Integration of Heterogeneous Classification Techniques to Predict the Risk Associated With Endometrial Cancer

A. Hency Juliet

Research & Development Centre, Bharathiar University, Coimbatore and Assistant Professor in Department of Computer Application, Mar Gregorios College, Chennai, India.

hencyjuliet@gmail.com

Dr.R. Padmajavalli

Research & Development Centre, Bharathiar University, Coimbatore and Associate Professor in Department of Computer Application, Bhaktavatsalam Memorial College for Women, Chennai, India. padmahari2002@yahoo.com

***Abstract*- Objective: Generating rules from the data will help to find the risk associated with the disease. Such rules can help the healthcare domain to identify the symptoms which are reasons for the disease. The clinical data for this study was collected from the International cancer institute. The insight on the data was analyzed with the help of data mining technique such as classification.**

Method: The decision tree classification technique was applied on the data and the rules were generated. The rules are given as input to the rule based classification and the new rule was generated. This rule was tested using statistical test such as chi-square test, it proves the variables which are associated with the risk of endometrial cancer. The accuracy was once again tested with the help of k-Nearest Neighbour algorithm. Finally, to improve the accuracy, the algorithms were tuned with hyper-parameters. The ensemble of heterogeneous classification technique was applied using Voting Ensemble Model.

Results: The Classification techniques proves the abnormal bleeding, obesity, diabetes, infertility, post menopausal, abnormal watery discharge and pelvic pain are leads to high risk of endometrial cancer. Abnormal bleeding and post menopausal and estrogen treatment leads to average risk of endometrial cancer. The Accuracy of J48, PART and IBK algorithms are 93.91%, 94.23% and 94.84% respectively. The accuracy of voting ensemble model is 95.68%. The J48 & PART algorithm generated 17 and 12 rules respectively.

Conclusion: The EC-Risk Predict is a user-friendly mobile app which uses the clinical attribute as input to predict the risk associated with endometrial cancer. Based on the input it will give either high risk or low risk or no risk. The Mobile app was uploaded in the play store for the patients' usage.

Keywords– J48; K-Nearest Neighbour; PART; Voting Ensemble; Chi-square test; Classification;

I. INTRODUCTION

Cancer is the second largest non-communicable disease. It is the leading cause of deaths in developed countries [1]. The incidence of cancer malignancy has increased worldwide. Endometrial Cancer (EC) is primarily a disease of the postmenopausal women and confirmed that 90% of patients with EC experienced postmenopausal bleeding [2]. EC is the fourth most cancer in Women. It arises on the endometrium, the inside layer of the uterus or womb. It is the consequence of the jarring growth of cells that have the ability to engage or widen to other parts of the body [3]. Hormones amend the endometrium, during the woman's menstrual cycle. In the early stage of the cycle, prior to discharge eggs from ovaries, the ovaries generate estrogen hormones [4]. Estrogen is the cause for the endometrium to condense so that it could cultivate an embryo if pregnancy happens [5]. A woman's hormone dependability is an important function in the maturity of EC [6]. The main tests for diagnosing EC are Trans Vaginal Ultra Sound (TVUS), Pelvic Examination, Hysteroscopy, Endometrial Biopsy and Dilation and Curettage [7]. First three tests insert a device transducer,

 DOI : 10.18231/2454-9150.2018.1392

speculum and a tube hysteroscope respectively, whereas for the last two tests tissue will be removed from uterine lining for laboratory analysis. In order to avoid the complexity and to help the cancer patients a simple prediction system has been proposed in this research. Early diagnosis is the key focus for medical experts in order to start the treatment earlier [8]. Most women are diagnosed at an early stage and have relatively good survival rates; however, women who are diagnosed with advanced-stage or recurrent disease have a poor prognosis [9]. To improve the effectiveness and efficiency of the diagnosis process in healthcare, the risk of EC in Postmenopausal Women with vaginal bleeding was predicted. The application of data mining in health care is of great importance because of its efficiency in disease prediction. Hence this research uses data mining algorithms such as Decision Tree, Rule-based Classification and k-Nearest-Neighbor (KNN) for the prediction of EC. The proposed prediction system accepts inputs such as age, menopausal status, vaginal bleeding, BMI, abnormal watery discharge, infertility, estrogen treatment and diabetes and gives the predicted result.

