Prediction of Prostate Cancer using Ensemble of Machine Learning Techniques

Oyewo O.A[1], Boyinbode O.K[2]
Department of Computer Science
Federal University of Technology, Akure
Ondo State, Nigeria

***Abstract*—Several diseases are associated with humans; some are synonymous to female and some to male. Example of diseases synonymous to the male gender is Prostate Cancer (PC). Prostate cancer occurs when cells in the prostate gland starts to grow uncontrollably. Statistics shows that prostate cancer is becoming an epidemic among men. Hence, several research works have tried to solve this problem using various methods. Although numerous medical research works are ongoing in the area, the need to introduce technology to battle the epidemic is paramount. Because of this, some researchers have developed several models to help solve issues of prostate cancer in men, but the area is still open to contribution. Recently, some researchers have adopted some well-established Machine Learning (ML) techniques to predict and diagnose the occurrence of prostate cancer, but issues of low prediction accuracy, inability to implement model, low sensitivity; among others still lingers. This paper approached these challenges by developing an ensemble model that combines three (3) ML techniques; Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron (MP) to predict PC in men. Our developed model was evaluated using sensitivity, specificity and accuracy as performance metrics, and our result showed a prediction accuracy of 99.06%, sensitivity of 98.09% and, specificity of 99.54%, which is a relative improvement on the existing systems.**

Keywords*—Prostate cancer; machine learning; support vector; machine; decision tree; multilayer perceptron; diseases

I. Introduction

Cancer is considered one of the most dangerous diseases in the world because it is responsible for around 13% of all deaths in the world[1]. Cancer usually starts as being primary to a specific organ in the body, which later metastasizes to other parts. A common type of cancer is the Prostate Cancer (PC).Prostate cancer is the most rampant and leading cause of cancer death among men in the world, second only to leukaemia [2]. Prostate cancer which is medically referred to as carcinoma of the prostate, begins when cells in the prostate gland starts to grow uncontrollably. Research in [3]explained that prostate cancer begins when healthy cells in the prostate gland change and grow out of proportion, thereby forming a mass called tumour. Recent development in artificial intelligence is now being applied to various fields in medicine and science generally. One of these fields is in the use of Machine Learning techniques to solve issues of prostate cancer. Although, several researchers have tried to predict and diagnose PC in men using several well established ML techniques individually, research in [4],[5], and[6], among others, shows that issues of low prediction accuracy and sensitivity still lingers. This research approached these challenges by combining three (3) well established machine learning techniques (Decision Tree, Support Vector Machine and Multilayer Perceptron), to form an ensemble model that aims to address the recurrent issues associated with the use of single Machine Learning techniques.

The rest of this paper is organized as follows: Section I introduces prostate cancer and justifies the need to carry out this research, Section II reviews related works that have attempted to predict and attend to issues relating to PC, Section III explains the methodology, Section IV presents the results and discussion, Section V concludes

II. Related Works

The prevalence of prostate cancer is increasing by the day. Statistics shows that almost one-third of men over 50 years old will be diagnosed with prostate cancer during their life time[7]. Author in[8], defined prostate cancer as the cancer that occurs in the prostate, a small walnut-shaped gland in men that produces the seminal fluid that nourishes and transports sperm. It is recommended that men have a prostate examination by age 50 [9]. Performing prostate test starts with a Prostate Specific Antigen (PSA) test, and a core biopsy is recommended should the patient have PSA value higher than normal [7]. Biopsy is the gold standard for cancer diagnosis.

Although several works have tried to contribute to Prostate cancer epidemic using various medical approaches, the advent of technology also brought about the development of some computer aided solutions. Example is in [7] where the authors developed a computer aided diagnostic tool that uses image processing techniques for efficient PC diagnosis and prognosis. The authors collected images of prostate gland as shown in Fig. 1, and then separate the images into various portions to diagnose prostate cancer.

