ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 2 Issue 1 January 2024 / Page No: 1-9
Paper Id: IJACT-V2I1P1011 / Doi: 10.56472/25838628/IJACT-V2I1P101

Original Article

Technical Evaluation of Machine Learning Models: An Empirical Study

Atta Yaw Agyeman[1], Samuel Gbli Tetteh[2]
[1,2]*D Jarvis College of Computing and Digital Media, DePaul University, Chicago, USA*

Received Date: 18 December 2023 Revised Date: 06 January 2024 Accepted Date: 22 January 2024

Abstract: *In the current era of technological advancement, the proliferation of diverse data sources has revolutionised decision-making processes across the globe. This exponential growth in data availability has reshaped decision-making paradigms and unlocked unprecedented opportunities for applying machine learning methodologies. Mainly, domains such as disease detection and intricate economic analysis have witnessed a significant transformation due to the advent of machine learning algorithms. Amidst these developments, the incidence of breast cancer continues to surge in both developed and developing nations, posing significant challenges to healthcare systems worldwide. In response to this pressing concern, this study endeavours to amalgamate these trends by comprehensively analysing major machine learning models to classify breast cancer tissues. Utilising the Wisconsin Breast Cancer Dataset as the foundational framework, this research aims to evaluate the efficacy of various machine learning algorithms in distinguishing between benign and malignant tissues. The repertoire of machine learning models under scrutiny encompasses Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors (KNN), as well as two variants of Support Vector Machine (SVM) — Radial Basis Function (RBF) and Linear classifier. Additionally, the study incorporates Decision Tree Classifier and Random Forest (RF) algorithms into its comparative analysis. The study's findings underscore the pivotal role of Random Forest (RF) and the diverse variations of Support Vector Machine (SVM) in achieving remarkable classification accuracy. Moreover, these models exhibit superior precision, recall, and f1-score performance metrics, highlighting their efficacy in breast cancer tissue classification tasks.*

Keywords: *Support Vector Machine (SVM), Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree Classifier, Random Forest (RF).*

I. INTRODUCTION

Recent technological advancements have revolutionised various aspects of human life, significantly increasing the volume and diversity of available data. This influx of data has profoundly influenced decision-making processes across industries, including business operations (Ozgur et al., 2015). The growing abundance of data underscores the importance of practical data analysis, which has proven invaluable in diverse fields such as disease detection, sales forecasting, economic analysis, and pattern recognition (Kaur et al., 2018). Consequently, the surge in data-driven insights has spurred extensive research in the realm of machine learning, offering powerful analytical tools and techniques.

Machine learning, a discipline within computer science, aims to enable computers to emulate human activities and continuously enhance their performance through self-improvement mechanisms (Wang et al., 2009). This field boasts many applications, including image recognition, sentiment analysis, news categorisation, video surveillance, prediction modeling, and recommender systems.

Breast cancer stands as one of the most prevalent malignancies worldwide, posing a significant threat to individuals in both developed and developing nations. Failure to detect breast cancer early can have life-threatening consequences, making early diagnosis imperative (Asri et al., 2016). Recent data indicate a notable rise in breast cancer cases, particularly in developing regions like Africa, underscoring the urgency of effective diagnostic methodologies (Walker et al., 2004). Consequently, researchers have turned to machine learning models to classify breast tissue samples intelligently as benign or malignant, leveraging datasets like the Wisconsin Breast Cancer Dataset (WBCD) (Seddik & Shawky, 2015). Various machine learning algorithms, including logistic regression, support vector machines, neural networks, and naive Bayes classifiers, have been explored for breast cancer classification, each offering unique insights into the disease (Edriss et al., 2016; Kharya & Soni, 2016; Al-Hadidi et al., 2017).

In light of the diverse opportunities presented by machine learning, this study endeavours to assess the performance of multiple machine learning models using the WBCD from the UCI data repository. The study aims to shed light on the efficacy of different machine learning algorithms in breast cancer diagnosis by employing rigorous evaluation metrics and

conducting detailed analyses. Noteworthy models under investigation include logistic regression, k-nearest neighbour, support vector machines (linear and radial basis function classifiers), naive Bayes, decision trees, and random forests.

The burgeoning interest in machine learning has fueled research efforts to enhance learning efficiency and explore novel applications across various domains (Loussaief & Abdelkrim, 2017; Boiy & Moens, 2009; Suleymanov & Rustamov, 2018). As depicted in Figure 1, machine learning encompasses diverse subfields and methodologies, reflecting the multidimensional nature of this rapidly evolving discipline (Dey, 2016). This research seeks to contribute to the ongoing discourse on machine learning applications, particularly in the context of breast cancer diagnosis and treatment.

