ISSN 2959-6157

Deep Learning for Accurate, Efficient, Economical, and Consistent Cancer Diagnosis Compared to Traditional Biopsy

Jiahong Lin

Hong Kong International School, Hong Kong, China
*251010@hkis.edu.hk

Abstract:

Early and precise cancer diagnosis is essential for enhancing the effectiveness of treatments. Traditional biopsy techniques, while reliable, are often time-consuming and economically inefficient. Furthermore, variations in diagnostic assessments among physicians introduce additional uncertainty in outcomes. This paper investigates the application of machine learning (ML) and deep learning (DL) methods to improve diagnostic accuracy and efficiency. It evaluates the advantages and disadvantages of feature-based versus image-based diagnostic approaches and introduces a new diagnostic workflow named AIStain. This workflow encompasses two pathways: one involving feature extraction followed by classical machine learning techniques, and the other using convolutional neural networks (CNNs) for deep learning analysis. Our analysis demonstrates that integrating machine learning can significantly enhance diagnostic speed, reduce costs, and improve consistency across evaluations without compromising accuracy. By leveraging advanced computational techniques, this approach aims to standardize cancer diagnostics and reduce the dependency on subjective human evaluation, potentially transforming cancer diagnosis practices.

Keywords: Deep Learning, Efficient, Economical, Consistent Cancer Diagnosis

1. Introduction

Cancer, characterized by the uncontrolled growth of malignant cells that outcompete healthy cells for resources and often lead to organ failure and death, impacts millions each year. These malignant cells may metastasize to other body parts in advanced stages, making treatment increasingly difficult. The World Health Organization states that „when cancer care is delayed or inaccessible, there is a lower chance of survival, greater problems associated with treatment, and higher costs of care“ [1]. Consequently, the development of methods for early detection of cancerous tumors is paramount. Techniques such as CT and MRI scans provide detailed internal body images, aiding physicians in identifying potential tumor sites. Nevertheless, these findings typically require confirmation through a biopsy.

Traditional biopsy involves extracting a small tissue sample from a suspected area, which is then manually examined under a microscope by a cytotechnologist. This method is prone to human error and varies significantly among physicians and institutions. Raab et al. reported that biopsy error rates are statistically significantly correlated with the institution, ranging from 1.79% to 11.8% [2]. Brouwer et al. highlighted that interobserver agreement varies among different discriminatory features, and the presence of these features does not consistently correlate with the final classification, even with physicians who have between 12 and 38 years of experience [2]. In this paper, the focus is on reducing interobserver variation in cancer diagnosis by reviewing various automated diagnostic methods for examining biopsy samples. Additionally, the paper proposes a cost-effective and accurate method for cancer diagnosis, leveraging some of the models analyzed within the study. This approach aims to enhance diagnostic reliability and reduce the discrepancies and delays inherent in traditional biopsy procedures.

2. Methodology

2.1 Datasets

2.1.1 The Wisconsin Breast Cancer Dataset

The first dataset used in this paper is the Wisconsin Breast Cancer Dataset, introduced by Street et al. It consists of 569 images analyzed by active contour models (“snakes”) to extract nuclear shape features automatically and precisely [3]. Users first indicate an approximate location and boundary, and the snake algorithm then contracts to find the exact cell boundary.

The extracted features are described below.
The radius is the average of the radial lengths. Radial lengths are measured from snake points to the centroid of the snake.
The perimeter is the total distance around the snake.
The area is measured by pixel counting.
The compactness is calculated by .
Smoothness is measured by the rate of change of radial lengths.
The concavity measures the quantity and degree of indentations.
The concave points statistic measures the number of concavities.
The symmetry statistic measures the length difference between lines on either side of the longest chord [4].
The fractal dimension gives an approximation of "roughness."
The texture is the variance of the grayscale pixel intensities.
The mean, extreme, and standard errors were calculated for each feature. Thus, there are 30 dimensions in the feature space.
The authors of the Wisconsin Breast Cancer Dataset achieved an impressive accuracy of 97% using plane separation of mean texture, worst area, and worst smoothness.

