# A Machine Learning Classification Technique for Predicting Prostate Cancer

Mohammed Ismail. B
Professor, Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation Deemed to be University, A.P, India
mdismail@kluniversity.in

Mansoor Alam
Professor and Dept Chair Electrical Engineering
Northern Illinois University College of Engineering and Engineering Technology, DeKalb, IL, USA
malam1@niu.edu

Mansour Tahernezhadi
Senior Associate Dean and Professor Electrical Engineering
Northern Illinois University College of Engineering and Engineering Technology, DeKalb, IL, USA
mtaherne@niu.edu

Hari Kiran Vege
Professor and Head, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation Deemed to be University, A.P, India
hari.vege@kluniversity.in

P. Rajesh
Associate Professor, Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation Deemed to be University, A.P, India
rajesh.pleti@kluniversity.in

***Abstract*—This paper presents and validates various classification techniques on supervised machine learning (ML) for predicting prostate cancer. A modified Logistic Regression (LR) classifier is proposed and implemented on patients who are susceptible to prostate cancer. The proposed classification technique uses both clinical and tumor stage characteristics. Clinical characteristics considered are BMI, age, cystitis infections, and smoking history. Tumor stage characteristics are stages of Tumor Node Metastasis (TNM), American Joint Committee on Cancer (AJCC) and Prostate Specific Antigen (PCA). Results obtained show improvement in accuracy and positive prediction value (PPV) as compared to existing classifiers. Results are compared and validated with performance measures of Specificity (Sp) and Sensitivity (Se), recording a minimum of 3% improvement in Pc prediction accuracy. The implemented ML classification technique also shows a clinical impact on Pc diagnosis with a 4 % improvement in Sp.**

***Keywords—Machine Learning,Classification methods,Prostate cancer,MRI, Specificity and Sensitivity***

## I. INTRODUCTION

Prostate cancer (Pc) is a regular and normal cause of cancer disease in men. In 2017 the new cases of pc reported were close to 1,61,000, with nearly 26,700 deaths occurred in the United States of America[1] itself. It is worldwide also reported as the seventh leading cause of male deaths [2]. The present popular test for Pc diagnosis is Prostate-Specific Antigen (PSA) [3] and later screening by needle biopsy. Nevertheless, this method of testing and screening recently has raised questions on its reduced efficacy [4].PSA also has an elevated benign condition, with 33% of men have false-positive results but tested true positive in needle biopsy[5].

Early detection of Pc can improve mortality rates, and less clinically significant cases can avoid over diagnosis and ineffective treatment. In this case of early detection, Magnetic Resonance Imaging (MRI) in specific Multi parametric (mp) MRI is increasingly being used for diagnosis and prediction. The technique of mpMRI, though it improves the accuracy of prediction dependents on diagnostic human reader experience. Hence the necessity of a computer-aided system that can help human readability arises. Machine learning (ML) or deep learning algorithms find such a solution wherein more effective predictions can be made through trained and validated learning models. For a machine learning model design, classifiers and data mining techniques play a significant role in classifying massive data into class labels. ML classifiers can analyze an enormous quantity of radiological data and label them more quickly and effectively than human radiomic study. ML classifiers assist in discriminating Pc of high and low grades to make critical clinical decisions and consequently avoid overtreatment. In this paper, a classification technique for machine learning based on logistic regression is proposed for predicting Pc from the existing clinical data [6].

## II. LITERATURE SURVEY

Earlier, for Pc assessments, Transrectal Ultrasound (TU) was the primary imaging method. It suffered from low Sensitivity (Se) and Specificity (Sp) rates. Currently, mpMRI is proving an accurate method [4] for regular clinical examination for Pc assessment. It improves further when combined with diffusion-weighted imaging (DWI) for the peripheral zone (PZ) lesions [7]. All these imaging diagnoses require human experts and specialized radiologists to suggest early Pc diagnosis to avoid over or under treatment. Hence learning algorithms using computers are used for early prediction by training the system with massive exposure to existing and new MRI data. Fig 1 shows a workflow for machine learning and deep learning approach in prostate cancer (Pc) MR Imaging.

