A Comparative and Predictive Analysis of Prostate Cancer Diagnosis and Treatment using Decision Tree, Neural Network, Support Vector Machine, Random Forest and K-Nearest Neighbor KNN Classification Algorithms

OGUOMA IKECHUKWU STANLEY[1], UKA KANAYO KIZITO[2], VICTORY CHIBUIKE ONUMAKU[3], ONU-NJOKU CHARLES ENYINNAYA[4], NNEDINMA CHRISTIANA NJOKU[5], CHUKWU ALPHONSUS CHEKWUBE[6]

[1] *Department of Computer Science, University of Agriculture and Environmental Science (UAES) Umuagwo, Nigeria*

[2]*Faculty of Physical Sciences, Department of Computer Science, Imo State University, Owerri, Nigeria*

[3] *Edge Hill University, Ormskirk, United Kingdom*

[4, 6] *Faculty of Medicine, Health and Life Sciences, Swansea University, Wales*

[5] *Faculty of Health, Medicine and Social Care, Anglia Ruskin University-Chelmsford, England.*

Abstract- ***Prostate cancer is an illness majorly found on men between age of 50 and above. It begins when the healthy cells in the prostate gland change and grow out of control, and grow a mass called tumor which might affect any part of the body whereby could become cancer cells, and then spread to other parts of the body. This study is centered on a comparative predictive analysis of prostate cancer diagnosis and treatment using Decision Tree, Neural Network, Support Vector, Random Forest and K-Nearest Neighbor KNN Classification Algorithms. The study tends to achieve the following objectives: are to design a more accurate and intelligent model for easy identification and diagnosis of prostate cancer disease on patients', to compare the accuracy of the results produced between the Five (5) algorithms in other to proffer a more long-lasting solution for prostate cancer disease prediction and hence lower mortality rate amongst patients. The research adopted a data mining methodology called classification algorithm by following the SEMMA (sample Explore modify model Access) approach while employing Five (5) machine learning algorithm as the modeling tool. The experiment on the collected prostate cancer dataset was analyzed with R Programming language while using JASP IDE for the experiment sourced from UCI machine learning repository. The result was able to implement a model that could easily and accurately predict the presence of prostate cancer in men efficiently and effectively with Decision Tree 80% test accuracy, Neural Network Algorithm 90% test accuracy, Support Vector Machine (SVM) 75% test accuracy, Random Forest Algorithm 80% test accuracy and Out-of-Bag (OOB) accuracy of 90% and K-Nearest Neighbors (KNN) algorithm 90%. Base on the comparison analysis conducted by this study, it was observed that Neural Network Algorithm and K-Nearest Neighbors (KNN) have the highest percentage accuracy towards the prediction of prostate cancer disease having 90% test accuracy each with KNN 1% validation accuracy. This research was able to show clearly how prostate cancer disease could be managed using prediction models on the tested 80% trained dataset on the various algorithms used for the experiment.***

Indexed Terms- ***Artificial Intelligence, Machine Learning, Prostate Cancer, Decision Tree, K-Nearest Neighbor, Vector Machine, Random Forest and Neural Network Algorithms***

I. INTRODUCTION

Men over the age of 50 are most commonly affected with prostate cancer. It starts when the prostate gland's healthy cells start to grow and change out of control, giving rise to a lump known as a tumor. Cancerous cells can develop from cells in almost any part of the body and spread to other areas. Because of this, medical professionals advise all males to get screened as soon as they approach or reach that age range in order to detect cancer early and treat it effectively. However, with the development of artificial intelligence, a more effective method of cancer detection and diagnosis has become feasible. Worldwide, especially in low-income nations like Nigeria, the use of artificial intelligence in the medical field and its practices has demonstrated a quick improvement in the treatment of illness. According to [1], in the diagnosis and treatment of illnesses associated to the disease, Cancer is a chronic illness that has drawn attention from all around the world in part due to its devastating effects and the vast amount of time, money, and human resources devoted to finding a long-term cure for this scourging scourge. A wider scope that catches all the relevant characteristics is needed because several medical trials for medications that were supposed to be able to cure the disease have failed at the last stage of testing. This is likely due to the initial data not being as comprehensive as supporters believed. However, a number of studies have demonstrated that the continuously rising death toll is not solely attributable to a single risk factor, but rather to a variety of factors that have been linked to an increased risk of the illness, including race, heredity, exposure to UV radiation, type of job, diet, and body mass index (BMI). The word 'Cancer' gets its name from the Latin word for crab because, like crabs, tumors can have very asymmetrical shapes and "grab on and do not let go." A new development that has the potential to infect nearby tissues, metastasis (spread to other organs), and may result in the patient's death is particularly referred to as cancer [2]. Cancers typically begin as primary tumors in one organ before spreading to other areas of the body. Prostate cancer, commonly known as carcinoma of the prostate in medical terminology, is one example of this [3]. Nevertheless, benign form of prostate (BNH) is treated with medication or transurethral surgery which happens to be a prostate as an organ of the male reproductive system that secretes the fluid that nourishes and protects the sperm cells in the seminal fluid (semen). It is situated directly in front of the rectum and below the bladder [4]. This can be distinguished from other disorders of the male reproductive system with the use of appropriate testing methods such serum proteome profiling, prostatic specific antigen testing, tumor markers, prostate imaging, and biopsies, which are medical procedures that involve removing tissues or cells for analysis.

Even though there are a number of testing methods for an accurate detection and diagnosis of prostate cancer, such as measuring the level of Prostate Specific Antigen (PSA) in blood or performing a Digital Rectal Exam (DRE), in which a doctor inserts a gloved finger into the rectum and feels the prostate for any abnormalities, Nigerian men have spoken out about the difficulties and health issues they face as a result of this deadly disease. Additional screening is carried out if the results of these tests are positive. In this regard, a number of health-related problems have been addressed by artificial intelligence (AI), which has resulted in the development of an accurate model that depicts the survival rate [5]. As a result, scientists and researchers have discovered the many benefits of data mining as a tool for data extraction, data prediction, and data discovery in problem solving. Data mining (DM) has been one of the underutilized big database extraction techniques in data prediction because of its enormous potential to help businesses seeking large-scale data discovery and knowledge warehouses globally [6]. Mining information from data has contributed to making it easier to undertake more research in the health sector, which has opened the door for the use of more new technologies in the fight against disease. The aim of this study is to involve a comparative predictive analysis of prostate Cancer disease using decision tree, K-Nearest Neighbor, vector machine and neural network algorithms to better understand the best predictive model and uncover a more accurate model for the diagnosis and treatment of prostate cancer in men while the researcher's objectives are to design a more accurate and intelligent model for easy identification and diagnosis of prostate cancer disease on patients', to compare the accuracy of the results produced between the Five (5) algorithms in other to proffer a more long-lasting solution for prostate cancer disease prediction

and hence lower mortality rate amongst patients. The study made use of an AI modeling approaches to uncover or create the models after a critical data analysis is done on the dataset. The structure of this document is as follows: The introduction provides a general overview of machine learning (ML), defines prostate cancer, and outlines the various types of prostate cancer test techniques. It also highlights the study's objectives and provides important information on the benefits of using data mining and machine learning tools in decision-making especially in health related issues. Literature Review: looks at generally the literature review on related works, machine learning modeling tools and technique, Methodology/Analysis: shows the adopted methodology for the study, analysis of the existing system, proposed system diagram, system algorithm while Results: present analysis using R Programming language while using JASP and RStudio as the IDE for the experiment output, the model rules in the Four (4) classification models, critical Comparative analysis of the results produced, conclusion and recommendation of the study.