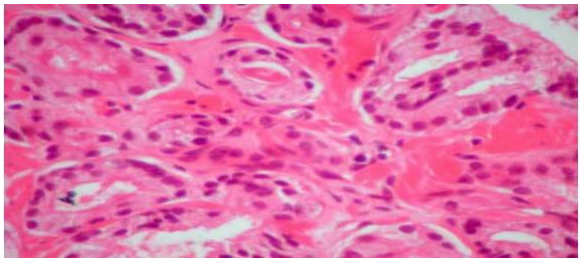

Fig 1. Test Image of Prostate Gland.

The introduction of imaging and machine learning techniques to acquire, process and analyse images from biopsies is of utmost importance[10], because some other diseases imitates prostate cancer. Example is the Benign Prostatic Hyperplasia (BNH), which occurs when the prostate begins to press against the urethra as a result of growth, thereby causing urinary problems[11]. However, the occurrence of prostate cancer is common among men aged 50 and above.

It is essential to trust prediction and diagnosis made using artificial intelligence[12]. Therefore, accuracy of ML predictions is very important. Research in [11] proposed the use of Artificial Neural Network to detect early signs of prostate cancer, but the model could not record perfect accuracy. Author in [13], also applied artificial neural networks (ANN) with back propagation to predict prostate cancer recurrence in patients, but the evaluation could not achieve optimal accuracy. Research in[14]also developed a model using Fisher Linear Classifier to predict recurrence of prostate cancer in men, but their model achieved an accuracy of 93%.Zhao *et al.*,[15]proposed a Penalized Logistic Regression Technique based on top-scoring pair (TSP) as a classification model to predict prostate cancer, but perfect accuracy was also not recorded. Authors in [16]proposed a prostate cancer predictive model using Decision Tree Algorithm. The research established Decision Tree as a useful data mining algorithm for predicting prostate cancer, but the model is not reliable due to low accuracy. Ge*et al.*,[17], proposed a prostate cancer predictive model using Logistic Regression and Artificial Neural Network, but the individual accuracy of the algorithms stood at 84.02% and 85.09% respectively. Takeuchi*et al.*,[18]proposed a prostate cancer prediction system on prostate biopsy using Multilayer Artificial Neural Network (ANN), but the system was able to predict with an accuracy of 71.6%, but this can be associated with the insufficient amount of dataset used for the research. In order to combat the recurrent issue of accuracy, our research proposes an ensemble model that combines three ML techniques. The method and functionality of our model is discussed in the next section.

III. Methodology

The architecture of the model is shown in Fig. 2. The architecture shows the components of the developed model. The functionalities of each component are explained in details below:

A. Datasets (Prostate and Non-Prostate Cancer)

The dataset used in the research is obtainable fromhttp://github.com/selva86/datasets/masters/prostate.csv. The obtained data contains about one thousand, nine hundred and forty (1,940) study patients which make up the instances of the data. Each of the instances consist of 10 attributes including class label indicating that an instance is either a Benign (0) or Prostate cancer sample (1). The attribute values are all numeric, Table 1 shows the description of the data attributes.

TABLE I. Description of Data Attributes

S/N	Data Attributes	Description
1	Icavol	Log of the Cancer volume
2	Iweight	Log of the prostate weight
3	Age	Age of the patient
4	Ibph	Log of the Benign prostatic Hyperplasia amount
5	Svi	Seminal Vesicle invasion
6	Icp	Log of the Capsular Penetration
7	Gleason	Gleason Score
8	Pgg45	Percentage Gleason score 4 or 5
9	Ipsa	Log of Prostate Specific Atigen

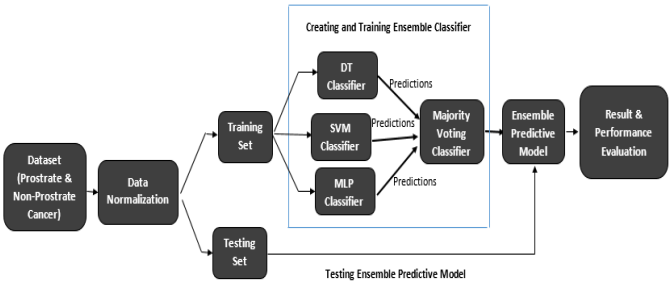

Fig 2. Architecture of Model.