II. LITERATURE REVIEW

Few of the important works in the literature review have been discussed in this section. A preferable model which fits the data better than some existing models that the average dose of estrogen for a case in a matched pair tends to be more than for control in the pair [10],[19]. A Meta-Analysis has been conducted for diabetes (largely type 2) and EC, based on 16 studies including 96,003 participants and 7,596 cases of EC, found that diabetes was statistically and significantly associated with an increased risk of EC [11], [17], [18]. A study on endometrial Carcinoma has been conducted using novel assays and applied Association rule mining algorithm and classification technique on EC data and it was proved that a woman who is more than 60 years of age, undergone menopause and under a non vegetarian diet, the occurrence of EC was high [12]. A nationwide population-based case-control study among postmenopausal women aged 50-74 years, and compared lean women with overweight women resulting in a 50% increase in risk for EC. Obesity and diabetes mellitus (Types 1 and 2) are associated with EC risk [13]. A method confirms that non-insulin-dependent diabetes is associated with the risk of EC. The association may be related to elevated estrogen levels in diabetic women [14]. The Norwich DEFAB risk assessment tool proposed an algorithm has been developed to predict the EC in PMB women, with high sensitivity of 81.9% [15].

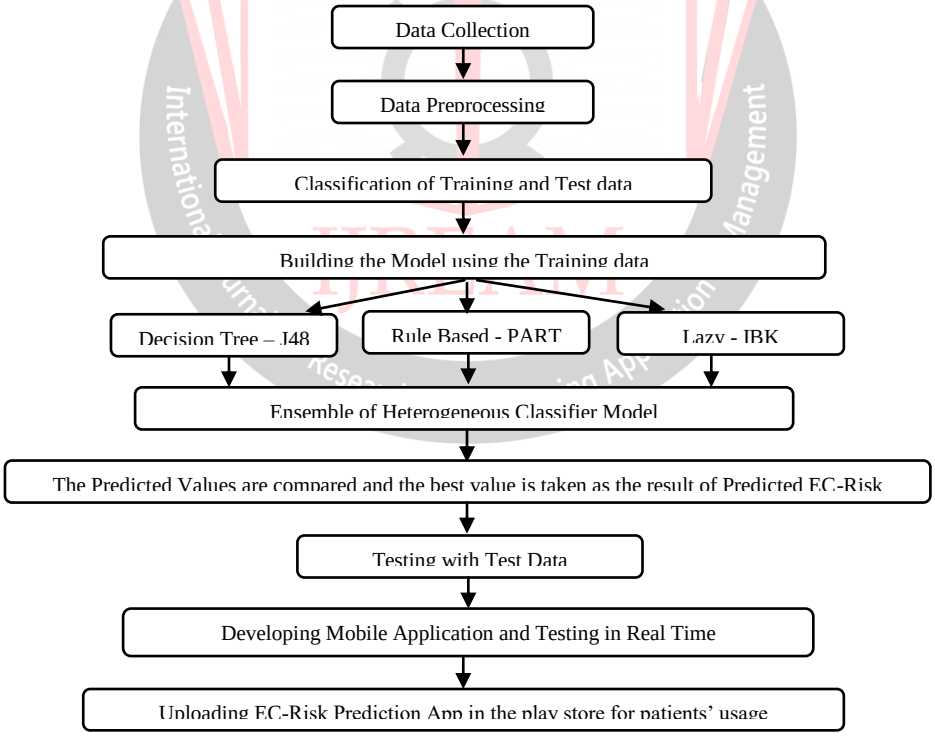

Fig. 1 Process Flow Diagram

III. RESEARCH METHODOLOGY

3.1 Process Flow

The proposed work is divided into six phases Data Collection, Data Preprocessing, Classification of training data, Building the model, Decision Tree, Rule-Based, Lazy Classifier and ensemble classification, testing the model with the test data and Testing in the real-time environment and Uploading the app in the play store for the public usage. Figure 1 represents the process flow diagram.

3.2 Data Description

The endometrial data set is collected from International Cancer Institute Neyoor. The data was retrieved from the

 DOI : 10.18231/2454-9150.2018.1392

patients' data sheet. The features such as age, height, weight, menopause, diabetes, estrogen treatment, abnormal bleeding, watery discharge, infertility, pelvic pain were collected. Body mass index and obesity is calculated by the mobile app with the help of height and weight. Body mass index (BMI) is based on height and weight that pertinent to grown men and women. BMI = (Weight in Kilograms / (Height in Centimeters) ^{2}) X 10000 for example, a person whose weight is 99.79 Kilograms and 190.50 centimeters tall has a BMI of 27.5. From the BMI Obesity was determined, obesity = BMI of 30 or above.

3.3 Data Preprocessing

After collecting the data, the data reduction technique applied to obtain a reduced representation of the data set that is smaller in volume yet produce the same analytical results. Using the dimensionality reduction number of random variables is reduced under consideration [16]. It includes wavelet transforms and principal component analysis which transfer the original data into smaller space. Selecting a subset from attribute is a method of dimensionality lessening, in which extraneous, feebly related or surplus attributes are detected and removed. So that we can obtain the original distribution with a minimum set of attributes. The best and worst attributes are determined using statistical significance. This research uses the Greedy method for subset selection [16]. The forward selection algorithm was shown in figure 2.