The machine learning (ML) algorithms learn and improve over time with trained data labelling, called supervised learning. This supervised machine learning uses classifiers for

978-1-7281-5317-9/20/$31.00 ©2020 IEEE

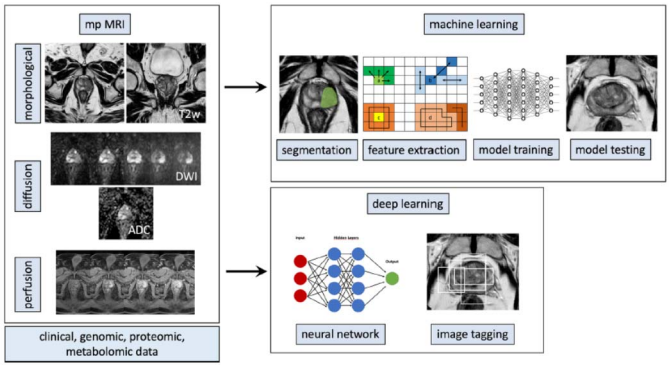


Fig. 1. Work flow for machine and deep learning approach in prostate cancer MR Imaging[8]

statistical analysis on a wide range of data mining class labels. The classification techniques in ML help in labelling the existing data prior to the start of training learning process. These classifiers are used to discriminate Pc stage or grade to make critical clinical decisions. ML methods usually do not need an internal understanding of Pc. It only requires obtaining input/output samples to train itself for improving prediction from clinical data having high non linearity.ML techniques use learning pipelines as shown in fig 1 for Pc and analyze the MRI in 8 steps as follows.

i. Examine mpMRI with weighted sequences
ii. Apply classifiers for the extracted image data and segment the distinct label classes.
iii. Pre-process image features and filter for creating a shrink or reduced data.
iv. Extract features from the MRI image pattern obtained.
v. Integrate radiology data with clinical attributes of the patient.
vi. Feature classify above integrated data to a labeled class of interest through a mining classifier.
vii. Train and Test the algorithm with obtained data in the ratio of 80 to 20, respectively.
viii. Cross validate the model on the external or newly arrived data.

In the above steps classification techniques are used in step ii and vi. The existing popular ML classifiers are Artificial Neural Networks (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) , Random Forest(RF), Logistic Regression (LR), Decision Tree(DT)etc.

All these methods are described and implemented in the following literature review with their pros and cons. Matsui [10] et al proposed ANN using mining techniques for clinical data in the Japanese population. It shows improvement in accuracy compared to LR and SVM with input characteristics of age, Gleason Score (GS), Prostate Specific Antigen (PSA) serum and density and tumour stage. It lagged the advance characteristics of Tumor, Node, Metastasis (TNM), American Joint Committee on Cancer (AJCC) stage disease progression.Olivier et a l[11] used ML method on SVM and function Bayesian to classify Pc pathology stage. Tsao et al [12] implemented ANN to data mine Pc stages in Taiwanese men. Body Mass Index (BMI), Gs biopsy was used as input parameters results were improved for LR and KNN. A fuzzy predicting system proposed by Maria et al reduced uncertainty in existing methods [13]. Castanho et al [14] implemented Genetic Algorithm (GA) with fuzzy logic for predicting Pc pathology stages. Mohammed et al [16-19] proposed a cuckoo search method for fractal compression. Classification based on decision Tree-is proposed by Jae Kwon Kim et al on the Korean population using Particle Swarm Optimization (PSO) Model[20] but the characteristics considered were limited retinal cystitis and ethnic origin of patients were not considered. It used the Gini Index analysis on binary recursion. All the contemporary literature discussed has some limitations on the clinical parameters selected and its classifier approach [21-23]. The proposed classifier based on the modified logistic regression technique tends to improve accuracy compared with existing classifiers and also has scope for considering more clinical attributes [24-26].