2.1.2 The LC25000 Dataset

The second dataset used in this paper is the LC25000, or Lung-Colon 25000 dataset, consisting of three classes of lung images and two classes of colon images with 5000 images (post-augmentation) in each class of dimension .

2.2 Dataset preprocessing

The Wisconsin dataset was balanced using SciKit Learn's resample function. After, the data was split in an train-test split stratified by the diagnosis class [5].
The author focused on the three lung classes of the LC25000 dataset, using Tensorflow's image_ dataset_ from _directory function to create a Tensorflow dataset with a batch size of 32. Images were resized to dimension with memory considerations. The dataset was split in an train-test split.

2.3 Training

Models were defined using SciKit Learn and Keras.
The paper evaluated ten separate algorithms for classifying the Wisconsin dataset, of which the first six were cross-validated with 15 randomized hyperparameter combinations [6].
K Nearest Neighbors (clustering) uses proximity to classify points according to given observations. Hyperparameters searched included n_ neighbors and algorithm.
Logistic Regression uses a learned multinomial logistic function to predict the probability of a point belonging to classes. The C hyperparameter (regularization strength) was searched.
Decision Trees split nodes according to entropy decrease so that deeper nodes refine the classification. Hyperparameters searched included split criterion, maximum depth, and minimum split proportion [7].
Random Forests use randomized feature selection (bootstrapping aggregation) and an ensemble of weak decision trees to make a collective decision. Hyperparameters searched included the number of estimators, split criterion, maximum depth, and minimum split proportion.
Adaptive Boosting uses a series of weak learners, giving more weight to the incorrect observations of previous models in the sequence to make a weighted prediction. Hyperparameters searched included the number of estimators and learning rate [8].
Gradient Boosting uses a series of weak learners predicting the error of previous models in the sequence. Hyperparameters searched included loss function, number of estimators, learning rate, minimum split proportion, and maximum depth.
XGBoost works similarly to Gradient Boosting, correcting the errors of previous models. It also implements techniques that make it robust against overfitting.
Soft Voting of models (1-7) by considering each model's prediction and prediction confidence.
Feedforward Neural Networks use layers of neurons with connected weights to make predictions. The author used eight layers of 60, 50, 40, 30, 20, 15, 5, and 2 densely connected units, respectively.
Adaptive Boosted Neural Network combines Adaptive Boosting and Feedforward Neural Networks with a series of weak neural networks that focus more on previous mistakes.
The author used 50 weak learners with six layers of 100, 75, 50, 25, 20, and two densely connected units, respectively, using AdaBoost.
For the LC25000 dataset, the author primarily used a Convolutional Neural Network. The specific architecture trained is shown below.
Convolution Filters = [16, 32, 32, 64, 16, 32].
Convolution Kernel Sizes = [1, 5, 2, 7, 3, 5].
Convolution Strides = [1, 2, 2, 6, 1, 2].
Pool Sizes = [2, 2, 2, 2, 2, 2].
Pool Strides = [1, 1, 1, 1, 1, 1].
Dense Units = [512, 3].
The CNN was trained over approximately 200 epochs spread out across multiple days of training. Early Stopping was enabled for initial training and architecture tuning monitoring validation loss, but it was later turned off [9].

A second model was tested, a Gradient Boosting model on the 512 intermediate and three final neuron values in the CNN. Intermediate neuron values were extracted using Keras' Functional API and stored in a Pandas DataFrame. The Gradient Boosting model was then initialized with the parameters previously found to be the best.

2.4 Model evaluation

Trained models were evaluated by interpolation (seen training data) and extrapolation (unseen validation/testing data). The models' prediction confusion matrices were plotted, with the heatmaps ideally displaying dark squares on the main diagonals. The models were also evaluated regarding accuracies and F1-scores (a balanced measure of precision and recall penalizing false negatives and false positives) [10].

3. Results

The results of the models on the Wisconsin dataset are shown in Figure 1, with f-score and loss plotted on a logarithmic scale.

