ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter127892020.pdf
https://doi.org/10.30534/ijeter/2020/127892020

A Machine Learning Based Improved LogisticRegression Method for Prostate Cancer Diagnosis

Mohammed Ismail B[1], P.Rajesh[2],Mansoor Alam[3]
[1]Professor, Department of Computer Science Engineering,
Koneru Lakshmaiah Education Foundation Vaddeswaram, A.P, India
mdismail@kluniversity.in
[2]Associate Professor, Department of Computer Science Engineering,
Koneru Lakshmaiah Education Foundation Vaddeswaram, A.P, India
rajesh.pleti@kluniversity.in
[3]Professor and Dept Chair Electrical Engineering
Northern Illinois University College of Engineering and Engineering Technology, DeKalb, IL, USA
malam1@niu.edu

ABSTRACT

The objective of this paper is to review and investigate machine learning methods and propose an improved Logistic regression method for Prostate Cancer (PrC) diagnosis and prediction. The paper compares proposed method with existing supervised classification techniques for a prostate cancer data set. An Improved Logistic Regression method is applied on patients vulnerable for PrC showing considerable improvement in prediction rate. The proposed methodincorporates clinical as well as tumor stages with patient ethnic characteristics. The comparative analysis of improved Logistic method show improvement on prediction accuracy rate and records a better Sensitivity and Specificity compared to other popular classification methods.

Key words: Machine Learning,Prostate cancer,Logistic Regression,Prediction rate,Specificity and Sensitivity

I.INTRODUCTION

Prostate Cancer (PrC) is a frequent and a regular cause of cancerous decease in males. In 2017the fresh cases of PrCregisteredare1,61,000causing 26,700 deaths in United States[1] alone.It is globally leading in seventh position for masculine deaths [2].PSA(Prostate Specific Antigen)[3] is the major screening test taken for PrC with needle biopsy having reasonable efficacy [4].Timely detection of PrC improves impermanence rates and leads to avoid over and ineffective treatment. Magnetic Resonance(MR) Imaging (MRI) is popularly used for cancer which purely depends on human examining experience. To assist and improve human readability machine learning computer aided methods are used. Trained and validated models are designed.

Machine learning techniques form an effective solution for prediction through training and testing phases. Classifiers play a vital role for analysing and classifying huge biological data into its class labels to discriminate PrC and a non PrC labels. Data mining rules with class labels improve learning model to grade critical decisions on diagnosis and perform better when assisted with examining through human radiomics. In the proposed work an improved classifier for learning model is implemented with logistic regression method for patient's prediction on PrC using data set [6].

II.LITERATURE REVIEW

Trans rectal Ultrasound was the popular method for prostate imaging, but had low Sensitivity and Specificity issues, later MRI is used for clinical examination [4] and assessment. This technique combined with diffusion weights for peripheral zones improved accuracy[7]. All the above methods require human intervention and specialists for premature analysis. Figure 1 depicts one such method workflow used with MRI scans. The machine learning and classification algorithms when supervised improve their prediction rate using statistical assessment with data labels. Modelling of a machine learning technique starts with labelling of existing data into different classes to categorize PrC stages and grade their severity in clinical decision making. The advantage of these methods is they are independent of internal understanding of PrC they only require valid samples to train and test for prediction. The technique uses machine pipelining for learning as shown in figure for a input sample MRI scan[8]. The approach used in this method uses a Multi parametric (Mp) MRI with machine and deep learning in 8 stages for validating the input MRI scan. The workflow of the model is stated with the following 8 stages.

