International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 10 (2018), pp. 1533-1544

http://www.irphouse.com

Early Warning System for Endometrial Cancer Prediction in PMB Women Using Novel Ensemble Model

A. Hency Juliet [1]

[1](Research & Development Centre, Bharathiar University, Coimbatore and Assistant Professor in Department of Computer Application, Mar Gregorios College, Chennai, India.
E-mail: hencyjuliet@gmail.com)

Dr. R. Padmajavalli [2]

[2](Research & Development Centre, Bharathiar University, Coimbatore and Associate Professor in Department of Computer Application, Bhaktavatsalam Memorial College for Women, Chennai, India.
E-mail: padmahari2002@yahoo.com)

ABSTRACT

It's urge to establish a reliable system with high accuracy and efficiency for early prediction of cancer. In the present study, a novel ensemble model is developed for the prediction of Endometrial Cancer (EC) in women presenting with post menopausal bleeding.

Methodology used: In this approach to improve the performance of the model ensemble model was formed. Four different machine learning models such as K-Nearest Neighbor, Navive bayes, Neural Network and Random Forest are combined to form the ensemble model to predict the accuracy. The performance of the model was compared between ensemble model and Random Forest for before diagnosis and after diagnosis data of the patients. **Findings**: The ensemble model increases the prediction accuracy. Novel ensemble model produces high accuracy, gini, AUC, specificity and sensitivity. The result shows that the proposed ensemble model is outperforming.

Keywords - Endometrial; ensemble; Naïve Bayes, specificity; sensitivity; Gini; Random Forest.

I. INTRODUCTION

The menopausal state and the obesity play an important role in the formation of endometrial cancer. It is classified into two types. Type I cancers are of the commonest form and associated with increased levels of circulating estrogen [3]. They tend to occur at young age and are not aggressive. Type II cancers are of high grade, aggressive and they tend to arise spontaneously. Burbos.N et.al [1] had proposed Norwich DEFAB risk assessment tool for the prediction of risk of EC in PMB women with vaginal bleeding, he got 81.9% as the high sensitivity. Nearly all women with endometrial cancer have a postmenopausal bleeding, although only a miniature group of women with PMB will be diagnose with endometrial cancer, a systematic review and meta-analysis has found[2]. It is primarily a disease of the post menopausal women, although 25% of the cases occur in premenopausal patients, with 5% occurring in patients younger than forty years of age [5]. Endometrial carcinoma is the fourth most common cancer among women in westernized countries and sixth most common in worldwide [7]. The incidence of endometrial cancer is increasing all over the world and large differences can be seen in incidence rates between countries [3]. The Indian registries of endometrial cancer from the year 1993-1997 shows Chennai- 2.5%, Bangalore-2.6%, Delhi-2.7%, Mumbai-2.9%, Trivandrum- 2.9% [1]. Kristen A et.al, analyzes the patient's data, in order to identify the endometrial cancer risk in postmenopausal bleeding (PMB) women [5]. These findings provide a foundation for evaluating early detection strategies for endometrial cancer and can support risk-informed decision making in clinical management of postmenopausal bleeding. Clarke.M.A et.al [6] analyzed 129 studies with 40,790 women; 34,432 unique patients had postmenopausal bleeding and 6,358 had endometrial cancer [8]. Among women who had endometrial cancer, the pooled prevalence of postmenopausal bleeding was 91%. Using R language, the ensemble model was built; the Random Forest Classification model and ensemble model gives high percentage of sensitivity, on patient's data that is before diagnosis and for the clinical data that is after diagnosis. In southern part of India, the Women used to get treatment in Neyyoor Cancer Hospital; their history was collected and analyzed. Analyzing of cancer data with Post Menopausal Bleeding details will lead to early detection of endometrial cancer.

II MATERIAL AND METHODS

2.1 Data Set and Features

This endometrial data set is colleceted from International Cancer Institute Neyoor. The data was retrived from the patients' data sheet. The features such as age, menopause, obesity, hypertension, diabetes, stage of the cancer, symptom and the other details were collected. The feature description was given in the Table 1. This work carried out in R. We have found a more preferable model using R tool, which fits the data better than some existing models [9],[10]**.** The feature menopause status was taken as target variable. The dataset was loaded in the model, 75% of data was taken to train the model and the remaining 25% of the data was used to test the model.