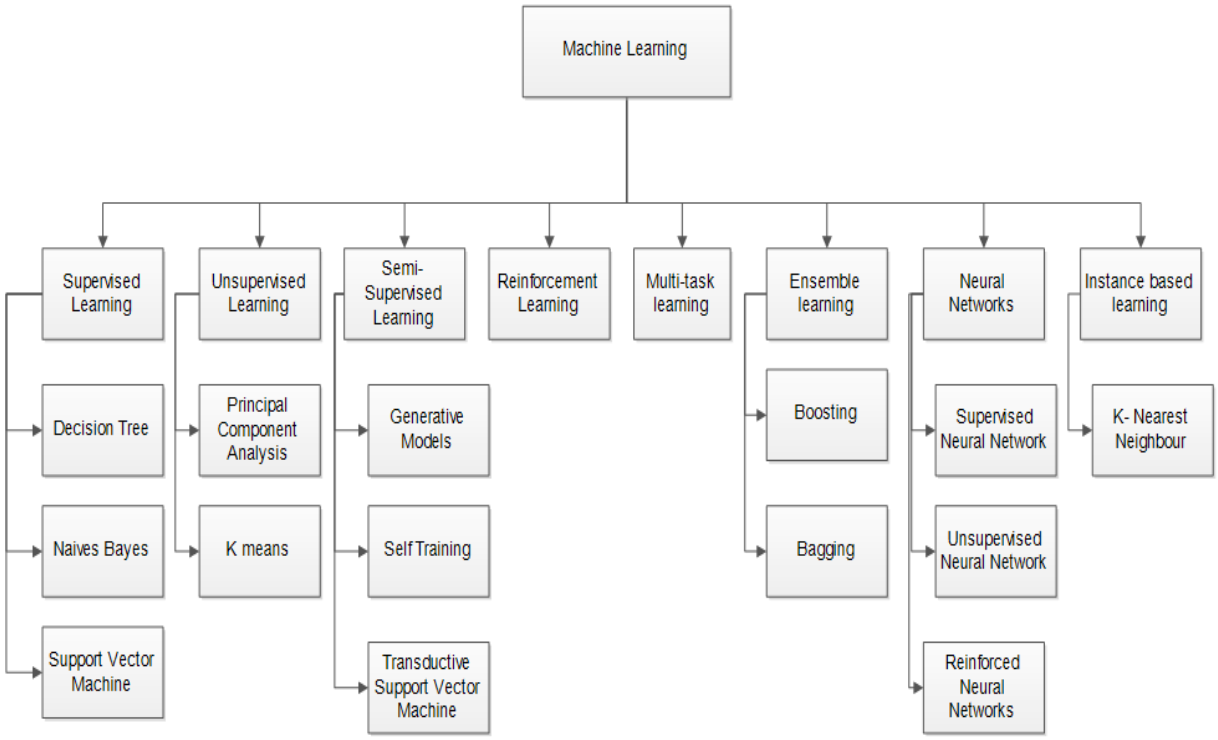

Figure 1: Types of Machine Learning (Dey, 2016)

The diagram illustrates the various types of machine learning: supervised, unsupervised, semi-supervised, reinforcement learning, multi-task learning, ensemble learning, neural networks, and instance-based learning. For this research, the premium is placed on the supervised learning technique and one borrowed concept of the instance-based learning technique, precisely the K-Nearest neighbour approach

Supervised machine learning is a machine learning technique that employs algorithms that mostly need external assistance. The dataset is categorised into two sets, namely, the training set and the testing set. Essentially, the training dataset, by its nature, has a corresponding output variable that needs to be classified or predicted. All learning algorithms can study patterns from the training set and subsequently apply them to the training dataset for prediction and classification (Kotsiantis, 2007). Figure 2 shows a diagram of supervised machine learning algorithms. This methodology motivated the adapted model that is further used in the methods of this paper.

As emphasised by (Kotsiantis, 2007), the initial step in any machine learning endeavour is the meticulous selection of the right data source. This is followed by a crucial data pre-processing phase to minimise noise and eliminate unnecessary features that could compromise data integrity in the subsequent supervised learning process. Algorithm selection ensues, followed by training based on standardised partitioning methods determined by the data scientist's discretion. Evaluation is conducted using test data, with optimal classifiers accepted, while non-ideal ones are subjected to parameter readjustments until an ideal classifier is attained.

The UCI Machine Learning Repository is a renowned repository for machine learning datasets, encompassing data generators, databases, and domain theories extensively utilised by the machine learning community for empirical analysis. It hosts the Wisconsin Breast Cancer Dataset, compiled from Fine Needle Aspirate (FNA) human breast tissue samples. This dataset comprises 699 clinical cases, with 458 (65.50%) identified as benign and 241 (34.50%) as malignant (Ahmed Medjahed et al., 2013). The dataset contains 16 missing observations, limiting the experiment to 683 clinical cases. Following

classical machine learning methodology, the performance evaluation of machine learning models necessitates partitioning the database into training and testing sets.

