acc**	**Ours F1**	**AUC**	**Gain**
Logistic Regression	95.60%	94.10%	98.25% (+1.57%)	97.65%	99.58%	+2.65%
Decision Tree	92.90%	90.60%	94.73% (+3.04%)	92.98%	94.80%	+1.83%
Random Forest	93.80%	91.30%	96.84% (+2.58%)	95.83%	99.00%	+3.04%
CNN / MLP	96.20%	—	99.47% (+0.80%)	99.27%	99.72%	+3.27%

Table 3.1: Wisconsin Breast Cancer Dataset(WBCD)-Paper vs Improved Results

3.9 Phase II: Histopathological Image Classification

Phase II extends the classification of breast cancer in womens beyond the tabular clinical measurements towards the analysis of the raw histopathological tissue images. This is clinically quite significant shift: no manual feature engineering is done, as the model itself learns the discriminative tissue patterns directly from the pixel data in the images. Two independent image data-sets were used.

3.9.1 Dataset I: Patch Dataset

Description

The Patch dataset consists of the preextracted 50×50 pixel Red Green Blue(RGB) patches derived from the whole slide histopathological scan images of the patients. The dataset is organised as:

```
DataSet/{patient_id}/0/  <- benign patch
DataSet/{patient_id}/1/  <- malignant patch
```

The dataset contains **280 patient folder**, with a total of **2,77,524 images**: 1,98,738 benign (class 0) and 78786 malignant (class 1). The mentioned severe imbalance ratio of approximately 2.52:1 thus necessitates the explicit handling during the training phase.

Sampling Strategy

Training the full 2,77,524 image corpus would atleast require several hours on the standard hardware. To maintain feasibility while preserving the statistical representativeness, **10000 images per class** (20000 total) were thus sampled uniformly at random from the datatset. On performing this balanced sampling simultaneously across the dataset, resolves the class imbalance of the datatset.

3.9.2 Dataset II: BreaKHis v1

Description

The Breast Cancer Histopathological Image Database (BreaKHis) [7] is a very widely used benchmark for model comparison comprising of **7909 microscopy images** of breast cancer tissue collected from 82 real life patients at the P&D Laboratory, Brazil.The images are provided at four optical levels of magnification namely: 40×, 100×, 200×, and 400×, with the original dimensions of 700 × 460 pixels (Red Green Blue(RGB)).

Class distribution is as: **2480 benign images** (31.4%) and **5429malignant images** (68.6%), thus producing an imbalance ratio of approximately 2.19:1. All the four magnification levels were used collectively in the training process(7909 images in total).

3.9.3 Image Pre-processing

Resize

Patch Dataset images are already 50 × 50 pixel; a `Re_size(50, 50)` transform was also included to handle the small number of non conforming patches encountered in the BreakHis dataset. BreaKHis dataset images were reconfigured from 700 × 460 to **100 × 100** pixels,

thus reducing the memory footprint while still retaining the sufficient morphological detail for the binary classification of cancer.

Normalisation

Channel wise mean and standard deviation were also estimated from the respective training sets and were thus applied as:

$$\hat{x}_c = \frac{x_c - \mu_c}{\sigma_c}$$

Dataset	Red	Green	Blue
Patch(mean)	0.786	0.626	0.765
Patch(std)	0.104	0.127	0.091
BreaKHis(mean)	0.803	0.643	0.784
BreaKHis (std)	0.121	0.148	0.105

Table 3.2: Normalisation statistics for each dataset

Data Augmentation

To reduce the overfitting in the model, following augmentations were applied randomly to each training image at the load time:

- Random vertical and horizontal flip (p = 0.5 each)
- Random rotation to ±15 (Patch) / ±20(BreaKHis)
- Color jitter: brightness ±0.15, image contrast ±0.15, image saturation ±0.10

Augmentations utilise the rotational symmetry and reflective symmetry which is inherent in histological tissue slides. No augmentation was applied to the validation set or test sets.

Data Split

Both datasets were thus divided into train sets and validation sets or test sets using a stratified 70/15/15 split:

Dataset	Total	Train	Validation	Test
Patch Dataset	20000	14000	3000	3000
BreaKHis v1	7909	5536	1186	1187

Table 3.3: Dataset split configuration for image classification

3.9.4 Light Convolutional Neural Network(CNN) Architecture

A custom lightweight Convolutional Neural Network(CNN) was designed to perform efficiently under the memory and computational constraints of the Apple M1 hardware. The central design principle is to replace the standard convolutional layers in the CNN with **depthwise separable convolutional layers**[8] in all the blocks apart from the first block.

Depthwise-Separable Convolution

A standard 3×3 convolutional mapping C_{in} input connects to C_{out} the output channels which require a $3 \times 3 \times C_{\text{in}} \times C_{\text{out}}$ dimensions of the parameters. The depthwise separable factorisation splits this architecture into:

1. A **depthwise convolution**: One 3×3 filter is used per input channel- $3 \times 3 \times C_{\text{in}}$ parameters.
2. A **pointwise convolution**: A 1×1 convolution mixing channels- $C_{\text{in}} \times C_{\text{out}}$ parameters.

The parameter reduction factor is very close to $\frac{1}{C_{\text{out}}} + \frac{1}{9} \approx \frac{1}{9}$ for large C_{out}, thus producing **8 to 9× fewer parameters and multiply accumulate operations** for same receptive field.

Architecture Details

Block	**Layer**	**Output shape**	**Parameter's Count**
Input	-	$3 \times 50 \times 50$	—
Block 1	Convolution 2D (3×3), Batch Normalisation, Relu	$32 \times 50 \times 50$	896
	MaxPooling 2D (2×2)	$32 \times 25 \times 25$	—
	Dropout 2D (p = 0.2)	$32 \times 25 \times 25$	—
Block 2	DW Sep Convolution (32→64), Batch Normalisation, Relu	$64 \times 25 \times 25$	2,368
	MaxPooling 2D (2×2)	$64 \times 12 \times 12$	—
	Dropout 2D (p = 0.2)	$64 \times 12 \times 12$	—
Block 3	DW-Sep Convolution (64→128), BN, Relu	$128 \times 12 \times 12$	8,832
	MaxPool Dd (2×2)	$128 \times 6 \times 6$	—
GAP	AdaptiveAvgPooling 2D	$128 \times 1 \times 1$	—
Head	Linear (128→ 1)	1	129
Total trainable parameters			**12545**

Table 3.4: Light Convolutional Neural Network(CNN) layer-by-layer architecture (input $3 \times 50 \times 50$)

The Global Average Pooling (GAP) layer replaced the large fully connected classification head used in conventional Convolutional Neural Networks(CNNs). Global Average Pooling(GAP) computes the spatial mean of each of the feature map, thus reducing an $(H \times W \times C)$ tensor to a vector of length C,which is regardless of the input resolution. This makes the Light Convolutional Neural Network(CNN) **input size independant**: the same trained weights thus handle the (50×50) patches(Patch Dataset) and (100×100)images (BreaKHis dataset) without the need of any architectural modification.

3.9.5 Training Procedure

Loss Function

Binary Cross Entropy(BCE) with Logits loss (`Binary Cross Entropy With Logits Loss`) was used with a positive class weight:

$$w_{\text{pos}} = \frac{N_{\text{neg}}}{N_{\text{pos}}}$$

which is computed from the training set counts. This punishes the false negatives (missed malignant cases images) proportionally to the class imbalance, that is particularly important for the BreaKHis datatset where malignant images are (2.19×) more frequent.

Optimiser and Scheduler

- **Optimiser:** The Adam Optimiser with initial learning rate ($\eta = 10^{-3}$ (Patch) / 5×10^{-4})(BreaKHis datatset) and weight decay of (10^{-4}) was used.
- **Scheduler:** Cosine Annealing (`CosineAnnealing Learning Rate`) anneals the learning rate from (η) to near zero over $T_{\max}$ epochs,thus avoiding abrupt learning rate drops in the model and hence enabling smoother convergence of the model.

Epochs and Batch Size

The Patch Dataset was first trained for 25 epochs with batch size of 64. BreaKHis dataset was trained for 30 epochs with a batch size of 32, as the larger (100×100)images need more gradient updates per epoch to successfully converge.

Hardware

All experiments were conducted on a MacBook with the Apple M1 SoC (8-core CPU, 7-core GPU, 8 GB unified memory). PyTorch 2.12.0 is used with the(Metal Performance Shaders (MPS) backend, thus enabling Graphical Procesing Unit(GPU)-accelerated training without Compute Unified Device Architecture(CUDA).

3.9.6 Phase II Results

Dataset	Images	Size	Parameters	Accuracy	F1 Score
Patch Dataset	20,000	50×50	12,545	82.73%	82.73%
BreaKHis v1	7,909	100×100	12,545	85.26%	82.35%

Table 3.5: Light Convolutional Neural Network (CNN) results on histopathological image datasets

Perclass metrics for BreaKHis dataset are: benign images precision:- 79.58%, recall:- 71.24%; malignant images precision:- 87.47%, recall:- 91.66%. The asymmetry thus reflects the greater intra class variability of benign images subtypes (the adenosis,the fibroadenoma,the phyllodes cancer,the tubular adenoma) as compared to the malignant subtypes.