General observation was made on the EC data to identify the percentage of patients are affected by EC with age greater than 50 and with post menopause bleeding on patients' history data, clinical data and on data of Cancer Genome Atlas. It was shown in Table 2. The process flow diagram was represented in Fig 1. The field menopausal status is taken as target variable. After partitioning train and test data, data model was build on training dataset, then using test data the model was predicted. Accuracies of all the models are calculated to evaluate the model. The results are analyzed with the help of Gini, AUC, sensitivity, specificity, and accuracies of the model. A step for the management of women with post menopausal bleeding was described in Fig 2.

Table 1. ICIEC – Patient Datasheet - Feature Description

Attribute Name	**Range**
AgeWhenCancerDiagnosed	Age Below 50 - 1, Age 51 - 60 - 2, Age Above 60 - 3
GeneralCondition	Good - 1, Fair – 2, Bad - 3
FirstSignofSymptom	Date of Onset
TotalDurationOfSymptom	Year, Months, Weeks
DelayBeforeConsulting_Months	Year , Months, Weeks
HistoryOfTrauma	Yes -1, No - 0
OccupationalFactor	Yes -1, No - 0
MenopausalAge	Below 50 - 1, Above 50 - 2
MenopauseStatus	PreMeno - 0 PostMeno - 1
PreviousTreatment	None- 0, Surgery -1, Radiotheraphy -2, OtherTreatment- 3
Lesion	New – 0, Healed – 1, Residual - 2 Recurrent - 3 Massive - 4
FamilyHistory	Positive - 1, Negative - 0
SpouseRelationship	Patternal side -1, Maternal side – 2, Both - 3
Parous	Nulliparous – 0, Otherwise – 1
Diet	Veg - 0 NonVeg - 1
Religion	Hindu - 1, Christian - 2, Muslim - 3
Area	Urban - 1, Rural - 2
Obesity	Yes – 1, No – 0

2.2 Machine learning methods

To get the better results, parameters of the models are needed to be tuned [11]. The model used in the present study is described in Table 3 with required packages and their tuning parameters [15], [16].

Table 2. Distribution of PMB Patient with Age >=50

Endometrial Cancer Data	Patient s' Age > = 50	Post Meno pause Patients	EC with PMB	EC with PreMB
ICIEC Patient Datasheet	75.57%	82.00%	-	-
ICIEC Clinical Data	86.07%	81.40%	84.00%	16.00%
TCGA Data[4]	98.60%	82.50%	82.50%	17.50%

Table 3. Machine Learning Model

Model	Method	Required Package	Tuning Parameter
Random Forest (RF) [14]	Rf	Random forest	mtry=2, ntree=500
Neural Network [17]	Nnet	Nnet	size=10
K-Nearest Neighbors [18]	Knn	Kncat	K=5:100
Naïve Bayes [19]	NB	Naïve bayes	None

