Assessing the Reliability of Machine Learning Predictions in Breast Cancer Diagnosis via Hypothesis Testing

Sonam Jawahar Singh
Department of Aerospace Engineering and Autonomous Systems
Defence Institute of Advanced Technology (DIAT)
Pune, India
singh.sj12@gmail.com

Pooja Agrawal
Department of Aerospace Engineering and Autonomous Systems
Defence Institute of Advanced Technology (DIAT)
Pune, India
poojaa.diat@gmail.com

doi: https://doi.org/10.37285/bsp.SACAD2025.66

***Abstract*—Medical research requires greater accuracy, verification, and validation of predictions, particularly in the era of precision medicine. Today's technology in the world is innovative and taking efforts towards precision medicine using Machine Learning (ML). Biomedical computing becomes easy by using an intelligent system. Biomedical computing includes developing novel medical devices and software for diagnosis and treatment, as well as using computers to analyze medical data for applications such as drug development, disease prediction, and medical imaging.**

Hypothesis Testing (HT) is used to study the predictions of intelligent systems that help to analyze and validate the predictions by mathematical approaches. This paper presents the parametric and non-parametric statistical methods of HT to study machine learning predictions on Breast Cancer (BC). The T-Test and Kruskal-Wallis Test (KWT) are applied to ML classifiers and studied the test statistic and p-value. The KWT gives a test statistic and p-value of 3.0000 and 0.3916, respectively. Hypothesis pruning, Type-I, and Type-II errors of the null hypothesis need to be studied for a much better way to verify medical predictions in the future.

Keywords—hypothesis testing, T-Test, Kruskal-Wallis Test, p-value, and precision medicine

I. INTRODUCTION

Hypothesis Testing (HT) is a statistical technique that assesses the possibility of a population assumption using sample data. A Test Statistic and p-value are determined by the Null Hypothesis (H0) and an Alternative Hypothesis (H1) [1]. The H0 and H1 use the collected sample schema for the simulation process. HT validates the Machine Learning (ML) performances on supervised, unsupervised, and test statistics during feature selection [2]. Genome datasets are too vast, making it challenging to extract them using both supervised and unsupervised methods [3]. HT provides a baseline to build semi-supervised methods to extract and encrypt genome datasets. H0 is a significance testing, utilizing a critical significance level of 0.05, is employed in a clinical decision support system [4]. The H0 implies the Type-I and Type-II hypothesis errors. The Type-I and Type-II hypothesis errors are symbolized as α and β, respectively, represented as diagrammatically shown in Fig. 1.

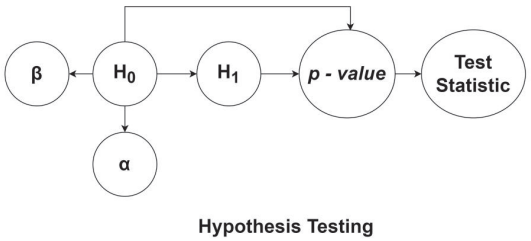

Fig. 1: Block Diagram of Hypothesis Testing

A scientific experiment is conducted to test the hypothesis that the study results won't show an effect. After testing, an H0 is either accepted or rejected, depending on several study characteristics and results, such as the p-value [5]. The fundamental social issue with the null hypothesis significance test is that it frequently fails to provide researchers with the information they believe it does [6]. Because of the way H0 significance testing is carried out, it is quite simple to mistake statistical significance for substantive or theoretical importance. This environment generates the requirement to study H1.

Although the p-value offers a magnitude of the strength of affirmation against H0, it does not specifically address the likelihood that H0 is correct [7]. HT fails in some cases; in such cases, hypothesis pruning can help with research [8]. Hypothesis pruning is a technique used in ML to reduce the number of viable hypotheses, or possible solutions. This approach is applied in domains such as computer vision to handle data contaminated by outliers, logic programming to enhance search efficiency, and natural language processing to optimize word alignment models.

This work presents the statistical importance of verifying and validating the performance of ML classifiers. Multilayer Perceptron (MLP), MultiClass, Random Forest (RF), and Sequential Minimal Optimization (SMO) are ML classifiers that measure performance based on Precision, Recall, F1 Score, Kappa Statistics, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) [9]. These performances are used as a schema for the simulation process for hypothesis result verification. These ML classifiers are applied on Breast Cancer [Dataset] [10] to predict the location of the tumor in the breast, represented as a feature *c_breast-quad* belonging to a multiclass classification challenge.