II. RELATED LITERATURE REVIEW

According to a study by [16] presented on advancing prostate cancer detection through a comparative analysis of two classifiers (PCLDA-SVM and PCLDA-KNN) for an enhanced diagnostic accuracy on the prostate cancer dataset sourced from National Cancer Institute's Cancer Data Access System and from their comparative analysis between the two algorithms shows a promising output on the accurate diagnosis and treatment of prostate cancer with PCLDA-SVM having an accuracy of 97.99%, with a precision of 0.92, sensitivity of 92.83% and with low error rate of 0.016 while PCLDA-KNN have an accuracy rate of 97.8%, precision of 0.93, sensitivity of 93.39% and an error rate of 0.006 respectively. The analysis of their study employed only two algorithms as a major comparison factor which is not a more clear point to present the best predictive algorithms for a more accurate prostate cancer prediction model. [7] While working on a review of the literature on the prevalence of prostate cancer in Germany, the study used the databases from PubMed, EMBASE, and the Cochrane Library to find papers that discussed the costs, health state utilities, and incidence and/or mortality rates of prostate cancer in the relevant settings. The study's conclusions demonstrated that during the previous 20 years, the incidence of prostate cancer has significantly increased in all settings. This increase has been partially attributed to a rise in the use of prostate specific antigen (PSA) screening, which has allowed for earlier tumor detection but has also increased overtreatment, which has increased the financial burden of the disease. Due to advancements in therapy and earlier discovery, mortality rates have decreased during this time. [8] conducted a study using three databases to examine prostate care and cancer from the viewpoints of men who have not received a diagnosis. The study's findings revealed that men frequently lack knowledge about screening, the anatomy of the prostate, or their risk of developing prostate cancer, and that concerns about being a man may discourage men from getting a checkup. According to a study by [9], the urban population of Nigeria is remarkably ignorant about prostate cancer. The majority of them which is, serum PSA testing and prostate cancer screening are unknown worldwide. Moreover, [10] discovered that 81.5% of them agreed to get screened for the illness by using that approach while a study by [11] conducted a systematic literature analysis and thematic synthesis of Black African and Black Caribbean men's post-treatment perspectives to focus on life after prostate cancer treatment. The authors concluded that that there are notable ethnic differences in prostate cancer prognosis and occurrence on their study on the utilization of prostate-specific antigen analysis for the early identification of prostate cancer. According to an empirical assessment conducted among urologists and general practitioners, prostate cancer (PCa) is the most common malignancy and the third largest cause of cancer-related death in males in Germany. [17] Carried a study on the prediction of prostate cancer using machine learning algorithms such as K-Nearest Neighbors, Support Vector Machines, Logistic Regression, Naive Bayes and Random Forest. When the Confusion Matrix of Various Algorithms was analyzed, logistic regression and random forest shows an accuracy of 70% and 90% on the prostate cancer dataset sourced from kaggle and the study could not involve decision tree algorithm in their analysis as decision tree helps for more clear and understandable analysis definition and accurate application by scientist in health sector for better treatment and diagnosis of diseases. Furthermore, [15]

worked on the comparison between K-Nearest Neighbor (KNN) and Decision Tree (DT) Classifier for Glandular Components. Their research was able to present a performance of the two algorithms which shows KNN as a better classifier than that of decision tree (DT) with an accuracy of 86.67%, sensitivity and specificity of 100% on both algorithms. The limitation of their study focuses only on the comparison of the KNN and DT and hence could not be used to draw conclusion on the best classification algorithm for the prediction of prostate cancer health issues. Methodology/Analysis: The study employed the SEMMA_DM methodological approach to achieve the proposed model while the analysis was done using a dataset extracted from the kaggle.com dataset repository.

III. ADOPTED ALGORITHMS

To produce the model, a supervised learning method was applied called classification algorithm which as stated by [13] as a learning approach where computers are programmed to learn from set of inputs as data and they systems uses the data to classify new observations through any applied algorithms in creating the model. The comparative algorithms employed for the study includes: Decision Tree, Neural Network, Support Vector Machine, Random Forest and K-Nearest Neighbor KNN Classification Algorithms and are all explained below:

- Decision Tree: As stated by [14], decision tree from the classification algorithm uses a classification or regression models to form a tree structure. The importance of the tree structure is to break down and provide a more detailed explanation of the developed model. The tree structure as stated by [14] represents the result of the experiment through the nodes and leaf nodes, where the decision nodes has two or more different branches while leaf nodes shows the classification or decision results.
- Neural Network: These classification algorithms which are often referred to as simulated neural networks (SNNs) or artificial neural networks (ANNs) which are a subset of machine learning. Their nomenclature and organization are derived from the human brain, emulating the communication patterns of actual neurons. Neural networks consist of a node layer with an output layer, an input layer, and one or more hidden layers. Every node, or artificial neuron, has a weight and thresholds that are connected to other nodes. Logistic
- Regression: The method of predicting the likely outcome of a discrete result given an input variable is known as logistic regression. A binary result, or something that can have two values, such as true or false, yes or no, and so on, is what most logistic regression models represent.
- Naïve Bayes: The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.
- Nearest neighbor (KNN): as a form of proximity search, is the optimization problem of finding the point in a given set that is closest (or most similar) to a given point. Closeness is typically expressed in terms of a dissimilarity function: the less similar the objects, the larger the function values.
- Support Vector Machine (SVM) Algorithm: SVM is an effective supervised method that performs best on complex but smaller datasets. Although Support Vector Machines, often known as SVMs, are useful for classification as well as regression applications, their performance is generally greatest in the former two major examples of SVM are (Linear SVM and Non-Linear SVM).
- Random Forest algorithm: A well-liked supervised machine learning approach for classification and regression issues in machine learning is the Random Forest approach. As we all know, a forest is made up of many trees, and the more trees it has, the more robust it is.

3.1.1 Analysis of the Proposed System on the Algorithms

In other for accurate analysis of the applied algorithms for better comparison of the prostate cancer diagnosis and treatment, a technique was developed shown in figure 1 below:

