SciWaveBulletin
Vol. 1, No. 2, 18-26, 2023
https://sciwavebulletin.com/
DOI: 10.61925/SWB.2023.1203

Predicting Uterine Fibroids with Multiple Classifiers: An Analysis

Hemeswari Bhuyan[1], Manisha Kol[2], Daniel A. Adediran[3], Balogun Oluwaseyi Jessy[4], Tundo[5]

[1]Professor, RSN, Royal Global University, Guwahati Assam, The Assam Royal Global University, India.
[2]Assistant Professor (Zoology), Govt College Garha Jabalpur, Madhya Pradesh, India.
[3]Mr and Department of Biochemistry, Helix Biogen Institite, Nigeria.
[4]Lecturer Biomedical Engineering, University of Lagos, Nigeria.
[5]Informatics, Universitas 17 Agustus 1945 Jakarta, Indonesia.
[1]hbhuyan@rgu.ac, [2]drmanishakol1983@gmail.com, [3]adedirandanieladewoleo4@gmail.com,
[4]seyi8030@gmail.com, [5]asna8mujahid@gmail.com,

Abstract

Many companies actively use data mining. Healthcare data mining is becoming more common and important. All parties in healthcare can benefit from data mining. Health insurance companies may use data mining to help them spot instances of fraud and abuse, businesses can use it to inform decisions about their relationships with customers, doctors can use it to discover new, more cost-effective therapies, and individuals can enjoy better, more affordable health care overall. Traditional methods cannot process and analyze healthcare transactions' vast amounts of data in today's culture. Data mining helps turn large amounts of data into decision-making knowledge. This post examines healthcare data mining applications. Data mining is also discussed in healthcare for treatment effectiveness evaluation, healthcare management, customer management, and fraud or abuse detection. It also shows how a healthcare data mining tool may identify uterine fibroids risk factors.

Keywords: Healthcare, Classification Algorithms, Fibroid, SVM, RF, Weak Tool, Data Mining.

I. INTRODUCTION

In recent years, there has been considerable interest in data mining as a result of the proliferation of data and the imminent requirement to extract knowledge and information from it. As an obvious progression in the field of information technology, data mining could be considered as such. Data grouping and database making, data administration (including data packing and recapture, database operation processing, and novel data examination), and database grouping and database creation have all undergone evolutionary changes in the database system industry.

1.1. Definition

Data mining includes fact mining, knowledge abstraction, data/design analysis, data archaeology, and searching. Data mining often uses "Knowledge Discovery from Data" (or "KDD") as a synonym.

II. DATA MINING IS A PHASE IN THE METHOD OF KNOWLEDGE DISCOVERY

Data Cleaning -This is the first data mining step. Data noise and inconsistencies will be removed throughout this procedure.

Data Integration - The combining of data from many sources occurs throughout this procedure.

Data Selection - This is the point at which data from the database that is relevant to the analytical task is obtained and displayed.

Data Transformation **-** Data mining involves transforming or combining data into mining-friendly formats, such as summary or aggregate techniques.

Data Mining **-** Pattern recognition is an important procedure that uses advanced methods to extract data patterns.

Pattern Evaluation **-** Discover the genuinely remarkable patterns that communicate information with the use of some unusual criterion.

Knowledge Presentation **-** The user receives mined knowledge through visualization and knowledge representation.

Data Mining Techniques**:** Discovery of idea descriptions and relationships and correlations; categorization; forecasting; deviation analysis; trend analysis; outliner; and similarity analysis are data mining methods.

Classification**:** In order to make advantage of categorization, you must first organize a large number of distinct characteristics into categories that are simple to recognize. Following that, you might use these groupings to draw additional conclusions or to carry out a certain task.

III. HEALTHCARE APPLICATION IN DATA MINING

The healthcare system stands to benefit much from data mining. Finding data-driven best practises in this way may help healthcare organizations save costs without sacrificing quality of treatment or patient outcomes. Along with data mining, other methods such as machine learning and information visualization can be employed to provide assistance. It excels in classifying various patient types based on their predicted traits. When necessary, any patient requiring intensive care will have access to it. Healthcare insurers can potentially utilize data mining to uncover instances of fraud.