B. Data Normalization

The obtained data was normalized using Z -score normalization in order to make training less sensitive to the scale of features.

Z score will convert the data into [0,1] distribution using equation (1)

$$x'_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where x'_i is the data value, x_i is the data value to be normalized, μ represents the mean of data values in the feature category

C. Data Training and Testing

The normalized data was divided into training and testing set using a 67% - 33% split ratio as shown in Table 2.The training set was used to train the classifiers, the testing set was used to evaluate predictive models.

TABLE II. Prostate Cancer Distribution

Class labels	Training set	Testing set
Non- Prostate	908	432
Prostate	391	209
Total	**1299**	**641**

The classification algorithm used in this research for predicting the presence of prostate cancer is an ensemble of three (3) classifiers: Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron. The ensemble algorithm predicts the presence of the three classifiers (SVM, DT and MP) predictions $P_1, P_2 and P_3$ respectively, to make final prediction P_f as follows:

Given training set of prostate cancer data is given as:

$$D = \{(x_i, y_i)\}_i^n = 1 \quad (2)$$

Where $D$is the training set of prostate data, x_i is an input for the i -th prostate data described by set of attributes $a_1 a_2 a_3 a_q, y_i \in \{0,1\}$ is its corresponding class label indicating whether the sample is a benign sample ($y_i = 0$) or a prostate cancer ($y_i = 1$), and n represents the total number of data samples.

The first classifier C_1 which is Linear SVM make prediction P_1 as either ($y_i = 0$) or ($y_i = 1$) , by creating decision boundaries (hyperplanes) that linearly separates the two classes using equation (3)

$$(w.x_i) + b = 0 \quad (3)$$

Such that

$$\text{Class } (x_i) = \{^{0,\ w.x_i+b\geq 0\ if y_i=0}_{1,\ w.x_i+b\leq 0\ if y_i=1} \quad (4)$$

Where x_i denotes an instance of a prostate cancer sample, w represents the weight vector, b is the bias.

However, associated with each hyperplane is a notion called margin, defined as the distance between the hyperplane ($w.x_i) + b = 0$ and the closest sample x_i which can be determined using equation (5)

$$\frac{|w.x_i+b|}{||w||} \quad (5)$$

The best choice of hyperplane depends on the hyperplane with maximum margin between both classes. This is achieved by minimizing weight vector $||w||$ using equation (6)

$$\min ||w||^2 , st. \{^{0,\ w.x_i+b\geq 0\ if y_i=+1}_{1,\ w.x_i+b\geq 0\ if y_i=-1} \forall i \quad (6)$$

Decision Tree

The second classifier C_2 make prediction P_2 by applying C4.5 algorithm, as it starts by selecting an attribute from the given set $a_1, a_2, a_3, \ldots, a_q$ to partition D into subsets ($d_1, d_2, d_3 \ldots, d_j$) using information gain presented in equation (7) and (8).

$$I(D) = -\sum^m P(y_i) \log_2 P(y_i) \quad (7)$$

$$IG(D, a_i) = I(D) - \sum_{v\in values(A)} \frac{|D_v|}{n} I(D_v) \quad (8)$$

Where v is a value in attribute a_i, value (a_i) represents all possible values in a_i, D_v represents instances for which a_i has v , n represents number of instances in $I(D)$ and $I(D_v)$ respectively, $P(y_i)$ represents the probability of class y_i in D, m is the distinct number of class values, and j is the number of outcome of test attribute a_i.