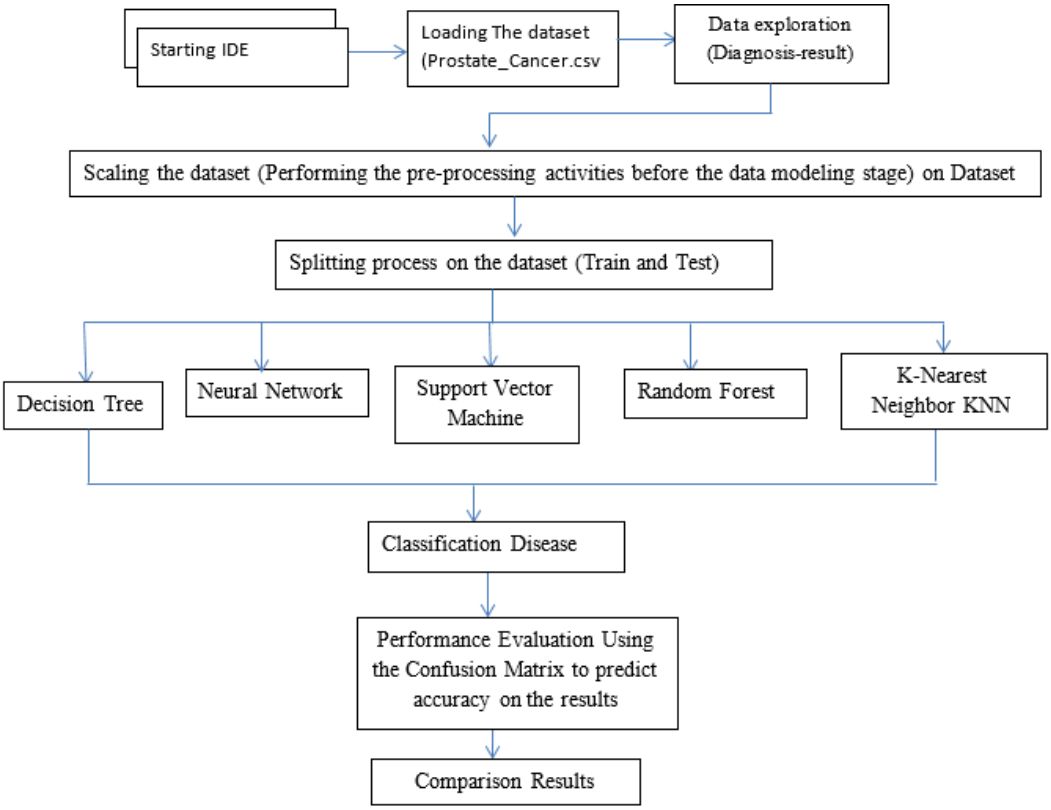

Figure 1: Analysis of the Proposed System on the Algorithms (Source: fieldwork 2023)

The above diagram illustrates how the proposed system loads the prostate cancer dataset before the processing on the dataset begins. After a successful dataset loading, the data exploration stage begins, the diagnosis on the dataset was done which shows a result through the Boxplot (malignant (labelled M)) and Boxplot (Benign tumor (labelled B). After the diagnosis was achieved, next was to processes with the Splitting process on the dataset which is (Train and Test) before data modelling was done. The experiment adopted more than five (5) algorithms which includes: Decision Tree, Random Forest, Support Vector Machine, KNN and Logistic Regression which helped in the prediction of the developed model which was accurately (Performance Evaluation) accessed by using the Confusion Matrix to predict accuracy on the results and hence compare the output of the experiments.

3.1.2 Dataset Overview/ Attribute Information

	id	diagnosis_result	radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
1	1	M	23	12	151	954	0.143	0.278	0.242	0.079
2	2	B	9	13	133	1326	0.143	0.079	0.181	0.057
3	3	M	21	27	130	1203	0.125	0.16	0.207	0.06
4	4	M	14	16	78	386	0.07	0.284	0.26	0.097
5	5	M	9	19	135	1297	0.141	0.133	0.181	0.059
6	6	B	25	25	83	477	0.128	0.17	0.209	0.076
7	7	M	16	26	120	1040	0.095	0.109	0.179	0.057
8	8	M	15	18	90	578	0.119	0.165	0.22	0.075
9	9	M	19	24	88	520	0.127	0.193	0.235	0.074
10	10	M	25	11	84	476	0.119	0.24	0.203	0.082
11	11	M	24	21	103	798	0.082	0.067	0.153	0.057
12	12	M	17	15	104	781	0.097	0.129	0.184	0.061
13	13	B	14	15	132	1123	0.097	0.246	0.24	0.078
14	14	M	12	22	104	783	0.084	0.1	0.185	0.053
15	15	M	12	13	94	578	0.113	0.229	0.207	0.077
16	16	M	22	19	97	659	0.114	0.16	0.23	0.071
17	17	M	10	16	95	685	0.099	0.072	0.159	0.059
18	18	M	15	14	108	799	0.117	0.202	0.216	0.074
19	19	M	20	14	130	1260	0.098	0.103	0.158	0.054
20	20	B	17	11	87	566	0.098	0.081	0.189	0.058
21	21	B	16	14	86	520	0.108	0.127	0.197	0.068

Figure 2: Prostate Cancer Dataset Overview

3.1.2.1 Attribute Information of the dataset

The dataset contains a total of 100 observations (rows) and 9 variables (10 columns), in which "Out-Come" is the dependent variable and other 8 variable are independent variable. The variables in the dataset are: Diagnosis_result: Shows the level of the cancer by indicating (B or M), Radius: Radius size (Rating level) of the Prostate, Texture: Indicating the texture size of the prostate, Perimeter: prostate perimeter size, Area: Area level of the prostate, Smoothness: Shows how smooth the prostate is displayed, Compactness: How compacted the prostate is in the patient's body, Symmetry: Indicate by showing the symmetric size of the prostate, fractal_dimension: This column shows the fractional dimension of the prostate

3.1.3 Splitting Method

The data splitting approach employed on the prostate cancer dataset was done in all the five (5) algorithms with a ratio of 80% train and 20% test shown in figure 2 below.

- Data Split Method

Figure 2 showing both the train and test data splitting

3.2 Experiments using the prostate cancer dataset on the Decision Tree, Neural Network, Support Vector Machine, Random Forest and K-Nearest Neighbor KNN Classification Algorithms

3.2.1 Decision Tree Classification Algorithm Result

Table 1: Decision Tree Classification Splitting Approach

Splits	n(Train)	n(Test)	Test Accuracy
20	80	20	0.800

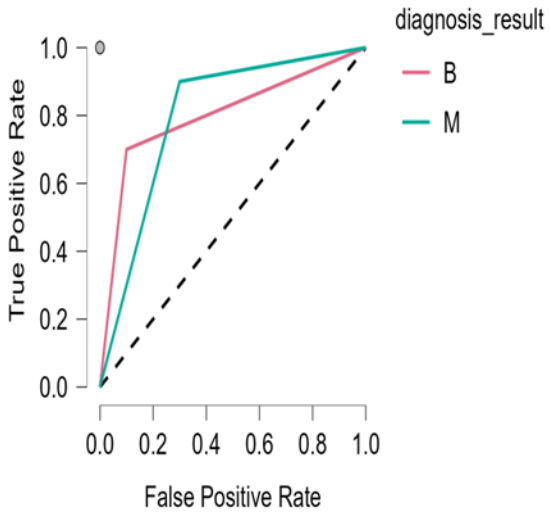

Figure 3: ROC Curves Plot

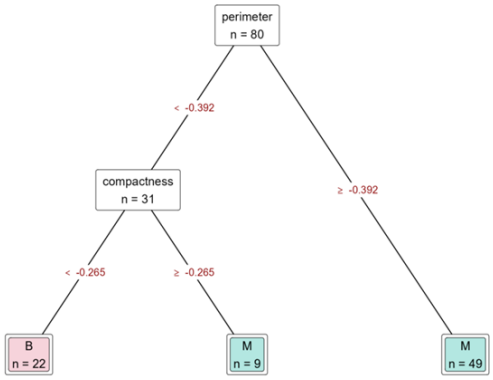

Figure 4: Decision Tree Plot

Table 2: Confusion Matrix

		Predicted	
		B	M
Observed	B	0.35	0.15
	M	0.05	0.45

Table 3: Class Proportions

Dataset	Training Set	Test Set
B	0.380	0.350	0.500
M	0.620	0.650	0.500

Table 4: Evaluation Metrics

	B	M	Average / Total
Support	10	10	20
Accuracy	0.800	0.800	0.800
Precision (Positive Predictive Value)	0.875	0.750	0.813
Recall (True Positive Rate)	0.700	0.900	0.800
False Positive Rate	0.100	0.300	0.200
False Discovery Rate	0.125	0.250	0.188
F1 Score	0.778	0.818	0.798
Matthews Correlation Coefficient	0.612	0.612	0.612
Area Under Curve (AUC)	0.800	0.800	0.800
Negative Predictive Value	0.750	0.875	0.813
True Negative Rate	0.900	0.700	0.800
False Negative Rate	0.300	0.100	0.200
False Omission Rate	0.250	0.125	0.188

Table 4: Evaluation Metrics

	B	M	Average / Total
Threat Score	1.400	1.286	1.343
Statistical Parity	0.400	0.600	1.000

Note. All metrics are calculated for every class against all other classes.