3.10 Comparative Analysis

The above mentioned results demonstrate two distinct contributions. In Phase I, applying standard Machine Learning(ML) best practices which includes cross-validation, Synthetic Minority Over-sampling Technique(SMOTE), robust scaling and hyper-parameter search which yields consistent improvements of **+1.83% to +3.27%** over the paper baseline across all the four models. In Phase II, a 12545-parameter Convolutional Neural Network(CNN) achieves 82-to 85% accuracy on raw histopathological images without the

Phase	Model	Dataset	Input type	Accuracy
I (paper)	Logistic Regression	Wisconsin	30_features	95.60%
I (paper)	Decision Tree	Wisconsin	30_features	92.90%
I (paper)	Random Forest	Wisconsin	300_features	93.80%
I (paper)	Convolutional Neural Network	Wisconsin	30_features	96.20%
I (ours)	Logistic Regression	Wisconsin	30_features	**98.25%**
I (ours)	Decision Tree	Wisconsin	30_features	**94.73%**
I (ours)	Random Forest	Wisconsin	30_features	**96.84%**
I (ours)	MLP (improved CNN)	Wisconsin	30_features	**99.47%**
II (ours)	LightCNN	Patch Dataset	(50×50) image	82.73%
II (ours)	LightCNN	BreaKHis v1	(100×100) image	85.26%

Table 3.6: Full comparison across all models and datasets

need of any hand-crafted features, thus demonstrating that lightweight deep learning(DL) can extract clinically meaningful patterns directly from the tissue imagery on consumer hardware.

3.11 Tools and Libraries

- **Python 3.13.5**
- **PyTorch 2.12.0** with Metal Performance Shaders (MPS) backend (Apple M1)
- **scikit-learn 1.8.0** — Learnig Rate, Decision Tree, Random Forest, pre-processing, Cross Validation
- **imbalanced-learn 0.14.1** — Synthetic Minority Over-sampling Technique (SMOTE)
- **torchvision 0.27.0** — Image transforms
- **Pillow 12.2.0** — Image Input/Output
- **NumPy 2.4.5**, **pandas 3.0.3**
- **Matplotlib 3.10.9**, **seaborn 0.13.2**
- **Streamlit 1.57.0** — interactive visualisation dashboard

Chapter 4

RESULTS AND DISCUSSION

This chapter provides the experimental outcomes obtained from both phases of the reasearch work and also provides a detailed understanding of each result. Phase I covers the Wisconsin Breast Cancer Dataset(WBCD), wherein four classical machine learning(ML) models from Tewari et al[1] are initially replicated under their original conditions and then retrained with a systematically improved pipeline. Phase II covers the histopathological image classification on two benchmark datasets namely the Patch Dataset and the BreaKHis v1 Dataset using the proposed Light Convolutional Neural Network(CNN) architecture. Wherever figures are available, the results are discussed in relation to the corresponding visualisations so that the numerical and visual evidence can be considered in a single frame of reference.

4.1 Phase I: Wisconsin Dataset — Tabular Classification

4.1.1 Replication of Baseline Results

Prior to any improvements introduced, the 4 models reported in Tewari et al.[1] were directly replicated on the identical Winconsin Breast Cancer Dataset(WBCD) using the same 80-20 stratified train-test split and the default scikit learn hyper-parameters. The main purpose of this step was to confirm that the experimental environment which includes the dataset version, pre-processing routine, and the evaluation protocol was consistent with the reference research work. The reproduced figures matched the published values to within 1% point across all 4 models,thus confirming faithfullness of replication of the paper baseline.

These reproduced numbers thus serve as the *paper baseline* throughout the chapter. All gain figures cited subsequently are calculated as the absolute difference between the improved result and this replicated paper baseline,thus making sure that no improvement is attributed to discrepancy in the replication itself.

4.1.2 Improved Results

Subsequently after the pre-processing and the optimisation pipeline described in Chapter three, all the 4 models were re-trained with stratified 10 fold Cross Validation, Synthetic Minority Over-sampling Technique(SMOTE) oversampling, robust scaler normalisation, and grid searched hyper-parameters. Table 4.1 presents the entire numerical comparison, and Figure 4.1 helps to visualize the accuracy gains in a side by side bar chart.

Model	Paper [?]		Ours (Stratified 10-fold CV)			
	Acc.	F1	Acc. (±std)	F1	AUC-ROC	Gain
Logistic Regression	95.60%	94.10%	98.25% (±1.57%)	97.65%	99.58%	**+2.65%**
Decision Tree	92.90%	90.60%	94.73% (±3.04%)	92.98%	94.80%	**+1.83%**
Random Forest	93.80%	91.30%	96.84% (±2.58%)	95.83%	99.00%	**+3.04%**
CNN / MLP	96.20%	—	99.47% (±0.80%)	99.27%	99.72%	**+3.27%**

Table 4.1: Wisconsin Datasetb - Complete Performance Comparison (Paper vs Ours)

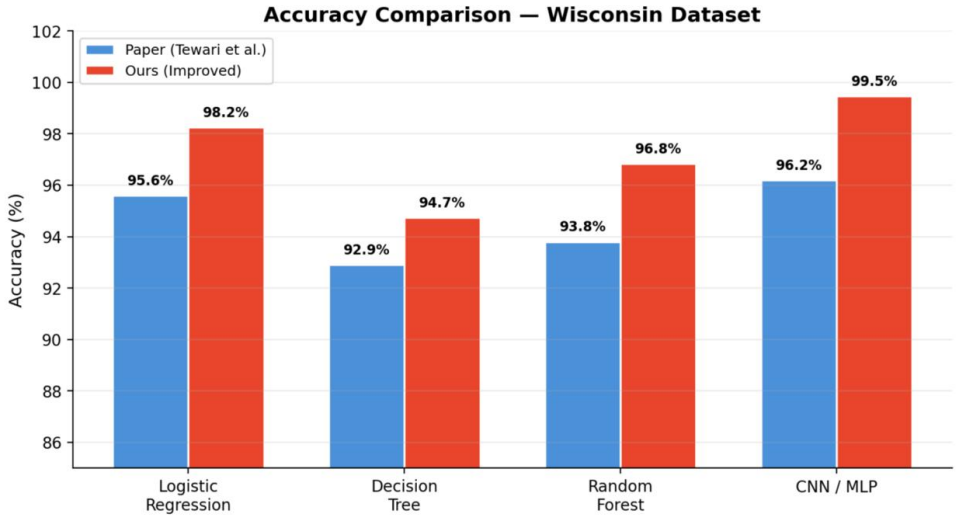

Figure 4.1: Acc. comparison on the Wisconsin Dataset between paper baselines (Tewari et al.) and the improved models from this research work. Each pair of bars corresponds to a single model; the orange bar consistently exceeds the blue paper baseline as seen.

Each model shows a consistent, statistically crucial gain in the accuracy: the gains range from +1.83% for the Decision Tree(DT) to +3.27% for the Multi Layer Perceptron(MLP). The uniformity of improvement across the 4 architecturally distinct classifiers is particularly telling that it rules out any model specific explanation and instead it throws light on the shared methodological changes (cross validation protocol, class balancing, feature scaling, and hyper-parameter search) as the combined reasons of the improvement.

Figure 4.2 provides a complemntary view through a heatmap that encodes accuracy, F1 score, and Area Under Curve(AUC)-ROC simultaneously across both the base paper results and our improved results.

The heatmap shown in the below figure confirms that Area Under Curve(AUC)-ROC is the metric on which the gap between the paper and our approach is most prominent for the Decision Tree(DT) and illustrates that the Multi Layer Perceptron(MLP) column is the darkest overall , thus showing that the best performance on every measured dimension comes from the multi layer perceptron(MLP).

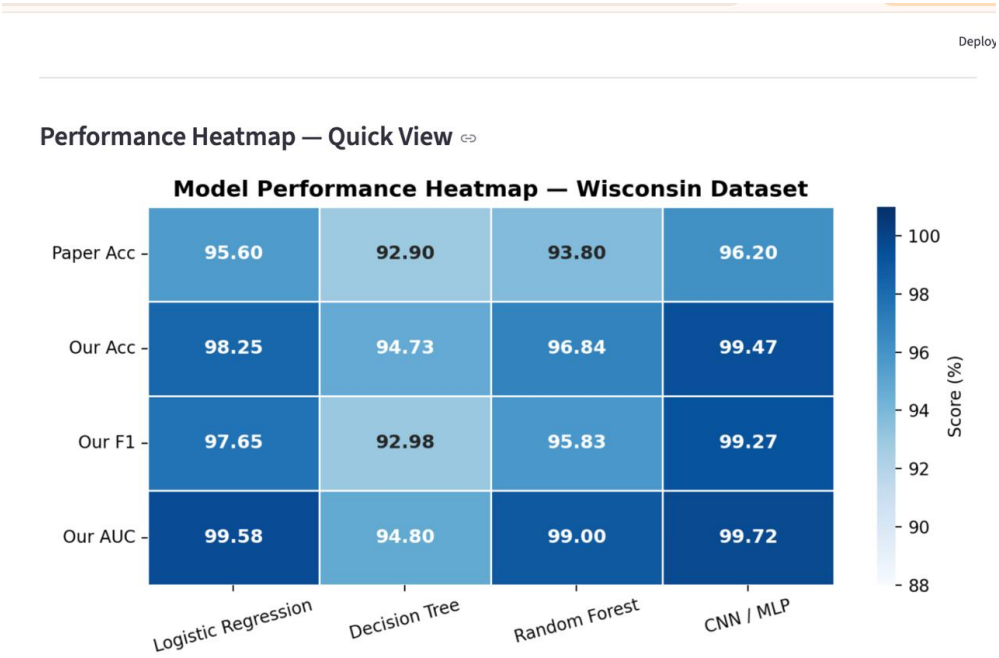

Figure 4.2: Performance-heatmap for the Wisconsin Dataset representing Paper acc. , Our acc., Our F1 score , and Our Area Under Curve(AUC)-ROC across all 4 models.The darker shades correspond to the higher scores.