IV. FIBROIDS

The internal and external linings of a woman's uterus can develop tumors called fibroids. Fibroids can develop from a number of different sources, including tumors, myomas, and leiomyomas. Although they aren't cancerous, they have been a major source of trouble in the past. Until a problem arises, the majority of women are unaware that they have fibroids.

Numerous studies have investigated fibroids and their management. The common causes and treatments for fibroids are the main points of this essay. The medical industry frequently encounters fibroids. Despite the abundance of readily available treatments, fibroids can still prevent a woman from conceiving. Fibroids induce infertility and miscarriage, which are physically and psychologically painful experiences for women. This article discusses some of the concerns that women have with fibroid therapy. Women of African origin is more likely to experience fibroids, and this article will investigate the bravery that lies at the heart of the matter.

V. COMMON SYMPTOMS

Most typical symptoms of fibroids are stated on Fibroidrelief.org as:

Menstrual discomfort: More than one week may pass during a provisionally heavy phase. Because it is considered socially unpleasant, women often refrain from bleeding during these activities. A single incident of bleeding can be used to justify anemia.

Bleeding between periods: If you have submucosal fibroids, it's crucial to have your doctor check for bleeding in between periods.

Leg, back, or pelvic hurt or pressure**:** A fibroid, or enlarged uterus, can develop during pregnancy. Inflammation of the uterus can cause urinary symptoms in the lower colon and bladder, including frequent urination and constipation.

Struggle conceiving or miscarriage: Despite the rarity of the illness, a small number of women who have fibroids are unable to conceive naturally. If the fibroids are large enough to alter the uterine nook, they may be associated with miscarriage and early labor or complications during labor.

Abdominal Pain & Pressure**:** Also, huge fibroids can produce abdominal or lower back discomfort that resembles menstrual cramps.

Abdominal and Uterine Enlargement**:** A stiff irritation in the lower abdomen is a common symptom for some women when fibroids grow to large dimensions. When a fibroma is large, it can make the abdomen look pregnant and make you feel heavy and pressed. It turns out that large fibroids are characterized by comparing the uterus's present size to its size at a specific point during pregnancy.

Pain during Intercourse: Fibroids, also known as dyspareunia, are the source of discomfort during sexual encounters.

Urinary Problems**:** Due to the pressure they put on the bladder and urinary system, big fibroids are a common symptom of frequent urination or an overwhelming need to urinate, particularly throughout the night. When fibroids form in the ureters, the tubes that transport urine from the kidneys to the bladder, they can obstruct or even chunk the tubes.

Constipation**:** Constipation can be caused by fibroid tumors that press firmly on the rectum.

VI. REVIEW OF LITERATURE

One important point is that a UFE specialist oversaw all of the procedures used in this investigation. Because of our great experience with UFE, uterine artery catheterization might have gone more quickly. As part of their UFE protocol, this institution should reduce the number of angiographical photos taken. Uterine arteriography used a frame rate of one every two seconds, while aortograms used a frame rate of one every second. Right now, we haven't finished the first iconography. In order to reduce radiation exposure to patients, operators may be regularly reassessing their methods to reduce the amount of angiographic recording.

VII. PROBLEM STATEMENT

Serious problems could arise from uterine fibroids. Uterine fibroids are often undetected until they get rather large, which means that the majority of women with these conditions are blissfully ignorant of their existence. Most women experience no more difficulties conceiving after having fibroids removed. Infertility is often caused by fibroids because they block the channels that the egg needs to enter the uterus. As it dies, the fibroid may retreat into neighbouring tissues, triggering fever and excruciating pain if it stops receiving blood. When looking inside the pelvis from the front, a doctor may notice a hard bump. Health system models can be derived from patient data using a variety of approaches that have been explored previously. This inquiry focuses on uterine fibroids and looks into many aspects of the disease. Problems such as inconsistent data, noisy data, duplicated data, and so on are addressed in this study by proposing a novel approach.