The process is continued over each d_i, where $1\leq i \leq j$, until all elements in each final subset falls under the same class.

Multilayer Perceptron

The third classifier C_3 makes its prediction P_3 as MLP accepts input vector x_i multiplied by a weight vector w_i, and added to a bias b to produce an output$\hat{y}$ using the following equation:

$$\hat{y} = f(\sum_{i=1}^n w_i x_i + b) \quad (10)$$

where n is the number of input-output pairs, and f is a non-linear activation function presented in equation (11).

$$f = \frac{1}{1+e^{-x_1}} \quad (11)$$

To determine the prediction error of MLP, the Mean Square Error (MSE) function is applied as follows:

$$E(\hat{y}, y) = \frac{1}{2}\sum_{i=1}^{n}(\hat{y} - y)^2 \quad (12)$$

Where $E$is the error function between the predicted class $\hat{y}$ and the target classy

Also, training the MLP by backward propagation involves computing the gradient of the error with respect to w is using chain rule of differentiation as follows:

$$\delta_i \leftarrow dE/w$$

Where δ_i is the gradient descent, w represents weight. Thereafter, w is updated in the direction via the gradient that helps minimize the loss.

1) Majority Voting Classification

This involves combining predictions P_1, P_2 and P_3, of the individual base classifiers C_1, C_2 and C_3 respectively to make a final prediction P_f, by predicting the class label that have been predicted most frequently using equation (13) and (14).

$$C_{r,y} = \{^{1, if p_i=y_i}_{0, if p_i\neq y_i} \quad (13)$$

$$P_f = \arg max\, i \sum_{i=1}^{q=3} C_{r,y} \quad (14)$$

Where C_r represents the decision of the $r-th$ classifier given class y_i , P_f represents the final prediction by the ensemble, and q is the number of the base classifiers.

2) Evaluation Measures

Our model was evaluated based on three metrics: Sensitivity, Accuracy and specificity. Sensitivity measures the proportion of positives (prostate cancer samples) correctly classified, Accuracy measures the proportion of the total number of correct predictions, specificity measures the proportion of negatives (Benign samples) correctly classified, using:

$$\text{Sensitivity} = \frac{AP}{AP+BN}$$

$$\text{Accuracy} = \frac{AP+AN}{AP+AN+BP+BN}$$

$$\text{Specificity} = \frac{AN}{AN+BP}$$

Where AP = True Positive, AN= True Negative, BP= False Positive, BN = False Negative.

Our ensemble model was implemented using Python 3.7, Spyder python editor via Anaconda Distribution, Excel spreadsheet package, Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz, 2501 Mhz, 2 Core(s), 4 Logical Processor(s).

IV. Results and Discussion

The Confusion Matrix result of the developed prostate cancer prediction model when applied on the test data is shown in Table III. From the study, it is shown that out of 209 actual prostate cancer data and 432 non-prostate cancers from the 641 test data, the model predicted 205 prostate cancer instances correctly, and predicted 4 incorrectly, while also predicting 430 non-prostate cancers correctly with 2 incorrectly. In all, 635 data was correctly classified, while 6 were incorrectly classified.

Table IV shows the total number of correct and incorrect classifications obtained after the developed prostate cancer predictive model had been tested. The total number of incorrect and correct classifications was computed by summing the number of AP, AN, BN and BP for incorrect classifications.

Correct Classification = AP + AN = 635

Incorrect Classification = BN + BP = 6

Table V shows the evaluation of the developed model using Accuracy, Sensitivity and Specificity.

$$\text{Sensitivity} = \frac{AP}{AP+BN} = \frac{205}{205+4} = 0.9809$$

$$\text{Accuracy} = \frac{AP+AN}{AP+AN+BP+BN} = \frac{205+430}{205+430+2+4} = 0.9906$$

$$\text{Specificity} = \frac{AN}{AN+BP} = \frac{430}{430+2} = 0.9954$$

In order to test the efficiency of the ensemble model, the dataset was tested with DT, SVM and MP individually, and the result is presented in Table VI.