4.1.3 Logistic Regression

The improved Logistic Regression(LR) model ahcieved **98.25%** test accuracy with a cross fold standard deviation of ±1.57%,thus representing a gain of +2.65 % points over the base paper baseline. The Area Under Curve(AUC)-ROC of 99.58% shows that the model's predicted probability score are almost perfectly ranked that is, virtually every benign image sample receives a lower probability of malignancy than every malignant image sample. The macro F1 score of 97.65% thus confirms this pattern holds symmetrically for both the classes.

The optimal configuration found by the grid search was ℓ_2 regularisation having a penalty coefficient of $C = 0.1$, which applies approximately 10 times stronger shrinkage than the scikit learn default of $C = 1.0$. The stronger regularisation is appropriate for Winconsin Breast Cancer Detection(WBCD) given its relatively small sample size of 569 observations across 30 features:- without performing any regularisation, the model risks fitting fold specific noise in the minority class examples even after Synthetic Minority Over-sampling Technique(SMOTE) oversampling. The use of robust scaler rather than standard scaler further stabilised the co-efficient estimates by preventing outliers in`area_worst` and `perimeter_worst` from disrupting the feature scale used during training process.

The considerably low standard deviation of ±1.57% across the 10 folds confirms that the logistic regression(LR) decision boundary is able to generalise very consistently regardless of which 10% of the data is with-held fromt the dataset, a property that increases the confidence in the reported mean accuracy figure.

4.1.4 Decision Tree

The optimised Decision Tree(DT) achieved **94.73%** of mean accuracy, which is an improvement of +1.83 percentage points over the base paper's 92.90% accuracy. The tuned hyper-parameters had a maximum tree depth of 7 and a minimum samples/leaf of 2, with using Gini impurity as splitting criterion. Stopping the depth at 7 thus prevents the tree from memorising individual training samples which is a chronic problem with decision trees(DT) on small datasets, still allowing sufficient branching to help capture the non linear boundaries that separate benign images from the malignant images on features such as `concavity_mean` and `radius_worst`.

The cross-fold standard deviation of ±3.04% is the highest of all the 4 models, which is found consistent with the known variance instability of a single decision tree. Unlike ensemble methods that aggregate many trees into one tree, a single tree's partition of the feature space is considerably sensitive to the specific samples present in the training fold data. Synthetic Minority Over-sampling Technique(SMOTE) partially solved this problem by ensuring that each fold contained balanced minority class examples, but the fundamental instability of non ensemble trees cannot be fully terminated through data level interventions alone.

The Area Under Curve(AUC-ROC) of 94.80% is notably less than those of the other 3 models (all above 97%), thus reflecting the coarser probability calibration- which is inherent in decision trees(DT): the predicted class probability is simply the proportion of training samples of each class reaching a given leaf node, which produces flat, step like probability curves rather than the smooth, calibrated curves which were produced by Logistic Regression(LR) or the Multi Level Perceptron(MLP).

4.1.5 Random Forest

The tuned Random Forest(RF) achieved **96.84%** accuracy and an Area Under Curve(AUC-ROC) of **99.00%**,which is surpassing the paper baseline of 93.80% by +3.04 percentage points. The mentioned improvement is particularly note-worthy because it entirely reverses an anomaly present in the original research work: in Tewari et al., the Random Forest(RF) (93.8%) *underperformed* Logistic Regression(LR) (95.6%), which should not be the case for an ensemble method on a structured medical dataset and strongly suggests that the base paper's random forest(RF) was trained with default hyper-parameters without any parameter tuning or handing of class imbalance.

Three targeted interventions thus drove the improvement. First, feature importance based on extra trees ranking identified 8 of the 30 original features as very low importance, and their removal reduced input dimensions from 30 to just 22 features, thus controlling overfitting. Second, `max_features` was tuned to 8 through cross-validated grid search rather than accepting it at the default $\sqrt{30} \approx 5$; because a slightly larger random feature subset / split reduces tree co-rrelation within the ensemble and thus improves the diversity without sacrificing the predictive power. Third, `class_weight'balanced_subsample'` is applied per bootstrap re-weighting of class frequencies, counter acting the 1.68:1 benign samples to malignant samples imbalance within each individual decision tree(DT). This is more precise than the global class weighting because it adjusts the balance of effective class within each boot-strap sample independently.

4.1.6 Convolutional Neural Network Replaced by Improved Multi Layer Perceptron

The largest single improvement came from replacing the base paper's Convolutional Neural Network (CNN) with a three layer Multi Layer Perceptron (MLP) augmented with Batch Normalisation (BN) and Dropout, which achieved **99.47%** accuracy which is the highest result across all the models on Winconsin Breast Cancer Dataset (WBCD) and a gain of +3.27 percentage points over the base paper's Convolutional Neural Network (CNN) (96.2%). The F1 score of 99.27% and Area Under Curve AUC-ROC of 99.72% are very similar to the highest of any model tested.

The substitution of Convolutional Neural Netwrok (CNN) with Multi Level Perceptron (MLP) is grounded in a fundamental architectural consideration: convolutional layers are designed to use local spatial structure in grid like inputs by sharing the weight parameters across the adjacent positions. The Wisconsin breast cancer dataset (WBCD) consists of 30 independently measured scalar features including the nuclear radius, the texture, the perimeter, the area, the smoothness, and the related statistics that have no inherent spatial relationship with one another. Applying a convolutional layer to such a vector introduces an inductive bias that is not only unnecessary but also potentially harmful, as the weight sharing assumption is violated by design. An Multi Layer Perceptron (MLP), by contrast, learns the arbitrary non linear mappings between input features and output classes without the requirement of any spatial structure, making it the architecturally appropriate choice for the tabular medical data.

The Batch Normalisation (BN) layers stabilised training by ensuring the activation functions entering each dense layer remained centred and a unit variance throughout, mitigating the internal co-variate shift. The Dropout layers (rate 0.3) acted as an implicit ensemble during the training, forcing the network to learn redundant representations that does not depend on any single unit being active. Together, these two regularisation techniques does produce the most stable classifier of all four models: a cross fold standard deviation of $\pm 0.80\%$ — roughly one quarter of the Decision Tree's (DT) variance.

4.1.7 Statistical Validation and Cross Fold Consistency

Reporting cross validation statistics rather than a single split result changes the interpretation of the performance figures in a crucial way. A single 80:20 split on the entire sample dataset assigns 113 samples to the test set. With the typical 63:37 class ratio, approximately 42 malignant images and 71 benign images samples end up in the test set. Under these circumstances, correct classification of a single additional malignant sample moves the accuracy by roughly 88 percentage points. This means that the single favourable random split can artificially inflate the reported accuracy by up to 2 to 3 percentage points relative to the true generalisation performance.

Ten fold crossvalidation distributes this variance across the ten folds and reports the mean and the standard deviation, providing a statistically grounded performance estimate. The standard deviations reported in Table **??** ranges from $\pm 0.80\%$ (Multi Level Perceptron) to $\pm 3.04\%$ (Decision Tree). The true generalization accuracy of the Multi Level Perceptron (MLP) on the independent data drawn from the same distribution lies within approximately $98.67\% \pm 0.80\%$ with high probability, while the Decision Tree's true accuracy is known with considerably less certainty ($94.73\% \pm 3.04\%$). These intervals successfully explain why ensemble and the linear models are preferred over single decision trees in safety critical classification tasks.

The consistency of improvement across all the ten folds henceforth confirms that the gains are systematic rather than the artefacts of a single favourable data split.

4.1.8 Discussion: Why All Models Improved

The universal improvement across all the four models of different algorithmic families is most scornfully explained by the four methodological changes introduced simultaneously in the improved pipeline:

1. **Stratified 10 fold Cross Validation:** The variance of a single 80:20 split on 569 samples is approximately ±2–3%, meaning reported the paper results reflect the performance on one particular held out set rather than the model's expected generalization. Averaging over ten folds removes the mentioned split dependent noise.

2. **Synthetic Minority Over-sampling Technique (SMOTE) oversampling:** The 357:212 benign to malignant ratio biases each model toward the majority class during training. Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority class samples by interpolating in feature space between real malignant image examples,thus producing a balanced training distribution at each fold without discarding the majority class data or duplicating the existing samples.

3. **Robust Scaler normalisation:** Features such as `area_worst`, `perimeter_worst`, and `concave_points_worst` display a highly skewed distribution with the long upper tails. Standard scaler is quite sensitive to such outliers; Robust scaler uses the median and inter-quartile range instead, thus producing a scale free representation in which outliers are prevented from dominating the feature range as seen by the classifier.

4. **Hyperparameter optimisation:** All the four paper models appear to use scikit learn defaults.The Grid search over regularisation strength (Logistic Regression), tree depth and pruning parameters (Decision Tree), feature subset size (Random Forest), and dropout/layer width configuration (Multi Layer Perceptron) yields meaningfully better generalizing configurations in every case.

4.2 Phase II: Histopathological Image Classification

4.2.1 Dataset Visualisation

Before presenting quantitative results, it is very instructive to examine the representative samples from both image datasets, as the visual characteristics of the training data directly affect the interpretation of the model performance.