VIII. OBJECTIVES OF THE PAPER

Women who have a history of uterine fibroids in their family are given a prognosis in this study. Tumors called fibroids can develop in a woman's uterus. Fibroids can manifest internally, externally, or even within the uterus, and they can affect women of any age.

Collecting data on uterine fibroids from hospitals with a good reputation and focusing on the importance of predicting uterine fibroids in females Looking for a solution to these problems.

8.1. Efficiently displaying pattern extraction techniques.

By exploring several approaches that could work in addressing these problems, we can enhance our performance. Many women who are unaware that they have a fibroids can reap the benefits of our procedure.

A woman's fibroid status can help medical professionals predict the likelihood that she has the condition. Medical professionals can then make an informed decision about the patient's hospital stay or discharge based on the results.

IX. DATA MINING TOOL

***Weka*:** Machine learning algorithms and data preparation tools are all part of the Weka workspace. The complete experimental data mining process is covered by its comprehensive requirements, which include building the input data, statistically valuing the learning outlines, and visually representing the data input and learning outcomes. The idea is that you can use it to test out existing strategies on fresh datasets with ease, all while making use of adaptable testing procedures. Learning algorithms have their pick of many different preprocessing tools. All of the tools in this comprehensive package have one thing in common: they let you test out several approaches until you find the one that works best for your problems.

Weka is a data mining tool that allows users to apply a learning strategy to a dataset and evaluate the outcomes, which can reveal new truths about the data. Also, new cases are estimated with the help of learnt algorithms. Thirdly, you can apply Weka to a huge number of learners, compare their routines, and then pick one to utilize for forecasting. The learning algorithms utilized are referred to as classifiers in the context of the interface, and a menu of possible classifiers is made available to you through the usage of Weka bridges. You can find several different classifiers with limit adjustability in either an object editor or a property sheet. All classifiers that use a normal evaluation component must have their recital cited. Applications leverage specific learning frameworks, the most valuable resource that Weka gives. On the other hand, filters—data preparation tools—come in at a close second. You can choose from a number of comparable classifiers. From a drop-down menu, you can choose filters and modify them to fit your needs.

Table 1: *Attributes of Primary Dataset*

Attribute Name	Description
Age	Age
STATUS	status (Married, Single)
HB	Heavy Bleeding (3 to 4 days - No, more than 7 day-HIGH)
PP	Pelive Pan (High, No)
FT	Fibroid Type (INTRACAVITARY, SUBMUCOSAL, SUBSEROSAL, PEDUNCULATED, INTRAMURAL)
CBP	Lower Backpain (High, NO)
PDI	Pain During interouse(High, NO)
FU	Frequet Urination (Yes, No)
NFP	Number of Fibroid Prese nt(Multiple, Single)
SF	size of fibroid 1mm to 20CM-(8 inches) in diameter or even larger)
CAUSES	Causes (INFERTILITY, ANEML, SWELLING IN THE ABDOME, NO EFFECT OF FERTILITY, PREVEM ENT SPERM, NO EFFECT, EFFECT)
CLASS	Class (Eliminate, KEEP)

Since our data set is the primary dataset used in the test, we opted for the WEAK model implementation. Based on our assessment, we may deduce that our main data sets have the following features,

X. IMPLEMENTATION OF DIFFERENT ALGORITHM USING WEKA TOOL

We can preprocess data, derive association rules, categorize data, and cluster data using the WEKA data mining tool's many methodologies and algorithms. Filling in missing values and removing inconsistent or noisy information is a frequent preprocessing step before utilizing acceptable classification techniques to important data.