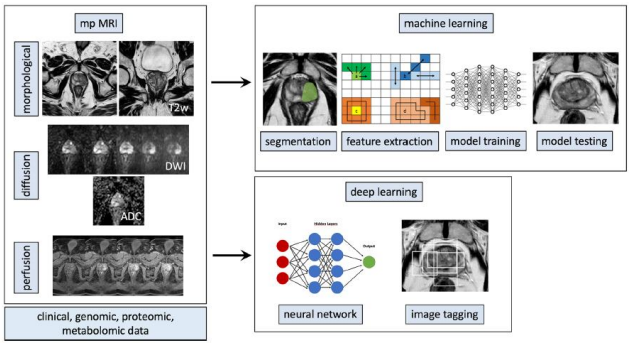

Figure 1: Work flow of Machine Learning Model[8]

a. Weighted sequence examination from MpMRI
b. Extraction and Segmentation of class labels
c. Pre-Processing and Filtering
d. Pattern Extraction of features
e. Integration of clinical and tumor stage data
f. Feature classification by suitable mining classifier
g. Training and Testing of model
h. Validate designed model for unknown data

Steps b and d uses machine learning classifiers for designing a Learning model, in which some of the classifiers like Random Forest, Nearest Neighbor, Decision tree and Logistic regression are popular.

The following literature explains the advantages and limitations of various classifying methods. Matsui[10] et al proposes Artificial Neural Network for Japan population data showing accuracy for patient inputs of age, PSA, tumor stages and Gleason score compared with Logistic and Decision tree methods. It did not have the staging advanced characteristics like TNM(Tumor Node Metastasis) and AJCC (American Joint Committee on Cancer) analysis. A support vector machine method was implemented by Olivier et al[11] for pathological stages based on Bayesian function. Another method on Taiwan population by Tsao et al[12]is implemented using Body Mass index (BMI) and biopsy for k means Nearest neighbor and logistic regression. Maria [13] proposed a fuzzy based system for prediction on PrC, Castanho et al proposed [14] Genetic Algorithm(GA) on PrC stages.Jae Kwon et al[16] used Korean data set for PrC prediction using PSO model but did not consider ethnic origin and cystitis disorders[17-20]. It used binary recursion with Gini Index for prediction. The literature discussed above has constraints sometimes on approach and sometimes on clinical and cancer stages data not giving better accuracy. The proposed Improved Logistic Regression(ILR) Classifier overcomes these constraints of accuracy and performance specifications with more data attributes[20-24].
Further the work presents proposed Improved Logistic Classifier implemented on data set in section 3 and 4. Section5 records obtained results and comparisons with existing methods. Section 6 concludes with future research directions.

III. PROPOSED IMPROVED LOGISTIC REGRESSION METHOD

Logistic Regression(L.R) classifier method has an advantage of performing better on data having less number of records[9] but losses its accuracy with huge and redundant sets. The proposed method helps L.R in optimizing and extracting features for selection during training phase of the learning model[25-28]. The distribution of this method is represented as shown in equation 1.

$$p(b|a) = \partial(\langle x, a\rangle)^b(1-\partial(\langle x, a\rangle)^{1-b}) \quad (1)$$

∂in equation 2 represents sigmoid function for a and b outputs

$$\partial(t) = \frac{1}{1+e^{-t}} \quad (2)$$

$x = \{x_1, x_2, x_3 \ldots .x_n\}$ presents a series of unknown coefficient learnt from existing set. Proposed ILR method takes event probability coefficient to obtain Positive Predictive Value and Specificity from the data available. The event occurrence indicates largest likelihood estimation probability from the data. Later a derivative loss function on negative likelihood is approximated by eliminatingless important features. It is labelled as Least Absolute Shrinkage and Selection Operator(LASSO) [18] coefficient which minimizes loss function shown in equation 3.

$$Minimum\ (Log\ Loss\ Function + \lambda \sum_{n=1}^{d} |x_n|) \quad (3)$$

Where

$$\log loss\ function = -\frac{1}{n}\sum_{i=1}^{n}\left(b_i log\ p(b_i) + (1-b_i)\log(1-p(b_i))\right) \quad (4)$$

The advantage of ILR method is it does not depend on shrinkage factor instead uses log loss function for closer level of data set.Figure 2 illustrates a learning model for proposed Improved LR method [29-31].