HT applied by using the T-Test and the KWT to study H0 and H1. These statistical analyses enable a decision support system in the medical domain to effectively manage chronic diseases. HT finds the most relevant ML classifiers used for prediction [11]. Mostly, the accuracy of prediction among all classifiers is considered the better-performing approach of classification. However, HT is used to study the overall performance of the approach, which indicates another aspect of classifier performance that gives them justice.

II. LITERATURE REVIEW

HT applied supervised and unsupervised approaches of ML for statistical analysis, feature selection, verification, and validation of predictions, and data types in [2]. The authors included the confusion matrix, specificity, sensitivity, accuracy, precision, recall, correlations between the features, and F1 score as performance measures. The limitations of HT and ML are discussed in [12], which states that by implementing a human brain and machine interface, the decision-making capacity can be collectively enhanced in chronic diseases such as cancer and heart disease. ChatGPT 4.0 generates a hypothesis on colorectal cancer is an example of the application of Artificial Intelligence (AI) used to study the research hypothesis on diseases like cancer [13]. The universal application of AI in research heralds a revolution of creativity, where AI and humane collaborate to solve complex problems in cancer detection.

Statistical analysis test performed and using p-value, breast lesion surgical evaluation is studied before and after COVID-19 [14]. The authors studied the impact of COVID-19 on patients' determination regarding whether to seek a second opinion from health experts or opt for surgical treatments. A statistically significant difference in lesion contrast-to-noise ratio between the two types of imagery was assessed using a paired T-Test in [15]. It determined whether there were statistically significant variations between the b-values selected by various evaluation techniques or clinical groupings. In [16], Fisher's Exact Test and T-Test are used to study the features associated with the predictions of biopsy mammogram images and screening examinations of abnormal and normal cell images.

The HT investigation focuses on determining how different kinds of human involvement in AI-ML co-operation impact the accuracy of demand predictions under various contextual circumstances, including product creativity or uncertainty [17]. AI research in the medical domain always looks for prediction validation. The importance of HT and p-value in medical research is discussed in [18] to determine the validity of scientific predictions, especially when analyzing the results of medical therapies, comprehending biological processes, or determining correlations between features. T-Test, KWT, p-value, and α discussed in [19] and stated statistical significance for predictions, especially in the medical area, to solve the problem through problems.

Mathematical model heuristics is a tactic that supports real-time problem-solving approaches that subdivide the big problem into subproblems and solve the subproblems. Approaches of heuristics are explained in [20]. The author supports real-time problems like disease studies in the medical domain are solved by applying heuristic approaches, e.g, to study the pattern of features, divide the complex problems into subproblems, work backward manner, etc. The authors proposed a graphical interface in the R environment [21] for non-parametric approaches like the KWT, Chi-Square, Fisher Exact, and Wilcoxon Test, etc., to use the hypothesis testing for interpreting the results.

A critical analysis of statistical significance testing in machine learning model comparisons suggests in [22] that a paradigm change away from the traditional dependence on p-values and T-Tests is urgently needed. Despite their foundational nature, these traditional methods have certain shortcomings, including a tendency to misread, results that are virtually useless because of sample size sensitivity, and an inability to adequately capture the breadth and unpredictable nature of observed effects. In the rapidly evolving field of ML, this over-reliance has resulted in a "statistical debt" that manifests as a reproducibility crisis where many published results may not withstand additional scrutiny.

III. METHODOLOGY

This work uses the T-Test in two conditions and the KWT to study the hypothesis. A hypothesis is an assumption about reality with validity that has not yet been established. HT is applied to the performance of ML classifiers. These hypothesis studies provide the baseline to validate the performance metrics. It will help to solve real-world challenges in chronic diseases (cancer, diabetes, and heart disease) by simplifying their complexity.

A. T-Test

It is a statistical approach of HT, a type of parametric test. For numerical data, parametric tests employ parameters such as the mean and make hypotheses about the statistical distribution of the population (e.g., health, disease, drug distribution). The T-Test investigates whether there is a significant difference between the means of two groups, illustrated in Equation (1).

$$T = \frac{\text{difference between means}}{\text{standard deviation } (\sigma) \text{ from mean}} \quad (1)$$

T-Test applied in two conditions, Paired and Independent T-Test, which are detailed below.

a) ***Paired T-Test:*** It compares the two paired samples. In the simulation schema classifier, MLP is paired with other classifiers for comparison. The H0 in the paired T-test contains the zero mean difference between the pairs, while in H1, the mean difference is not zero.

b) ***Independent T-Test:*** This test represents a significant difference between the separate samples or groups. Each classifier with its performance metric compares independently. The H0 in the Independent T-Test considers that the mean value in both groups is are same. There is no difference between the two groups. On the other hand, in the case of H1, the mean value in both groups is not equal. There are differences between the two groups.