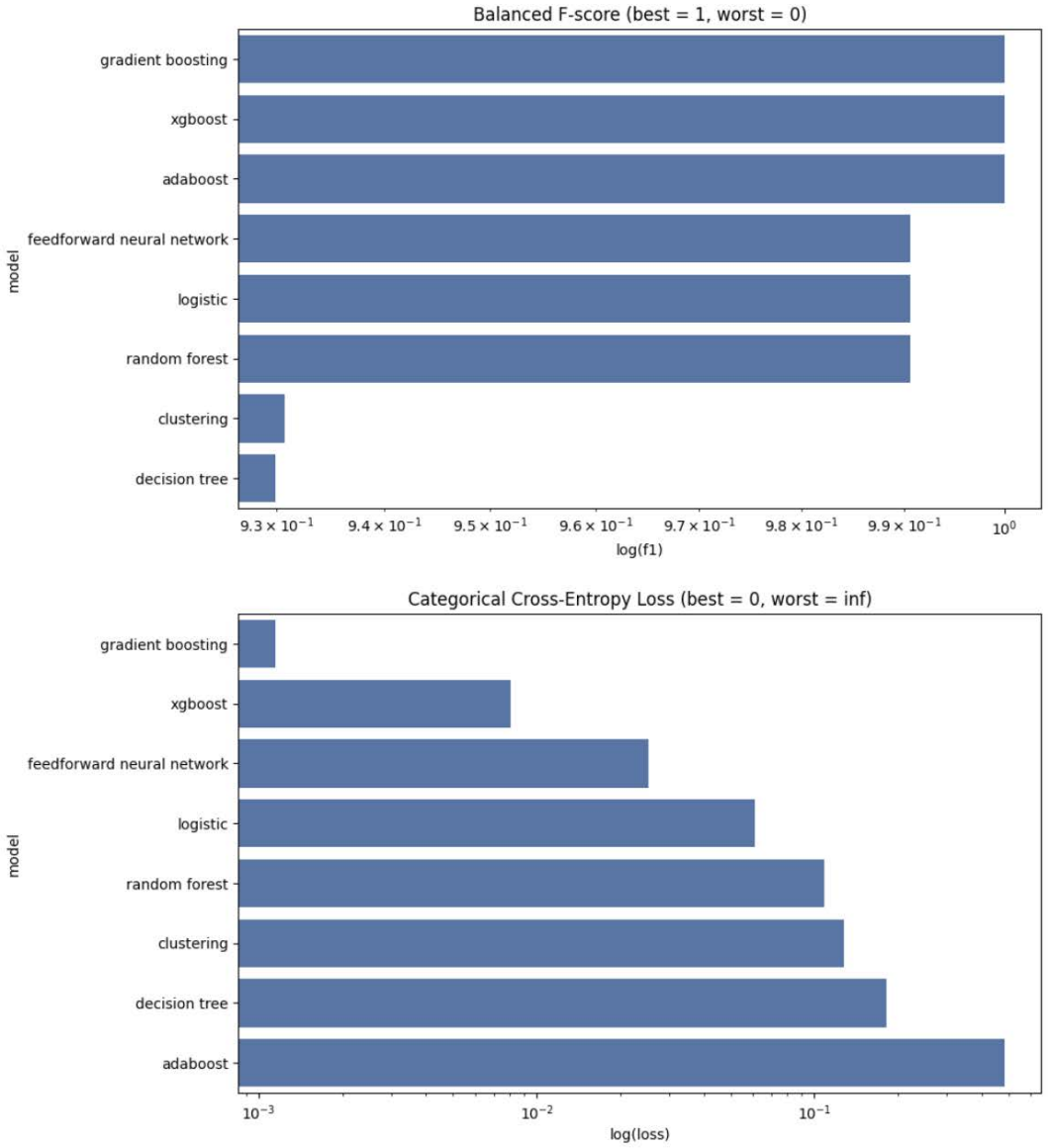

Fig. 1 Wisconsin dataset trained model f-score and loss performance (Photo/Picture credit: Original).

The gradient boosting and XGBoost algorithms consistently did best, with neural network, logistic regression, and random forest in the middle and clustering and decision tree performing worst. Interestingly, although AdaBoost had a high f score, indicating accurate results without false negatives or false positives, its loss was

high, indicating lower confidence levels. The result of the boosted neural network is shown in Figure 2, along with the weak network results in Figure 3.