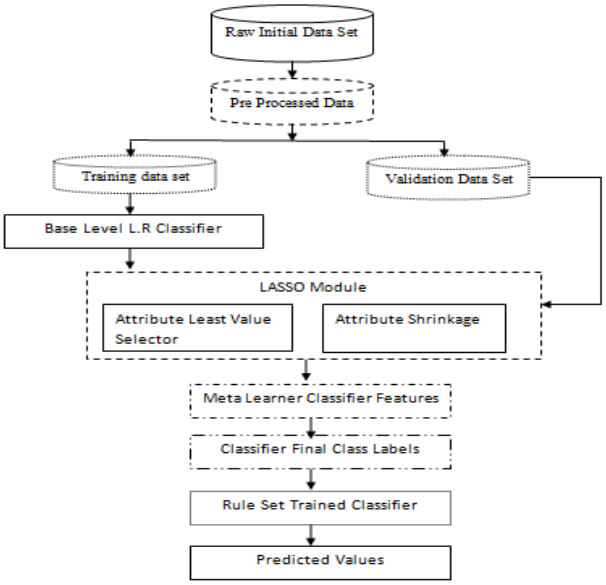

Figure 2:.Learning Model for proposed Improved LR Method

IV.MATERIALS AND DATA SET

Improved Logistic Regression(ILR)method is implemented on PrC classification for the data partially collected Zhou W. et al[6] for Prostate prediction in males with risk of PrC [32-33]. Sometimes high risk of PrC may occur due to Cytos is disorders, family history or personal health habits. The utilized data set is of 378in which 186 have PrC and 192 set have no PrC [6].

Table 1 shows the attributes of the data implemented and analyzed. For each PrC subject 10characteristics are taken where 2 of them describe cancer stage and other 8 describe PrC risk aspects like BMI, Age Factor, Smoking, Origin of ancestors (ethnic), Nutrition intake, PrC family history, PSA Test on Blood and cystitis disorders. The cancer stage attributes considered are Tumor NodeMetast as is (TNM) and American Joint Committee on Cancer (AJCC) stage.
TNM stage defines tumor node, Number of lymphs and Meta size body part spread.AJCC stage defines progression of cancer spread according to American joint committee norms.PSA blood test on antigen on prostate gland[22].

Table 1: PcR Data Attributes Descriptors

Attributes	**Description**
Age	PcR risk Patient age
Smoking habit	Smoking or non smoking
BMI	Male Body Mass Index
Ethnic	Geographical region or Origin of US and European countries
Food intake	Food fat intake low (20%) Moderate (20–30)% and High more than 30%
TNM	Notation for Tumour Node Metastasis description
PCA	Family history on PrC
AJCC	American Joint Committee description of Four classifications II A &B,III and IV stages
PSA	Three Level Prostatespecific antigen, <10, 10 to 20 and > than 10 ng/mL
Cystitis disorders	Urinary tract infections

V.EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed Improved Logistic Regression method with other comparative classification techniques is implemented on partial data set [6]for classifier methods of Decision Tree, K Nearest Neighbour, Random Forest and Logistic Regression using Matlab R2018 on Intel i7 processor with 8 GB memory. Table 2 presents the data set division and results with performance predictors are depicted in table 3.Comparative analysis of performance metrics of the proposed Improved Logistic Regression method is tabulated in table 3with other existing techniques. Parameters of Accuracy (Ac),Sensitivity(Sen) and Specificity(Spe) are represented by equ(4),(5) and (6) respectively. *T.P* represents True Positive, *F.P* represents False Positive, *F.N* represents False Negative and *T.N* represents True Negative values. Prediction is done through two values called as, Positive Prediction (P.P) and Negative Prediction (N.P) shown in equ (7) and (8).

$$Ac = \frac{T.P+T.N}{T.P+T.N+F.P+F.N} \quad (4)$$

Table 2 : Division of Dataset

Training Data Samples	
Classification Label	*Sample No*
Normal Data Set	192
Cancerous Data Set	186
Total Data Set	378