Table 5: Feature Importance

	Relative Importance
perimeter	28.014
area	26.207
compactness	19.778
smoothness	9.192
radius	6.816
fractal_dimension	6.558
symmetry	2.943
texture	0.491

3.2.2 Neural Network Classification Algorithm Result

Table 6: Neural Network Classification

Hidden Layers	Nodes	n(Train)	n(Test)	Test Accuracy
1	1	80	20	0.900

Note. The model is optimized with respect to the *sum of squares* .

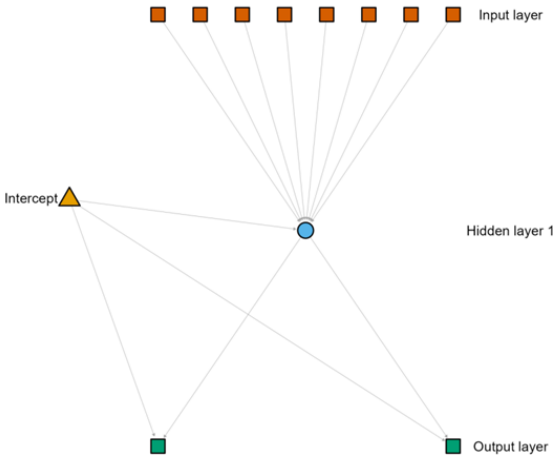

Figure 5: Network Structure Plot

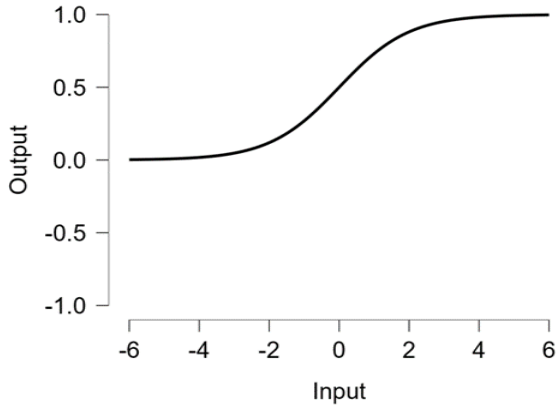

Figure 6: Logistic Sigmoid Activation Function

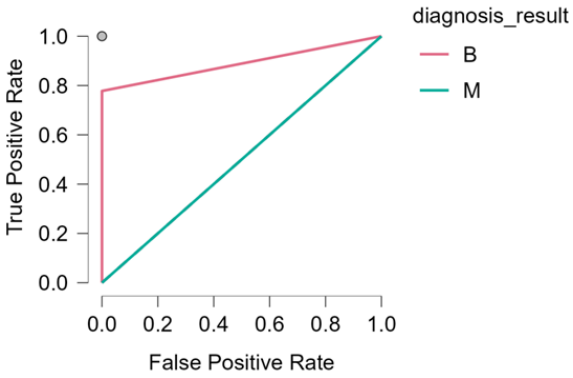

Figure 7: ROC Curves Plot

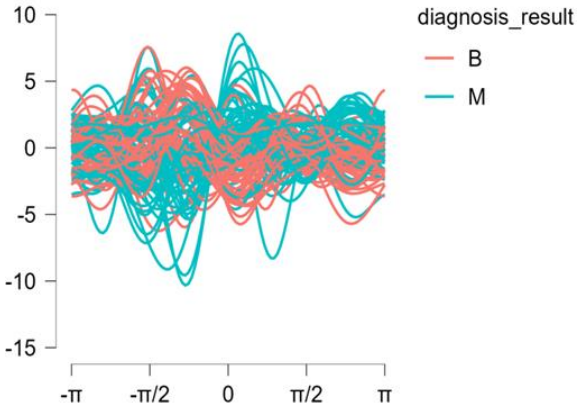

Figure 8: Andrews Curves Plot

Table 7: Evaluation Metrics

	B	M	Average / Total
Support	9	11	20
Accuracy	0.900	0.900	0.900
Precision (Positive Predictive Value)	1.000	0.846	0.915
Recall (True Positive Rate)	0.778	1.000	0.900
False Positive Rate	0.000	0.222	0.111
False Discovery Rate	0.000	0.154	0.077
F1 Score	0.875	0.917	0.898
Matthews Correlation Coefficient	0.811	0.811	0.811
Area Under Curve (AUC)	0.944	0.889	0.917
Negative Predictive Value	0.846	1.000	0.923
True Negative Rate	1.000	0.778	0.889
False Negative Rate	0.222	0.000	0.111
False Omission Rate	0.154	0.000	0.077
Threat Score	3.500	2.750	3.125
Statistical Parity	0.350	0.650	1.000

Note. All metrics are calculated for every class against all other classes.

Table 8: Network Weights

Node	Layer		Node	Layer	Weight
Intercept		→	Hidden 1	1	-3.192
texture	input	→	Hidden 1	1	-0.520
radius	input	→	Hidden 1	1	0.501
perimeter	input	→	Hidden 1	1	-1.799
area	input	→	Hidden 1	1	-2.398
smoothness	input	→	Hidden 1	1	-0.877
compactness	input	→	Hidden 1	1	-2.116
symmetry	input	→	Hidden 1	1	-0.374
fractal_dimension	input	→	Hidden 1	1	0.194
Intercept		→	M	output	-1.431
Hidden 1	1	→	M	output	2.423
Intercept		→	B	output	1.236
Hidden 1	1	→	B	output	-2.282

Note. The weights are input for the logistic sigmoid activation function.