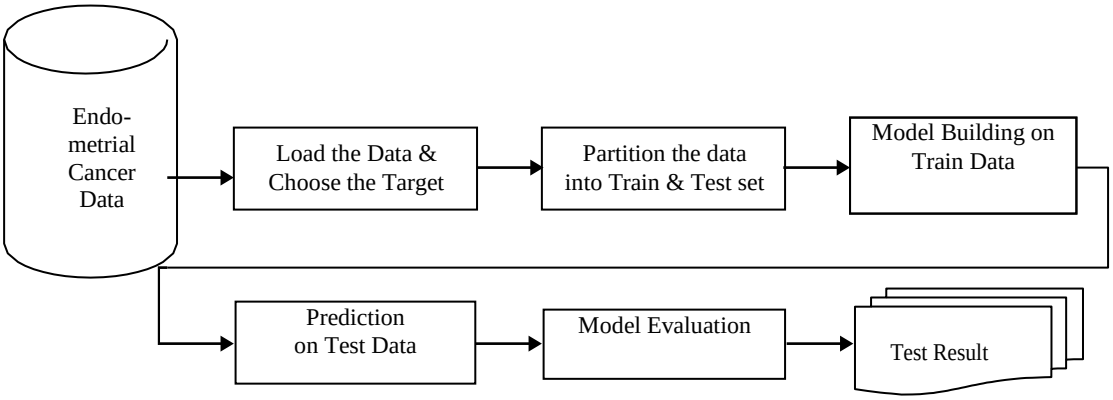

Fig. 1 Process Flow Diagram

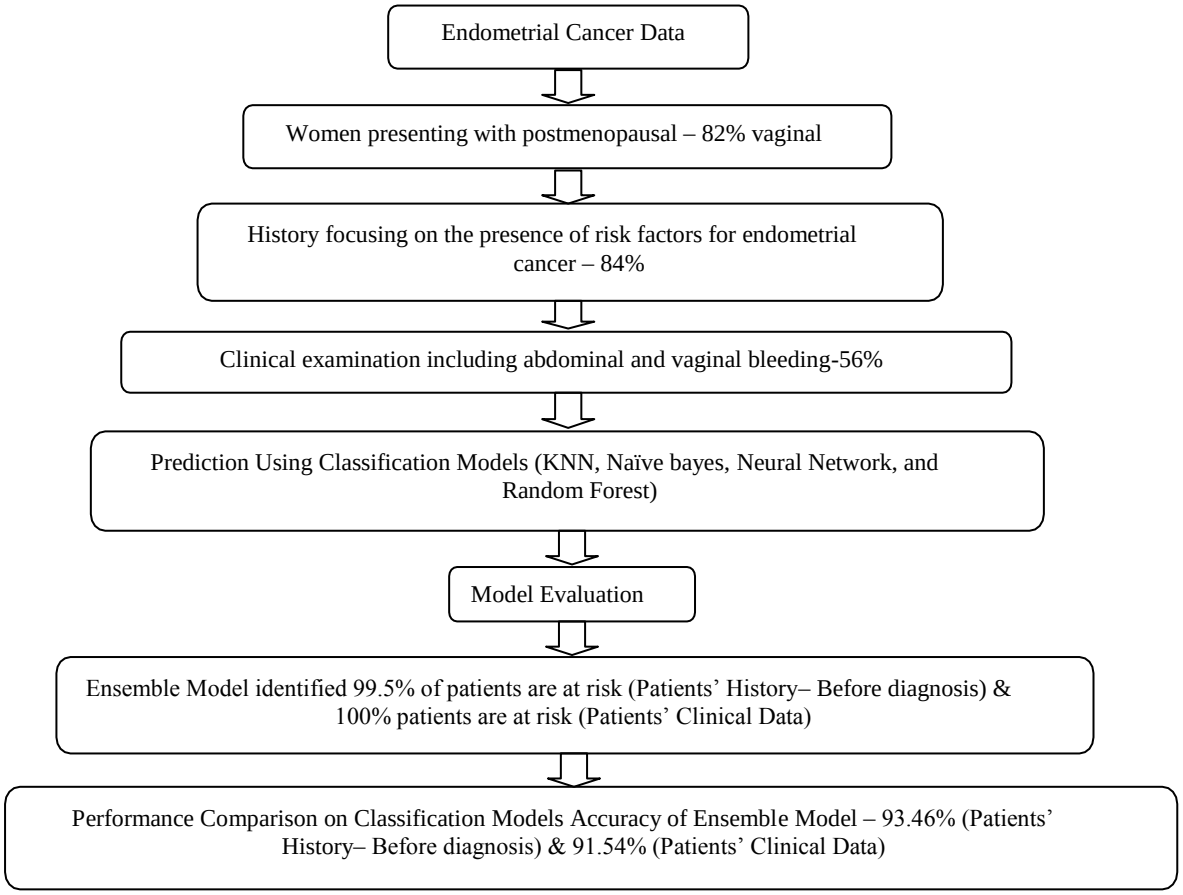

Fig. 2 Steps for the management of women with post menopausal bleeding

2.3 Predictive ensemble model

Ensemble is used to deal with the worst case of model prediction. The present work focus on the false prediction as well as true prediction of the model and ensemble model is used to deal with false and true predictions [13]. Four models i.e. KNN, NB, Neural Network and Random Forest are combined to get better accuracy. The models are trained on 75% of the dataset and 25% is used for testing. The predictive ensemble model is explained in Fig 3.