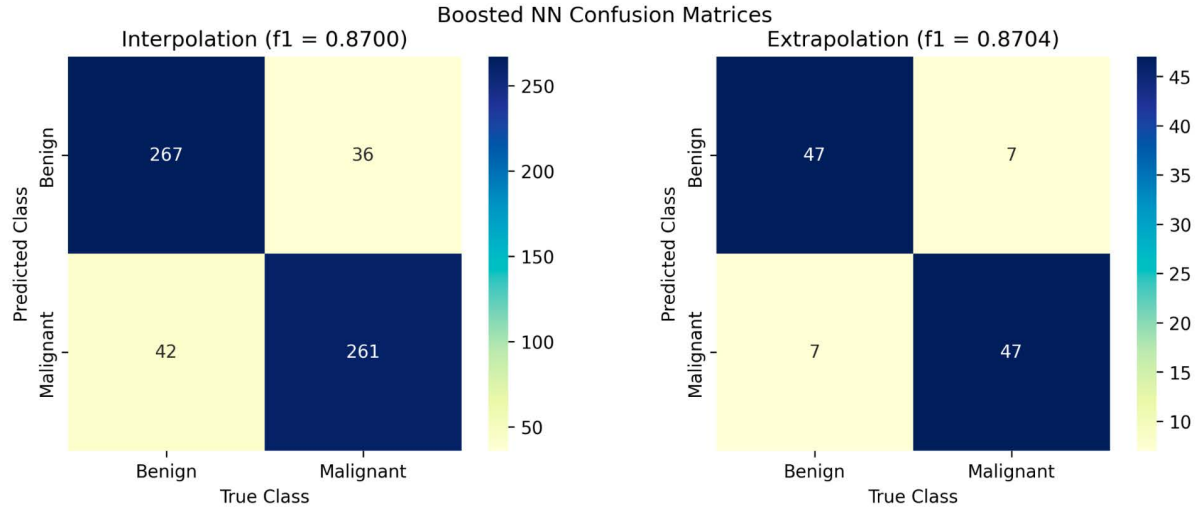

Fig. 2 Boosted Neural Network confusion matrices and f-score performance (Photo/Picture credit: Original).

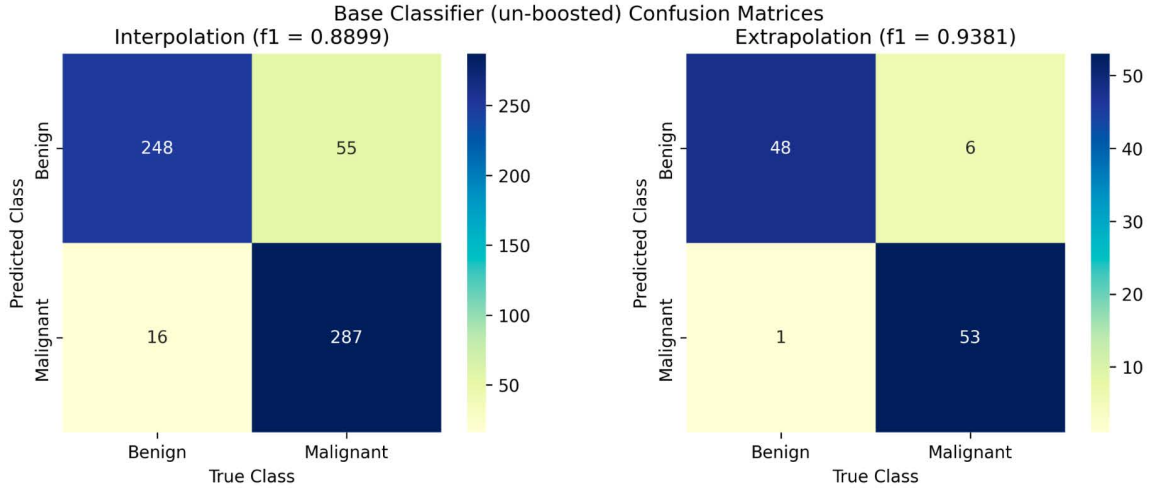

Fig. 3 Weak Neural Network confusion matrices and f-score performance (Photo/Picture credit: Original).