The paper is further arranged in the following sections. Section III depicts the proposed modified LR Classifier. Section IV presents the materials and data set used for implementation. Results and discussions with comparative analysis are presented in section V. Finally, section VI gives conclusions and future scope on the investigated methods and obtained results.

## III. Proposed Method

Learning algorithms based on data mining classifiers play an essential role in any machine learning model during the training phase. These classifiers help in parameter optimization and feature selection of the training data set. One such popular method for classification is Logistic Regression (LR) performing better on datasets with few records. It is an analytical method for modeling event probability. Its distribution is represented in equation 1.

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

Where $\partial$ is a sigmoid function given by equation 2 for x and y output and target variables respectively

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

$q = \{q_1,q_2,q_3 \ldots .q_n\}$ represents a set of unknown coefficients which will be learned from existing data.

The proposed method uses the event probability to find PPV and Sp. The possibility of event occurrence is indicated by the probability of the highest likelihood estimation. Next, a derivative function of negative likelihood is estimated by removing unimportant features. A coefficient called Least Absolute Shrinkage and Selection Operator (LASSO) [22] is used to minimize the Loss function from the data equation 3.

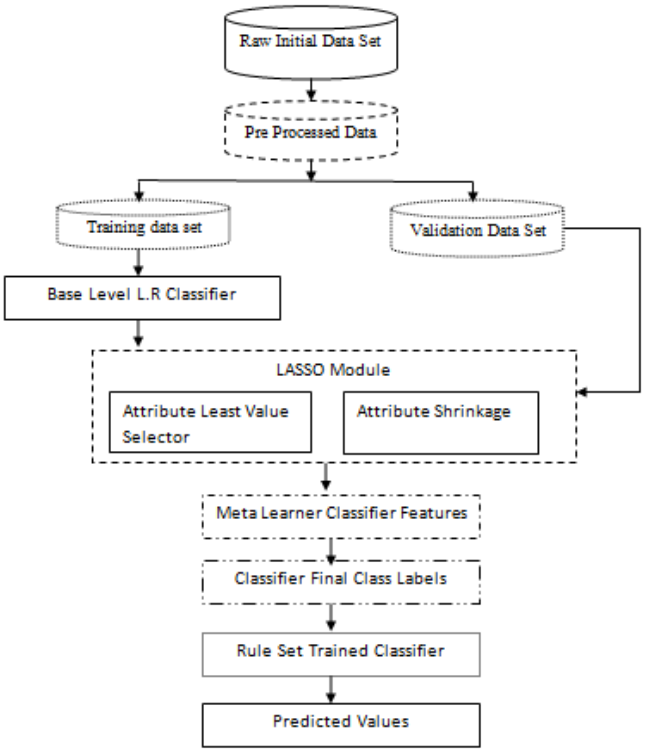


Fig. 2. Process learning flow for the proposed LR classification technique

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

Where,

$$Loss\ Function = -\frac{1}{n}\sum_{i=1}^{n}\left( y_i log\ p(y_i) + (1-y_i)\log(1-p(y_i))\right) \quad (4)$$

The proposed method does not depend on the shrinkage factor as that of conventional LR. It uses log loss function to minimize and has no closed levels of data attributes. Fig 2 shows a learning process for the proposed modified LR technique.

## IV. Materials and Methods

The proposed Modified Logistic Regression (MLR) classification technique is performed on the prostate cancer identification from the data set Zhou W. et al [6]. This identification study is beneficial for males having a high risk of Pc due to personal health issues or family cancer history. The men data samples used for classification are 387 out of 188 have Pc, and 190 samples do not have Pc[6].