The result showed that the developed ensemble model had the highest number of correctly classified instances with 635 instances with number of incorrectly classified instances as zero (6) instances. However, MP also showed to be effective as it correctly classified 626 instances and misclassified 15 instances. From the study, SVM result was not suitable for the purpose of this research work as it correctly classified all non-prostate cancer instances as it predicted all the 432 non-prostate cancer correctly, but wrongly classified all prostate cancer instances with the number of AP recorded as zero (0).

The graphical representation is presented in Fig. 3.

Table VII shows the Accuracy, Sensitivity, and Specificity of the developed ensemble model and the base models. Our ensemble model shows to be the most effective model with an Accuracy of 99.06%, Sensitivity of 98.09%, and Specificity of 99.54% as compared to the result from other models displayed in table. Figure 4 shows graphical representation of evaluation of the proposed ensemble model with the base models.

In order to evaluate the performance of the developed ensemble system, our results were compared with some existing works as shown in Table VIII, in which the developed model shows to be a better model for the prediction of prostate cancer based on its high Accuracy. Fig. 5 shows graphical representation of the comparison.

TABLE III. Confusion Matrix Result of the Developed Ensemble Prostate Cancer Detection Model

		Predicted Class	
		Non-Prostate Cancer	**Prostate Cancer**
Actual Class	**Non-Prostate Cancer**	AN 430	AP 2
	Prostate Cancer	BN 4	BP 205

TABLE IV. Number of Correct and Incorrect Classification Obtained by the Developed Prostate Cancer Prediction Model

Number of Test Data	**Correct Classification**	**Incorrect Classification**
641	**635**	**6**

TABLE V. Evaluation of Developed Model using Sensitivity, Accuracy and Specificity

Accuracy	**Sensitivity**	**Specificity**
0.9906	0.9809	0.9954

TABLE VI. Comparison of our Ensemble Model with Individual Base Algorithms

Models	**AN**	**AP**	**BN**	**BP**	**Correct Classification**	**Incorrect Classification**
MLP	426	6	9	200	626	15
DT	432	0	45	164	596	45
SVM	432	0	209	0	432	209
Developed Ensemble Model	**430**	2	4	**205**	**635**	6

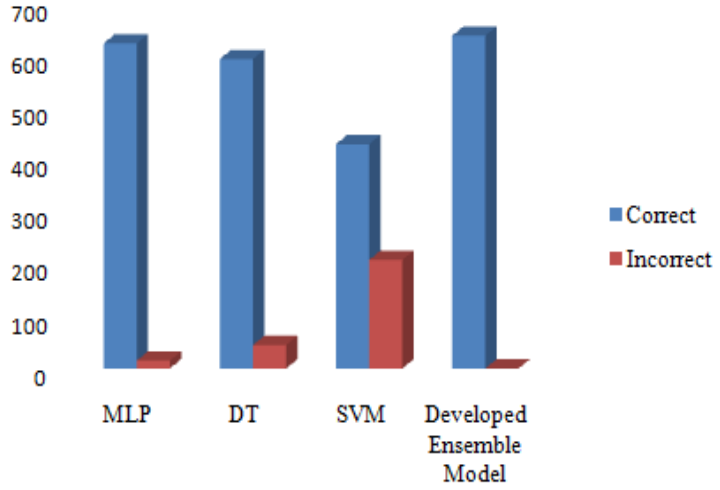

Fig 3. Representation of Correct and Incorrect Classification Ensemble Model with Individual base Models.