Initially, Neural Network, KNN, Naïve Bayes and Random Forest models are trained with 75% of dataset and generated predictions from 25% of dataset. Then, the models NN, KNN and NB are combined to train the random forest model that provides the last predictions.

In this approach, true predictions as well as false predictions are refined to get accurate proposed model. The data is travelled through four models because of this; the models perfectly learnt the data to provide reliable and accurate results.

III MODEL EVALUATION

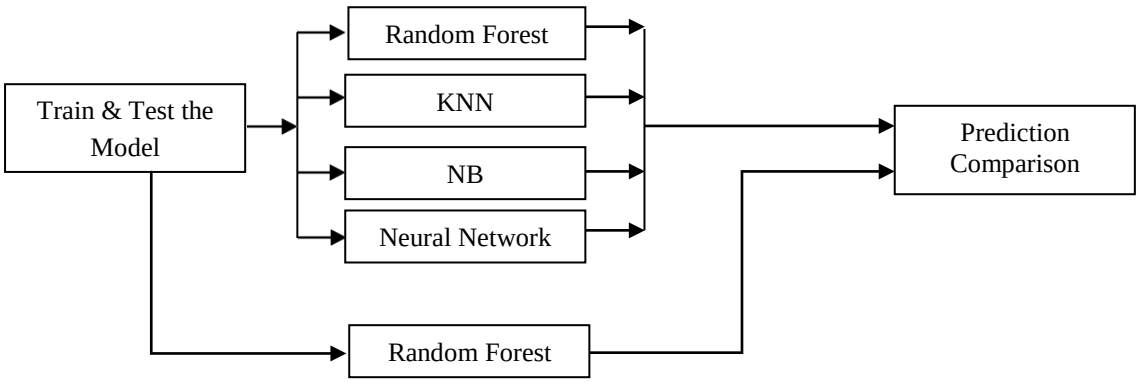

Fig. 3 Predictive Ensemble Model

Various parameters such as GINI, Accuracy, AUC, Specificity and sensitivity are calculated to evaluate the performance of model using R tool. Repeated k-fold cross validation is performed to test the robustness of model. The evaluation results are represented in table 4 and 5.

3.1 Performance evaluation of machine learning models GINI Coefficient

Inequality in the distribution is measured through Gini coefficient. The range of Gini value is between 0 and 1. Like, model M has Gini value 60% and model D has Gini value 45% then Model M is considered as an efficient model as compare to model D [11].

AUC

Area under the curve (AUC) is calculated to measure the quality of classier. The amount of area under the receiver operating characteristics (ROC) curve is AUC. The model scoring high AUC as compared to other models is considered as efficient model. Its value is between 0 and 1. The quality of model is good if it has AUC value near to 1[12].

Accuracy:

Accuracy is calculated to measure the correctness of classifier. The accuracy can be calculated as, Accuracy = TP + TN / (Total Data)

Sensitivity:

Sensitivity is also known as **recall** or **true positive** rate. It is the proportion of actual

positives which are correctly identified as positives by the classifier and is computed as, Sensitivity = TP / (TP + FN)

Specificity:

It is also known as true negative Rate. It relates the classifiers ability to identify negative results and is computed as,

Specificity = TN / (TN + FP)

TN: True negative, FP: False positive, TP: True positive and FN: False negative. For the validation of model, Patient History data and Clinical data are compared using various parameters such as Gini, AUC, accuracy, specificity and sensitivity.

3.2. Repeated K-fold Cross Validation

The large number of comparisons is always preferred, to compare the performance of model. To run K-fold cross validation multiple time or increase the number of comparisons, repeated K-fold cross validation is useful. In the K-fold cross validation only k comparisons are acquired. In cross validation, in each fold, random data is provided to do the comparisons. Here, 10-fold cross validation is repeated for 3 times. Table 4 and 5 specifies the evaluation results.