Additionally, some of the models supported feature importance. Rescaling the feature importances to the interval [0, 1] and then taking the mean, the below importance values were obtained.

The authors of the Wisconsin dataset used mean texture, worst area, and worst smoothness as three important features. Their strategy correlates with a limited extent to the trained models in this paper, which identified the worst perimeter, the worst concave points, and the worst area as the three most informative features, with the worst perimeter more than twice as important as any other variable. The full results are shown in Figure 4. From a biological standpoint, it makes sense that the worst statistic would be more critical since cancer cells tend to be outliers regarding characteristics.

These results also somewhat contrast with the typical features that pathologists look for in cancer tissue, namely that cancer cells might be larger or shorter and have uneven shapes compared to normal cells (Kumar et al.).

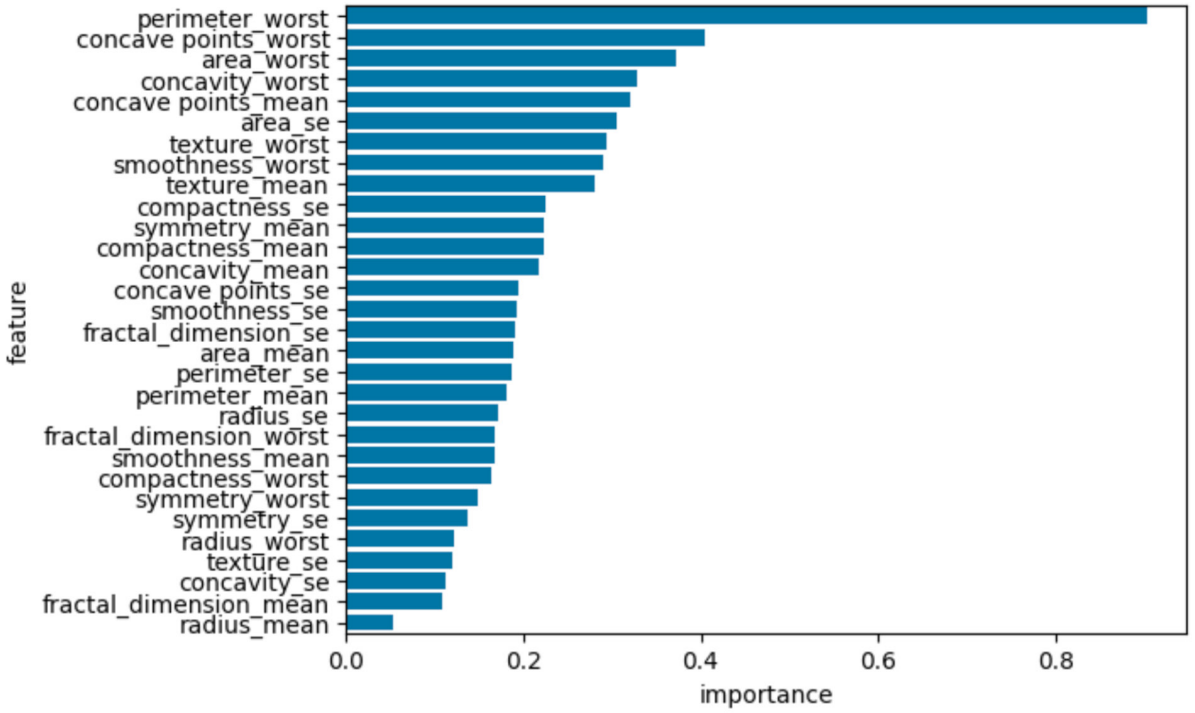

Fig. 4 Aggregated feature importances from all supported models (Photo/Picture credit: Original).

The results of the best iteration of the CNN trained on the LC25000 dataset are shown in Figure 5, along with a few examples of images classified in Figure 6.