Procedure for the stepwise forward selection:

1. *Initial attribute set {A_1, A_2, A_3, A_n} and Set of attribute set as the reduced set.*
2. *Take an empty set { }*
3. *Find the best original attribute and add to the reduced set using statistical significance. Ex. {A_1} then {A_1, A_4}*
4. *Repeat step 3 to add the best attribute to the reduced set.*

Fig.2. A Stepwise forward selection algorithm

The final reduced attribute set {A_1, A_4, …. }

The most significant attributes have been finalized using the entropy. The total number of an instance is 1040 in which High-Risk instances are 280, Average Risk instances are 373 and No risk instances are 387.

Entropy of (decision) = $\sum$-P(i) X $\log_2$P(i) ----> (1)

Entropy of (decision) = - P(High Risk) X $\log_2$P(High Risk) – P(Average Risk) X $\log_2$P(Average Risk) – P(No Risk) X $\log_2$P (No Risk) = 1.57083

3.4. Data Analysis

The most significant attributes have been finalized using the entropy. The dataset was divided into two sets one is the training set with 728 instances and test datasets of 312 instances. A gain for each attribute is calculated using equation 2. A Gain ratio is calculated using equation 3.

Gain (Decision, Menopause) = Entropy (Decision)

$$-\sum P \text{ (Decision / Menopause) ---> (2)}$$

Gain Ratio = Gain (A)/Split Info (A) ----------> (3)

Where

Split info (A) = $\sum$|Pj| / |D| X log2 |Dj| / |D|----> (4)

The Gain ratio with greater value will be taken as root node, from the above calculation the split Info (A) as Vaginal Bleeding. The work is carried-out in WEKA, it is written in the language Java, has been formed to demonstrate the ideas called the Waikato Environment for Knowledge Analysis.

3.5 Classification Technique

3.5.1 Decision Tree Classification

Tree-based classification J48 has been applied on the dataset which can be converted to Classification IF-THEN rules. J48 adopt a greedy approach in which the decision tree is constructed in a top-down recursive divide and conquer manner. Some splitting sequences are replicated into the tree [16]. It is known as the "replication problem". The Decision tree – J48 algorithm was described in figure 3 and the decision tree generated by J48 was shown in figure 3.

Few extracted rules from a J48 pruned tree:

R1: *If Vaginal_Bleeding = Yes and Menopause = Post and Infertility = Yes watery_discharge=Yes and Pelvic_Pain = Yes and Obesity = Yes then High Risk*

R2: *If Vaginal_Bleeding = Yes and Menopause = Post and Infertility = Yes watery_discharge=Yes then Average Risk*

R3: *If Vaginal_Bleeding = Yes and Pelvic_Pain = No and Menopause = Peri then No Risk*