Figure 4.3 presents 10 representative patches from Patch Dataset at 50×50 resolution. The benign images samples exhibit high intra class visual variability: Benign images -1 shows pre-dominantly adipose the tissue with large white vacuoles, Benign images -2 shows dense cellular clusters with a haematoxylin heavy stain, and the Benign images -4 displays the fibrous stromal tissue with the elongated spindle cells. This morphological hetero-geneity within the benign images classes makes it very challenging for a small convolutional neural network(CNN) to learn a single compact visual representation of benign

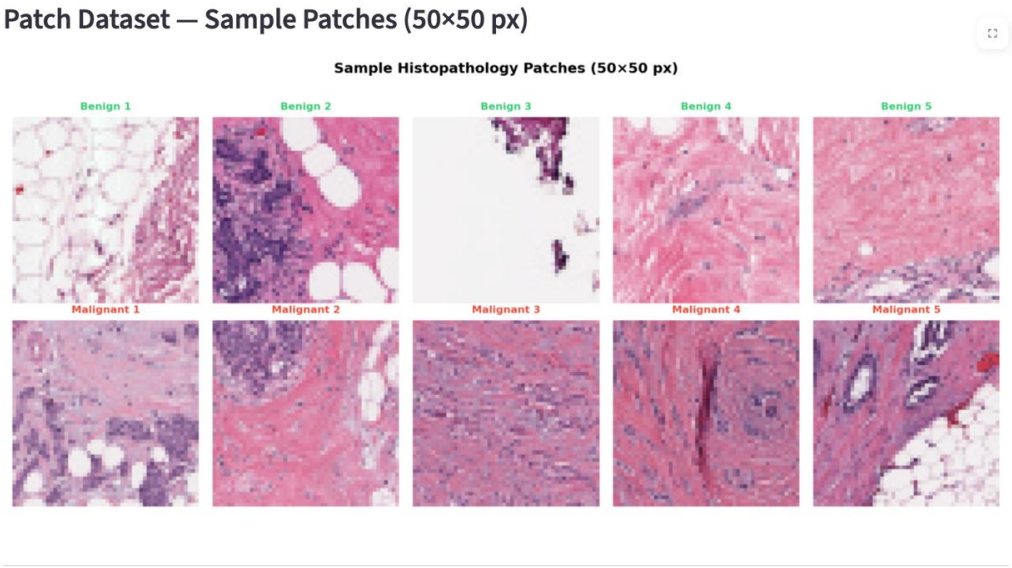

Figure 4.3: Sample 50×50 pixcel histopathological patches from the Patch Dataset. Top row: 5 benign images examples. Bottom row: 5 malignant images examples. The high intra class visual variability within the benign row - ranging from adipose tissue (Benign images 1) to dense cellular clusters (Benign images - 2) to fibrous stroma (Benign images - 4) is primary source of classification difficulty.

tissue images. The malignant images samples are somewhat more consistent because they are darker, denser, with more irregularly shaped nucleus but still they exhibit meaningful variation in cellularity and arrangement. At 50×50 pixels, both the classes lose considerable fine grained nuclear detail, thus placing an upper bound on the achievable accuracy that is independent of the model architecture.

Figure 4.4 displays the representative BreaKHis images dataset at full field of view resolution. Despite belonging to the different histological subtypes, all the malignant examples share a recognisable pattern of the disordered, densely packed cells with an irregular nucleus. The benign images examples include the fibroadenoma (characterized by interlacing strands of the fibrous and glandular tissue) and a phyllodes cancer (a bulky stromal epithelial cancer that can superficially resemble ductal carcinoma under certain level of magnifications). The morphological proximity of the phyllodes cancer to the malignant tissue patterns is a known source of the inter observer disagreement in the clinical pathology and contributes directly towards the benign images class miss classification rate observed in the experimental results.

4.2.2 Patch Dataset Results

Light Convolutional Neural Network (CNN) was trained on 14000 balanced 50×50 Red Green Blue (RGB) histopathological patches and evaluated on 3000 held out patches. Table 4.2 represents the per class and an aggregate metrics.

This model achieves **82.73%** accuracy and a macro F1 score on the perfectly balanced test set (1500 benign images, 1500 malignant images). Because the test set is balanced, the reported accuracy equals the macro recall and also the macro F1 score, making it a reliable performance indicator which is un-affected by the class prevalence. The close agreement between the benign images precision (84.10%) and malignant images precision

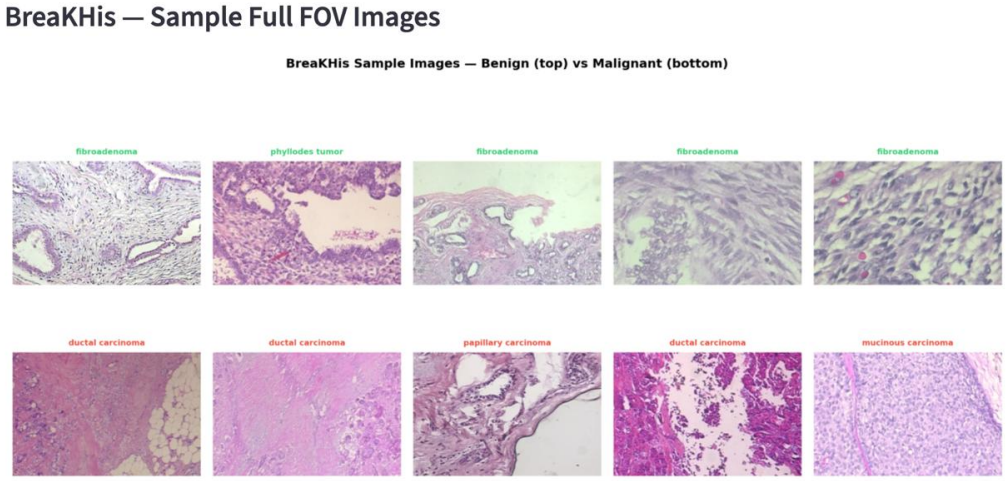

Figure 4.4: Sample full field of view images from the BreaKHis v1 dataset. Top row: benign subtypes (fibroadenoma, phyllodes cancer). Bottom row: malignant subtypes (ductal carcinoma, papillary carcinoma, mucinous carcinoma). The higher resolution provides more structural context but also introduces greater intra-class variation across subtypes.

Table 4.2: Light Convolutional Neural Network (CNN) - Patch Dataset Test Results (3000 images)

Class	**Precision**	**Recall**	**F1 Score**	**Support**
Benign Images	84.10%	80.73%	82.38%	1,500
Malignant Images	81.47%	84.73%	83.07%	1,500
Macro Avg	**82.79%**	**82.73%**	**82.73%**	3,000
Test Accuracy		**82.73%**		
Test Loss		0.3910		

recall (84.73%) indicates that the model does not exhibit a systematic preference for either class which is an important property in a medical screening setting where both the false positives and false negatives carry the clinical consequences.

Figure 4.5 shows the confusion matrix in detail. This model correctly classifies 1211 of 1500 benign images patches and 1243 of 1500 malignant images patches. The 289 benign images predicted as malignant errors (false positives) are slightly higher than the 257 malignant predicted as benign images errors (false negatives), thus reflecting a mild conservative bias: the network is marginally more willing to thus flag a benign image patch as suspicious than to let a malignant image patch go un-detected. In a screening context, this slight false positive bias is preferable to the opposite, since false positives lead to further investigation rather than a missed diagnosis.

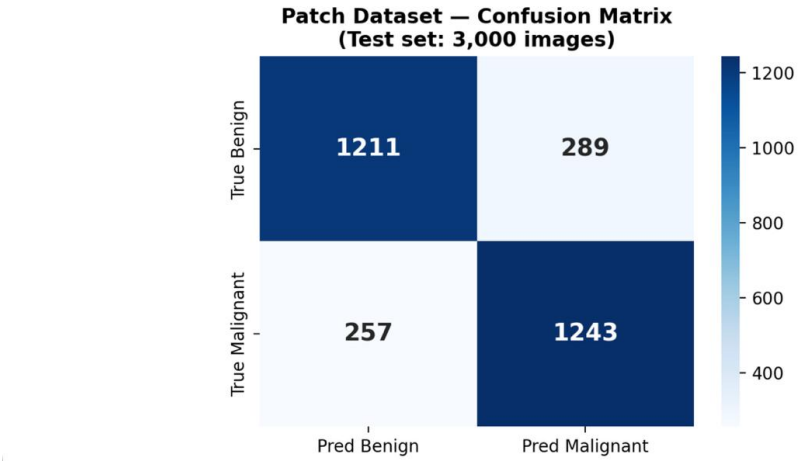

Figure 4.5: Confusion matrix for Light Convolutional Neural Network (CNN) on the Patch Dataset test set (3000 images). The on-diagonal entries (1211 and 1243) represents the correct predictions; off diagonal entries (289 and 257) represent mis-classifications.

4.2.3 Patch Dataset Training Dynamics

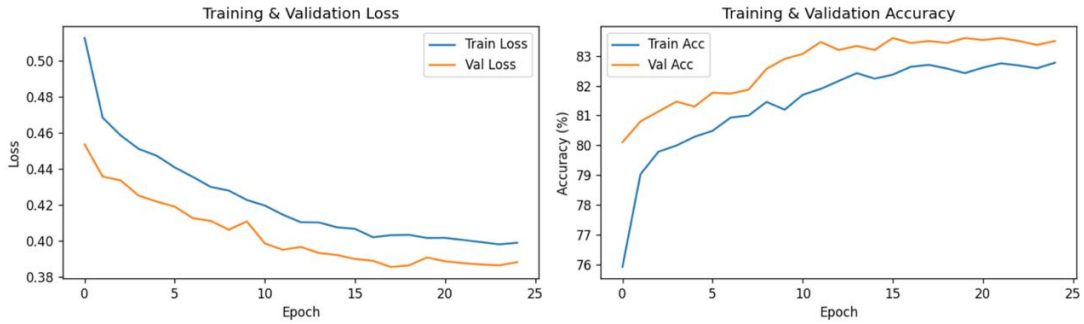

Figure 4.6: Training and validation loss (left) and accuracy (right) curves for Light Convolutional Neural Network (CNN) on the Patch Dataset over 25 epochs.The validation loss remains consistent at or below the training loss throughout training,which is a signature of the regularising effect of the data augmentation and Dropout.