Table 1 shows the description of the dataset attributes used here for analysis. For every Patient, 12 attributes are considered out of which 3 describes cancer stages, and other 9 describes Pc risk factors. Factors considered here are age, smoking habits, BMI, ethnic origin, Diet intake, Pc family history, PSA (Prostate Specific Antigen) blood test, cystitis history and mtDNA-CN (mitochondrial DNA copy number). The three attributes which strongly describe the cancer stages used in our analysis are Tumor, Node, Metastasis (TNM), American Joint Committee on Cancer (AJCC) stage disease progression and Gleason score (Gs) factor.

Tumor Node Metastasis (TNM) stages describe tumor spread on body parts with lymph nodes. N represents the number of lymph nodes with Cancer and M represents Meta size of cancer spread from the primary part to other parts of the body. AJCC stage is an important cancer staging system to determine the progression of disease proposed by the American joint committee. Prostate Specific Antigen (PSA) is also the most common blood test done for initial analysis to check protein production by malignant prostate gland.

TABLE I. PC DATASET CHARACTERISTIC DESCRIPTION

| Characteristic | Description |
|---|---|
| Age | Male Patient's age in years |
| Smoking habits | Smoking ever or never smoking |
| BMI | Body Mass Index in kg/m$^2$ |
| Ethnic origin | Clinical Pc depending on origin or geographical regions of USA and Europe Yes or No |
| Diet intake | Percentage of fat intake energy per day classified as low for 20% Moderate for range of 20% – 30% and high for 30% |
| TNM | Tumor, Node, Metastasis. notional system describing tumour stage |
| PCA Family history | Family history of prostate cancer |
| AJCC | Four stage II A , II B ,III and IV Classification on cancer progression by American Joint Committee On Cancer. |
| PSA | Prostate-specific antigen 3 level, less than 10,between 10 to 20 and greater than 10 ng/mL |
| Cystitis history | Cystoscopy or urinary tract infection (UTI) history Yes or No |
| mtDNA-CN | Mitochondrial DNA (mtDNA) copy number is a critical component of overall mitochondrial health |
| Gs | Grading system called Gleason score number with 3 stages first stage between 5 to 6, second at 7 and third between 8 to 10 with increasing order of aggressiveness |

Gleason score is a biopsy test with a grading range of low to high abnormal tissue deciding aggressiveness of pc. Along with Age, BMI and diet of Patient, the data considered here includes Cystitis history of urological or body infections. The other important characteristic considered in the study is the ethnic geographical origin. Breslow N et al [9][15] says that Pc occurrence rate is high in the US and Northern region of Europe compared to the south or East Asia and increases if these people move to the USA. Finally, mtDNA also plays a critical role in the data classification characteristic representing ancestral origin.

## V. Experimental Results and Discussions

The data split from the dataset [6] used for implementation is as per Table 2. The classifiers of DT, ANN, KNN, SVM, RF, LR and proposed MLR are implemented for the above data using Matlab R2018b on Intel core i7 processor with 16 GB RAM. Obtained results and tabulated, as shown in Table 3.The performance measures [12] of the proposed modified L.R (M.L.R)* classifier is validated with indicators of

TABLE II. DATA SEPARATION FROM PC DATASET

| **Training Data Set** | |
|---|---|
| *Classification division* | *Sample Number* |
| Normal Set | 190 |
| Cancer Affected Set | 188 |
| Total | 378 |
| | |

TABLE III. COMPARATIVE PERFORMANCE MEASURES AND PREDICTORS

| **Classifier** | **Acc(%)** | **Se(%)** | **Sp(%)** | **PPV(%)** | **NPV(%)** |
|---|---|---|---|---|---|
| DT | 77.95 | 68.49 | 83.03 | 68.47 | 83.05 |
| ANN | 93.86 | 88.34 | 96.75 | 93.45 | 94.06 |
| KNN | 78.75 | 73.46 | 82.11 | 72.32 | 82.94 |
| SVM | 89.77 | 85.74 | 92.16 | 86.65 | 91.59 |
| RF | 92.84 | 85.83 | 96.69 | 93.45 | 92.54 |
| L.R | 91.99 | 83.05 | 96.92 | 93.71 | 91.20 |
| | 96.86 | 95.50 | 98.39 | 96.91 | 96.83 |