 DOI : 10.18231/2454-9150.2018.1392

Fig.3. J48 (C4.5) Algorithm

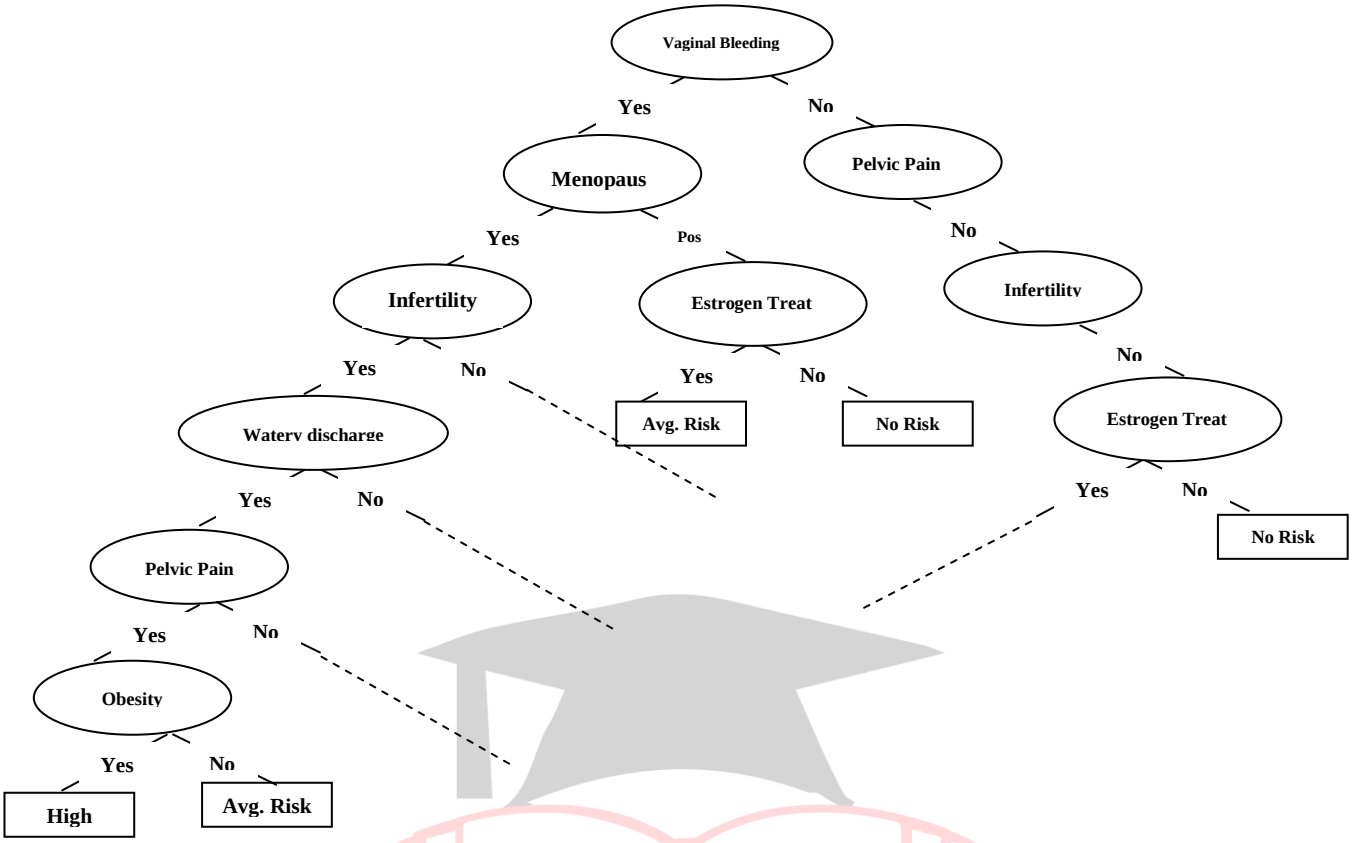

Fig.4. Decision Tree generated by J48 Algorithm

3.5.2 Rule-Based Classification

The PART algorithm combines the divide-and-conquer strategy with the separate-and-conquer strategy of rule learning. It builds a partial decision tree from the instance and creates rules from decision tree that is leaf with the largest coverage is made into a rule. From the unpruned decision tree, we can create a rule [17]. The unpruned tree generated by J48 was shown in figure 7 is taken for rule creation of PART algorithm. The tree searches the average purity on the leaves when it splits a node [18]. The separate and conquer on the other hand tries to exploit the purity of one leaf only when it tries to create a rule. Compared to the predictive association rule algorithms [19], they do not suffer from the redundancy of the induced rules. The scheme is smooth to construct the least set of rules which permits classifying accurately a new instance. It enables to handle the problem of collision about rules when an instance activates two or several rules which lead to inconsistent conclusions [20].

Separate and Conquer - The separate and conquer algorithms are based on the sequential covering principle [21]. The separate and conquer steps are shown in figure 5 and a part algorithm was shown in figure 6.

1. Define a rule which accurately predict one value of the target attribute is referred as conquer.
2. Removing the covered instance from the learning set is referred as separate.
3. Iterate step 1 and 2 until all the training instances are covered.

Fig.5 The Separate and Conquer procedure

The objective is to find out a prediction rule from data. In the supervised learning structure, the attribute to the termination part is the target attribute and a rule is associated with the target attribute. But the target attribute may be alarmed by several rules.

PART Algorithm:

1. Build a partial decision tree on the current set of instances
2. Create a rule from the decision tree.
The leaf with the largest coverage is made into a rule
3. Discarded the decision tree
4. Remove the instances covered by the rule
5. Go to step one