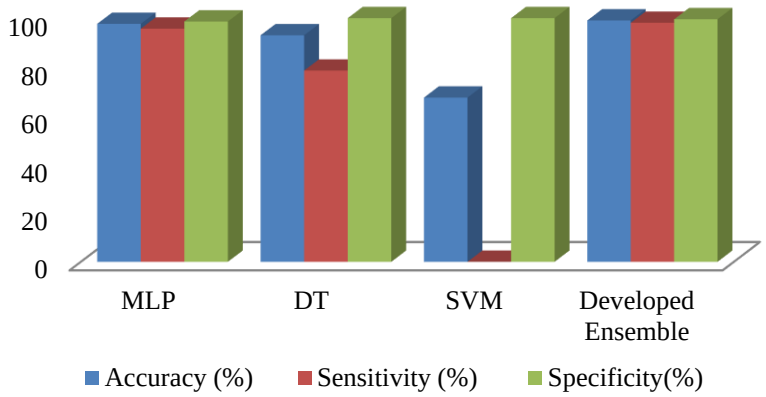

Fig 4. Representation of Evaluation of the Proposed Ensemble Model with Individual base models.

TABLE VII. COMPARISON OF THE DEVELOPED MODEL WITH INDIVIDUAL BASE MODELS

Models	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
MLP	97.65	95.69	98.61
DT	92.97	78.47	100.00
SVM	67.39	0.00	100.00
Developed Ensemble	99.06	98.09	99.54

TABLE VIII. COMPARISON OF DEVELOPED ENSEMBLE MODEL WITH EXISTING MODELS

Author(s)	**Method /Technique used**	**Accuracy (%)**
Goa and Chen (2015)	Logistic Regression (LR) and ANN	85.09
Xiao et al., (2016)	Random Forest Model	83.10
Takeuchil et al., (2018)	ANN	71.6
Developed Model (2019)	Ensemble of DT, MLP, and SVM	99.06

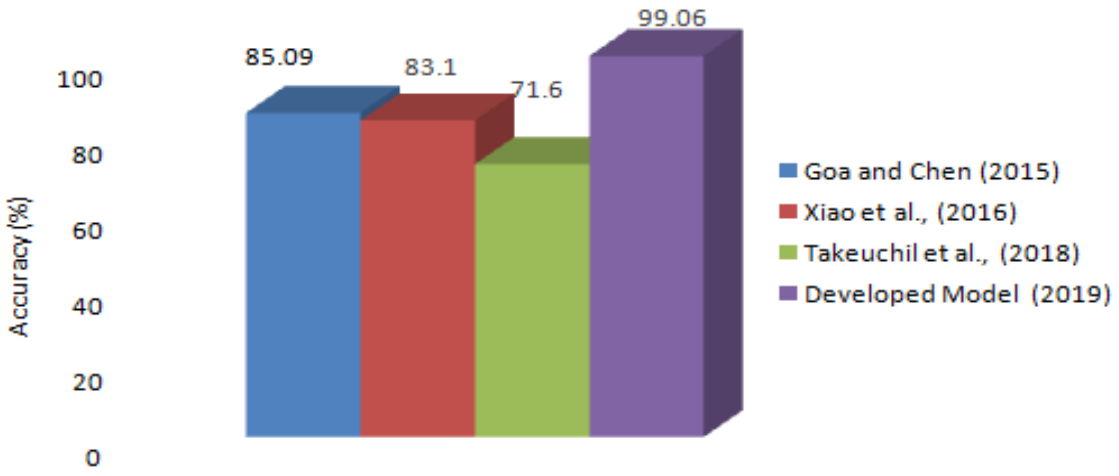

Fig 5. Comparison of Developed Ensemble Model with Existing Systems.

V. Conclusion

The developed model is revealed to be effective in detecting both non-prostate and prostate instances. Using sensitivity, specificity and accuracy as performance metrics, our result has shown a prediction accuracy of 99.06%, sensitivity of 98.09% and, specificity of 99.54%, which is a relative improvement on the existing systems. In other words, we have been able to significantly tackle issues of accuracy and sensitivity in the prediction of prostate cancer in men, using this ensemble model, which shows a relative improvement when compared to the individual base algorithms and some existing models.