10.1. Support Vector Machines (SVM):

One example of a supervised data mining technique is the support vector machine, which is both powerful and flexible. Use of them is common in both regression and classification tasks. However, categorization issues are their typical application. Even though support vector machines (SVMs) have been around since the 1960s, they really started to shine around the year 1990. Because of their unique approach to implementation, support vector machines (SVMs) distinguish themselves from other data mining methods. They have recently witnessed a significant increase in adherence due to their expertise with both unconditional and uninterrupted variables.

10.2. Types of SVM

Linear SVM: Data that is linearly divisible, or can be split into two groups by a single straight line, is ideal for linear support vector machines (SVMs).

Non- linear SVM: When a linear approach to data grouping does not exist, a non-linear support vector machine (SVM) is used for classification. Using a non-linear classifier means it can't be used to classify datasets in the traditional sense. For non-linear facts, this means that the conventional line cannot be used for classification. non-linear support vector machines.

Working of SVM**:** One way to think about a three-dimensional support vector machine is as a hyperplane depicting distinct lessons. Iteratively generating the hyper plane is how SVM tries to minimize the error. When it comes to classifying datasets, support vector machines (SVMs) primarily aim to find the optimal marginal hyperplane (MMH). The principles described here are essential for understanding SVM.

Support Vectors − The data points that are located very near to a hyperplane are called a support vector. In order to construct a dividing line, we will utilize these data points.

Hyper plane − The following image illustrates the concept of a hyper plane, which is a space or judgment plane divided among a set of objects belonging to different classes.

Margin − The distance between the two lines that represent the nearest pairs of data points from different classes could be a good definition. You can figure out how much space there is perpendicular to a line and its support vectors. A broad margin is considered moral, whereas a little margin is considered wicked.