Fig.6. PART Algorithm

 DOI : 10.18231/2454-9150.2018.1392

```
J48 unpruned tree
Vaginal_Bleeding = Yes
|  Menopause = Post
|  |  Infertility = Yes
|  |  |  Watery_Discharge = Yes
|  |  |  |  Pelvic_Pain = Yes
|  |  |  |  |  Obesity = Yes: High Risk (293.0/14.0)
|  |  |  |  |  Obesity = No: Average Risk (2.0/1.0)
|  |  |  |  Pelvic_Pain = No: Average Risk (6.0)
|  |  |  Watery_Discharge = No
|  |  |  |  Obesity = Yes
|  |  |  |  |  Age <= 49
|  |  |  |  |  |  Diabetes = Yes: No Risk (5.0)
|  |  |  |  |  |  Diabetes = No
|  |  |  |  |  |  |  Pelvic_Pain = Yes: Average Risk (13.0)
|  |  |  |  |  |  |  Pelvic_Pain = No: No Risk (3.0)
|  |  |  |  |  Age > 49: Average Risk (137.0/8.0)
|  |  |  |  Obesity = No: No Risk (8.0)
|  |  Infertility = No
|  |  |  Watery_Discharge = Yes: Average Risk (183.0/5.0)
|  |  |  Watery_Discharge = No
|  |  |  |  Obesity = Yes
|  |  |  |  |  Diabetes = Yes: Average Risk (19.0)
|  |  |  |  |  Diabetes = No
|  |  |  |  |  |  Age <= 50: No Risk (3.0)
|  |  |  |  |  |  Age > 50: Average Risk (15.0/5.0)
|  |  |  |  Obesity = No: No Risk (4.0)
|  Menopause = Pre
|  |  Estrogen_Treatment = Yes: Average Risk (6.0/2.0)
|  |  Estrogen_Treatment = No: No Risk (11.0)
|  Menopause = Peri: No Risk (50.0)
Vaginal_Bleeding = No
|  Pelvic_Pain = Yes: No Risk (232.0)
|  Pelvic_Pain = No
|  |  Infertility = Yes: No Risk (27.0)
|  |  Infertility = No
|  |  |  Estrogen_Treatment = Yes
|  |  |  |  Diabetes = Yes: No Risk (4.0)
|  |  |  |  Diabetes = No: Average Risk (3.0/1.0)
|  |  |  Estrogen_Treatment = No: No Risk (16.0)

Number of Leaves: 21     Size of the tree:  40
```

Fig.7. The J48 Un pruned Tree

From the unpruned tree with the largest coverage, the rules are created. Then the instances covered by the rule can be removed from the training set. Repeat the process until stopping criteria met [19].

A few decision lists generated by PART:

R1: *Watery_Discharge = Yes and Vaginal_Bleeding = Yes and Pelvic_Pain = Yes and Menopause=Post and Obesity = Yes then Cancer_Risk=High Risk*

R2: *Watery_Discharge = No and Obesity = Yes and Menopause = Post and age >49 and Vaginal_Bleeding = Yes then Cancer_Risk = Average Risk*

R3: *Vaginal_Bleeding = No and Pelvic_Pain = Yes then Cancer_Risk = No Risk*

3.5.3 Lazy Classification

The lazy classifier IBK is a K Nearest Neighbour classifier that uses the distance metric. The number of nearest neighbor can either be specified or determined using leave one out cross validation method [22]. To speed up the task of finding the nearest neighbour different kinds of search algorithms can be used. A Linear search, KD-Tree, Ball-Tree, and Cover-Tree are some search algorithms [23]. The distance can be measured using the Euclidean, Chebyshev, Manhattan and Minkowski distances [23].

The performance of the machine learning model was improved by tuning the algorithm parameters using Weka Experimenter. The Experimenter is used to execute experiments or statistical check on the dataset. A proscribed testing was intended to tune the hyper-parameters of a machine learning algorithm, and the outcome of tuning an experimentation were interpreted using statistical significance with confidence 0.05 [24]. The tuning process started for KNN with K=1, 3 and 7 and distance = Euclidean and then with K=1, 3 and 7 with distance = Manhattan

IBk, k=1, distance Function=Euclidean
IBk, k=3, distance Function=Euclidean
IBk, k=7, distance Function=Euclidean
IBk, k=1, distance Function=Manhattan
IBk, k=3, distance Function=Manhattan
IBk, k=7, distance Function=Manhattan

Table 1 – Experimenter Results of IBK algorithm with tuning parameters

Dataset	(1) lazy.IBk	(2) lazy.	(3) lazy.	(4) lazy.	(5) lazy.	(6) lazy.
ECIC	(100) 94.84	94.73 *	94.73 *	94.83	94.72 *	94.72 *
	(v/ /*)	(0/0/1)	(0/0/1)	(0/1/0)	(0/0/1)	(0/0/1)

In general k=1 with Euclidean distance performed well than the others. Also, the accuracy was increased to 94.84%

3.5.4 Ensembles of Heterogeneous Classification

Voting is a trendy ensemble method. Voting integrates the decision from different models based on a combinational rule which happens to be a numerous combination of likelihood estimates. The idea used in voting method is very much straight forward. The voting means a class or learner which gets the labels as inputs from diverse resources and uses likelihood estimates to make a concluding decision [25]. The likelihood estimates which are associated with voting are average of probability, majority voting, a product of probability, maximum and minimum of probability and median [25].

In this study, voting integrates the decisions from the decision tree, rule-based and lazy classifiers. The voting algorithm is tuned using hyper-parameters k and distance. The different k values such as k=1, 3, and 7 and the Euclidean distance are applied to the algorithm and got the improved accuracy of 95.68%. The performance comparison results were shown in figure 8.