References

[1] S. L. Win, Z. Z. Htike, F. Yusof, and I. A. Noorbatcha, "Cancer Recurence Prediction using Machine Learning," Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 2, pp. 11–20, 2014.

[2] A. Jemal et al., "Cancer statistics 2006," CA Cancer J. Clin., vol. 56, pp. 106–130, 2006.

[3] T. O. Akinremi, A. Adeniyi, A. Olutunde, A. Oduniyi, and C. N. Ogo, "Need for and relevance of prostate cancer screening in Nigeria," pp. 6–11, 2014.

[4] E. Alexandratou, V. Atlamazoglou, T. Thireou, and D. Yova, "Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading Eleni Alexandratou *, Vassilis Atlamazoglou and Trias Thireou George Ag," Int. J. Comput. Intell. Syst. Biol., vol. 1, no. 3, pp. 298–315, 2010.

[5] T. Takeuchi and K. R. Hospital, "Prediction of prostate cancer by deep learning with multilayer artificial neural," no. August, 2018.

[6] P. Leydon, F. Sullivan, and F. Jamaluddin, "Machine Learning in Prediction of Prostate Brachytherapy Rectal Dose Classes at Day 30," in Proceedings of the 17th Irish Machine Vision and Image Processing Conference, 2015, pp. 105–109.

[7] M. Gao, P. Bridgman, and S. Kumar, "Computer Aided Prostate Cancer Diagnosis Using Image Enhancement and JPEG2000," in Proceedings of SPIE Annual Meeting, 2003, vol. 5, no. August.

[8] S. W. D. Merriel, G. Funston, and W. Hamilton, "Prostate Cancer in Primary Care," Advances in Therapy, vol. 35, no. 9. Springer Healthcare, pp. 1285–1294, 2018.

[9] D. G. Bostwick and J. . Eble, "Urologic Surgical Pathology, Mosby – year book." 1997.

[10] O. Saidi, C. Cordon-Cardo, and J. Costa, "Technology insight: will systems pathology replace the pathologist," Nat. Clin. Pract. Urol., vol. 4, p. 39–45., 2007.

[11] O. E. Ernest, O. Awodele, and O. Ebiesuwa, "Early Detection and Diagnosis of Prostate Cancer using Artificial Intelligence Concept," Int. J. Comput. Appl., vol. 149, no. 6, pp. 42–46, 2016.

[12] A. O. Ige, A. O. Akingbesote, and A. O. Orogun, "Trust e-Market Environment : : A Review," Can. open Inf. Sci. Internet Technol. J., vol. 1, no. 1, pp. 1–9, 2019.

[13] L. E. Peterson, "Artificial neural network analysis of DNA microarray-based prostate cancer recurrence," in Computational Intelligence in Bioinformatics and Computational Biology, 2005, pp. 1–8.

[14] N. Iizuka, "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection," Lancet, no. 361, pp. 923–929, 2003.

[15] H. Zhao, S. Qi, and Q. Dong, "Predicting prostate cancer progression with penalized logistic regression model based on co-expressed genes," in 5th International Conference on BioMedical Engineering and Informatics, 2012.

[16] K. H. Gülkesen, İ. T. Köksal, S. Özdem, and O. Saka, "Prediction of prostate cancer using decision tree algorithm," Turkish J. Med. Sci., vol. 40, no. February 2009, pp. 681–686, 2010.

[17] P. Ge, F. Gao, and G. Chen, "Predictive models for prostate cancer based on logistic regression and artificial neural network.," in IEEE International Conference on Mechatronics and Automation (ICMA), 2015.

[18] K. Takeuchi, T., Hattori-Kato, M., Okuno, Y., Iwai, S., & Mikami, "Prediction of prostate cancer by deep learning with multilayer artificial neural network." 2018.