Table 9: Confusion Matrix

		Predicted	
		B	M
Observed	B	7	2
	M	0	11

Table 10: Class Proportions

	Data Set	Training Set	Test Set
B	0.380	0.362	0.450
M	0.620	0.637	0.550

3.2.3 Support Vector Machine Classification Algorithm Result

Table 11: Support Vector Machine Classification

Support Vectors	n(Train)	n(Test)	Test Accuracy
34	80	20	0.750

Table 12: Support Vectors

Row	radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
3	0.031	0.533	-0.455	-0.572	0.360	0.446	0.027	0.529
4	0.031	-0.622	0.305	0.244	-0.391	0.038	-0.298	-0.453
6	0.236	1.304	0.009	0.032	-0.801	-0.911	-1.013	-0.943
7	-0.584	-0.622	1.488	1.314	-0.391	1.951	1.521	1.633
8	1.465	0.533	0.263	0.298	-1.416	-0.976	-1.305	-0.943
10	-0.379	-0.044	-1.342	-1.223	0.701	-0.747	2.626	0.651
12	1.260	1.496	-0.793	-0.819	0.565	0.414	1.196	1.633
19	0.441	1.304	-0.920	-0.860	-1.143	-1.254	-1.402	-0.698
21	1.260	1.689	-0.075	-0.125	-0.869	-0.666	-0.785	-0.698
26	1.465	-0.429	-0.455	-0.438	-1.416	-1.091	-0.493	-1.066
30	-1.404	-0.044	-0.962	-0.941	0.497	-0.535	-0.298	0.651
31	-0.994	0.919	-0.033	-0.159	0.701	0.119	0.612	-0.207
34	-0.379	0.533	-0.413	-0.494	0.087	0.283	0.124	0.406
35	0.236	-1.007	-0.751	-0.725	-0.733	-0.960	-0.688	-0.698
36	1.670	-1.392	-0.540	-0.710	1.111	1.853	0.319	2.124
37	0.441	0.726	-0.413	-0.409	-1.757	-1.075	-1.890	-0.575
39	-0.174	1.689	-0.117	-0.187	-0.323	-0.208	-0.168	-0.085
41	-0.174	-0.815	-0.455	-0.572	0.360	0.005	0.124	0.406
42	-0.174	0.148	-0.582	-0.638	0.633	-0.011	-0.070	0.161
44	0.031	-1.392	-0.413	-0.428	-0.323	-0.747	-0.135	-0.821
45	-1.404	-0.429	-0.075	-0.056	-0.255	-0.895	-1.110	-0.698
47	0.851	1.111	-0.962	-0.907	-0.869	-0.846	-1.013	0.161
51	1.056	-0.429	-0.582	-0.616	-0.255	-0.518	-0.688	-0.575
52	-1.404	-1.200	0.136	0.079	-0.733	-0.371	-0.688	-0.453
59	1.670	-1.392	-0.413	-0.444	-1.006	-0.813	-0.395	-0.943
60	0.236	-1.200	-0.117	-0.219	-0.323	-0.273	-0.135	-0.453
63	0.441	1.689	-1.047	-1.038	1.384	-0.077	-0.103	0.529
64	-0.379	-0.044	-0.286	-0.391	1.111	0.626	0.872	1.265
67	-1.404	-0.622	-0.498	-0.534	-0.391	-0.355	-0.590	-0.330
68	0.646	-0.815	-0.455	-0.525	0.906	-0.061	0.644	0.406
71	-1.609	1.496	-1.596	-1.435	-0.323	0.430	-0.103	3.105
74	0.851	-0.044	0.305	0.360	-0.733	-0.698	-0.428	-1.312
76	1.670	0.533	-0.835	-0.813	-0.391	-0.895	0.482	-0.575
78	-1.609	-1.007	1.530	1.949	2.750	-0.780	-0.395	-0.943

Table 13: Confusion Matrix

		Predicted	
		B	M
Observed	B	0.25	0.2
	M	0.05	0.5

Table 14: Class Proportions

	Data Set	Training Set	Test Set
B	0.380	0.362	0.450
M	0.620	0.637	0.550

Table 15: Evaluation Metrics

	B	M	Average / Total
Support	9	11	20
Accuracy	0.750	0.750	0.750
Precision (Positive Predictive Value)	0.833	0.714	0.768
Recall (True Positive Rate)	0.556	0.909	0.750
False Positive Rate	0.091	0.444	0.268
False Discovery Rate	0.167	0.286	0.226
F1 Score	0.667	0.800	0.740
Matthews Correlation Coefficient	0.504	0.504	0.504
Area Under Curve (AUC)	0.732	0.732	0.732
Negative Predictive Value	0.714	0.833	0.774
True Negative Rate	0.909	0.556	0.732
False Negative Rate	0.444	0.091	0.268
False Omission Rate	0.286	0.167	0.226
Threat Score	0.833	1.111	0.972
Statistical Parity	0.300	0.700	1.000

Note. All metrics are calculated for every class against all other classes.

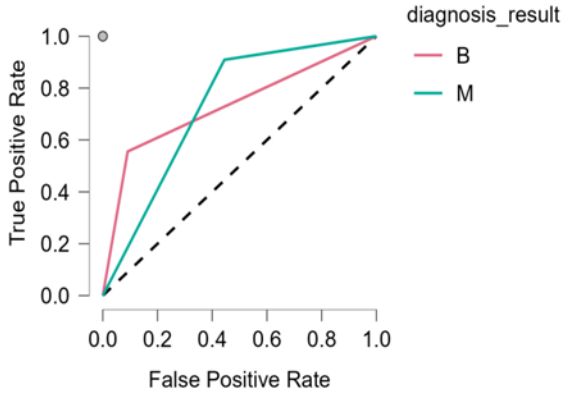

Figure 9: ROC Curves Plot

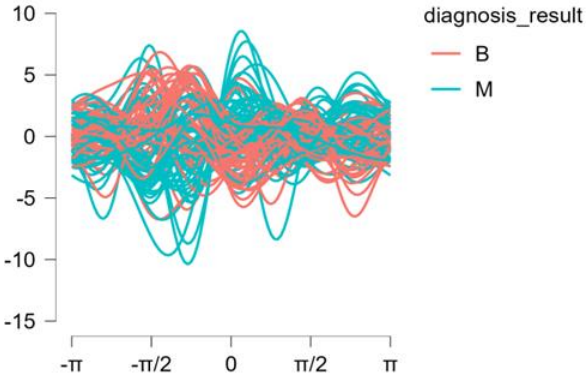

Figure 10: Andrews Curves Plot

3.2.4 Random Forest Classification Algorithm Result

Table 16: Random Forest Classification

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
27	2	64	16	20	1.000	0.800	0.902

Note. The model is optimized with respect to the *out-of-bag accuracy* (OOB).