Table 4. Evaluation Result of Models on Clinical Data

Classifier Model	Gini	AUC	ER	Sens	Spec	Precision	Recall	TP	FP	TN	FN	Accuracy
Random Forest	0.291	0.645	0.154	1	0	0.846	1	220	40	0	0	84.62
Proposed Ensemble	0.897	0.948	0.085	1	0.522	0.907	1	214	22	24	0	91.54

Table 5. Evaluation Result of Models on Patients' Data Before Diagnosis

Classifier Model	Gini	AUC	ER	Sens	Spec	Precision	Recall	TP	FP	TN	FN	Accuracy
Random Forest	0.753	0.877	0.131	0.96	0.265	0.897	0.96	217	25	9	9	86.92
Proposed Ensemble	0.988	0.994	0.065	0.995	0.66	0.93	0.995	212	16	31	1	93.46

IV RESULT ANALYSIS, COMPARISON AND DISCUSSION

True Positives is the number of patients who have post menopausal and were identified as at-risk. According to Fig 5, the predicted value of TP by all the models are high than TN. From that, we can identify the patients those who have PMB are at risk. The same result was given by the clinical data. True Negatives is the numbers of patients who have pre menopausal and were not identified as at-risk. From the Table 6, we can identify the patients' with pre menopause are not at risk. False Negatives (type II error) is the number of patients who have post menopausal but were not

identified by the models as at risk. All the models are performed well on clinical data.

False positives (type I error) is the number of patients who have pre menopausal but were identified by the models as at risk.

Sensitivity gives the positive Result that is Patient is at Risk

Accuracy (Post Menopausal) = True Positive / No. of Post Menopause Patients.
= True Positive / (True Positive + False Negative)

Specificity gives the Negative Result that is Patient is not at risk

Accuracy (Pre Menopausal) = True Negative / No. of Pre Menopause Patients
= True Negative / (True Negative + False Positive)

For the models, a negative result means the patient is not at risk, a positive result means the patient is at risk and has to go for the treatment. Thus it will alert the patient and the healthcare team to do further treatment. Table 3 describes the machine learning models that are trained on the dataset with optimum tuning parameters. The dataset is partitioned into two parts 75% and 25%. The prepared models are unknown to the 25% of dataset.

The proposed model is combination of seven models that makes it an ensemble model. Fig. 4 represents the Area under the ROC curve for Random Forest. Random Forest identified 96% patients are at risk (Patients' History– Before diagnosis) and 100% patients are at risk (Patients' Clinical Data - After Diagnosis). Proposed ensemble model identified 99.5% are at risk (Patients' History– Before diagnosis) and 100% are at risk (Patients' Clinical Data - After Diagnosis). Table 6 shows the performance comparison. Performance Comparison on Classification Model is carried out using accuracy of Random Forest and ensemble model.

Table 6. Performance Comparisons on EC Data

Evaluation Parameters	Before Diagnosis		After Diagnosis	
	RF	**Ensemble Model**	**RF**	**Ensemble Model**
Sensitivity% At-Risk Patient	96%	99.5%	100%	100%
Specificity%	26.5%	66%	0.00%	52.2%
TP	217	212	220	214
FP	25	16	40	22
TN	9	31	0	24
FN	9	1	0	0
ACC	86.92	93.46	84.62	91.54%

The Accuracies are represented in Table 7 and 8. The model random forest produced the accuracy 86.92% on Patients' History– Before diagnosis and 84.62% on Patients'

Clinical Data - After Diagnosis. The proposed ensemble model produced the accuracy 93.46% on Patients' History– Before diagnosis and 91.54% on Patients' Clinical Data - After Diagnosis. Fig 4 represents the ROC curve. Fig 5 & 6 represents the performance difference between random forest and proposed ensemble model.

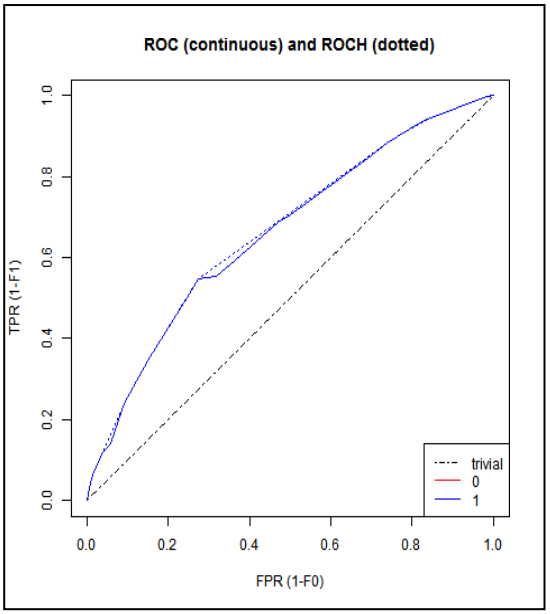

Fig. 4 Area under ROC curve for Random Forest is 0.901

Table 7. Patient data – Before Diagnosis

Model	Precision	Recall	Accuracy
Random Forest	89.70%	96.00%	86.92%
Proposed Ensemble	93.00%	99.50%	93.46%