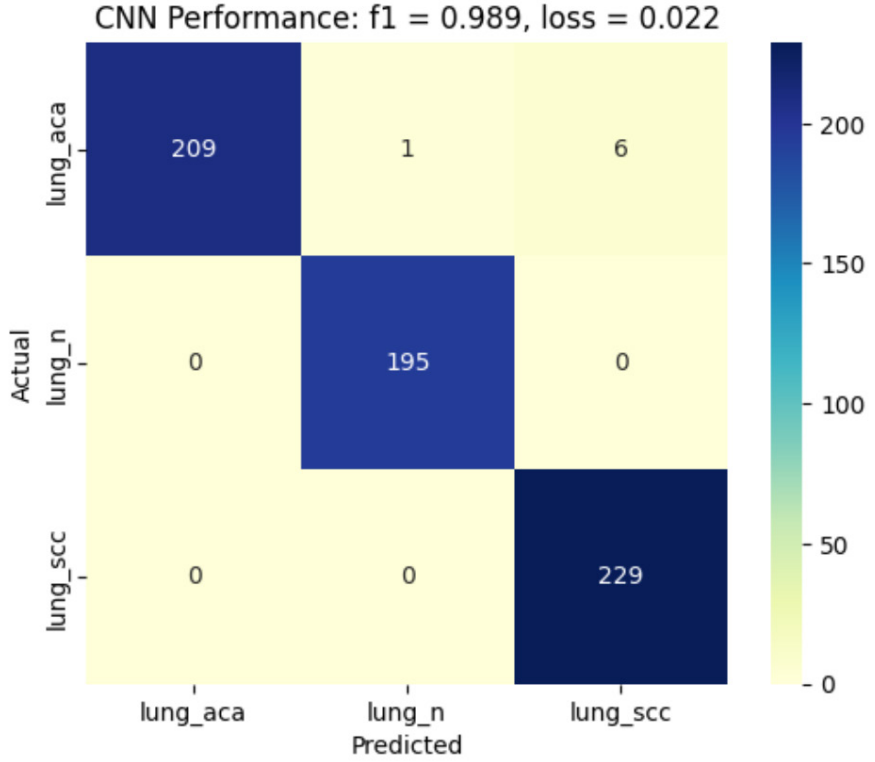

Fig. 5 CNN confusion matrix and f-score performance (Photo/Picture credit: Original).

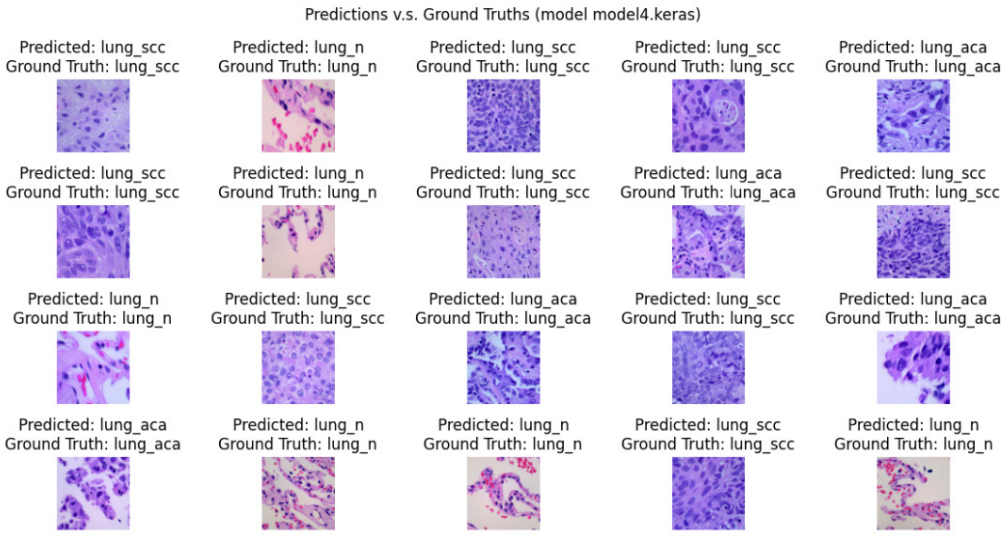

Fig. 6 Predicted and ground truths for CNN (Photo/Picture credit: Original).

The CNN did very well, achieving nearly 99% accuracy, a high f-score, and minimal loss.