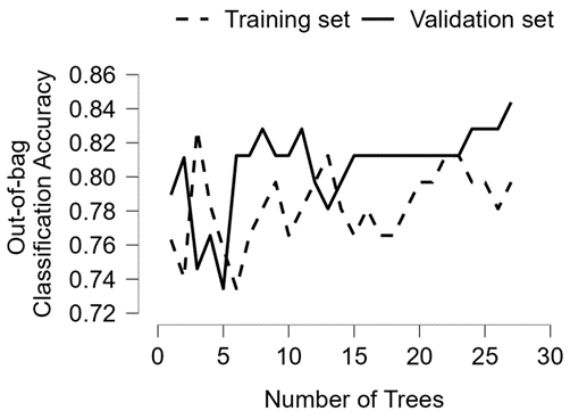

Figure 11: Out-of-bag Classification Accuracy Plot

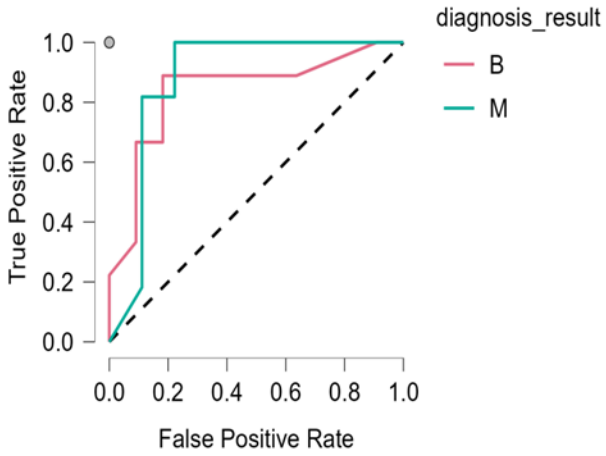

Figure 12: ROC Curves Plot

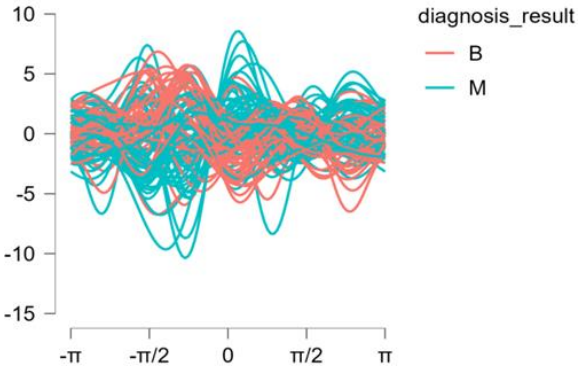

Figure 13: Andrews Curves Plot

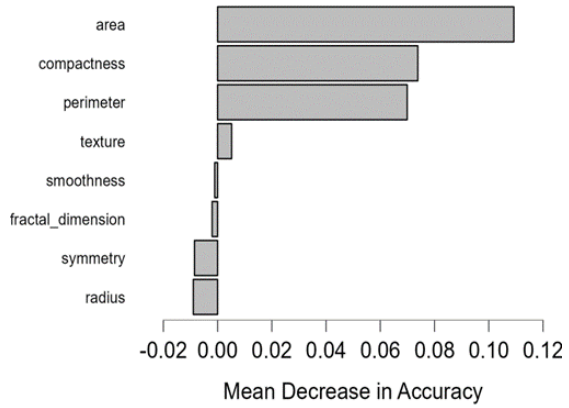

Figure 14: Mean Decrease in Accuracy

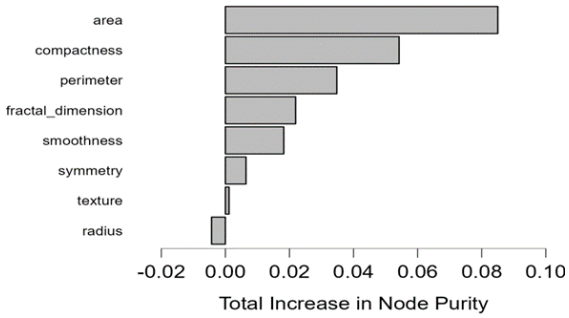

Figure 15: Total Increase in Node Purity

Table 17: Confusion Matrix

		Predicted	
		B	M
Observed	B	0.35	0.1
	M	0.1	0.45

Table 18: Class Proportions

	Data Set	Training Set	Validation Set	Test Set
B	0.380	0.359	0.375	0.450
M	0.620	0.641	0.625	0.550

Table 19: Evaluation Metrics

	B	M	Average / Total
Support	9	11	20
Accuracy	0.800	0.800	0.800
Precision (Positive Predictive Value)	0.778	0.818	0.800
Recall (True Positive Rate)	0.778	0.818	0.800
False Positive Rate	0.182	0.222	0.202

Table 19: Evaluation Metrics

	B	M	Average / Total
False Discovery Rate	0.222	0.182	0.202
F1 Score	0.778	0.818	0.800
Matthews Correlation Coefficient	0.596	0.596	0.596
Area Under Curve (AUC)	0.828	0.884	0.856
Negative Predictive Value	0.818	0.778	0.798
True Negative Rate	0.818	0.778	0.798
False Negative Rate	0.222	0.182	0.202
False Omission Rate	0.182	0.222	0.202
Threat Score	1.167	1.500	1.333
Statistical Parity	0.450	0.550	1.000

Note. All metrics are calculated for every class against all other classes.

Table 20: Feature Importance

	Mean decrease in accuracy	Total increase in node purity
area	0.109	0.085
compactness	0.074	0.054
perimeter	0.070	0.035
fractal_dimension	-0.002	0.022
smoothness	-0.001	0.018
symmetry	-0.009	0.006
texture	0.005	0.001
radius	-0.009	-0.004

3.2.5 K-Nearest Neighbors (KNN) Classification Algorithm Result

Table 21: K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
7	rectangular	Euclidean	64	16	20	1.000	0.800

Note. The model is optimized with respect to the *validation set accuracy* .

Table 21: K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy

- Data Split

Figure 16: K-Nearest Neighbor data split percentage

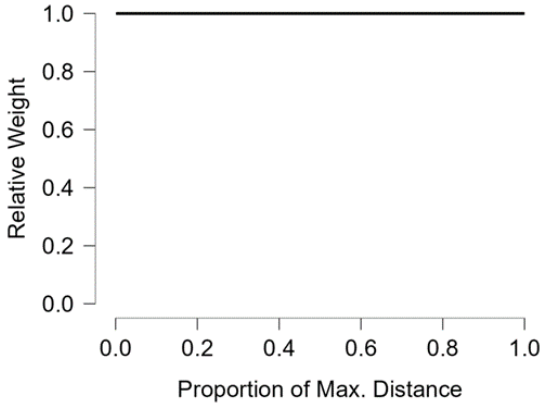

Figure 17: Rectangular Weight Function

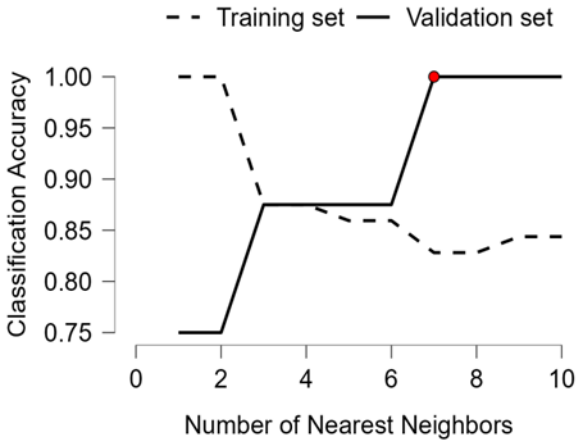

Figure 18: Classification Accuracy Plot

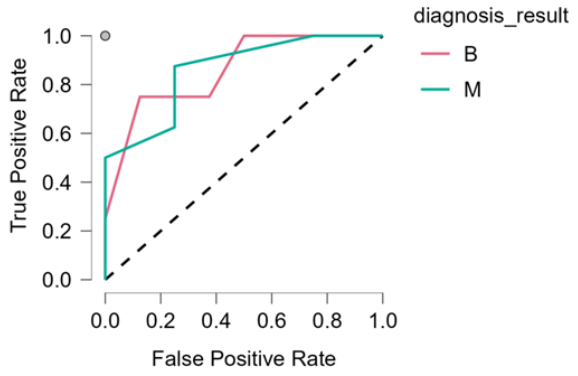

Figure 19: ROC Curves Plot

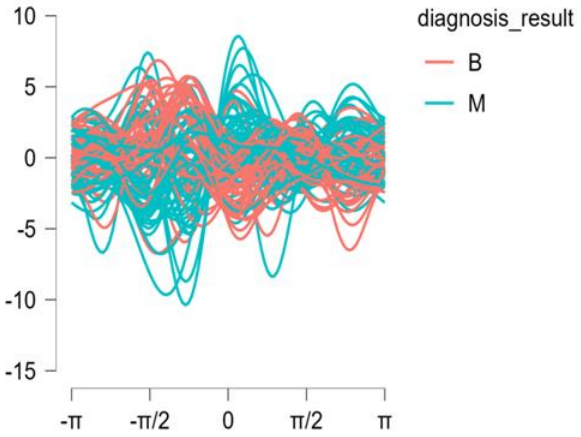

Figure 20: Andrews Curves Plot

Table 22: Confusion Matrix

		Predicted	
		B	M
Observed	B	0.15	0.05
	M	0.15	0.65

Table 23: Class Proportions

	Data Set	Training Set	Validation Set	Test Set
B	0.380	0.453	0.313	0.200
M	0.620	0.547	0.688	0.800