Accuracy (Acc), Sensitivity (Se) and Specificity (Se) represented by equations (4),(5) and (6) respectively.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Se = \frac{TP}{TP+FN} \quad (5)$$

$$Sp = \frac{TN}{TN+FP} \quad (6)$$

*TP, FP, FN, TN* represents true positive, false positive, false negative and true negative respectively.

$$PPV = \frac{TP}{TP+FP} \quad (7)$$

$$NPV = \frac{TN}{TN+FN} \quad (8)$$

The predictors, Positive Prediction Value (PPV) and Negative Prediction Value (NPV) calculation are done as per the equations (7) and (8). The above measures are compared for the proposed and existing ML classifiers. Obtained results are tabulated in Table 3 and are compared with the other existing current methods. Table 3 signifies Se the ratio of correctly classified positive samples from all existing positive samples, and Sp indicates the ratio of correctly classified negative samples from all existing negative samples. Similarly, Acc computes the ratio of all correctly classified samples from all existing samples. Table 4 shows confusion matrix for the proposed modified L.R. The result obtained implies that in 193 total men tested 184 had Pc and out of 194 showing some symptoms 191 did not have Pc. The obtained results depicts the level of Se as 96% and Sp as 98% which improves the accuracy of classifier by at least 3% as compared to the existing methods. Figure 3 and 4 show comparative charts for Accuracy and Specificity.

## VI. CONCLUSIONS AND FUTURE SCOPE

This paper has implemented a modified LR classifier, which predicts Pc better by a minimum of 3 % compared to seven popular machine learning classifier methods.

TABLE IV. CONFUSION MATRIX FOR PROPOSED MLR CLASSIFIER

| | **Pc** | **Normal** | **Total** |
|---|---|---|---|
| **Pc** | 9 | 184 | 193 |
| **Normal** | 191 | 3 | 194 |
| | Se 96% | Sp 98% | |

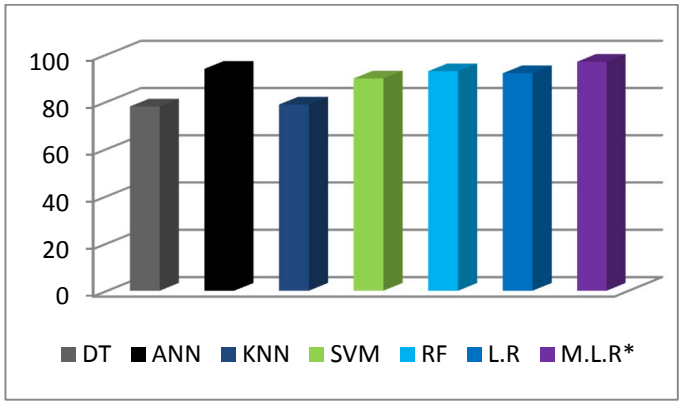


Fig. 3. Comparaitive chart for Accuracy (Acc%)

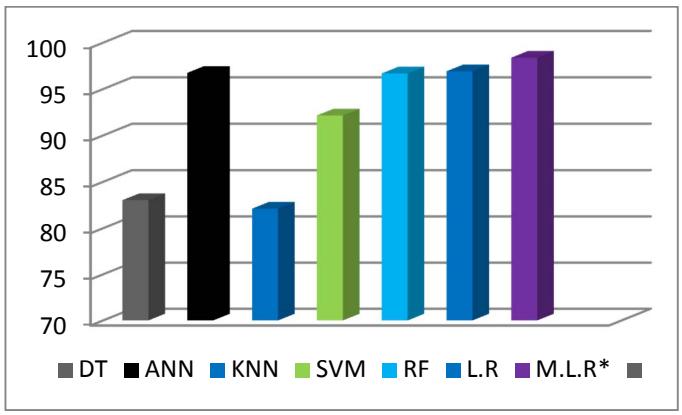


Fig. 4. Comparative chart for Specificity (Sp%)