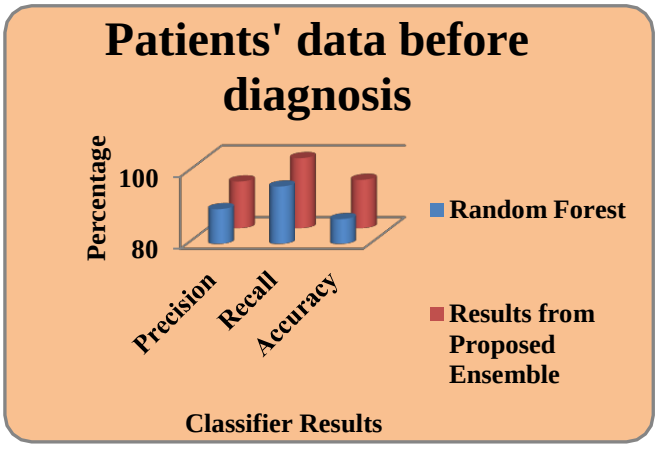

Fig 5. Classifier Result on Patient data - before diagnosis

Table 8. Patient data – After Diagnosis

Model	Precision	Recall	Accuracy
Random Forest	84.60%	100%	84.62%
Proposed Ensemble	90.70%	100%	91.54%

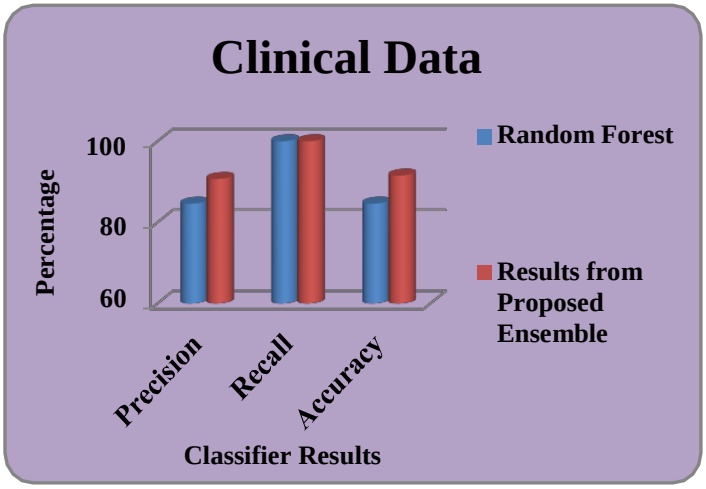

Fig 6. Classifier Result on Patient data - After diagnosis

V CONCLUSION

The ensemble model increases the prediction accuracy of the data. In the present study, four models such as Naïve Bayes, Neural Network, K-Nearest Neighbor and Random Forest are used to create the model. This model was developed for risk prediction with high accuracy, Gini, AUC, Specificity and sensitivity. This study predicted the women presenting post menopausal with vaginal bleeding are at risk. This prediction was conducted on patients' data before diagnosis and after diagnosis. But both data gives almost same prediction. Using R language, the model was built; the proposed ensemble model gives 99.5% of sensitivity, on patient's data that is before diagnosis and 100% of sensitivity on clinical data. In Future, Android App can be developed for the early prediction, so that the patient can identify their health issue and go for the necessary treatment in time.

Early detection tactic spotlight on women with PMB has the potential to detain as many as 90% of endometrial cancers; however, most women with PMB will not be identified with endometrial cancer. These results can aid in the assessment of the latent clinical value of new early detection makers and clinical management scheme for endometrial cancer and will help to inform clinical and epidemiologic risk prediction models to support decision making.