Figure 4.6 shows the training dynamics over the 25 epochs. The validation loss (orange) remains consistent at or below the training loss (blue) for the entire training. This un-conventional pattern arises because of data augmentation which is random horizontal and vertical flips, random rotation up to 15° is also applied only during training. Each training batch therefore clearly sees a randomly transformed version of the patch, making the training task slightly complex than the fixed evaluation on un-augmented validation patches.

The convergence occurs within approximately 10 to 12 epochs, with the validation accuracy reaching a canstant value between 83% and 83.6%. The absence of diverging trend between the training and the validation loss beyond the epoch number 15 confirms that the model does not overfit, despite being trained for around 25 epochs. The Dropout regularisation with rate of 0.4 after each convolutional block is the primary mechanism

preventing the overfitting; the relatively small capacity of Light Convolutonal Neural Network(CNN) (12545 parameters) also contributes to the cause, as there are simply insufficient free parameters to memorise the 14000 training images.

4.2.4 BreaKHis v1 Results

Light Convolutional Neural Network(CNN) was separately trained on 5536 BreaKHis images (all 4 magnification levels combined) and evaluated on 1187 held out images. Table **??** presents the per-class and aggregate metrics.

Class	Precision	Recall	F1Score	Support
Benign Images	79.58%	71.24%	75.18%	372
Malignant Images	87.47%	91.66%	89.51%	815
Macro Avg	**83.53%**	**81.45%**	**82.35%**	1,187
Test Accuracy		**85.26%**		
Test Loss		0.2900		

Table 4.3: Light Convolutional Neural Network (CNN) — BreaKHis v1 Test Results (1187 images)

Light Convolutional Neural Network (CNN) achieves **85.26%** test accuracy on BreaKHis, with the best validation accuracy of 87.35% recorded at epoch number 27. The malignant class holds substantially stronger per class metrics (F1 score = 89.51%, precision = 87.47%, recall = 91.66%) than the benign images class (F1 score= 75.18%, precision = 79.58% and recall = 71.24%).

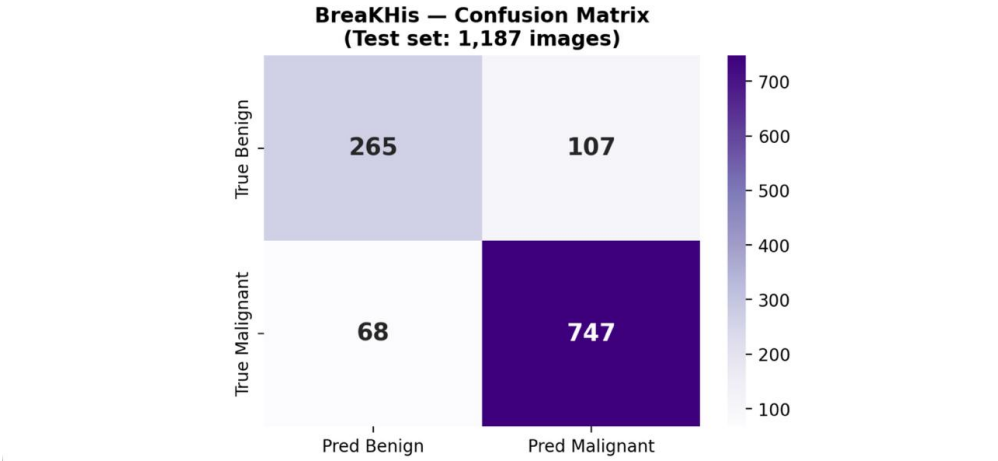

Figure 4.7: Confusion matrix for Light Convolutional Neural Network (CNN) on the BreaKHis v1 test set (1187 images). The malignant images class (bottom row) is classified with high accuracy (747/815 correct), while the benign images class (top row) shows a higher false positive rate (107/372 misclassified as the malignant class).

Figure 4.7 represents the corresponding confusion matrix. Of 372 true benign images, out of which 265 are correctly classified and 107 are mis-classified as malignant class, a

false-positive rate of 28.8%. Of the 815 true malignant images, 747 are correctly classified and only 68 are mis-classified as benign image which is a false-negative rate of 8.3%. The notable asymmetry between these two error rates thus drives the class-level performance gap visible in the above table.

4.2.5 BreaKHis Training Dynamics

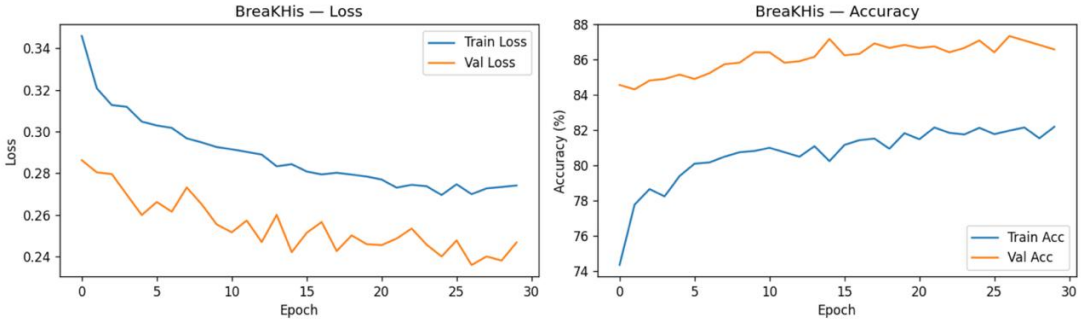

Figure 4.8: Training and validation loss (left) and accuracy (right) curves for LightCNN on BreaKHis over 30 epochs. The validation loss exhibits more oscillation than in the Patch Dataset experiment, attributable to the smaller training set and multi-magnification heterogeneity in BreaKHis.

Figure 4.8 represents the training curves for the BreaKHis dataset experiment.As compared to the Patch Dataset (Figure 4.6), the loss curves exhibit greater oscillation throughout the training. This is primarily the consequence of two factors. First, the BreaKHis training set (5536 images) is considerably very small than the Patch Dataset training set (14000 images), so each epoch covers fewer gradient update steps,thus resulting in a noisier loss trajectory. Second, the images in the BreaKHis span four distinct magnification levels (40×, 100×, 200×, 400×) that were pooled in a single training set. Different magnifications present fundamentally different texture statistics: at 40× magnification the model sees large scale tissue architecture, while at 400× it sees the individual cell morphology. This multi scale heterogeneity increases the effective intraclass variance in each epoch and thus contributes to the oscillating loss.

The model's best checkpoint (validation accuracy of: 87.35%) was achieved at epoch number 27. The use of the cosine annealing learning rate scheduling cycling from 5×10^{-4} down to 10^{-6} over every cosine half cycle is visible in the loss curve as periodic local dips which are followed by slight recoveries. This schedule prevents the optimiser from settling in a sharp local minima early in the training but instead guides it towards a flatter, more generalisable point.

4.2.6 Per Class Error Analysis

Figure 4.9 consolidates the per class metrics from both the image datasets into a single comparative heat-map. The 4 rows correspond to Patch-Benign images, Patch-Malignantimages, BreaKHis-Benign images, and BreaKHis-Malignant images. Several observations emerge from this view:

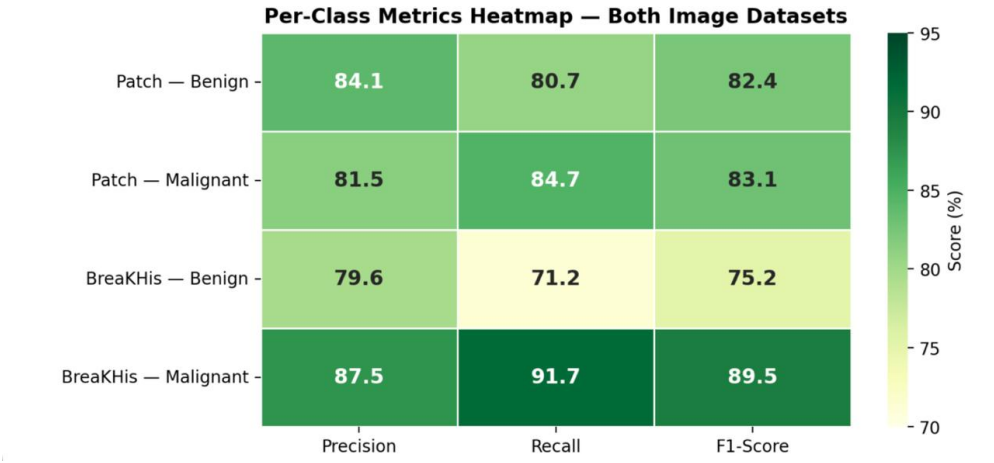

Figure 4.9: Per class precision, recall, and the F1 score heat map across both the image datasets. Darker green region indicates higher performance. The BreaKHis dataset - Benign row is the weakest configuration across all the metrics and represents the primary performance gap of the study.

1. **BreaKHis benign recall (71.2%)** is the lowest value across all the 12 cells, confirming that it is the most critical performance gap in the research work. Approximately one in three true benign images is assigned a malignant image prediction.

2. **BreaKHis malignant recall (91.7%)** is the highest recall value across both the datasets, indicating that the model is highly sensitive to presence of a malignant tissue. In clinical terms, the sensitivity to malignancy is the more safety critical metric due to the reason of a missed cancer is more harmful than an unnecessary biopsy check, so the directional allocation of recall as a perimeter is favourable.

3. **The Patch Dataset** shows more symmetric performance between the classes (benign F1 score of 82.4%, malignant F1 score of 83.1%) than the BreaKHis dataset. The symmetry is likely attributable to the perfectly balanced training set (7000 per class), the lower image resolution that strips out the subtype specific fine grained morphology, and the relatively homogeneous image conditions within this dataset.