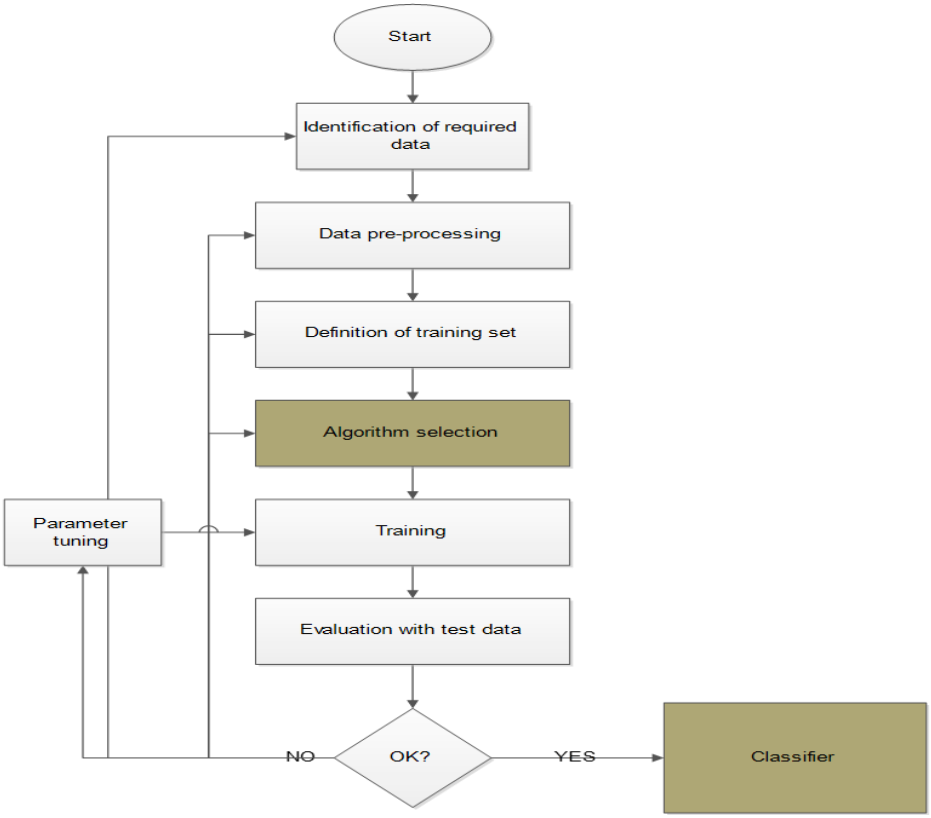

Figure 2: The Workflow of Supervised Machine Learning Algorithms (Kotsiantis, 2007)

The dataset features of the Wisconsin Breast Cancer Dataset include radius, texture, perimeter, area, smoothness, compactness, concavity, and concave points. The rise in breast cancer cases globally has spurred the development of numerous machine-learning models aimed at facilitating its detection through binary classification (Seddik & Shawky, 2015; Shravya et al., 2019; Nguyen et al., 2013; Kharya & Soni, 2016). With the advent of robust machine learning platforms, there is an opportunity to implement and evaluate the performance of diverse machine learning models, enabling more profound analysis, especially on breast cancer datasets sourced from repositories like UCI.

The surge in breast cancer incidences, coupled with the need for decentralised and accurate diagnosis, along with the evolution of machine learning paradigms, underscores the rationale behind this research endeavour. The study focuses on six machine learning models: Logistic Regression, K Nearest Neighbor, Support Vector Machine (Linear Classifier), Support Vector Machine (RBF Classifier), Gaussian Naive Bayes, Decision Tree Classifier, and Random Forest Classifier. The implementation environment of Google's cloud-based machine learning laboratory, Colab, was chosen due to its extensive machine learning and data science libraries and robust online computing power, which are essential for conducting computationally intensive machine learning implementations.

II. MODELS IN MACHINE LEARNING

The inspiration and mathematical underpinnings of the machine learning models, as considered in this paper, are highlighted in the following sub-headings.

A. Logistic Regression Model

Logistic regression fits well in situations where there is the need to create a model relationship between two sets of variables, namely, a categorical outcome variable and a set of predictor variables (Seddik & Shawky, 2015).
Mathematically, logistic regression represents a binary output Y that is expressed as:

$$Y = \pi(X) + \varepsilon \qquad (1)$$

Where (X) is a vector that has $x_{i,}\ i = 1,2,3 \ldots \ldots \ldots . . n$ independent predictors. $\pi(X)$ Represents the conditional probability of experiencing the event $Y = 1$ given the independent variable vector X with ε as a random error term. $\pi(X)$ is expressed as:

$$\pi(X) = P(Y = 1|X) = \frac{e^{X^T\beta}}{1 + e^{X^T\beta}} \quad (2)$$

β represents the model's parameter vector and alternatively could be written as:

$$In\left(\frac{\pi}{1-\pi}\right) = (\beta_0 + \beta_1 x_1 + \dots\dots\dots\dots\dots + \beta_n x_n) \quad (3)$$

The above function is called the Logit Link function. It is observed that whereas the right side is linear for β the left-hand side is not linear regarding π. Alternatively, the logit function could be represented in the odds ratio:

$$\left(\frac{\pi}{1-\pi}\right) = \frac{P(Y=1)}{P(Y=0)} = e^{X^T\beta} \quad (4)$$

The odds ratio indicates the likelihood of event Y=1 occurring. The impact of x_i, the independent variable, on the odds ratio is measured by the term e, signifying the change in the odds ratio for a one-unit increase in the independent variable x_i while holding another variable constant. A higher value of the term implies a more significant effect on the predicted probability of the resulting output. This helps in ranking predictors based on their influence on the outcome (Seddik & Shawky, 2015) and utilises the Pearson goodness-of-fit to assess the suitability of the constructed model for the observed data points. Alternatively, Deviant Statistics can also be employed to evaluate the fit quality.