B. Kruskal-Wallis Test (KWT)

It is employed to determine whether many independent groups vary from one another. It works well with a tiny schema. The KWT is an analysis of variance. In the schema, every variable has a rank. First place goes to the greater value, second place goes to the lower value, and so forth. Just rank the first group first, followed by the second, and then sum up the ranks for the other groups. Next, add up all of the ranks in

every group. The H0 is that there is no difference between the independent groups, and H1 is the a difference between the independent groups. It is a non-parametric test.

Key Terms: The performances of the HT are studied by using the test statistic and p-value, which are discussed below.

a) ***Test Statistic:*** In a hypothesis test, it is a standardized number that is computed from sample data to ascertain how far the sample data deviates from the H0. It assists statisticians in determining whether to reject the H0 by quantifying the difference between the observed data and the H0. To determine statistical significance, the test statistic's value is either compared to a critical value or utilized to compute a p-value.

b) ***p-value:*** It is the hypothesis's statistical significance. If the H0 proves true, it is known as the possibility of the observations plus even more severe outcomes.

IV. DATA PREPARATION

MLP, MultiClass, RF, and SMO classifiers predict the location of tumors in the breast that belong to multiclass classification [9]. The success metrics are Precision, Recall, F1 Score, Kappa Statistics, MAE, and RMSE. For the simulation purpose, these predictions are depicted in Table I. The ML classifiers are used on the Breast Cancer [Dataset] [10] to predict the multiclass classification. As Table I illustrates, the SMO gives higher predictions than other classifiers; however, to verify the predictions, the T-Test and KWT are applied to the classifiers' performance.

TABLE I. DATA USED FOR HYPOTHESIS TESTING [9]

SN	*Success Metrics*	*MLP*	*MultiClass*	*RF*	*SMO*
1	Precision	0.337	0.350	0.254	0.375
2	Recall	0.326	0.372	0.256	0.442
3	F1 Score	0.329	0.360	0.253	0.405
4	Kappa Statistics	0.0136	0.0563	-0.1041	0.1211
5	MAE	0.2738	0.2762	0.2871	0.287
6	RMSE	0.4641	0.4009	0.4216	0.3817

The UCI repository provides five different kinds of datasets of the disease BC [23]. Breast Cancer [Dataset] [10] is one of them, which is used for the simulation to predict the location of tumors. This dataset contains 10 features of cancer named as *'c_age', 'c_menopause', 'c_tumor-size', 'c_inv-nodes', 'c_node-caps', 'c_deg-malig', 'c_breast', 'c_breast-quad', 'c_irradiat' and 'c Class'*. The feature *'c_breast-quad'* is considered a target class for the predictions of tumor location.

V. RESULTS

T-Test applied on classifiers' predictions to verify the predictions; however, for the T-Test, the sample size is minimal, as depicted in TABLE II and TABLE III. In both conditions of the T-Test, gets test statistic and p-value are NaN. This motivates the HT by using a non-parametric approach, i.e., the KWT. The test statistic and p-value of the KWT are 3.0000 and 0.3916, respectively, represented in TABLE IV. KWT applied to MLP, MultiClass, RF, and SMO classifiers discussed in Section IV. HT aims to study whether their performance whether there is a significant difference is present or not and to study the various hypotheses of the predictions. These tests are flexible for a small amount of data on several samples or groups. The test statistic and p-value of the KWT are the same, indicating that there is no significant difference between the performance of classifiers, which indicates there is no difference between the clinical prediction of the classifiers.

TABLE II. PERFORMANCE OF THE T TEST: CONDITION 1 - PAIRED T-TEST

SN	*Outcomes of T Test: Condition 1*	*Test Statistic*	*p-value*
1	MLP vs MultiClass Classifier	NaN	NaN
2	MLP vs Random Forest	NaN	NaN
3	MLP vs SMO	NaN	NaN

Note: NaN indicates that the test statistic and p-value could not be computed due to insufficient data. For the Kappa Statistics, invalid conditions occurred for the paired t-test.