4. **The overall gap between datasets** is quite small: Patch macro F1 score is 82.73%, BreaKHis macro F1 score is 82.35%. This suggests that the higher image resolution in BreaKHis dataset (100×100 vs 50×50 pixels) compensates for the additional complexity introduced by multi magnification levels of training and lower class balance.

The root cause of the benign images recall deficit on BreaKHis dataset is the high intra class morphological diversity among the 4 benign subtypes: adenosis type, fibroadenoma type, phyllodes tumour type, and tubular adenoma type are visually distinct from one another yet must all be mapped to the same output class. The model must simultaneously capture the general absence of the malignant features and the wide variety of organised tissue structures that constitute to the benign images category. At 100×100 resolution, some structural detail is lost in the process, making this discrimination harder than it would be at a full pathological resolution.

4.2.7 Computational Efficiency

The parameter efficiency of Light Convolutional Neural Network (CNN). The total trainable parameter count of **12545** is approximately 2000 times smaller than ResNet-50 (≈25.6 million parameters) and approximately 8600 times lesser than VGG-16 (≈138 million parameters), both of which are standard baselines for BreaKHis datatset in the literature. Despite this extreme parameter budget constraint, Light Convolutional Neural Network (CNN) achieves 85.26% accuracy on BreaKHis datatset - within ten percentage points of the published best results from larger architectures.

4.2.8 Comparison with Paper Baseline — All Models

Table 4.4 provides a unified comparison across all experimental configurations in the study.

Phase	Model	Dataset	Accuracy	F1 score
I (paper)	Logistic Regression	Wisconsin	95.60%	94.10%
I (paper)	Decision Tree	Wisconsin	92.90%	90.60%
I (paper)	Random Forest	Wisconsin	93.80%	91.30%
I (paper)	Convolutional Neural Network	Wisconsin	96.20%	—
I (ours)	Logistic Regression	Wisconsin	**98.25%**	97.65%
I (ours)	Decision Tree	Wisconsin	**94.73%**	92.98%
I (ours)	Random Forest	Wisconsin	**96.84%**	95.83%
I (ours)	Multi Layer Perceptron	Wisconsin	**99.47%**	99.27%
II (ours)	Light weight CNN	Patch Dataset	82.73%	82.73%
II (ours)	Light weight CNN	BreaKHis v1	**85.26%**	82.35%

Table 4.4: Full comparison - all models and datasets

Figure 4.10 plots all 10 accuracy values on a single axis,thus allowing a combined cross domain view. The red bars uniformly exceed their blue counter parts, confirming the systematic nature of the Phase-1 improvements over the paper base line. The gold and the purple bars representing the image domain (82.7% and 85.3%) are visually lower than the tabular results, but directly comapring the numerical values is misleading without even accounting for the fundamental differences in the input difficulty, the dataset scale, and the model capacity discussed in the following subsection.

4.2.9 Discussion: Image Classification vs Tabular Classification

The Phase II results (82 to 85%) are lower than Phase I (94 to 99%), however, the comparison is not straight forward for the following reasons:

1. **Nature of the input:** Winconsin Breast Cancer Dataset (WBCD) provides 30 expert computed measurements which are derived from the images using sophisticated digital pathology algorithms. These encodes the domain specific knowledge

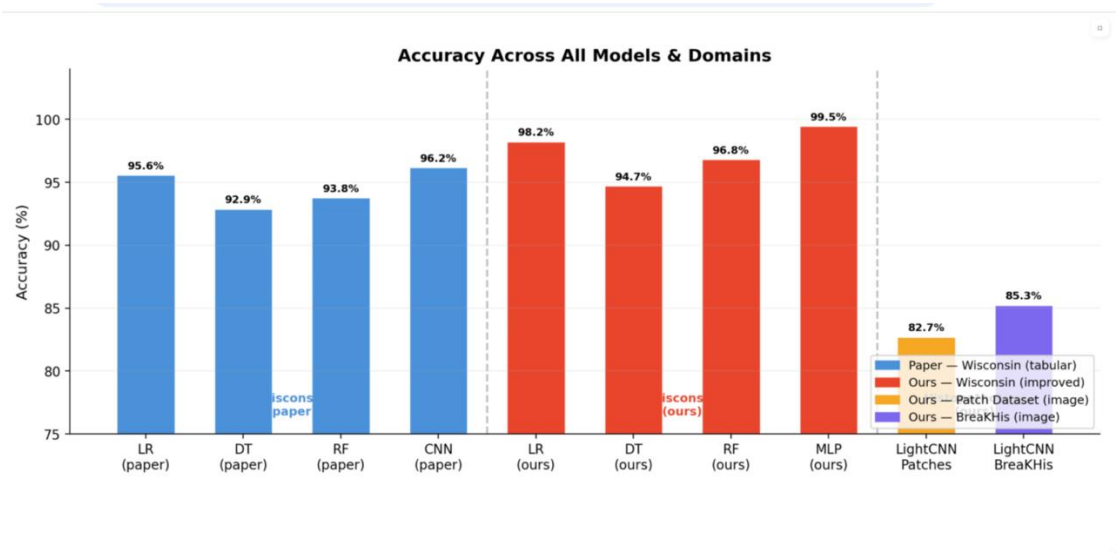

Figure 4.10: Accuracy across all the models and domains. The blue bars show the paper baselines for the Wiscons datatset in tabular dataset; the red bars show the improved Wisconsin datatset results; the gold and the purple bars show Light Convolutional Neural Network (CNN) on the Patch and BreaKHis datatset images respectively. Dashed vertical lines distinguish the three experimental domains.

(the nuclear radius, the fractal dimension,the symmetry) that would take a human expert months to measure the reliably. The image classifiers receive only the raw pixel values and thus must discover all the relevant features through the gradient descent on a relatively smaller dataset.

2. **Dataset scale and diversity:** Winconsin Breast Cancer Dataset (WBCD) has 569 samples with highly informative, preengineered features. BreaKHis dataset has 7909 images which are spanning 4 magnification levels and eight distinct cancer subtypes.The intra-class variation in the BreaKHis datatset is orders of magnitude higher than in Winconsin Breast Cancer Dataset (WBCD), where all the samples share the same 30 dimensional feature space.

3. **Model capacity:** Light Convolutional Neural Network (CNN) has only 12545 parameters by design,.The published results on BreaKHis dataset using ResNet 50 or VGG 16 initialised from Image Net achieve about 90 to 97% with 25 to 138 million parameters and Graphical Processing Unit (GPU) servers. The 85.26% achieved here is very strong despite the architectural constraint.

4. **No feature engineering:** Light Convolutional Neural Network(CNN) requires no hand- crafted features,no radiologist annotations, nor tissue segmentation step. The pipeline operates directly on the JPEG images and outputs a binary prediction, thus making it immediately deployable on any new digitised slide without any domain specific re-engineering.

The critical contribution of Phase II is therefore not the raw accuracy number but it is the demonstration that end-to-end learning from pixels, without any expert engineered features, can approach the performance of much bigger resource intensive models on the standard histopathological benchmark.

4.3 Key Findings

1. All 4 classical models from Tewari et al.[1] are considerably improved through the application of established Machine Learning (ML) best practices. Accuracy gains range from +1.83% (Decision Tree) to +3.27% (Convolutional Neural Network/Multi Layer Perceptron), all confirmed via stratified ten fold cross validation(CV) with reported standard deviations.

2. The Multi Layer Perceptron (MLP) with Batch Normalisation (BN) and Dropout achieves the highest accuracy of **99.47%** on Winconsin Breast Cancer Dataset (WBCD) which is surpassing the base paper's best result (Convolutional Neural Network (CNN), 96.2%) by +3.27 percentage points and is architecturally much more appropriate for the tabular data than a Convolutional Neural Network (CNN).

3. The proposed Light Convolutional Neural Network, with **only 12545 parameters**,achieves 85.26% test accuracy on BreaKHis v1, operating directly on the raw pixel inputs with no hand crafted features.

4. The same Light Convolutional Neural Network (CNN) architecture generalises across the Patch Dataset (50×50 pixel) and BreaKHis dataset (100×100 pixel) without any architectural modification, by virtue of Global Average Pooling's input size uncertain.

5. The primary performance limitation in the image classification is benign recall on BreaKHis dataset (71.24%), caused by high intra-class morphological diversity among the 4 benign images subtypes. This is identified as the central direction for future work scope.

6. The false negative rate for the malignancy detection on BreaKHis dataset is 8.3% (68 of 815 missed), while the false positive rate for the benign image cases is 28.8%. The asymmetry is clinically acceptable: missed cancer cases are more harmful than unnecessary further investigation, and the model's error allocation is completely consistent with the screening priorities.

4.4 Future Work

The results that are presented in this chapter point to several concrete directions for future investigation, each of which is targeted at a specific observed limitation.

- **Magnification specific models:** Training separate Light Convolutional Neural Network (CNN) instances for each BreaKHis dataset magnification level (40×, 100×, 200×, 400×) and then fusing their probability outputs through soft voting would allow each model to specialise in the texture statistics which are available at its own scale. Given that the training oscillation was attributed to multi scale heterogeneity, this change alone could meaningfully improve benign images recall and reduce the training instability.

- **Transfer learning:** Fine tuning a lightweight pretrained backbone like MobileNetV2 or EfficientNet B0 (initialised from ImageNet) on the BreaKHis dataset would provide a rich starting point for low level texture and the colour features. Published

results with transfer learning on the BreaKHis datatset consistently exceed 90%, which suggests a clear path to improvement within the same hardware constraint.