B. Gaussian Naïve Bayes

Naïve Bayes forms part of a statistical classification algorithm grounded on Bayes theorem. The training stage is characterised by data point estimates of a class label using class probabilities and conditional probability. In cases of two classed datasets, data point classification is done with a premium to the higher-class probability.

The Bayes theorem calculates the posterior probability of samples in the c class using equation (5)(Güzel & Engineering, 2013). Naïve Bayes classifiers are mathematically represented as:

$$p(c|x_1, \dots\dots, x_n) = \frac{p(x_1, \dots\dots.., x_n|c)p(c)}{p(x_1, \dots\dots, x_n)} \quad 5$$

$$p(c|x_1, \dots\dots, x_n) = \frac{p(c)\Pi_{i=1}^{n} \quad p(x_i|c)}{p(x_1, \dots\dots, x_n)} \quad 6$$

The Naive Bayesian classification model is simple and assumes that the classification features are independent and have no correlation between them. However, it has become evident that the notion of attribute independence is not entirely true for all cases, and since then, there has been some innovation to increase performance (Karthika & Sairam, 2015).

The Naive Bayesian algorithm based on the description provided is given as follows:

- Let G be the training set with N tuples where each tuple is represented as a 'k' dimensional attribute vector M, where $M = \{M_1, M_2, \dots M_k\}$
- Let there be 'p' classes $C_1, C_2, \dots C_p$. According to this Naive Bayesian classifier, a tuple T belongs to the class C_x only when it has a higher conditional probability than any other class C_y, where $x \neq y$.

  $P(C_x|T) > P(C_y|\text{T})$ and $P(C_x|T) = (P(T \mid C_x * P(C_x)) / P(T)$
- Since class conditional independence is assumed,

  $$P(M \mid C_x) = \prod_{i=1}^{k} P(Mk \mid C_x) = P(M_1) * P(M_2) * P(M_3) \dots * P(M_k)$$
- Class C_x is predicted as the output class when

  $P(M \mid C_x) * P(C_x) > P(M \mid C_y) * P(C_y), where\ 1 \leq x, y \geq p\ and\ x \neq y$

C. K-Nearest Neighbour

K-Nearest Neighbour is one of the well-known and most used machine learning areas. It is an instance-based methodology that does not require a learning phase like the other models. The training sample selection is associated with the choice function, and the distance function is motivated by the nearest neighbours in the model. A similarity measure is a basis for comparison before any classification is done(AhmedMedjahed et al., 2013).

The algorithm is summarised as:

- Choose a value for the parameter k

  Input: Give a sample of N examples and their classes

  The class of a sample x is $c(x)$

  Give a new sample y
- Determine the k-nearest neighbor of y by calculating the distances

- Combine classes of these y examples in one class c
  Output: the class of y is $c(Y) = c$

The various distance measures of the KNN are City-block distance (1-distance), Euclidean distance (2-distance), Minkowski distance (p-distance), and others.

D. Support Vector Machine

Support Vector Machines (SVMs) offer a novel algorithm for data classification and regression. This allows the expansion of data provided by a training set to be expressed as a linear combination of a subset of data within the training set (Fields, 1998). SVMs develop a hyperplane that separates data by class using a training tuple called support vectors, transforming the training dataset into a higher dimension (Agarwal, 2014). During training, the convex cost function optimises without local minima, simplifying the learning process. Model evaluation involves using support vectors to classify the test dataset, with performance based on determining the error rate as the test dataset size approaches infinity. Extensive literature details SVM algorithms' mathematical formulation and underpinnings for data classification and regression (Campbell, 2002).

A critical consideration is selecting a suitable kernel function to transform non-separable data into a new feature space where they become separable. Standard kernel functions include Linear, Polynomial, Radial Basis Function, and Sigmoid (Agarwal, 2014). The SVM process, as summarised by Agarwal, involves several key steps:

1. Nonlinear mapping transforms original training data into a higher-dimensional space, known as kernelling, with kernel functions selected based on research requirements. An optimal linear separating hyperplane, the decision boundary, is sought within this transformed dimension.
2. With an appropriate nonlinear mapping to a sufficiently high dimension, a hyperplane can always separate data from two classes.
3. The SVM locates this hyperplane using support vectors (essential training tuples) and margins defined by the support vectors.

