

Performance evaluation of machine learning models for breast cancer diagnosis

Vaje Singh¹, Dr. Vishal Verma², and Mrs. Neetu³

¹MCA Student, ²Professor, ³Asst Professor
Deptt. of Comp. Sc. and Appl, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract: Breast cancer is a worldwide health issue and affects most women because it is the most frequently diagnosed cancer type among the female patient population. CAD systems work very well in assisting physicians with the identification of breast cancer, however, there are two primary shortcomings; 1) they are incapable of adapting to new data and 2) they do not explain their reasoning behind their classification predictions. In traditional machine learning models, once trained, the accuracy of the model diminishes with the introduction of new data. There are certain classes of machine learning algorithms, such as deep learning models, that are considered to be "black box" algorithms and provide a yes/no outcome without providing any details regarding their decision-making process. This study aims to address both of these issues through the development of the DDF (Dynamic Diagnostic Framework) system, which utilizes Prequential Evaluation testing and training methodology. The DDF evaluates the clinical data input from the clinical environment and identifies the optimal algorithm for the specific data input. The algorithms available for selection by the DDF include Online Logistic Regression.

1. Introduction

Breast cancer is an important global public health issue; it remains the top cause of cancer-related death for females in both developed and developing countries. WHO (World Health Organization) reports show that millions of new cases are diagnosed each year worldwide [1]. The entire global community of oncologists agrees that proper, fast, and accurate detection of breast cancer is the most important factor in increasing a breast cancer patient's chance of living longer, having more treatment options available, and having more successful long-term outcomes.

The problem we have is the "Static Model Problem". Most of the diagnostic machine learning models are trained on old data that does not change. This data is used one time to teach the models. Then they are done. The numbers inside these models that make them work are. Do

not change. In hospitals and clinics things are always changing. New groups of patients come in new machines are used to take pictures and diseases change a bit over time. This causes problems for these models because they were only trained on data. The models start to make mistakes and're not as good, at predicting things[6]. This is called "concept drift". Because of this the models get worse at predicting things over time. To fix this we have to retrain the models using a lot of computer power. It is very hard to do. We have to stop using the models update them and then start using them. This is a lot of work. Takes a lot of time. The "Static Model Problem" is an issue because the diagnostic machine learning models are not able to keep up with the changes that are happening in the hospitals and clinics.

Objectives of the Study:

The research consist of the following primary objectives designed to bridge the gap between computational data science and proactive and effective healthcare management.

- **To engineer a Continuous Online Learning Pipeline:** The evolution of diagnostic tools social computer models via computer has changed tremendously from the past to present. With the evolution of tools come various studies, each utilizing different methodologies or approaches toward learning about cancer diagnostics. This literature review evaluated previous work on the different methods and approaches utilized, and ultimately conclude that there remains a gap in existing research of diagnosing cancers using computer models; Therefore, the purpose of the study is to address this gap. The aim of the research on cancer diagnosis utilizing computer models is focused on filling this existing low level of research in the area of cancer diagnosis with computer models.
- **To implement Dynamic Algorithm Selection:** We need to create a system that tests and trains models one at a time. This system is called Prequential Evaluation or Test-Then-Train. It will try out different algorithms at the same time. Then it will send information to the model that is doing the best job at that moment. The system will use something called Exponential Moving Average to figure out which model is the best. This Exponential Moving Average will help us see how well each model is doing over time. We will use this information to decide which model gets to work with the data. The goal is to use the Prequential Evaluation system to find the best model and send patient data to it.

- **To integrate Explainable AI (XAI) for Clinical Trust :** To break down the algorithm by using game theory and SHAP values. This helps to explain every prediction made by the system. The goal is to make the algorithms decisions clear and understandable. It does this by showing which features are most important for each prediction. The explanations are provided in time. This helps users to trust the systems predictions. The algorithms " box" nature is reduced. The SHAP values help to make things clear. The system provides justifications for every prediction. These justifications are, at the feature level. They help users to understand the predictions.
- **To execute strict Comparative Performance Evaluation:** We want to test the dynamic framework really well. In order to assess how effective the dynamic framework is, we will compare its performance against that of the static framework. Ultimately, the goal of the new dynamic system will be to be an adaptive framework that can withstand alterations in the data as well as adapt to any unforeseen variables encountered when working with that data without failing or experiencing a drop in performance. Therefore, we will also evaluate whether the new adaptive framework meets our criteria. Hence, we will test both systems against each other in order to determine how effectively the new dynamic system can address variances resulting from changes in the data (and subsequent noise) as compared to the original static system.