On the other hand, the gradient-boosted CNN features did not do well, with its confusion matrix shown in Figure 7 nearly homogenous in color (identical to that of a random guess), which may indicate the incompatibility between the deep learning intermediate weights and classical machine learning due to the different types of features selected by the two methods.

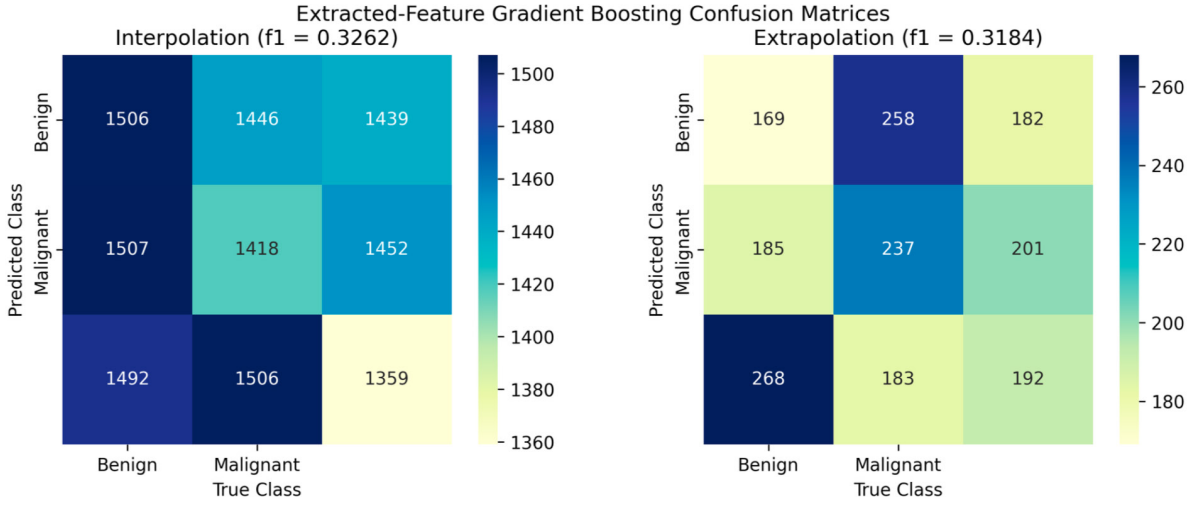

Fig. 7 Gradient boosting on CNN features confusion matrices and f-score performance (Photo/Picture credit: Original).

4. Conclusion and Discussion

4.1 A comparison of feature-based and image-based diagnosis

In general, both types of diagnosis did very well, exceeding the human accuracy rate. In addition, machine learning or deep learning techniques are more consistent and reliable: the same set of learned weights will make identical predictions, eliminating interobserver variation.

While the feature-based Wisconsin dataset yielded the best result (gradient boosting: accuracy = 1, f1 = 1) across all models and datasets surveyed, the data collection is time-consuming, with each cell needing to be analyzed manually to some extent (placing initial cell boundary estimates). On the other hand, the image-based LC25000 dataset offered lower costs (with image and weight load-

ing being simple and predictions taking around 288 milliseconds, ignoring the initial training overhead) while maintaining a high accuracy rate of 98.9% on unseen data. A side-by-side comparison of the two approaches is shown in Table 1.

Table 1. side-by-side comparison of feature-based and image-based diagnosis.

	Feature-Based Dataset	Image-Based Dataset
Pros	High accuracy: use a boosting model for the most confident and accurate results.	Comparatively lower accuracy; still very high (98.9% on validation).
Cons	High cost Tissue requires staining Manual preliminary image analysis required for each cell for feature extraction.	Comparatively lower cost Tissue requires staining High initial (fixed) cost on model training Low computational cost of prediction: loading trained model is nearly instantaneous; after, the prediction took)