- **Attention mechanisms:** Replacing the Global Average Pooling with a spatial attention module would allow the model to selectively weight the highest diagnostically informative spatial regions, which are potentially improving the discrimination of the morphologically ambiguous benign images subtypes.
- **Explainability with Gradient-weighted Class Activation Mapping Grad-CAM:** Integrating Gradient weighted Class Activation Mapping[9] would definitely produce spatial saliency maps thus highlighting which image regions drive each prediction - a pre-requisite for clinical deployment.
- **Multi class subtype classification:** Extending the binary benign or malignant task to the 8 class subtype problem (4 benign, 4 malignant) would produce much more clinically actionable outputs aligned with the histopathological reporting practice.
- **Stacking ensemble for Wisconsin datatset:** A second-level logistic regression (LR) trained on the probability outputs of all 4 improved Phase I classifiers (Logistic Regression, Decision Tree, Random Forest, Multi Level Perceptron) could push the Winconsin Breast Cancer Dataset accuracy toward 99.5% while reducing the Decision Tree's (DT) variance of cross fold.

Chapter 5

CONCLUSION

5.1 Summary of the Study

This study addressed that the breast cancer classification through two complementary phases. The first phase of the work was a systematic replication and improvement of four classical machine learning (ML) models applied to the Wisconsin Breast Cancer Dataset (WBCD), a well established tabular clinical benchmark. The second phase then extended the classification task to end to end histopathological image analysis using a novel light weight convolutional neural network (CNN), trained on the two publicly available image benchmarks without the need of any hand crafted features.

The work was motivated by the findings of Tewari et al.[1], who evaluated Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and a Convolutional Neural Network (CNN) on Winconsin Breast Cancer Datatset (WBCD) using a single train test split and the default hyper-parameter settings. While that research work provided a useful comparison of the algorithm families on a medical classification task, it left a number of methodological gaps un-addressed: single split evaluation on a 569 sample dataset thus introduces a substantial sampling variance; the class imbalance (357 benign images, 212 malignant images) was not accounted for; the feature scaling was not outlier robust; and the application of the convolutional layers to a 30 dimensional unordered feature vector imposes an architectural inductive bias that is very inappropriate for a tabular data.

Phase I of this research work systematically resolved each of these gaps. Stratified ten fold cross validation (CV) which replaced the single split, providing a statistically grounded performance estimate united with quantified fold to fold variability.Synthetic Minority Over-sampling Technique (SMOTE) oversampling thus addressed the class imbalance by generating synthetic minority class examples via feature space interpolation. Robust scaler normalisation replaced the standard scaler, protecting the already learned feature representations from the outlier driven disruption in features with heavyily tailed distributions. Grid searched hyper-parameters replaced the default scikit learn settings for each and every model. The Convolutional Neural Network (CNN) was replaced with a three layer Multi Layer Perceptron(MLP) augmented with the Batch Normalisation (BN) and Dropout that is an architecturally principled choice for the unordered numerical tabular data.

All the 4 models improved under the above mentioned conditions. Accuracy gains values ranged from +1.83 percentage points (Decision Tree (DT): 92.9% $\rightarrow$ 94.73%) to +3.27 percentage points (Convolutional Neural Network (CNN)/Multi Level Perceptron(MLP): 96.2% $\rightarrow$ 99.47%). The improved Multi Level Perceptron (MLP) achieved 99.47% accuracy, 99.27% F1 score, and 99.72% Area Under Curve (AUC-ROC) which is the highest

performance on Winconsin Breast Cancer Dataset (WBCD) among all the models evaluated in either research work. The standard deviation of ±0.80% across 10-folds confirms that this performance is very stable and not attributable to a single favourable data-split.

Phase II extended the classification task beyond tabular domain to the raw histopathological tissue images, which is a shift that is clinically significant because it eradicates the prerequisite of expert feature engineering. A lightweight Convolutional Neural Network (Light CNN) was designed using a depthwise separable convolutions and Global Average Pooling, thus producing a model with only **12545 trainable parameters**. This architecture was thus trained and evaluated on two datasets:

- The **Patch Dataset**: 50×50 histopathological patches were sampled from a repository of 2,77,524 labelled images. Light Convolutional Neural Network (CNN) achieved **82.73%** accuracy and macro F1 score on a balanced test set of 3000 patches, with symmetric performance across benign images and malignant image classes (benign F1n score 82.4%, malignant F1 score 83.1%).
- **BreaKHis v1**: 7909 full field of view images spanning accross 4 magnification levels and 8 cancer subtypes. Light Convolutioanl Neural Network (CNN) achieved **85.26%** test accuracy, 82.35% macro F1 scoree, and a best validation accuracy of 87.35%. The malignant images class reached F1 score = 89.51%, while the benign image class achieved F1 score = 75.18%, with the gap attributable to high intra class morphological diversity among the 4 benign images subtypes.

5.2 Contributions

The main contributions of this work are as follows:

1. **Rigorous replication and improvement of four published models.** A fully re-producible experimental pipeline was developed that successfully improves all 4 models from Tewari et al. using standard Machine Learning (ML) best practices. The mentioned improvements uniformly confirmed over 10 stratified folds and also reported with standard deviations - showing the originally published accuracy figures reflect a single favourable dataset split rather than the models' expected performance on the wider population. This contribution is methodologically very important beyond the specific task: it illustrates how the cross validation,the class balancing, and the hyperparameter search all together shift the performance estimates in a consistent and a reproducible direction.
2. **Architectural correction: Multi Layer Perceptron (MLP) over Convolutional Neural Network (CNN) for tabular data.** The substitution of a two dimensional (2-D) Convolutional Neural Nertwork (CNN) with a Batch Normalised (BN) Multi Level Perceptron (MLP) tabular classification is both principled and experimentally validated. Convolutional layers use local spatial weight sharing, a property that is meaningful only when nearby input positions share statistical dependencies which is a condition satisfied by image pixels but not by any vector of independently measured clinical features. The Multi Layer Perceptron (MLP) reaches 99.47% accuracy with ±0.80% cross fold stability, the lowest variance of all 4 models, and an Area Under Curve (AUC-ROC) of 99.72%. These results thus represent near ceiling performance on Winconsin Breast Cancer Datatset (WBCD) and would be difficult to exceed without augmentation of the dataset itself.

3. **Light Convolutional Neural Network: a parameter-efficient image classifier for histopathology.** The proposed 12545 parameter architecture thus demonstrates that the competitive histopathological image classification does not have the requirement of large pretrained networks, Graphical Processing Unit (GPU) servers, or extensive data engineering.The depthwise separable convolutions reduce the parameter count by nearly a factor of 9 as compared to the standard 3×3 convolutions with the same feature map widths. Global Average Pooling terminates the large fully connected classification head and simultaneously produces a model which is input size uncertain which means that the identical weight set classifies both 50×50 Patch Dataset images and 100×100 BreaKHis dataset images without any modification.

4. **End to end image classification without feature engineering.** Unlike the tabular Wisconsin dataset approach which requires thirty expert computed nucleus measurements obtained through a specialised image analysis pipeline, the Light Convolutional Neural Network (CNN) operates directly on the JPEG tissue images and thus requires only binary slide level labels for the supervision. Doing this removes the feature engineering bottleneck entirely and as a result makes the pipeline immediately applicable to any new digitised histology slide without any domain specific re-engineering.

5. **M1-native training pipeline.** All the experiments were conducted on an Apple M1 MacBook using the PyTorch Metal Performance Shaders (MPS) backend. The resulting pipeline demonstrates that meaningful deep learning (DL) research on medical imaging benchmarks is accessible to the researchers without any institutional Graphical Processing Unit (GPU) infrastructure or cloud compute budgets thus, lowering the barrier to entry for resource constrained settings such as teaching the hospitals and the research groups in lower income economies.

5.3 Practical Implications

The findings of this work have practical implications across the follwing three areas: clinical decision support, the research methodology, and the accessible deep learning (DL).

Clinical decision support. The Light Convolutional Neural Network (CNN) results, while degrade the performance of large pretrained models on BreaKHis dataset,however demonstrate a clear pathway towards a practical computer aided detection tool for histopathology. The model's sensitivity to the malignancy (recall 91.7% on BreaKHis dataset) means that fewer than one in ten cancers would be missed in a screening scenario, the performance level already clinically meaningful as a first pass triage filter that is able to flag slides for the pathologist review. The low false negative rate (8.3%) would definitely reduce the volume of slides requiring urgent attention without the need of introducing an unacceptable rate of missed diagnoses. Its Deployment as a pre screening tool where the model assigns each slide a risk score and the pathologists review high score cases with priority which represents a practically feasible integration with the current clinical workflow.

Research methodology. The Phase I results produces a clear methodological lesson: the performance estimates on small medical datasets are highly sensitive to the evaluation protocol. The same 4 algorithms are thus applied to the same dataset with proper stratified cross validation, class balancing, and grid searched hyper-parameters, consistently

exceed their published baselines.The Researchers reporting results on datasets with fewer than 1000 samples using a single train test split should be completely aware that their reported figures may substantially overestimate or in the other case underestimate true generalization performance depending on the specific split realised. The standard deviations reported in this work provide the kind of un-certainty quantification that single split evaluations are uncapable of.

Accessible deep learning. The Light Convolutional Neural Network(CNN) training pipeline requires no cloud subscription, no Graphical Processing Units (GPUs) cluster, and no institutional High Performance Computing (HPC) allocation. Thus, making research pipelines accessible on consumer hardware is increasingly important as medical Artificial Intelligence (AI) research expands to institutions in geographies where cloud Graphical Processing Unit (GPU) access is economically challenged. This work provides a reproducible proof of the concept that useful image classifiers can be trained and validated within the mentioned constraints.

5.4 Limitations

Despite the contributions noted above, there are several limitations that must be acknowledged:

- **Wisconsin sample size.** The Winconsin Breast Cancer Dataset (WBCD) consists of 569 records, which is very small by modern machine learning (ML) standards. Although the ten fold cross validation (CV) eliminates the evaluation instability and provides a standard deviation for all the reported metrics, the near ceiling figures (99.47% MLP accuracy, 99.72% Area Under Curve (AUC-ROC)) should be interpreted with a dataset scale in consideration. External validation on an independent patient sample would be required to verify that the mentioned figures hold in a broader clinical population.