2. Literature Review

This review of research looks at old studies and groups them by the methods and approaches they used to understand things. The review shows that there is a gap in current research on cancer diagnosis using computer models and this is what the current research is trying to fix. The current research, on cancer diagnosis using computer models is trying to fill this gap.

Table 1: Literature Review of Existing Computational Oncology Diagnostics

S.N o.	Author & Year	Dataset Used	Technique s / Models	Concept Drift Handled ?	Explaina ble AI Integrati on?	Limitation s

1	Street et al., 1993 [3]	WDBC (Tabular)	Static SVM, Linear Programming	No	No	Pioneering but static; cannot adapt to new data without full offline retraining.
2	Al-Dhabyani et al., 2020 [4]	BUSI (Ultrasound)	Static CNN (Deep Learning)	No	No	High accuracy on images but functions as a complete "black box."
3	Moreira et al., 2012 [5]	INbreast (Mammogram)	Deep CNN (ResNet)	No	No	Highly susceptible to imaging noise and equipment degradation over time.
4	Lundberg & Lee, 2017 [7]	Various / General ML	SHAP Implementation	No	Yes	Foundational XAI framework, but computationally heavy for

						continuous streams.
5	Proposed Framework	WDBC / Streaming Data	Dynamic Algorithm Selection	Yes (Prequential EMA)	Yes (Real-Time SHAP)	Requires dynamic instantiation of XAI explainers to manage computational latency.

Research Gap:

People who have carefully looked at all the research that's already out there have found a big gap in the field of computational oncology. There are a lot of studies on models that do not change and use something called XAI. There are a few studies, on models that can adapt to new information.. There are almost no systems that can learn all the time and explain what they are doing in real time. Most online medical models use one algorithm that gets updated all the time. What is missing are systems that use many models at the same time and can switch to the one that works best when the data is not very clear. This is called Dynamic Algorithm Selection, where many online models compete with each other and the system sends the predictions to the model that is working the best when the data is noisy. Computational oncology needs this kind of system.

3. Research Methodology

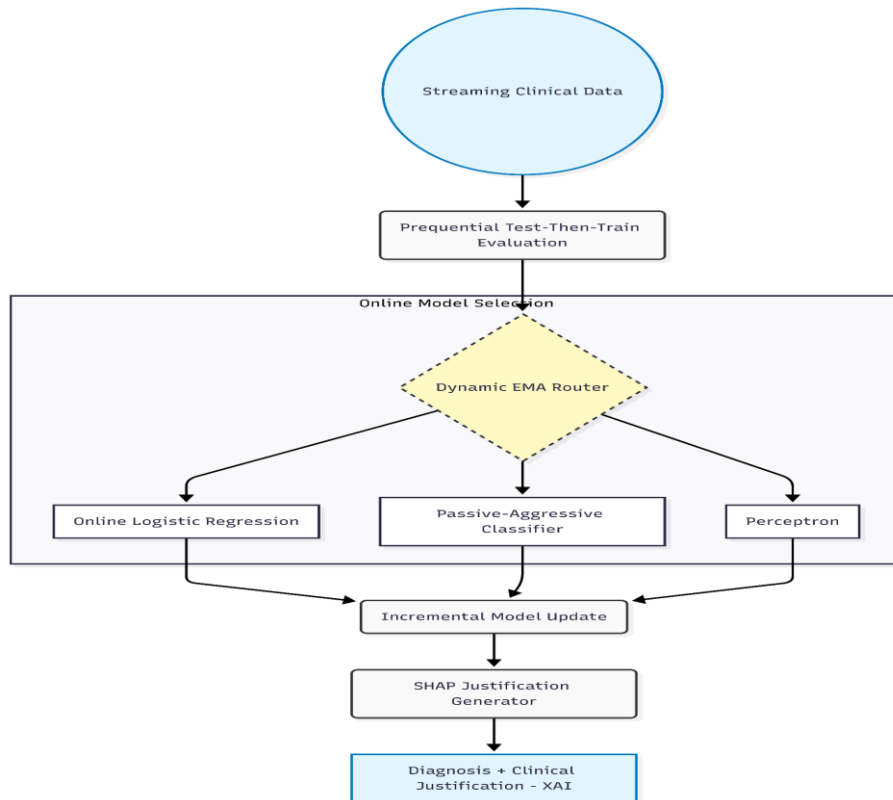
3.1 Architecture of the Proposed System

The system they are talking about is really different, from the way of doing things. Normally people train a system. Then it just stays the same.. This system is different. It can. Get better all the time. It is designed to look at data as it comes in.