SVMs offer advantages such as accuracy in high-dimensional spaces and memory efficiency through the use of support vectors. However, they are prone to overfitting, significantly when the number of features exceeds the sample size, and do not provide probability estimates, which are often desirable in classification problems. SVMs may also exhibit inefficiencies with small datasets.

E. Random Forest

A Random Forest (RF) is a classification algorithm in Machine Learning that employs multiple decision trees. It operates as an Ensemble of Classifiers, where decision tree attributes are chosen randomly. RF is conceptualised as an ensemble technique inspired by the concept of randomised tree ensembles. The core unit of an RF is a binary tree formed through recursive partitioning. The construction of an RF, as delineated by (Nguyen et al., 2013), involves the following steps:

1. Generate n-tree bootstrap samples from the original dataset.
2. Construct a tree for each bootstrap dataset. At each tree node, randomly select and try variables for splitting. Grow the tree to ensure each terminal node contains fewer cases than the specified node size.
3. Aggregate information from the n trees to predict new data, such as employing majority voting for classification.
4. Compute an out-of-bag (OOB) error rate using the data excluded from the bootstrap sample.

F. Decision Tree

Decision Tree classifiers are integral to supervised classification methodologies (Zhao & Zhang, 2007). They draw inspiration from the structure of a typical tree, comprising roots, nodes, branches, and leaves, with its construction primarily centred on nodes (Ali et al., 2012).

The application of Decision Tree classifiers has notably advanced pattern recognition in classification tasks. Early research focused on character recognition and image classification (Safavian & Landgrebe, 1991). Decision Trees are superior in handling complex classification problems due to their adaptability and computational efficiency.

The underlying principle of a decision tree is to delineate all potential decision paths in a tree-like structure. The fundamental steps involved in constructing a decision tree include:

1. Selecting an attribute from the dataset.
2. Assessing the significance of the attribute in data partitioning.
3. Partitioning the data based on the value of the selected attribute.
4. Iterating the process from step 1 to further refine the decision tree.

III. TECHNICAL PERFORMANCE ANALYSIS AND JUSTIFICATION

The underlying methodology of this paper is an adaptation of supervised learning (Kotsiantis, 2007). Figure 3 summarises the step-by-step approach to the experiment.

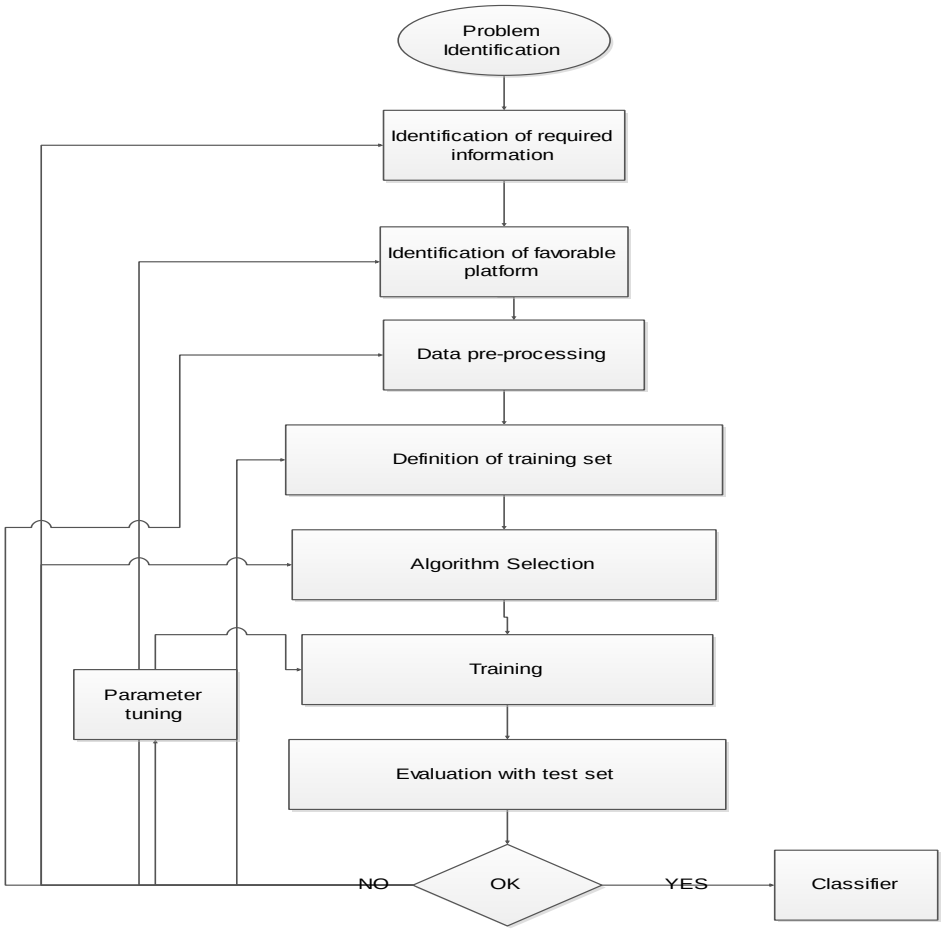

Figure 3: An Adaptation of the Supervised Machine Learning Model on WBCD (Kotsiantis, 2007)

As previously mentioned, the other processes remain consistent with the original flowchart apart from platform identification. Machine learning researchers utilise various evaluation metrics to assess the performance of machine learning model experiments on the designated dataset. This paper gives particular attention to the Classification Report provided by the implementation platform.