Table 2 – Experimenter Results of Vote algorithm with tuning parameters

Dataset	(1) meta.Vot	(2) meta.	(3) meta.
ECIC	(100) 95.68	95.52	95.38
	(v/ /*)	(0/1/0)	(0/1/0)

 DOI : 10.18231/2454-9150.2018.1392

3.5.5 Statistical Test

The statistical chi-square calculation was conducted on the finally selected rules in order to identify the strong association between the attributes and risk. The level of significance is 0.05, the degrees of freedom is (r-1) x (c-1) the test statistics used is a Chi-Square test.

$$\text{Chi-Square value} = \frac{\sum(\text{Observed} - \text{Expectation})^2}{\text{Expectation}} \quad (5)$$

The Chi-Square value is 514.7592, an expected value of chi-square for 4 degrees of freedom and 5% level of significance, that is the P value is <0.0001, the result is significant at $p<0.05$, to reject the Null hypothesis. Therefore the attributes watery discharge, vaginal bleeding, menopause, obesity and infertility are associated with high risk. Similarly, Chi-square value was calculated for the average risk rule with 3 degrees of freedom and 5% level of significance and found the result is significant at $p<0.05$. Therefore the attributes such as vaginal bleeding, obesity, watery discharge, and menopause are associated with average risk.

Table 3. Results of test data for Decision Tree, Rule-Based, KNN and Ensemble Classification Algorithms

Evaluation Parameters	Classifier Model											
	Decision Tree - J48			Rule Based – PART			K Nearest Neighbor - IBK			Ensemble Classifier Vote		
	High Risk	Average Risk	No Risk	High Risk	Average Risk	No Risk	High Risk	Average Risk	No Risk	High Risk	Average Risk	No Risk
TP	77	109	107	77	113	104	77	111	106	77	111	107
TN	230	186	189	228	183	195	229	185	192	229	186	192
FP	5	7	7	7	10	1	6	8	4	6	7	4
FN	0	10	9	0	6	12	0	8	10	0	8	9
Sensitivity	1	0.91	0.92	1	0.95	0.89	1	0.93	0.91	1	0.93	0.96
Specificity	0.97	0.96	0.96	0.96	0.95	0.99	0.97	0.95	0.98	0.97	0.96	0.98
Precision	0.93	0.94	0.93	0.91	0.91	0.99	0.92	0.93	0.96	0.92	0.94	0.96
Accuracy	93.91%%			94.23%			94.23%			94.87%		
After tuning with hyper-parameters							94.84%			95.65%		

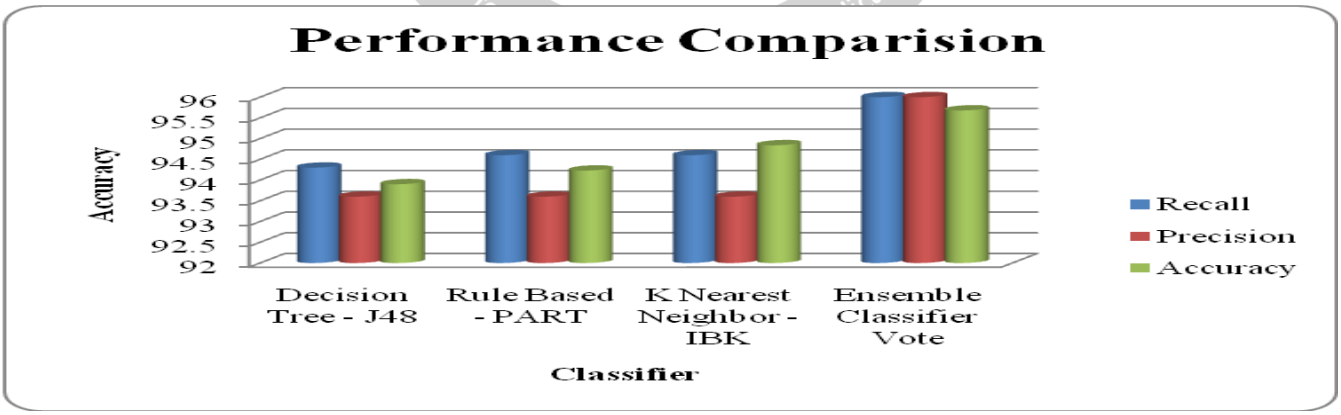

Fig.8. Performance comparison between the evaluated parameters

3.5.6 The Proposed Integrated EC-Risk Prediction system

In this research, a combination of Decision Tree and Rule-Based algorithms has been considered to form the rules for predicting the
endometrial cancer risk. The proposed method has been implemented as Mobile App "EC-Risk Prediction" which is based on the rules generated by J48 and PART algorithm. The real-time analysis was conducted for 100 patients' data. And the result was shown in table 4. Screenshot of EC-Risk prediction a mobile app was shown in figure 5.