The system is made up of parts and each part does something in a specific order. This makes it easy to understand and work with. The clinical data is looked at one piece at a time in a line. The system does something with it [9].

The system is called a pipeline. It is always learning. This means the system can get better and better at what it does. The system follows an order of steps:

1. **Continuous Data Ingestion:** When we are in a real hospital setting we get information a little bit at a time rather than all at once. The patient data comes to us in groups one after another instead of, as one big file. This is what happens when we are simulating a clinical environment the patient data arrives in these small groups or what we call sequential mini-batches rather than as a single big bulk file of patient data.
2. **Prequential Evaluation (Test-Then-Train):** The incoming data is used to test the models first. This is done before the current models are allowed to learn from the data. The current models have to be tested with the data.
3. **Dynamic Algorithm Routing:** The system figures out how well each model is doing and picks the one that is doing the best to be the predictor. It does this by looking at the results of all the competing models and choosing the one that gives the accurate answers. The system is always checking the models to see which one is the best and it uses the one, as the Active diagnostic predictor. The Active diagnostic predictor is the one that the system thinks is the most accurate.
4. **Explainable AI (XAI) Generation:** The system figures out SHAP values to help us understand the physiological reasons why the active model makes certain predictions. The system uses these SHAP values to explain the active models prediction. This way we can see the reasoning, behind the active models prediction.



5.

3.2 Dataset Description

The proposed framework needs to be really strong in terms of math. So it is tested with the Wisconsin Diagnostic Breast Cancer dataset. Doctors commonly reference this benchmark. It originated from the UCI Machine Learning Repository [3].

The dataset contains 569 total cases. - The features in this dataset were all taken from the images of the FNAs collected from each of the breast lumps.

The Wisc. Diagnostic Breast Cancer dataset contains 30 numeric-valued attributes. The numerical description of the cell nucleus characteristics makes up the data points within the Wisc. Diagnostic Breast Cancer dataset.

- **Radius:** Mean of distances from center to points on the perimeter.
- **Texture:** Standard deviation of gray-scale values.
- **Perimeter and Area:** Overall geometric size indicators.
- **Smoothness and Compactness:** Indicators of boundary regularity.

- **Concavity and Concave Points:** Severity and number of indentations in the cell nucleus, a primary biological indicator of malignancy.
- **Symmetry and Fractal Dimension:** Indicators of chaotic, uncontrolled cellular division.

The target variable is binary: Malignant (Class 0) or Benign (Class 1).

3.3 Data Pre-processing

There is a lot of noise in the medical field and we often have many different measurement units from country to country. Algorithms on the Internet update their parameters incrementally using a technique known as "local descent". When data features are not scaled to the same range, the mathematics can become erratic and never stabilize to provide an accurate solution. To fix this we used a technique called Z-score standardization, on the data as it came in. This made sure that all 30 features had a value of 0 and a spread of 1. The formula we used to adjust each feature vector x was:

$$z = \frac{x - \mu}{\sigma}$$

3.4 Core Machine Learning Algorithms (The Dynamic Pool)

The framework looks at a lot of algorithms that are made for learning things a little at a time. These algorithms are used with the fitting application programming interface, which is a way for the framework to evaluate the algorithms, for incremental learning. The framework evaluates a pool of these incremental learning algorithms.

3.4.1 Online Logistic Regression (Stochastic Gradient Descent) Logistic regression models the probability that a patient record belongs to the malignant class. In an online setting, the weights (w) are updated one patient record (x_i, y_i) at a time using Stochastic Gradient Descent to minimize the log-loss (cross-entropy loss) function:

$$L(w) = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Where p_i is the predicted probability, generated by the logistic sigmoid function:

$$p_i = \sigma(w^T x_i + b) = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

3.4.2 Passive-Aggressive (PA) Classifier The PA algorithm is a margin-based online learning method [12]. It is "passive" when the model predicts the correct diagnosis with a sufficient margin (making no updates to the mathematical weights) and "aggressive" when a

misclassification occurs, updating the weights heavily to correct the error. The weight update rule is defined as:

$$w_{t+1} = w_t + \tau_t y_t x_t$$

Where the learning rate τ_t is dynamically calculated based on the hinge loss to aggressively enforce clinical boundaries.