Table 24: Evaluation Metrics

	B	M	Average / Total
Support	4	16	20
Accuracy	0.800	0.800	0.800
Precision (Positive Predictive Value)	0.500	0.929	0.843

Table 24: Evaluation Metrics

	B	M	Average / Total
Recall (True Positive Rate)	0.750	0.813	0.800
False Positive Rate	0.188	0.250	0.219
False Discovery Rate	0.500	0.071	0.286
F1 Score	0.600	0.867	0.813
Matthews Correlation Coefficient	0.491	0.491	0.491
Area Under Curve (AUC)	0.859	0.859	0.859
Negative Predictive Value	0.929	0.500	0.714
True Negative Rate	0.813	0.750	0.781
False Negative Rate	0.250	0.188	0.219
False Omission Rate	0.071	0.500	0.286
Threat Score	0.429	2.600	1.514
Statistical Parity	0.300	0.700	1.000

Note. All metrics are calculated for every class against all other classes.

IV. PREDICTED RESULTS COMPARISON AND DESCRIPTION

The experiment was done using the R programming language and RStudio frameworks. First, the dataset explanation and overview which was sourced from kaggle.com is done which contains a total of 100 observations (rows) and 9 variables (10 columns), with 8 variable as independent variable. The dataset was Split into two with the percentage of (80% = training and 20%=testing) respectively, using R language library which was applied for the splitting before application of the five (5) algorithms was applied on both train and test dataset to achieve the model then after the prediction result by the computer on the dataset and model is made, there is need for more accuracy of the prediction on both dataset and model by applying a confusion matrix each indicating the result accuracy as produced by the various models. Decision Tree Classification Algorithm Result: the result produced by decision tree algorithm on the ROC Curves Plot of the decision tree algorithm shows a predictive model on the False Positive Rate (M) of 90% accuracy while True Positive Rate (B) of 70% accuracy on the predicted model using decision tree algorithm. On the other hand, Confusion Matrix

predicted accuracy on the True Positives Rate (B) 35% while False Positives Rate (M) predicted 15% accuracy. The True Negatives Rate (B) also shows 5% while False Negatives Rate (M) shows 45% accuracy on the dataset. In summary, the predicted test accuracy of the decision tree algorithm produced 80% accuracy rate shown in table 1 above.

Neural Network algorithm Classification result: The ROC Curves Plot of the Neural Network algorithm shows a predictive model on the False Positive Rate (M) of 11% accuracy while True Positive Rate (B) of 7% accuracy on the predicted model using decision tree algorithm. On the other hand, Confusion Matrix predicted accuracy on the True Positives Rate (B) 8% while False Positives Rate (M) predicted 2% accuracy. The True Negatives Rate (B) also shows 1% while False Negatives Rate (M) shows 0% accuracy on the dataset. Figure 6 shows the logistic Sigmoid Activation Function with output of 0% and input of 1% prediction. In summary, the predicted test accuracy of the neural network algorithm shows 90% accuracy rate while figure 5 above shows the Network structure plot of the predictive output.

Support Vector Machine (SVM) Classification Algorithm Result: The ROC Curves Plot in figure 9 of the Support Vector Machine (SVM) algorithm shows a predictive model on the False Positive Rate (M) of 2% accuracy while True Positive Rate (B) of 25% accuracy on the predicted model using SVM algorithm. On the other hand, Confusion Matrix predicted accuracy on the True Positives Rate (B) with 2% while False Positives Rate (M) predicted 1% accuracy. The True Negatives Rate (B) also shows 5% while False Negatives Rate (M) shows 5% accuracy on the dataset. In summary, the predicted test accuracy of the SVM algorithm shows 75% accuracy rate while figure 5 above shows the Support Vector Machine (SVM) predictive output. Andrews Curves Plot also shows the prostate cancer diagnosis result on True Positive Rate of (B) 6% and False Positive Rate (M) of 7% accuracy shown in figure 10 above and in summary, table 15 shows all calculated metrics for every class against all other classes.

Random Forest Classification algorithm Result: From the analytical result of the Random Forest algorithm prediction, it was clear enough to showcase the predicted results from various model outputs with much emphasis on the model optimized with respect to the out-of-bag accuracy (OOB) with 90% accuracy shown in figure 11 while ROC Curves Plot in figure 12 predicted a model with True Positive Rate (B) 77% and False Positive Rate (M) of 80% When the Andrews Curves Plot shown in figure 13 predicted a model with result of True Positive Rate (B) of 60% accuracy and False Positive Rate (M) of 80% accuracy. On the Mean Decrease in accuracy of the variables, figure 14 shows area having the highest level percentage decrease of 80% while smoothness the lowest level of 1%. On the other hand, figure 15 shows a total increase in Node Purity with area of 83% and texture 0% respectively. In summary table 17 shows the confusion matrix of the predicted Random Forest Classification Algorithm predicted True Positive Rate value (B) of 35%, False Positive Rate (M) of 1% while True Negatives Rate (B) of 1% and False Negatives Rate (M) of 45% accuracy.

K-Nearest Neighbors (KNN) Classification Algorithm Result: The data split for this algorithm was done three places namely: train of 64%, 16% validation and 20% test making a total of 100% showed in figure 16 above. After the experiment, the predicted model on the rectangular weight function was between the relative weight and the proportion of max distance and it gave a 92% rate accuracy on the predicted result shown in figure 17 while figure 19 presented the Classification Accuracy Plot on the training set and validation set between the numbers of nearest neighbors dataset with a percentage of 98% and 99% respectively. The confusion matrix of the predicted KNN Classification Algorithm predicted True Positive Rate value (B) of 15%, False Positive Rate (M) of 1% while True Negatives Rate (B) of 15% and False Negatives Rate (M) of 65% accuracy presented in table 22 of this study.

4.1 Table 25: Performance Evaluation of the Five (5) Classification Algorithms

Algorithms Used	Test Accuracy %	M (False Positive Rate)	B (True Positive Value)
Decision Tree Algorithm	80%. accuracy	90%	70%
Neural Network Algorithm	90% accuracy	11%	7%
Support Vector Machine (SVM)	75% accuracy	2%	25%
Random Forest Algorithm	80% accuracy and OOB accuracy of 90%	80%	77%
K-Nearest Neighbors (KNN)	90% Test accuracy Validation Accuracy 1%	1%	15%

4.2 Conclusion

As earlier stated, that the aim of this research is to carry out a comparative predictive analysis of prostate cancer diagnosis and treatment using five (5) classification algorithm which includes: Decision Tree, Neural Network, Support Vector, Random Forest and K-Nearest Neighbor KNN.