The Scikit-learn library offers a range of convenient reporting tools tailored for classification problems, providing insights into the model's accuracy across multiple measures. The classification_report() function available on the Google Colab platform utilises a set of algorithms to return specific metrics for evaluating machine learning models. These metrics include:

- Precision: This indicates the percentage of correct predictions, focusing on the accuracy of positive predictions.
- Recall: Representing the ratio of true positives to the sum of true positives and false negatives, recall identifies the correct percentage of positives.
- F1-score: The F1 score is a weighted harmonic mean of precision and recall, with a perfect score of 1.0 and a minimum score of 0.0. F1 scores incorporate precision and recall into their calculation, resulting in values lower than accuracy measures.
- Support: Support refers to the number of occurrences of each class in the specified dataset.

IV. RESULTS AND DISCUSSION

After platform identification and data preprocessing, which was characterised by eliminating null and unnecessary attributes, a random partitioning of the dataset into 75% for training and 25% for testing was agreed upon. It should be noted that all data preparation processes, including the labelling of categorical and dependent attributes, were done. The attribute correlation was also considered to enhance understanding of the feature attributes. The training accuracies of the dataset after training are summarised in the bar graph in Figure 3.

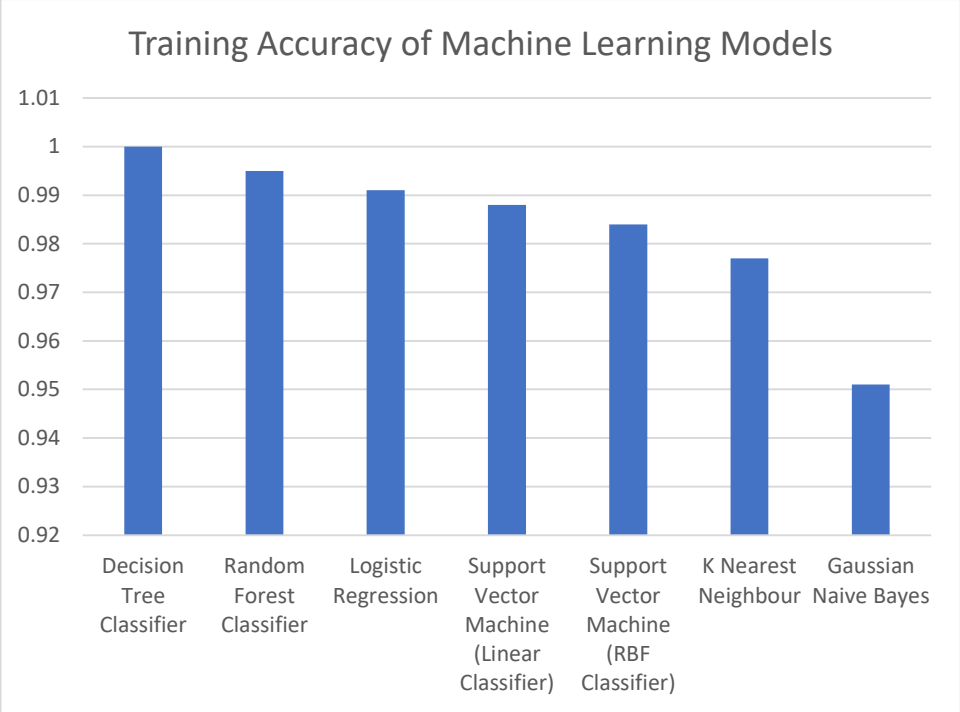

Figure 4: Training Accuracy of Machine Learning Models

It became extremely evident that the Decision Tree Classifier (DTC) obtained a perfect score of 1.00, which ideally confirms the superb performance of DTC on relatively more minor datasets like the WBCD. Following closely in terms of training performance are the random forest classifier, logistic regression, and the two variations of SVM, specifically the linear kernel and the RBF kernel. The Classification report during the testing phase generated the following outcomes of interest, as illustrated in Table 1 below:

Tab 1: Classification Report Summary of the Machine Learning Models

Machine learning Model		Precision	Recall	F1-score	Support
Logistic Regression	0	0.96	0.96	0.96	90
	1	0.92	0.92	0.92	53
K Nearest Neighbour	0	0.95	0.99	0.97	90
	1	0.98	0.91	0.94	53
SVM (Linear Classifier)	0	0.98	0.97	0.97	90
	1	0.94	0.96	0.95	53
SVM (RBF Classifier)	0	0.97	0.98	0.97	90
	1	0.96	0.94	0.95	53
Gaussian Naive Bayes	0	0.93	0.94	0.94	90
	1	0.90	0.89	0.90	53
Decision Tree Classifier	0	0.99	0.93	0.96	90
	1	0.90	0.98	0.94	53
Random Forest Classifier	0	0.98	0.97	0.97	90
	1	0.94	0.96	0.95	53