Comparison of performance measures like Se and Sp with Predictive parameters PPV and NPV prove that the technique implemented here is better for predicting a patient's susceptibility for getting Pc. Our method is implemented considering a variety of patient characteristics, including clinical data, prostate stages, and urinary tract disorders that avoids the risk of more regular biopsies and tests. The proposed method can successfully decrease unnecessary over and under diagnosis of Pc. The proposed method's implementation shows the highest accuracy rate of 96.6%, leading to specificity (sp) of 98%. This method can be further improved and used on mpMRI image stages to enhance the diagnosis techniques guiding radiologists to analyze and caution patients or doctors regarding advance Pc stages and it's prematurity. The applicability of this method can lead to further development of a computerized tool or a GUI on medical devices, which can be faster, better, and easier to use regularly as a clinical practice by the patient himself.

## REFERENCES


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

[2] Fitzmaurice C, Allen C,Global Burden of Disease Cancer Global, regional, and national cancer incidence, mortality, years of 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. 2017,3(4):524-548.

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

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

[5] Schroder FH, van der Cruijsen-Koeter I, de Koning HJ, Vis AN, Hoedemaeker RF, Kranse R. Prostate cancer detection at low prostate specific antigen. J Urol. 2000;163(3):806-812.

[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 et al (2016) Synopsis of the PI-RADS v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use. Eur Urol 69:41–49

[8] Renato Cuocolo, Maria Brunella Cipullo, 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 (2019) 3:35

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

[10] Matsui Y, Egawa S, Tsukayama C, et al. Artificial neural network analysis for predicting pathological stage of clinically localizedprostate cancer in the Japanese population. Jpn J Clin Oncol. 2002;32(12):530-535

[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. Artif Intell Med. 55(1):25-35.

[12] Tsao CW, Liu CY, Cha TL, et al. Artificial neural network for predicting pathological stage of clinically localized prostate cancer in a Taiwanese population. J Chin Med Assoc. 2014;77(10):513-518.

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

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

[15] Sreemanth Pisupati, Mohammed Ismail.B "Image Registration Method for Satellite Image Sensing using Feature based Techniques" International Journal of Advanced Trends in Computer Science and Engineering 2020;9(1);490-593

[16] Ghousia Anjum, T.Bhaskara Reddy, Mohammed Ismail, Alam, M., Tahernezhadi, M. "Variable Block Size Hybrid Fractal Technique for Image Compression" Proceedings IEEE 6th International Conference on Advanced Computing & Communication Systems 2020

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

[18] Moulana Mohammed, M. Venkata Sai Sowmya, Y. Akhila, B. Naga MeganaVisual Modeling of Data using Convolutional Neural Networks International Journal of Engineering and Advanced Technology 9 (1) 2019 ;4938-4942

[19] 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 2020 ,8 (5) ;2199-2204

[20] Jae Kwon Kim, Mi Jung Rho,Jong Sik Lee,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 2017, Vol. 16(6); 740–748

[21] 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' 8 (8),2019 ;1529-1532

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

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

[24] Mohammed Ismail .B, Dr. Mahaboob basha shaik, Dr. B. Eswara Reddy "Improved Fractal Image Compression Using Range Block Size" Proceedings of IEEE International Conference on Computer Graphics, Vision and Information Security; 2015

[25] Rajendra Prasad, K., Mohammed, M. & Noorullah, R.M. Correction to: Visual topic models for healthcare data clustering. Evol. Intel; 2019

[26] Mohammed Ismail .B,T. Bhaskara Reddy,B. Eswara Reddy "Spiral Architecture Based Hybrid Fractal Image Compression" IEEE International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques; 2016.