Table 3: .Relative Performance Metrics and Predictors

Classifier Technique	**Ac(%)**	**Sen(%)**	**Spe(%)**	**PPV(%)**	**NPV(%)**
Decision Tree	75.13	66.92	79.43	63.04	82.08
K-Nearest Neighbour	78.83	74.48	81.54	71.52	83.70
Random Forest	79.31	77.24	80.60	71.33	85
Logistic Regression	79.62	79.72	79.56	71.51	85.91
Improved Logistic Regression*	81.74	82.27	81.36	76.02	86.47

$$Sen = \frac{T.P}{T.P+F.N} \quad (5)$$

$$Spe = \frac{T.N}{T.N+F.P} \quad (6)$$

$$P.P.V = \frac{T.P}{T.P+F.P} \quad (7)$$

$$N.P.V = \frac{T.N}{T.N+F.N} \quad (8)$$

The proposed Improved logistic regression method gives Sen of 82 % and Spe of 81 % with enhanced accuracy of 2.2%Figure 3 and 4 shows comparison plots for Ac and Spe.

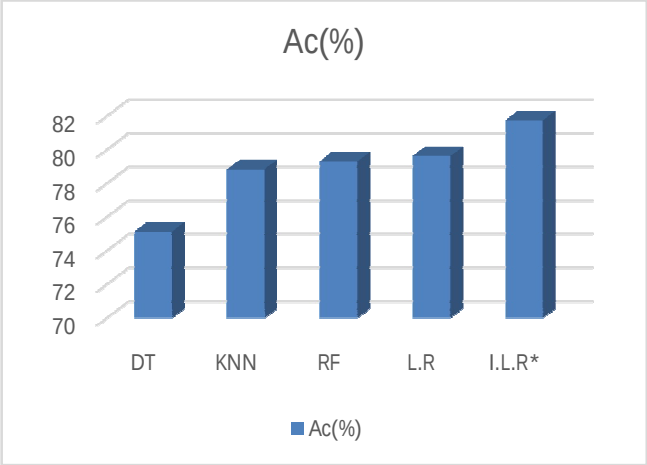

Figure 3: ComparisionPlotsof Accuracy(Ac%)

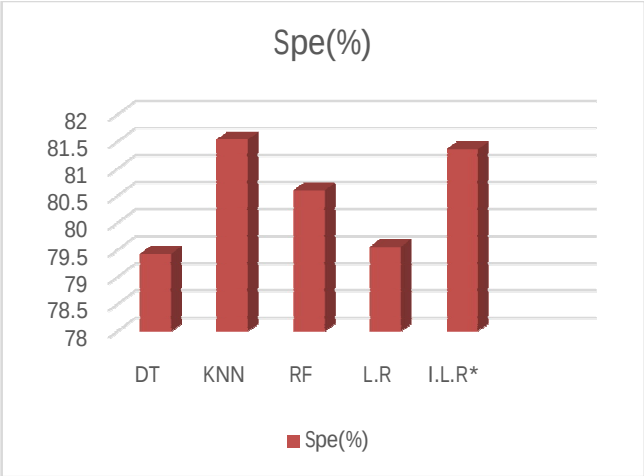

Figure 4: Comparative Plotsof Specificity(Spe%)

VI.CONCLUSIONS AND FUTURE ENHANCEMENTS

This work shows successive implementation of an Improved LR method based on a Machine Learning classifier for a PcR data set. Results of proposed method is compared with 5 popular methods of classification. The implementation results prove a better predicting parameter for the proposed LR with LASSO coefficient. The positive predictive value P.P.V is improved by 5% with the conventional LR method proving the proposed method performs better in predicting Prostate cancer. The significance of proposed method is its data attributes forming a blend between clinical and tumour stages used in prediction of PcR. By improving accuracy and prediction this method gives a better chance to assist the radiologists with reduced risk of over or under diagnosis. This technique can further be combined with MRI scan optimization methods to avoid advance stages in PcR by predicting in the premature stages itself. The usability of this technique can further be enhanced on medical devices interfaced to a GUI for a regular self-examining by patient.