It should be noted that the labelling of the dependent attributes transformed Malignant to a 1 value and benign tissues to a zero value. Table 1 gives the classification report's precision, recall, and f1-score values. It is observed that RFC and the two variations of the SVM (RBF and linear classifier) had the best averages in terms of precision, recall, and f1-score in the region of 0.96, emphasising its impressive performance in the testing phase. KNN had a fantastic average precision of 0.965 and recorded lower scores for recall and F1-score. The Decision Tree Classifier was similar to logistic regression for average recall and f1-score values. The Gaussian Naïve Bayesian model recorded the lowest performance in terms of averages for precision, recall, and f1-score. The testing phase recorded the following accuracies, as shown in the graph in Figure 4.

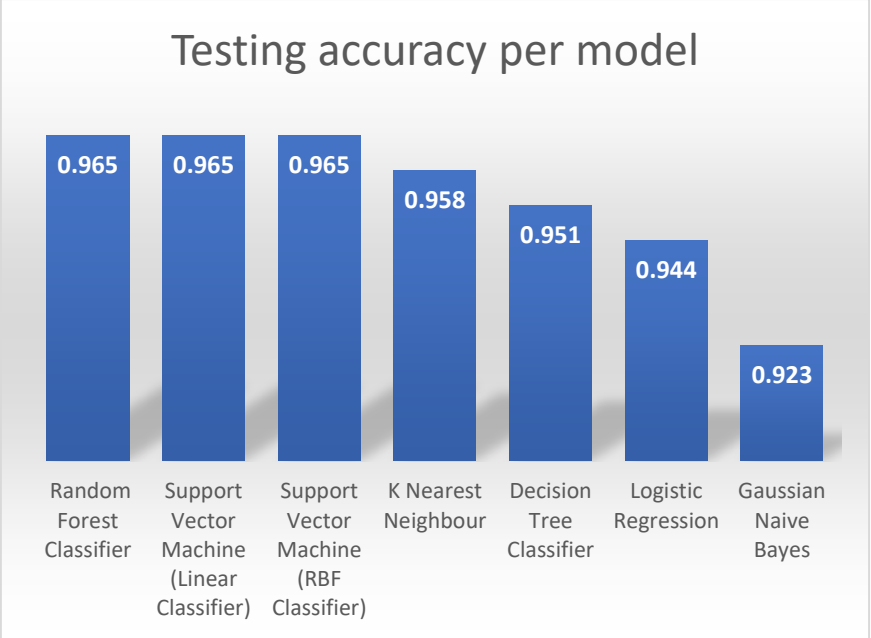

Figure 4: Testing Accuracy per Model

It became evident that unlike the testing phase, where the decision tree had a perfect score, a very different drift was seen in the outcome of the testing. It is observed that the Random Forest Classifier and the two variations of the Support Vector Machine had the most impressive performance, with a classification accuracy of 0.965. KNN also gave an impressive showing during the testing phase of the experiment and ranked 2nd after RFC and SVM. This is after ranking 6th in training. The Decision Tree classifier also ranked 3rd in testing after ranking 1st in the training phase of the research. It is generally observed the Gaussian Naïve Bayes model does not do too well in both phases of the experiment, ranking last in both scenarios.

V. MODEL CHALLENGES AND DEFICIENCIES

The performances of these machine learning models must be put in the proper perspective. Even though there were quite impressive accuracies in some models, it must be noted that the application area needs a predictive model that guarantees prediction or diagnosis with the most minimal margin of error. This has inspired the development of ensemble models that are carefully engineered to have better and more accurate classification accuracy along with various evaluation metrics.

As established in this paper, some challenges of the machine learning models include the following.

1. Decision Tree algorithms are generally inadequate in the application of regression and prediction of continuous values
2. Random forest is generally noted to have slower prediction rates, which may trigger challenges in real-time applications
3. The decision on which type of distance to use and which attribute to employ to obtain better results is unclear in KNN implementation.
4. One known disadvantage of Logistic regression is that it struggles with its restrictive expressiveness, and due to this, other models may have better performances.
5. One widespread disadvantage of Gaussian Naive Bayes classifiers is the apparent assumption of independence of class features, which makes it nearly impossible to find a data set of that kind.

VI. CONCLUSION

The diverse array of machine learning models, coupled with the expanding sources of datasets, underscores the importance of comprehending model behaviour. Disease detection and diagnosis, within the realm of machine learning and artificial intelligence, are pivotal for achieving swift and reliable diagnoses in the future. Using a pertinent evaluation metric, known as the Classification report, notable testing performances were observed in the Random Forest Classifier and the SVM (RBF and Linear kernels). These models exhibited commendable precision, recall, and f1-score values when applied to the Wisconsin Breast Cancer Dataset from the UCI repository. The aggregate performance of the machine learning models underscores the imperative to develop more hybrid and intelligent models to enhance the efficiency of these fundamental algorithms, including ensemble techniques. Additionally, conducting similar tests on larger datasets could yield valuable

insights into the performance of these models. These interventions, among others, are crucial additions to ongoing efforts to achieve optimal efficiency and classification accuracy of machine learning models.