TABLE III. PERFORMANCE OF THE T TEST: CONDITION 2 - INDEPENDENT T-TEST

SN	*Outcomes of T Test: Condition 1*	*Test Statistic*	*p-value*
1	MLP vs MultiClass Classifier	NaN	NaN
2	MLP vs Random Forest	NaN	NaN
3	MLP vs SMO	NaN	NaN

The significance level is measured by the p-value. If the p-value is equal to or less than 0.05, then H0 is rejected. In the simulation approach, KWT gives a p-value of 0.3916, illustrated in TABLE IV, which is greater than the significance level. The H0 is not rejected, although there is no significant difference. Also, the p-value indicates the ideal situation for the H1, but H1 is false because 0.3916 is much higher. In this situation, hypothesis pruning, α, and β need to be studied in the future. Additionally, it highlights the importance of non-parametric approaches in medical research.

TABLE IV. PERFORMANCE OF KWT

SN	*Outcomes of KWT*	*Test Statistic*	*p-value*
1	Precision	3.0000	0.3916
2	Recall	3.0000	0.3916
3	F measure	3.0000	0.3916
4	Kappa Statistics	3.0000	0.3916
5	MAE	3.0000	0.3916
6	RMSE	3.0000	0.3916

The medical diagnostic metrics, which are crucial for the identification of breast cancer, show that SMO has a minor clinical advantage in sensitivity and F1 score. The clinical predictions of the machine learning classifiers will be discussed in Section VI.

VI. DISCUSSION

This research demonstrates the application of statistical hypothesis testing methods to validate the performance of machine learning classifiers used for breast cancer prediction. The study compared four classifiers—MLP, MultiClass, RF, and SMO — using performance metrics such as Precision, Recall, F1 Score, Kappa Statistics, MAE, and RMSE. The results from the KWT produced a p-value of 0.3916, which is greater than the significance level $\alpha = 0.05$. Therefore, the H0 cannot be rejected, indicating that there is no statistically significant difference in the performance of the classifiers. In this case, the H1, which suggests a significant difference between classifier performances, is not supported. This outcome implies that all evaluated classifiers perform comparably on the breast cancer dataset.

From a medical diagnostic perspective, this finding highlights the importance of statistical validation in machine learning research. While classifier metrics such as Precision, Recall, and F1 Score reflect individual performance tendencies, hypothesis testing ensures that these differences are not due to random variation. Such statistical reliability is crucial in sensitive applications like cancer diagnosis, where false positives or negatives may directly impact patient outcomes. Future work may extend this study by incorporating larger and more diverse datasets, optimizing model architectures, and applying additional non-parametric tests to strengthen statistical inference. Integrating hypothesis testing with explainable AI could also enhance clinical trust and decision-making reliability in cancer diagnostics.

The statistical outcomes indicate that the machine learning classifiers for BC predictions performed comparably, as discussed.

1) No single classifier is superior: It cannot claim that, for example, MLP is better than Random Forest for detecting breast tumors based on these metrics.

2) Decision-making: Since classifiers perform similarly, one might choose a classifier based on computational efficiency, ease of implementation, or interpretability rather than performance.

3) Sensitivity (Recall) importance: In medical diagnostics, high sensitivity is crucial to minimize false negatives (missing actual cancer cases). If all classifiers have similar sensitivity, they are equally reliable in catching positive tumor cases.

4) Null hypothesis stands: Because H0 is not rejected, we cannot support the idea that any classifier is statistically better. For future studies, we need larger datasets or more discriminative features to detect meaningful differences.

VII. CONCLUSIONS

This study presents a statistical evaluation of ML classifiers for predicting breast tumor locations. Both parametric (T-Test) and non-parametric (KWT) hypothesis testing methods were employed to assess classifier performance. The KWT demonstrated superior flexibility in evaluating the null and alternative hypotheses compared to the t-test. The obtained test statistic (3.0000) and p-value (0.3916) indicate that the H0 cannot be rejected. However, KWT states the no significant difference between ML classifiers' predictions regarding the tumor locations. All classifiers are comparable, and if collectively considering the predictions of classifiers, that motivates the screening test of patients to identify the tumors in the real world for precautionary purposes. The KWT is suitable for small-scale data with multiple classifier groups. Future work will focus on hypothesis pruning and the analysis of Type I (α) and Type II (β) errors to enhance medical decision support systems.

REFERENCES

[1] K. E. Meyer, A. van Witteloostuijn, and S. Beugelsdijk, "What's in a p? Reassessing best practices for conducting and reporting hypothesis testing research," JIBS Special Collections, pp. 77–110, 2019.