REFERENCES

[1] Siegel RL, Miller KD, **Jemal A. Cancer Statistics,2017.** CA Cancer J Clin.;67(1):pp. 7-30 2017

[2] Fitzmaurice C, Allen C,**Global Burden of Disease Cancer Global, regional, and national cancer incidence, mortality, yearsof life lost, years lived with disability, and disability-adjusted life years for 32 cancer groups, 1990 to 2015: a systematic analysis for the Global Burden of Disease Study**. JAMA Oncol. ;Volume 3(4):pp .524-548.2017

[3] Mettlin C, Jones G, Averette H, Gusberg SB, Murphy GP. **Defining and updating the American Cancer Society guidelines for thecancer-related checkup: prostate and endometrial cancers**. CACancer J Clin. Volume 43(1):pp.42-46 1993

[4] Pinsky PF, Prorok PC, Yu K, **Extended mortality results forprostate cancer screening in the PLCO trial with median follow up of 15 years. Cancer**. Volume 123(4):pp.592-599 2017

[5] Sagar Imambi S, Vidyullatha.P, **Explore Big Data and forecasting future values using univariate ARIMA model in R**, International Journal of Engineering and Technology(UAE), 2018

[6] Zhou, W., Zhu, M., Gui, M., Huang, L., Long, Z., Wang, L.,Chen, H., Yin, Y., Jiang, X., Dai, Y., Tang, Y., He, L., Zhong, K.:**Peripheral blood mitochondrial DNA copy number is associated with prostate cancer risk and tumor burden. PLoS** 2014

[7] Barentsz JO, Weinreb JC, Verma S **Synopsis of the PI-RADS v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use.** Eur Urol 69 pp. 41–49 2016

[8] Renato Cuocolo, Maria BrunellaCipullo, Arnaldo Stanzione , Lorenzo Ugga, Valeria Romeo, Leonardo Radice, Arturo Brunetti and Massimo Imbriaco **Machine learning applications in prostate cancer magnetic resonance imaging** European Radiology Experimental 3 pp :35 2019

[9] Breslow N, Chan CW, Dhom G. **Latent carcinoma of prostate at autopsy in seven areas.** The International Agency for Research on Cancer, Lyons, France. Int J Cancer Volume 20(5):pp..680-8 Nov 1977

[10] Matsui Y, Egawa S, Tsukayama C **Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the**

Japanese population. Japan Journal Clin Oncol. Volume 32(12):pp.530-535 2002

[11] Olivier RC, John M, Robert L, Thomas L, Sam M, James N.**Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers**. ArtifIntell Med. Volume 55(1) pp.25-35.

[12] Tsao CW, Liu CY, Cha TL. **Artificial neural network for predicting pathological stage of clinically localized prostate cancer in a Taiwanese population**. J Chin Med Assoc. volume 77 pp.513-518.2014

[13] Maria J, de PC, Laecio C, de B, Akebo Y, Laercio LV. **Fuzzy expert system: an example in prostate cancer**. Appl Math Comput.Volume 202(1) pp.78-85 2008

[14] Castanho MJP, Hernandes F, De R´e AM, **Fuzzy expert system for predicting pathological stage of prostate cancer**. Expert Systems Appl. Volume 40(2):pp 466-470 2013

[15] Vidyullatha, P, Rajeswara Rao D, **Machine learning techniques on multidimensional curve fitting data based on r- square and chi-square methods**, International Journal of Electrical and Computer Engineering, Volume 6, Issue 3, pp. 974-979.June 2016

[16] Jae Kwon Kim, Mi Jung Rho,Jong SikLee,Yong Hyun Park, Ji Youl Lee, In Young **Choi Improved Prediction of the Pathologic Stage of Patient With Prostate Cancer Using the CART–PSO Optimization Analysis in the Korean Population**, Technology in Cancer Research & Treatment, Volume. 16(6) 740–748 2017

[17] Mohammad Ismail, V.Harsha Vardhan, V.Aditya Mounika, K.Surya Padmini **An Effective Heart Disease Prediction Method Using Artificial Neural Network** International Journal of Innovative Technology and Exploring Engineering' Volume 8 pp.1529-1532, June 2019

[18] G. James, D. Witten, T. Hastie, **An introduction to Statistical Learning with Applications in R**, Springer, New York, (2013).