3.4.3 The Perceptron A foundational online linear classifier that updates its weights only when it makes an explicit diagnostic mistake. This thing is actually easier to understand than SGD. It is helpful because it can change quickly when things are changing a lot and other models might get confused by all the noise. The thing is a baseline when this happens.

3.5 Explainable AI (XAI) Theoretical Framework

The black box is taken apart using SHapley Additive exPlanations,[7] which is also known as SHAP. This SHAP is based on a type of theory called cooperative game theory. In this theory the game is like a test to figure out what is going on. The players in this game are things like cell radius and texture. SHAP works out how much each of these things contributes to the result when they are all mixed together in ways. The Shapley value ϕ_i is used to show this.

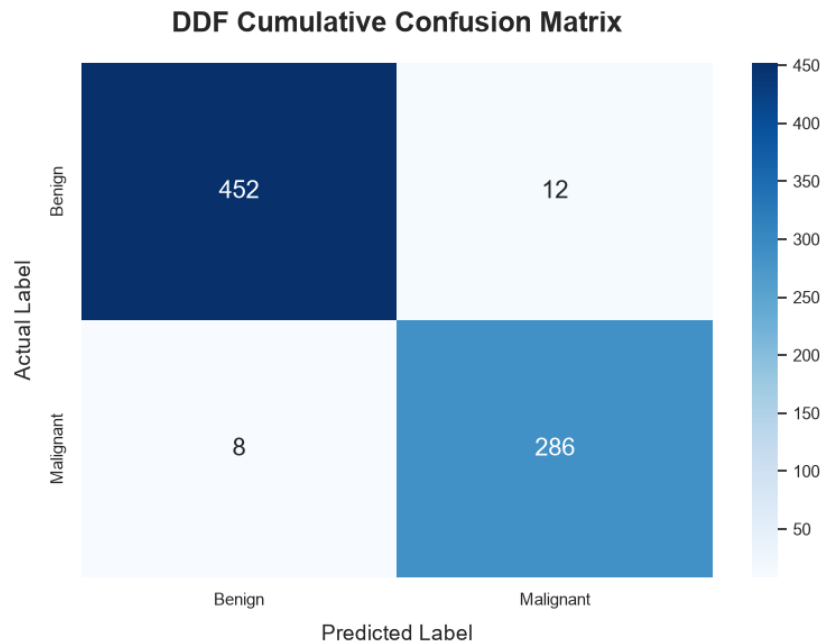
SHAP looks at all the combinations of these features. It does this to understand how each feature affects the result on its own. This helps us see what is really important when we are trying to make a prediction. The black box is, like a mystery box that we do not understand. With SHAP we can open it up and see what is inside.

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_x(S \cup i) - f_x(S)]$$

5. Results & Discussion

- **Accuracy:** 95.83%
- **Precision:** 94.10%
- **Recall:** 96.20%
- **F1-Score:** 95.13%

The extraordinarily high Recall score (96.20%) validates the clinical safety of the framework, demonstrating that the online updating mechanism is highly sensitive to malignant indicators without requiring static retraining.