 DOI : 10.18231/2454-9150.2018.1392

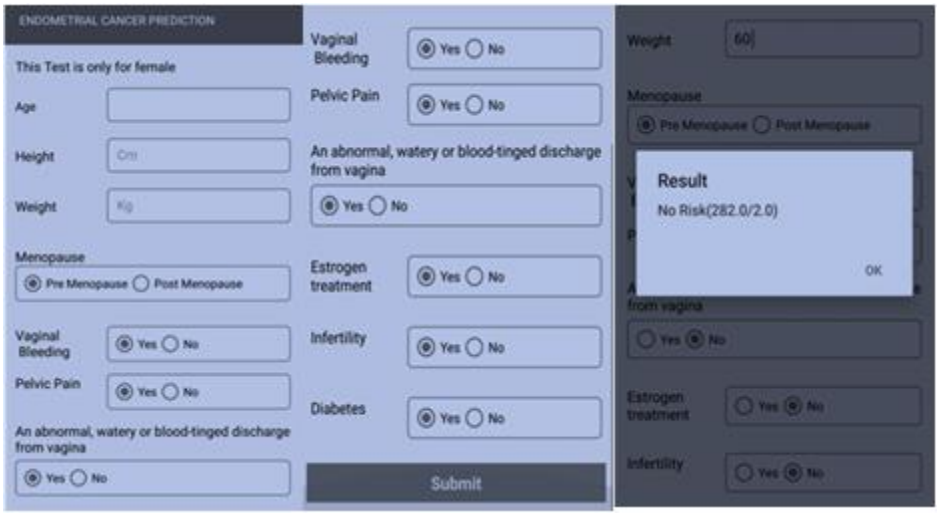

Fig. 5 Screenshots of Mobile app "EC–Risk Prediction"

IV. MODEL EVALUATION

Various parameters such as Accuracy, Specificity and sensitivity are calculated to evaluate the performance of the classifier model.

TP – True Positive: Refers positive tuples that were correctly classified by the classifier.

TN – True Negative: Refers negative tuples that were correctly classified by the classifier.

FP – False Positive: Refers negative tuples that were incorrectly labeled as positive by the classifier.

FN – False Negative: Refers positive tuples that were mislabeled as negative by the classifier.

Accuracy = (TP +TN) / (P+N)

Error Rate = (FP+FN) / (P+N)

Sensitivity (or) True Positive Rate (or) recall= TP/P

Specificity (or) True Negative Rate = TN / N

Precision = TP / (TP + FP)

In order to improve the accuracies of the decision tree; rule-based and lazy classifier the ensemble voting algorithm is used. In this study J48, PART and IBK algorithms are integrated to form ensemble voting. The accuracies are further improved by tuning the hyper-parameters of KNN and ensemble voting algorithm.

V. RESULT ANALYSIS, COMPARISON AND DISCUSSION

Different classifiers are applied to endometrial cancer data and the results are shown in table 3. The accuracy of Decision Tree, Rule-based classification, K-nearest neighbor and ensemble voting algorithms are 93.91%, 94.23%, 94.23%, and 94.87% respectively.

The IBK and Vote algorithm are tuned using hyper-parameters to improve the performance. The analysis was done in Weka experimenter. For the K-Nearest Neighbor algorithm, for different k values such as k=1, 3 and 7 and for the different distance measures such as Euclidean and Manhattan distance, the algorithm was tuned, and the outcome of tuning an experimentation were interpreted using statistical significance with confidence 0.05. After tuning the algorithm, the accuracies of IBK and ensemble Vote algorithms are 94.84% and 95.68% respectively. When compared with previous accuracy it was improved.

In this study the outcome of decision tree rules with un pruned tree is given as input to PART algorithm. The decision list generated by the PART algorithm is tested using Chi-square statistical test, and the association between the variables are found. The statistical analysis found the attributes are significantly related with each other. The rules generated by J48 and PART algorithm were given as input to EC-Risk Prediction mobile app to predict the risk of patients. EC-Risk Prediction app is a simple and user friendly mobile app. Using EC-Risk Prediction mobile app 100 patients' data were tested in real time and the results are shown in Table4. The new proposed mobile app is predicting the risk correctly. These finding will provide a foundation for early detection of EC and can support risk-informed decision making in clinical management.