This research was able to show clearly how prostate cancer disease could be diagnoses using five different classification algorithms with train dataset of 80% and 20% respectively in some of the algorithms for the different analysis. For more accurate model prediction, each of the analysis conducted evaluation metrics was involved, confusion matrix, Out-of-bag Classification accuracy plot on the random forest algorithm, Andrews Curves Plot model, ROC Curves Plot model, Class Proportions of the dataset was done for more understanding of the dataset proportions, logistic sigmoid activation function on the weights of the input involved for the prediction and network structure Plot model was also conducted. Base on the comparison analysis conducted by this study, it was observed that Neural Network Algorithm and K-Nearest Neighbors (KNN) have the highest percentage accuracy towards the prediction of prostate cancer having 90% test accuracy with KNN 1% validation accuracy. Below are the outlined findings of the analysis.

4.2.1 Decision Tree Classification Algorithm Result

False Positive Rate (M) of 90% accuracy while True Positive Rate (B) of 70% accuracy on the predicted model using decision tree algorithm while the predicted test accuracy produced was 80%.

4.2.2 Neural Network algorithm Classification result

False Positive Rate (M) of 11% accuracy.
True Positive Rate (B) of 7% accuracy
Confusion Matrix predicted accuracy on the
True Positives Rate (B) 8%.
False Positives Rate (M) predicted 2% accuracy.
The True Negatives Rate (B) also shows 1%.
False Negatives Rate (M) shows 0% accuracy

4.2.3 Support Vector Machine (SVM) Classification Algorithm Result

Predictive model on the False Positive Rate (M) of 2% accuracy
True Positive Rate (B) of 25% accuracy on the predicted model using SVM algorithm

Confusion Matrix predicted accuracy on the
True Positives Rate (B) with 2%.
False Positives Rate (M) predicted 1% accuracy.
The True Negatives Rate (B) also shows 5%.
False Negatives Rate (M) shows 5% accuracy on the dataset
SVM algorithm shows 75% accuracy rate

4.2.4 Random Forest Classification algorithm Result:
True Positive Rate (B) 77% and False Positive Rate (M) of 80% . Andrews Curves Plot shown in figure 13 predicted a model with result of True Positive Rate (B) of 60% accuracy and False Positive Rate (M) of 80% accuracy. Confusion matrix of the predicted Random Forest Classification Algorithm predicted True Positive Rate value (B) of 35%, False Positive Rate (M) of 1% while True Negatives Rate (B) of 1% and False Negatives Rate (M) of 45% accuracy.

4.2.5 K-Nearest Neighbors (KNN) Classification Algorithm Result:
The confusion matrix of the predicted KNN Classification Algorithm predicted True Positive Rate value (B) of 15%, False Positive Rate (M) of 1% while True Negatives Rate (B) of 15% and False Negatives Rate (M) of 65% accuracy presented. The rectangular weight function was between the relative weight and the proportion of max distance and it gave a 92% rate accuracy

4.3 Recommendation
The study has been able to achieve its stated aim and objective and it was also able to proffer solution to the medical professionals for a more accurate and reliable outcome after diagnosis through the predictive model and intelligent analysis produced by the prostate dataset. The comparison was able to showcase the most perfect algorithm for more accurate predictive model on the diagnosis and treatment of prostate cancer. The researcher therefore recommends the following:

1. Neural Network Algorithm and K-Nearest Neighbors (KNN) should be employed in the prediction of prostate cancer and other health related diagnosis for more accuracy in data prediction.
2. Full adoption of machine learning tools should be used in solving real life challenging problems more especially in health related problems more especially in Nigeria.
3. Other organizations should be encouraged to apply machine learning tools for easy decision making.

REFERENCES

[1] Onuiri Ernest E., Awodele Oludele and Ebiesuwa Oluwaseun (2016) Early Detection and Diagnosis of Prostate Cancer using Artificial Intelligence Concept, International Journal of Computer Applications (0975 – 8887) Volume 149 – No.6,

[2] Hruban, R. (2012). What are tumors? Retrieved 2022, from http://pathology.jhu.edu/

[3] National Cancer Institute. (2014). Prostate cancer.

[4] American cancer society (2015). Prostate cancer prevention and early detection. Retrieved 22 sept 2021, from http://www.knowledge.scot.nhs.uk/ecomscormplayer/ADRmodule6//index.html.

[5] Vogelzang, N., & Shore, N. (2015). upfront chemotherapy in prostate cancer. Retrieved 2015, from cancer network: http://www.cancernetwork.com/prostate-cancer/upfront-chemotherapy-prostate-cancer?

[6] Pravarti Jain And Santosh Kr Vishwakarma (2017) A Case Study on Car Evaluation and Prediction: Comparative Analysis using Data Mining Models, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.9

[7] Smith-Palmer, C. Takizawa,and W. Valentine(2019) Literature review of the burden of prostate cancer in Germany, France, the United Kingdom and Canada, accessed from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6421711/, Doi: 10.1186/s12894-019-0448-6

[8] Ashwini Kannan, Maggie Kirkman, Rasa Ruseckaite(2019) Prostate care and prostate cancer from the perspectives of undiagnosed men: a systematic review of qualitative research, retrieved form https://bmjopen.bmj.com/content/9/1/e022842/ http://orcid.org/0000-0003-2962-8400Sue M Evans1

[9] Ajape AA, Babata A, Abiola OO.(2009), Knowledge of prostate cancer screening among native African urban population in Nigeria. Nig Q J Hosp Med.19(3):145–7. Retrieved from [PubMed] [Google Scholar]

[10] Oladimeji O, Bidemi YO, Olufisayo JA, Sola AO.(2010) Prostate Cancer Awareness, Knowledge, and Screening Practices among Older Men in Oyo State, Nigeria. Int Q

Community Health Educ.30(3):271–86. Retrieved from [PubMed] [Google Scholar]

[11] Bamidele, H. McGarvey , B.M. Lagan, N. Ali (2017) Life after prostate cancer, retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/ecc.12784

[12] Sanny Kappen, Verena Jürgens, Michael H Freitag, Alexander Winter (2019) Early detection of prostate cancer using prostate-specific antigen testing: an empirical evaluation among general practitioners and urologists retrieved from https://www.dovepress.com/early-detection-of-prostate-cancer-using-prostate-specific-antigen-tes-peer-reviewed-fulltext-article-CMAR#

[13] Sidana (2017) A Case Study on Car evaluation and Prediction: Comparative Analysis using Data mining Models, International Journal of Computer applications (0975 – 8887) Volume 172 – No.9

[14] BITS Pilani, Dubai, BITS Pilani, Dubai and Ronak Sumbaly (2015) Diagnosis of Diabetes Using Classification Mining Techniques accessed from International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1)

[15] Hun, C. C., Yazid, H., Safar, M. J. A., & Ab Rahman, K. S. (2022, February). Comparison Between K-Nearest Neighbor (KNN) and Decision Tree (DT) Classifier for Glandular Components. In Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications: Enhancing Research and Innovation through the Fourth Industrial Revolution (pp. 292-297). Singapore: Springer Singapore.

[16] Dubey, P., & Kumar, S. (2023). Advancing prostate cancer detection: a comparative analysis of PCLDA-SVM and PCLDA-KNN classifiers for enhanced diagnostic accuracy. Scientific Reports, 13(1), 13745.

[17] Srivenkatesh, M. (2020). Prediction of prostate cancer using machine learning algorithms. Int. J. Recent Technol. Eng, 8(5), 5353-5362.