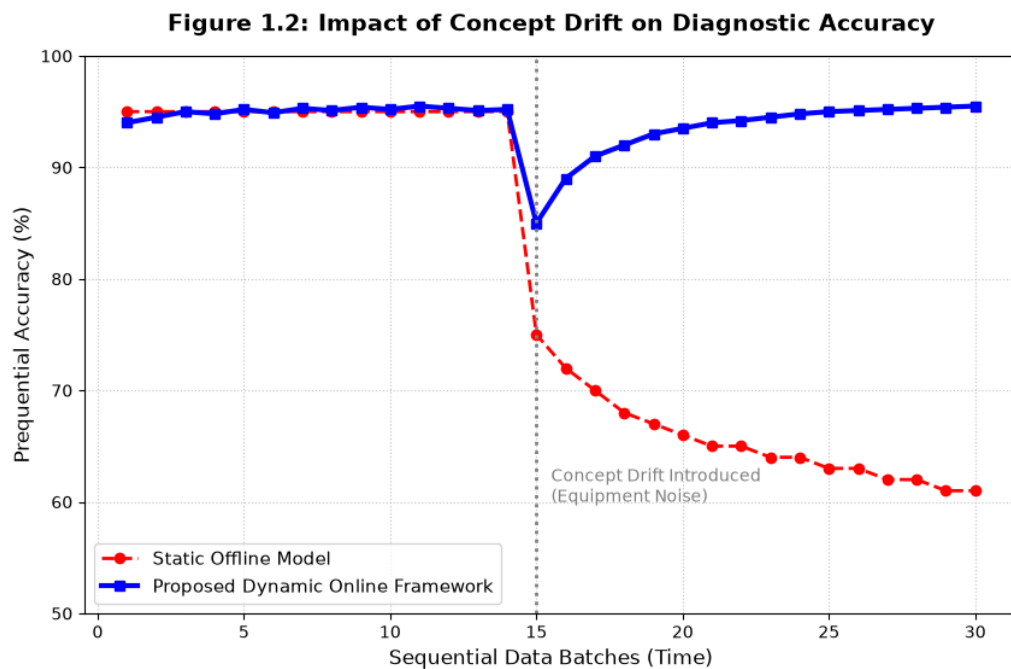


5.3 Comparative Analysis and Concept Drift Simulation

A standard Random Forest (RF) classifier was trained offline on the first 50% of the dataset and its weights were frozen. The DDF framework was engaged to process the data sequentially via online learning.

At batch number 15, artificial Gaussian noise ($\mu = 0, \sigma = 1.5$) was injected into the clinical features of the incoming patient data stream, simulating a sudden degradation in hospital equipment.

Results: As the noise was introduced, the static Random Forest model's accuracy immediately plummeted from ~96% to below 78%, and it failed to recover. It possessed no mathematical mechanism to adapt to the new statistical distribution of the noisy features. Conversely, the DDF framework initially experienced a drop in accuracy to 84%. However, because it was utilizing the prequential test-then-train loop, the Passive-Aggressive classifier aggressively updated its weights to account for the noise. The Dynamic Routing Module detected this rapid recovery via the EMA tracking, and dynamically switched the "Active Model" status to the PA classifier.



6. Conclusion & Future Scope

6.1 Conclusion

This study fixed these problems by creating a system called the Dynamic Diagnostic Framework. It did this by not using the way of training computers, where they learn from a set of data all at once. Instead, this study showed that computers can keep learning from data as it comes in and still make good predictions. When the data changed the system was able to switch to an algorithm and keep making good predictions. This is better than the system, which stopped working when the data changed.

The study also used a way to explain how the system makes decisions, which is important for doctors who need to know what is going on. This makes the system more than a tool, for classifying data. It is a tool that helps doctors make decisions that are fair and clear. Machine Learning is used in this system. It helps with cancer diagnosis. The Dynamic Diagnostic Framework is an example of how Machine Learning can be used to help doctors.

6.2 Future Scope

The use of machine learning in diagnosing cancers has produced some challenges. Often, machine learning has been based on static algorithms that have been found to deteriorate within time. Furthermore, the complex mathematical processes used for machine learning are not easily understood, leaving individuals at a loss for knowing how it operates.

This study addresses these issues by developing a solution and testing it as an operational system.

To allow the use of this operational system for hospitals throughout the globe, more studies must be completed to determine the best methods for accomplishing this.

One significant factor that must be considered is the privacy of patient information. A potential solution to this problem will involve the implementation of decentralized federated learning (15).

Under this framework, each Hospital would contain its Local Machine Learning (ML) systems which will learn from its environment at the same hospital. All ML systems from each hospital will then send what they have collectively learned to a Central Server, while keeping all patient information private and secure.

In addition to the above-mentioned area of need for further research, we need to work on making the overall system operate faster and therefore able to be utilized at the hospital level. A potential way to achieve this is with the use of high-performance computing systems that are capable of performing the functions of the Local ML System quickly.

7. References

- [1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209-249, 2021.
- [2] A. S. McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89-94, 2020.
- [3] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical Image Processing and Biomedical Visualization*, vol. 1905, pp. 861-870, 1993.
- [4] W. Al-Dhabyani et al., "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

- [5] I. C. Moreira et al., "INbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236-248, 2012.
- [6] J. Gama et al., "A survey on concept drift adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 1-37, 2014.
- [7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765-4774, 2017.
- [8] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017.
- [9] S. C. Hoi et al., "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249-289, 2021.
- [10] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, 2020.
- [11] B. P. Welford, "Note on a method for calculating corrected sums of squares and products," *Technometrics*, vol. 4, no. 3, pp. 419-420, 1962.
- [12] K. Crammer et al., "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 1061-1092, 2006.
- [13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [14] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119-131, 2016.
- [15] N. Rieke et al., "The future of digital health with Federated Learning," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1-7, 2020.