Table 4. The Real-time results of patients using a mobile app

 DOI : 10.18231/2454-9150.2018.1392

Age	Menop	Obesity	Diabete	Vaginal	Watery	Estroge	Pelvic_	Infert	Cancer_Ri
55	Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High Risk
59	Post	Yes	No	Yes	Yes	No	Yes	No	Average R
48	Post	Yes	No	Yes	Yes	No	No	No	Average R
50	Peri	Yes	No	No	No	No	No	No	No Risk
65	Peri	No	No	No	Yes	No	Yes	No	No Risk
44	Peri	No	No	Yes	No	Yes	No	No	No Risk
48	Peri	No	No	No	No	No	No	No	No Risk
49	Post	No	No	No	No	No	No	No	No Risk
47	Peri	No	No	Yes	No	No	No	No	No Risk
52	Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High Risk
54	Post	Yes	No	Yes	Yes	No	Yes	No	Average R
56	Post	Yes	No	Yes	Yes	No	No	No	Average R
51	Peri	Yes	No	No	No	No	No	No	No Risk
53	Peri	No	No	No	Yes	No	Yes	No	No Risk

VI. CONCLUSION

The application of data mining in health care is of great importance because of its efficiency in disease prediction. The "EC-Risk Prediction" is a novel ensemble classification model to predict the risk of EC. It is a simple, user-friendly method implemented as a mobile application, requiring only nine features for prediction, utilizing Data Mining techniques (Decision Tree, Rule-based, lazy and ensemble classification method). The methods used in this app are straightforward and flexible giving scope for future refinement such as increasing the predictive value and intimating the physician based on the urgent situation. These findings provide a foundation for evaluating early detection strategies for EC and can support risk-informed decision making in clinical management. This data analysis has also specified interesting outcome like High Risk and mandatory to consult a doctor, Average risk - consult a doctor and No Risk.

REFERENCES

[1] Cancer – World Health Organization. Available: https://www.who.int / News / Fact sheets / Detail

[2] Post ablation Endometrial Carcinoma, NCBI, NIH. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/ *PMC5444558/*

[3] Endometrial Cancer – Wikipedia. Available: https://en.wikipedia.org/wiki/Endometrial_cancer

[4] Endometrial Cancer – American Cancer Society. Available: https://cancer.org/cancer/endometrial-cancer/about/what-is-endometrial-cancer.html

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

[6] A detailed guide – endometrial cancer. Available: http://cancer. org/ cancer/ endometrial cancer/cancer-risk-factors.

[7] Early Detection. Diagnosis and Staging. Test for Endometrial Cancer. Available: https://www.cancer. org/cancer/endometrial-cancer/detection-diagnosis-staging/how-diagnosed.html

[8] Huiqiao Gao and Zhenyu Zhang, "Systematic Analysis Of Endometrial Cancer-Associated Hub Proteins Based On Text Mining", Biomed Research International Volume 15 (2015), Article Id 615825

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

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

[11] E. Friberg & N. Orsini & C. S. Mantzoros & A. Wolk, "Diabetes Mellitus and Risk of Endometrial Cancer: A Meta-Analysis", Springer-Verlag 2007.

[12] Sridhar R, "Association Rule- Spatial Data Mining Approach For Exploration Of Endometrial Cancer Data", International Journal Of Advanced Research In Computer Science And Software Engineering, Volume 3(10), October 2013

[13] Weiderpass E, "Body Size In-Different Periods of Life, Diabetes Mellitus, Hypertension, and Risk of Postmenopausal Endometrial Cancer", Cancer Causes and Control - Published By Springer.

[14] Parazzini F, "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.

[15] 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

[16] Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, 3rd Edition, The Morgan Kaufmann Publications. Pages 99 – 105 & 490 - 520

[17] The Cancer Genome Atlas. A pilot project of the National Cancer Institute.

[18] Mack, Pike, Henderson B.E., Pfeffer R.I., Gerkins V.R., Arthur B.S. & Brown S.E(1976). "Estrogen and endometrial cancer in a retirement community". New England Journal of Medicine 294(23): 1262-1267.

[19] Rule based Classification, Available in http://staffwww .itn.liu.se/~aidvi/courses/06/dm/lectures/lec4.pdf

[20] Predictive Association Algorithm. Available in http://www.comp.nus.edu.sg/~dm2/

[21] Separate and Conquer, Available in ttp://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Rule_ Induction.pdf.

[22] K–Nearest NEighbours, available in https://gerardnico.com/data_mining/knn#classification.

[23] Wikipedia, the free encyclopedia, available in https://en.wikipedia.org/wiki/Ball_tree#mw-head

[24] Jason Brownlee, How to Tune Machine Learning Algorithms in Weka Machine Learning, August 1, 2016

[25] Ensemble of Heterogeneous Classifier Model. Available in www. shodhganga.inflibnet.ac.in/ bit stream/10603/36558/9/chapter%205.pd

 DOI : 10.18231/2454-9150.2018.1392