[19] Mohammed Ismail .B,T. BhaskaraReddy,B. Eswara Reddy**Spiral Architecture Based Hybrid Fractal Image Compression** IEEE International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)Dec 2016

[20] Devisetty, Vidyullatha. P,**Sentiment analysis of tweets using rapid miner tool,** International Journal of Innovative Technology and Exploring Engineering,2019

[21] Nikhila, Sagar Imambi, S .**Sequence labeling using deep neural nets** , International Journal of Advanced Trends in Computer Science and Engineering 2019

[22] S. Rizwana, S.SagarImambi **Enhanced biomedical data modeling using unsupervised probabilistic machine learning technique** International Journal of Recent Technology and Engineering, volume 7.No 6. pp. 579-582 2019

[23] S.SagarImambi , K.Bhagavan, ShahanaBano, **A compressive survey on different image processing techniques to identify the brain tumor** , International Journal of Engineering and Technology Volume 7.No 2.7 pp:1081-1084 2018

[24] Rahul Shahne, Mohammed Ismail, CSR Prabhu **Survey on Deep Learning Techniques for Prognosis and Diagnosis of Cancer from Microarray Gene Expression Data** Journal of computational and theoretical Nanoscience Volume 16 (12), pp.5078-5088, Dec 2019

[25] K.Naga Lakshmi, Y. Kishore Reddy, M. Kireeti, T.Swathi, Mohammad Ismail,**Design and Implementation of Student Chat Bot using AIML and LSA** International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 (6)**,** pp**.**1742-1746, April 2019.

[26] Mohammad Ismail, V.Harsha Vardhan, V.Aditya Mounika, K.Surya Padmini **An Effective Heart Disease Prediction Method Using Artificial Neural Network** International Journal of Innovative Technology and Exploring Engineering volume 8 (8),pp. 1529-1532, **June 2019**

[27] K.Srinivas, Mohammed Ismail.B **Testcase Prioritization With Special Emphasis On Automation Testing Using Hybrid Framework** Journal of Theoretical and Applied Information Technology volume 96(13) 4180-4190 July 2018

[28] P R Anisha, B Vijaya **Babu EBPS: Effective Method for Early Breast Cancer Prediction using Wisconsin Breast Cancer Dataset** International Journal of Innovative Technology and Exploring Engineering Volume-8 (2) December, 2018

[29] Sowjanya V, Divyambica CH, Gopinath P, Vamsidhar M, B.Vijaya Babu, **Improved Prediction of Diabetes based on Glucose Levels in blood using Data Science** Algorithms International Journal of Engineering and Advanced Technology, Volume-8 Issue-4, April 2019

[30] Amarendra K, Venkata Naresh Mandhala, B.Chetangupta, G.Geetha Sudheshna, V.Venkata Anusha **Image Steganography Using LSB** International Journal Of Scientific & Technology Research Volume 8, Issue 12, pp 906-909 December 2019

[31] Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail **Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms** International Journal of Emerging Trends in Engineering Research Volume 8. No. 5, pp 2199-2204 **May 2020**

[32] Babitha D , Mohammed Ismail, Subrata Chowdhury **Automated Road Safety Surveillance System using Hybrid CNN-LSTM Approach** International Journal of Advanced Trends in Computer Science and Engineering Volume 9 No.2, pp 1767-1773**April 2020**

[33] Mohammad Ismail, Sreemanth Pisupati **Image Registration Method for Satellite Image Sensing using Feature based Techniques** International Journal of Advanced Trends in Computer Science and Engineering Volume 9(1),pp. 490-593 **Feb 2020**