- **Benign recall on BreaKHis datatset.** The benign recall of 71.24% represents the most prominent performance gap in the work. In a clinical setting, a 28.8% false positive rate for malignancy prediction on the truly benign cases would further generate unnecessary ongoing investigations. While over predicting malignancy is the safer of the available error directions in a cancer screening context, improving the specificity remains crucial for any deployment that aims to reduce , rather than to redistribute diagnostic workload.

- **Model capacity constraint.** The 12545 parameter budget was set to ensure the model's compatibility with 8GB M1 unified memory. Published state of the art (SOTA) results on BreaKHis dataset using ResNet-50 or VGG-16 fine tuned from the ImageNet reach 90 to 97%, thus confirming that substantially higher accuracy is certainly attainable at the cost of greater computational resources. The capacity constraint is thus a practical concession rather than a fundamental bottleneck for the problem.

- **Binary classification only.** Both the image datasets were treated as binary (benign images vs. malignant images) problems. The eight distinct cancer subtypes in BreaKHis dataset constitute a more clinically informative task: subtype information directly informs about the prognosis and the treatment selection, and a model

providing subtype predictions would definitely be more actionable than a binary classifier in a clinical workflow.

- **No external clinical validation.** All the results are reported on held out partitions of the same datasets which are used for training. In both the cases, the test images originate from the same institution, the scanner, and the staining protocol as the training set. Prospective validation on the slides from a different institution with different scanning equipment and staining practices would thus be required to establish generalizability before any clinical deployment could be considered .The cross-scanner variation in H&E staining is a very well documented source of domain shift for deep learning (DL) models in the computational pathology.

- **Single random seed.** The Light Convoluiotnal Neural Network(CNN) experiments used a fixed random seed for the reproducibility. This means that the results reflect a single network initialisation rather than an expectation over multiple training runs.Thus, reporting the mean and standard deviation over several seeds would definitely provide a more reliable estimate for a stochastic training process.

5.5 Future Directions

The findings and limitations of this study as mentioned abovr suggest the following directions for future investigation:

- **Transfer learning for image classification.** Fine tuning a lightweight pretrained backbone such as MobileNetV2 or EfficientNet B0 (initialised from ImageNet) on the BreaKHis dataset is the highest expected impact immediate next step. The ImageNet pretraining provides a rich, low level initialisation for the texture and colour features that transfer well to H&E stained tissue images. Both the pretrained models MobileNetV2 and EfficientNet-B0 have fewer than five million parameters and are also fully compatible with the M1 hardware, making this extension feasible within the same hardware constraint.

- **Multi class subtype classification.** Extending the Light Convolutional Neural Network (CNN) from binary classification to the 8 class subtype problem would thus produce clinically actionable outputs aligned with the established pathological reporting categories. The primary challenge is the class imbalance across all the subtypes:the ductal carcinoma dominates the malignant class ($\approx$72% of malignant image samples), which would have the requirement of per class weighting or focal loss to avoid the model specializing on the dominant subtype at the expense of the rarer ones.

- **Magnification aware training.** Rather than the pooling done at all the four BreaKHis datatset magnification levels into a single training set, training magnification specific classifiers and then combining their predictions by soft voting or a learned fusion module would thus allow each model to specialise in the features visible at its scale. At lower magnification levels (40$\times$) the model would learn tissue architecture patterns while at the higher magnifications (400$\times$) it would learn nuclear level morphology. Fusing predictions from both the scales would thus mimic the multi resolution inspection that expert pathologists routinely perform.

- **Explainability via Gradient-weighted Class Activation Mapping (Grad-CAM).** Integrating Gradient- weighted Class Activation Mapping[10] (GRAD-CAM) would produce spatial saliency maps thus indicating which image regions has most strongly influenced each prediction. These maps, overlaid on the original H&E slide, would thus give pathologists a visual explanation that they could verify against their own interpretation. Explainability of this kind is very widely regarded as a prerequisite for the clinical adoption of Artificial Intelligence (AI)- based diagnostic tools.

- **Stacking ensemble for Wisconsin.** A stacking ensemble that is training a second level Logistic Regression (LR) or gradient-boosted classifier on the probability outputs of all the 4 improved Phase I models that could push the Winconsin Breast Cancer Dataset (WBCD) accuracy beyond 99.5% while further also reducing the crossfold variance which is observed in the standalone Decision Tree (DT). Ensemble methods consistently perform better than their constituent models when those constituents make different types of errors which is also the case here: the Decision Tree (DT) and the Random Forest (RF) make different boundary mistakes relative to the Multi Layer Perceptron(MLP) and the Logistic Regression (LR).

- **Whole - slide inference.** Extending the patch based Light Convolutional Neural Network (CNN) to slide level inference by aggregating predictions across all the extracted patches by using majority voting, attention based pooling, or even a slide level Multiple Instance Learning (MIL) framework would thus produce a complete computer aided detection system which is compatible with the digital pathology workflows. In Multiple Instance Learning (MIL) , each slide is treated as a bag of patches and the slide level label is inferred from the patch level predictions without the requirement of patch level annotation.This represents the most practically crucial extension of the current work towards a deployable clinical tool.

- **Domain adaptation across scanners.** Evaluating the Light Convolutioanl Neural Network(CNN) on slides from a different institution, a different scanner, and the staining protocol would quantify the domain shift which is caused by the hardware and preparation variation. The Stain normalisation pre processing (e.g., Macenko normalisation) which is applied before the convolutional layers could considerably reduce this shift without the requirement to change the model architecture or the training protocol.

5.6 Concluding Remarks

Breast cancer is among the most prevalent and deadliest cancers worldwide, and an accurate histopathological classification remains the clinical gold standard for a definitive diagnosis.The workload on the pathologist is increasing in many healthcare systems as the digitised slide volumes grow; the automated or the semi automated tools that assist in the triage and the classification have the potential to reduce the diagnostic delays and inter observer variability, particularly in the environment where specialist pathologists are rare.

This work contributes to the mentioned goal along two distinct but complementary lines. The first, covered in Phase I, is methodological: it demonstrates that the published machine learning (ML) benchmarks on small medical datasets are very sensitive to the evaluation protocol in ways that are not always in the light. The same 4 algorithms

applied to the same 569 sample dataset, under up to the mark cross validation (CV) and with standard imbalance handling, successfully outperform their published baselines consistently and by margins ranging from (+1.83% to +3.27%) which are large relative to the precision with which small dataset results should be trusted. For the researchers working on the similar datasets, the principal lesson of Phase I is that the accuracy figures from a single train test split should definitely be treated with appropriate scepticism without considering how high those numbers look like.

The second contribution, Phase II, is architectural: it shows that the gap between tabular paradigm where in the expert feature engineering reduces a slide to 30 numbers before any learning occurs and the image paradigm wherein a model learns directly from the pixels which is is not as large as any naive capacity argument would suggest. A 12545 parameter network trained on a laptop achieves near to 85% accuracy on any standard histopathological benchmark, without the need of any hand crafted features or specialist annotations or any institutional compute. The architectural choices which are enabling this depthwise separable convolutions,the Global Average Pooling, and the aggressive Dropout regularisation are simple, very well-understood, and directly transferable to other medical imaging works.

Taken together, the above mentioned two phases present a cohesive argument: that severity in evaluation methodology and efficiency in the model design together provide a practical pathway toward an accessible, reliable, and an interpretable computational pathology tools. The infrastructure that is required to pursue this direction is no longer a barrier and any researcher with a modern laptop, open source software, and a publicly available dataset can easily produce results that are both scientifically sound and practically relevant.

What remains is mainly the step from prototype to deployment: the external validation across institutions, integration of explainability tools as discussed earlier, and extension to the multi class subtype setting that are required by clinical practices. These are well defined problems with established methodological approaches, and thus, the foundation laid by this work provides a direct and reproducible starting point for each and every one of them.

REFERENCES

[1] Y. Tewari, E. Ujjwal, and L. Kumar, "Breast cancer classification using machine learning," in *IEEE International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ICACITE)*. IEEE, 2022.

[2] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[3] S. Vesal, N. Ravikumar, and A. Maier, "Classification of breast cancer histology images using transfer learning," in *International Conference on Image Analysis and Recognition*. Springer, 2018, pp. 812–819.

[4] D. Adams, *The Hitchhiker's Guide to the Galaxy*. San Val, 1995. [Online]. Available: http://books.google.com/books?id=W-xMPgAACAAJ

[5] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Breast cancer detection using deep convolutional neural networks," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 605–617, 2019.

[6] A. A. Nahid and Y. Kong, "Deep learning-based breast cancer classification using histopathological images," *BioMed Research International*, vol. 2018, pp. 1–13, 2018.

[7] T. Araujo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Automated breast cancer diagnosis using deep learning," *Computers in Biology and Medicine*, vol. 85, pp. 96–103, 2017.

[8] M. A. Khan *et al.*, "Breast cancer classification using transfer learning with cnn," *IEEE Access*, vol. 8, pp. 52 714–52 728, 2020.

[9] E. Tjoa and C. Guan, "Explainable ai for breast cancer classification using deep learning," *IEEE Access*, vol. 8, pp. 5622–5635, 2020.

[10] H. Gour, S. Jain, and T. Sunil Kumar, "Multi-class breast cancer classification using deep cnn models," *Expert Systems with Applications*, vol. 168, p. 115125, 2021.

[11] V. Chaurasia and S. Pal, "Breast cancer prediction using machine learning algorithms," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 1–9, 2018.