ACKNOWLEDGEMENTS

- The author would like to thank Mr. Julius Gnanamoney, Radiotherapy Clinical Specialist, Kent Oncology Centre, Maidstone Hospital, Hermitage Lane, Maidstone, ME16 9QQ, London, Who supported a lot and provided the detailed information on Endometrial Cancer.
- The author also would like to thank Dr. V.G.Sudhakaran, Head Department of Radiation and oncology, and Dr. Arul Pirakasam, International Cancer Institute, Neyyoor, 629802, for the support and guidance.

REFERENCE

[1] Shridhar A, Association Rule- Spatial Data Mining Approach for Exploration of Endometrial Cancer Data, International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October - 2013, pp. 1111-1116

[2] Burbos N, Predicting the risk of endometrial cancer in postmenopausal women presenting with vaginal bleeding: the Norwich DEFAB risk assessment tool, British Journal of Cancer (2010) 102(8),1201 – 1206

[3] Yamamoto K, Tomizawa S (2012) Statistical Analysis of Case-Control Data of Endometrial Cancer Based on New Asymmetry Models. J Biomet Biostat 3:147. doi:10.4172/2155- 6180.1000147

[4] The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Endometrial Carcinoma. Nature. May 2, 2013. DOI: 10.1038, nature12113

[5] Kristen A. Matteson, Opportunities for Early Detection of Endometrial Cancer in Women with Postmenopausal Bleeding," Women and Infants Hospital, Providence, Rhode Island.

[6] Clarke MA, Long BJ, Del Mar Morillo A, Arbyn M, Bakkum-Gamez JN, Wentzensen N. Association of Endometrial Cancer Risk With Postmenopausal Bleeding in Women A Systematic Review and Meta-analysis. *JAMA Intern Med.* Published online August 06, 2018. doi:10.1001/jamainternmed 2018.2820

[7] E. Friberg & N. Orsini & C. S. Mantzoros & A. Wolk, Diabetes Mellitus And Risk Of Endometrial Cancer: A Meta-Analysis, Springer-Verlag 2007.

[8] Weiderpass E, Persson I, Adami Ho, Magnusson C, Lindgren A, Baron Ja, Body Size In Different Periods Of Life, Diabetes Mellitus, Hypertension, And Risk Of Postmenopausal Endometrial Cancer, Cancer Causes And Control - By Springer.

[9] I.P.Constantinou, C. A. Koumourou, M. S. Neofytou, V. Tanos,C. S. Pattichis, E.C. Kyriakou, "An Integrated Cad System Facilitating The Endometrial

Cancer Diagnosis", Information Technology And Applications In Biomedicine, Itab 2009, Larnaca, Cyprus, 5-7 November 2009.

[10] Parazzini F, La Vecchia C, Negri E, "Diabetes and Endometrial Cancer: An Italian Case-Control Study". International Journal for Cancer, 1999, 81:539–542, Article In International Journal Of Cancer · May 1999.

[11] Luca Giannella, A Risk-Scoring Model for the Prediction of Endometrial Cancer among Symptomatic Postmenopausal Women with Endometrial Thickness>4 mm, BioMed Research International, Volume 2014, Article ID 130569.

[12] F. Marbouti, Models for early prediction of at-risk students in a course using standards-based grading, Elsevier Computers & Education 103 (2016) 1e15

[13] Divya Khanna , Multilevel ensemble model for prediction of IgA and IgG antibodies, Elsevier - Immunology Letters, 0165-2478/© 2017 European Federation of Immunological Societies.

[14] A. Liaw, M. Wiener, Classification and regression by randomforest, R news 2(3) (2002) 18–22.

[15] S.S. Keerthi, E.G. Gilbert, Convergence of a generalized smo algorithm for svmclassifier design, Machine Learn. 46 (1-3) (2002) 351–360.

[16] P.D. Berger, A. Gerstenfeld, A.Z. Zeng, How many suppliers are best? adecision-analysis approach, Omega 32 (1) (2004) 9–15.

[17] B.Ripley, W.Venables, M.B. Ripley, Package 'nnet'.

[18] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, B. Hofner, F. Sobotka, F.Scheipl, M. B. Hofner, Package 'mboost'.

[19] C. K. Williams, A. Engelhardt, T. Cooper, Z. Mayer, A. Ziem, L. Scrucca, Y. Tang,C. Candan, M. M. Kuhn, Package 'caret'.