VII. REFERENCES

[1] Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

[2] AhmedMedjahed, S., Ait Saadi, T., & Benyettou, A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, *62*(1), 1–5. https://doi.org/10.5120/10041-4635

[3] Al-Hadidi, M. R., Alarabeyyat, A., & Alhanahnah, M. (2017). Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm. *Proceedings - 2016 9th International Conference on Developments in ESystems Engineering, DeSE 2016*, *August*, 35–39. https://doi.org/10.1109/DeSE.2016.8

[4] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278.

[5] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*(Fams), 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

[6] Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, *12*(5), 526–558. https://doi.org/10.1007/s10791-008-9070-z

[7] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery 2, 121-167, 1998*, *2*, 121–167. https://doi.org/10.1111/sms.12977

[8] Campbell, C. (2002). Kernel methods: A survey of current techniques. *Neurocomputing*, *48*(1–4), 63–84. https://doi.org/10.1016/S0925-2312(01)00643-9

[9] Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com

[10] Edriss, E., Ali, E., & Feng, W. Z. (2016). Breast Cancer Classification using Support Vector Machine and Neural Network. *International Journal of Science and Research (IJSR)*, *5*(3), 1–6. https://doi.org/10.21275/v5i3.nov161719

[11] Güzel, C., & Engineering, F. (2013). Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation. *AWERProcedia Information Technology & Computer Science*, *04*(May), 401–407.

[12] Houts, P. S., Lenhard, R. E., & Varricchio, C. (2000). ACS cancer facts and figures. *Cancer Practice*, *8*(3), 105–108. https://doi.org/10.1046/j.1523-5394.2000.83001.x

[13] Karthika, S., & Sairam, N. (2015). A Naïve Bayesian classifier for educational qualification. *Indian Journal of Science and Technology*, *8*(16). https://doi.org/10.17485/ijst/2015/v8i16/62055

[14] Kaur, P., Sharma, M., & Mittal, M. (2018). Big Data and Machine Learning Based Secure Healthcare Framework. *Procedia Computer Science*, *132*, 1049–1059. https://doi.org/10.1016/j.procs.2018.05.020

[15] Kharya, S., & Soni, S. (2016). Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection. *International Journal of Computer Applications*, *133*(9), 32–37. https://doi.org/10.5120/ijca2016908023

[16] Kotsiantis, S. B. (2007). *Supervised Machine Learning : A Review of Classification Techniques*. *31*, 249–268.

[17] Loussaief, S., & Abdelkrim, A. (2017). Machine learning framework for image classification. *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2016*, *3*(1), 58–61. https://doi.org/10.1109/SETIT.2016.7939841

[18] Max, W. (2011). *A First Encounter with Machine Learning*. *11*(1), 24–32. https://doi.org/10.1145/134304.134306

[19] Naku Ghartey Jnr, F., Anyanful, A., Eliason, S., Mohammed Adamu, S., & Debrah, S. (2016). Pattern of Breast Cancer Distribution in Ghana: A Survey to Enhance Early Detection, Diagnosis, and Treatment. *International Journal of Breast Cancer*, *2016*. https://doi.org/10.1155/2016/3645308

[20] Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *06*(05), 551–560. https://doi.org/10.4236/jbise.2013.65070

[21] Ozgur, C., Kleckner, M., & Li, Y. (2015). Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015584379

[22] Safavian, S. R., & Landgrebe, D. (1991). A Survey of Decision Wee Classifier Methodology. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, 21*(3).

[23] Seddik, A. F., & Shawky, D. M. (2015). Logistic regression model for breast cancer automatic diagnosis. *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, 150–154. https://doi.org/10.1109/IntelliSys.2015.7361138

[24] Shravya, C. H., Pravalika, K., & Subhani, S. (2019). Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6), 1106–1110.

[25] Suleymanov, U., & Rustamov, S. (2018). Automated News Categorization using Machine Learning methods. *IOP Conference Series: Materials Science and Engineering*, *459*(1). https://doi.org/10.1088/1757-899X/459/1/012006

[26] Walker, A. R. P., Adam, F. I., & Walker, B. F. (2004). Breast cancer in black African women: A changing situation. *Journal of The Royal Society for the Promotion of Health*, *124*(2), 81–85. https://doi.org/10.1177/146642400412400212

[27] Wang, H., Ma, C., & Zhou, L. (2009). A brief review of machine learning and its application. *Proceedings - 2009 International Conference on Information Engineering and Computer Science, ICIECS 2009*. https://doi.org/10.1109/ICIECS.2009.5362936

[28] Zhao, Y., & Zhang, Y. (2007). Comparison of decision tree methods for finding active objects. *Advances of Space Research*.