4.2 Diagnosis workflow

Both approaches still suffer one significant costly step: the staining of tissue. This process is time-consuming (between 6-24 hours) and sometimes may require specialized dyes that cost large sums; these factors contribute to slow and costly diagnosis. Fortunately, Latonen et al. offer a solution via the virtual staining of pathological images. One could imagine a diagnosis workflow as follows.
The patient extracts a small tissue sample for biopsy.
The sample is photographed in situ.
The photograph is virtually stained (Latonen et al.).
From here, two pathways are available.
Pathway 1: the image is analyzed by hand for approximate cell boundaries. The snake algorithm automatically extracts the cell's features, like shape and texture. The features are then passed to a trained Gradient Boost model, which returns a highly accurate and confident prediction.
Pathway 2: the image is directly passed to the CNN, and an accurate and confident prediction is made.
Both pathways are approximately equal in confidence and accuracy, but pathway 1 requires time-consuming manual analysis (refer to Table 1). For this reason, the author recommends pathway 2 as the more practical solution.

4.3 Implications for cancer diagnosis

As discussed earlier, automating analysis and diagnosis has the quality of consistency and reliability, eliminating institutional-dependent accuracies and up to an 11% error rate as found by Raab et al. Additionally, automated analysis in combination with virtual staining has a significantly reduced cost. It takes much less time, making cancer diagnosis and treatment more accessible and immediate without sacrificing, and arguably improving accuracy. Automated analysis also does not require prior expert knowledge, especially for image-based diagnosis, as abstract features are learned automatically during training.

4.4 Future steps

This paper specifically addresses breast cancer using feature-based diagnosis and lung cancer through image-based techniques. It is conceivable that the datasets employed here could be extended to include a broader range of cancer types. Additionally, there is potential for future developments to automatically determine cancer stages and grades, a process currently marred by significant interobserver variation (Chowdhury). Future research could also explore the application of transfer learning, utilizing pre-trained biomedical versions of architectures such as ResNet. This approach may enhance the accuracy and generalizability of diagnostic models across various forms of cancer, contributing to more standardized and precise cancer diagnostics.

References

[1]Brouwer, NPM & Lord, Amy & Terlizzo, M & Bateman, AC & West, Nicholas & Goldin, R & Martinez, A & Wong, Newton & Novelli, Marco & Nagtegaal, Iris & Brown, Gina. (2021). Interobserver variation in the classification of tumor deposits in rectal cancer – is the use of histopathological characteristics the way to go?. Virchows Archiv. 479. 10.1007/s00428-021-03197-0.

[2]Chowdhury, Nilotpal et al. "Interobserver variation in breast cancer grading: a statistical modeling approach." Analytical and quantitative cytology and histology vol. 28,4 (2006): 213-8.

[3]Kumar, Rajesh et al. "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features." Journal of Medical Engineering vol. 2015 (2015): 457906. doi:10.1155/2015/457906

[4]Latonen, Leena, et al. "Virtual Staining for Histology by Deep Learning." Trends in Biotechnology, Elsevier BV, Mar. 2024. Crossref, doi:10.1016/j.tibtech.2024.02.009.

[5]Raab SS, Grzybicki DM, Janosky JE, Zarbo RJ, Meier FA,

Jensen C, Geyer SJ. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer. 2005 Nov 15;104(10):2205-13. doi: 10.1002/cncr.21431. PMID: 16216029.

[6]Street, Nick & Wolberg, William & Mangasarian, O. (1999). Nuclear Feature Extraction For Breast Tumor Diagnosis. Proc. Soc. Photo-Opt. Inst. Eng.. 1993. 10.1117/12.148698

[7]Faust, O., Acharya, U. R., Meiburger, K. M., Molinari, F., Koh, J. E., Yeong, C. H., ... & Ng, K. H. (2018). Comparative assessment of texture features for the identification of cancer in ultrasound images: a review. Biocybernetics and Biomedical Engineering, 38(2), 275-296.

[8]Zhu, X., Xu, H., Zhao, Z., & others. (2021). An Environmental Intrusion Detection Technology Based on WiFi. Wireless Personal Communications, 119(2), 1425-1436.

[9]Bhatt, H., Shah, V., Shah, K., Shah, R., & Shah, M. (2023). State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: A comprehensive review. Intelligent Medicine, 3(03), 180-190.

[10]Javeed, A., Khan, S. U., Ali, L., Ali, S., Imrana, Y., & Rahman, A. (2022). Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: A systematic review and future directions. Computational and Mathematical Methods in Medicine, 2022.