[2] G. Alkhatib, "Exploring Machine Learning Hypothesis Testing," International Journal of Computers, vol. 9, 2024.

[3] S. J. Singh and P. Agrawal, "Machine learning based genome cancer dataset approaches: Review," AIP Conference Proceedings, vol. 3217, no. 1, 2024.

[4] P. M. Sedgwick, A. Hammer, U. S. Kesmodel, and L. Pedersen, "Current controversies: Null hypothesis significance testing," Acta Obstetricia et Gynecologica Scandinavica, vol. 101, no. 6, pp. 624–627, 2022.

[5] J. C. Travers, B. G. Cook, and L. Cook, "Null Hypothesis Significance Testing and p Values," Learning Disabilities Research & Practice, vol. 32, no. 4, pp. 208–215, 2017.

[6] J. Gill, "The Insignificance of Null Hypothesis Significance Testing," Political Research Quarterly, vol. 52, no. 3, pp. 647–674, 1999.

[7] J. Uttley, "Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research," LEUKOS, vol. 15, no. 2–3, pp. 143–162, 2019.

[8] A. Cropper and R. Morel, "Learning programs by learning from failures," Machine Learning, vol. 110, no. 4, pp. 801–856, 2021.

[9] S. J. Singh and P. Agrawal, "Multi-class Prediction and Anomaly Detection for Breast Cancer using Machine Learning–Data Mining Approach," Proc. 9th Int. Conf. on Computing, Communication, Control and Automation (ICCUBEA), Pune, IEEE, 2025.

[10] M. Zwitter and M. Soklic, "Breast Cancer," UCI Machine Learning Repository, 1988. [Online]. Available: https://doi.org/10.24432/C51P4M.

[11] M. C. Gonc¸alves and R. Silva, "The Effect of Statistical Hypothesis Testing on Machine Learning Model Selection," Lecture Notes in Computer Science, pp. 415–427, 2023.

[12] G. Habboub, M. M. Grabowski, M. L. Kelly, and E. C. Benzel, "The embedded biases in hypothesis testing and machine learning," Neurosurgical Focus, vol. 48, no. 5, p. E8, 2020.

[13] L. Yao et al., "Generating research hypotheses to overcome key challenges in the early diagnosis of colorectal cancer – Future application of AI," Cancer Letters, vol. 620, p. 217632, 2025.

[14] G. Vanni et al., "Breast Cancer and COVID-19: The Effect of Fear on Patients' Decision-making Process," In Vivo, vol. 34, no. 3 suppl, pp. 1651–1659, 2020.

[15] M. R. DelPriore et al., "Breast Cancer Conspicuity on Computed Versus Acquired High b-Value Diffusion-Weighted MRI," Academic Radiology, vol. 28, no. 8, pp. 1108–1117, 2020.

[16] A. Akselrod-Ballin et al., "Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms," Radiology, vol. 292, no. 2, pp. 331–342, 2019.

[17] E. Revilla, M. J. Saenz, M. Seifter, and Y. Ma, "Human–Artificial Intelligence Collaboration in Prediction: A Field Experiment in the Retail Industry," Journal of Management Information Systems, vol. 40, no. 4, pp. 1071–1098, 2023.

[18] M. Krishnakumar, S. Dachepalli, M. Nair, G. Gutjahr, and R. Sankaran, "Statistics for Doctors: Hypothesis Testing and P Values," Amrita Journal of Medicine, vol. 21, no. 1, pp. 43–47, 2025.

[19] H. M. Salman and N. Aleem, "An Easy Way to Understand the Research Tools, Scales and Parametric & Non-Parametric Tests," SSRN Electronic Journal, 2024.

[20] L. C. Larson, Problem-Solving Through Problems, New York, NY, USA: Springer-Verlag, 1983.

[21] P. Kumar-M and A. Mishra, "Inferential Statistics for the Hypothesis Testing of Non-parametric Data," R for Basic Biostatistics in Medical Research, pp. 157–197, 2024.

[22] M. Nadim et al., "Statistical Significance Testing in ML Model Comparisons: Beyond p-values and t-tests," International Journal of Testing 2025. [Online]. Available: https://www.researchgate.net/publication/392727623

[23] UCI Machine Learning Repository, "Breast Cancer Dataset," UCI.edu, 2018. [Online]. Available: https://archive.ics.uci.edu/datasets?search=Breast%20Cancer