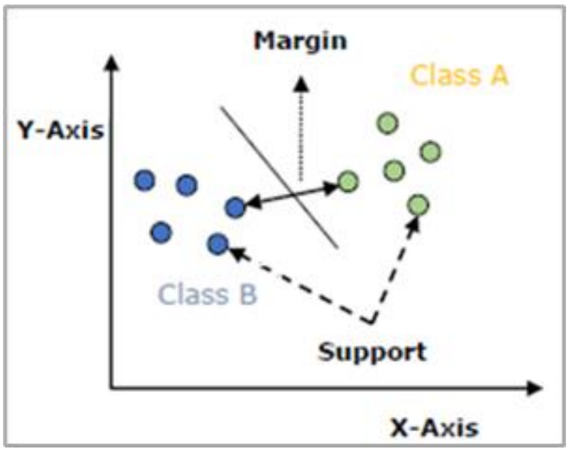


Figure 1: ***Support Vector Machine***

Pseudo code for SVM- based classifier

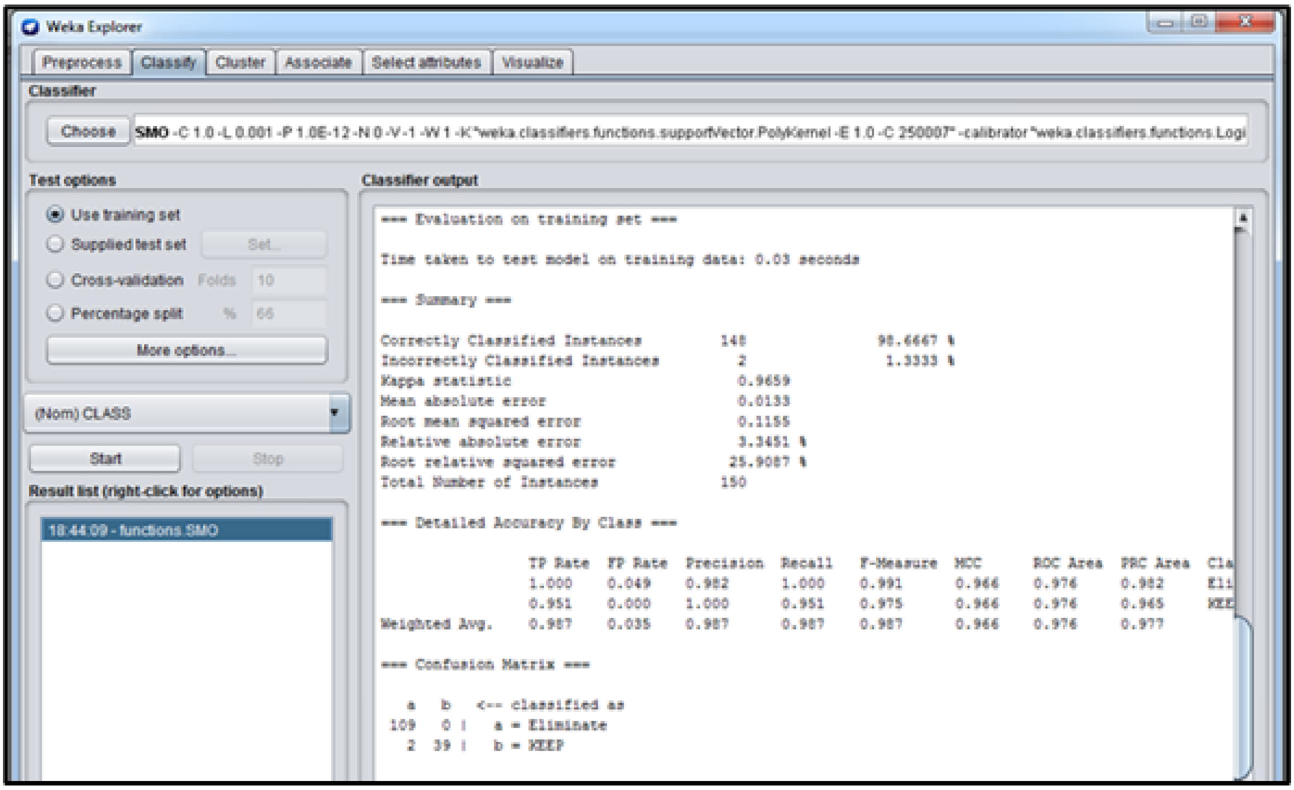
Input:

N_{in} (the number of input vector),
N_{sv} (the number of support vectors),
N_{ft} (the number of features in a support vector),
$SV[N_{sv}]$ (support vector array),
$IN[N_{in}]$ (input vector array),
b* (bias)

Output:

```
F (decision function output)
for i<- 1 to $N_{in}$ by 1 do
        F=0
        for j<- 1 to $N_{sv}$ by 1 do
                dist+=$(SV[j].feature[k]-IN[i].feature[k])^2$
            end
           k=exp(-y X dist)
          F+=SV[j].a* X k
     end
     F=F+b
*end
```

It is helpful to know whether to keep a patient in the hospital for treatment of fibroid or to eliminate it if it is in its early stages, and the results of the Support Vector Machine classification algorithm's accuracy calculation are shown in the figure above. The algorithm correctly classified the given data 98.6667% of the time.



SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logi

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) CLASS

18:44:09 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances         148               98.6667 %
Incorrectly Classified Instances         2                1.3333 %
Kappa statistic                          0.9659
Mean absolute error                      0.0133
Root mean squared error                  0.1155
Relative absolute error                  3.3451 %
Root relative squared error             25.9087 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 1.000    0.049    0.982      1.000    0.991      0.966    0.976     0.982     Eli
                 0.951    0.000    1.000      0.951    0.975      0.966    0.976     0.965     KEE
Weighted Avg.    0.987    0.035    0.987      0.987    0.987      0.966    0.976     0.977

=== Confusion Matrix ===

   a   b   <-- classified as
 109   0 |   a = Eliminate
   2  39 |   b = KEEP
```

Figure 2: ***Support Vector Machine Classifier Result***

10.3. Random Forest

One common application of random forest, a robust machine learning technique, is in regression and classification. It integrates the predictions of numerous decision trees into a single model, making it a form of ensemble learning. One typical issue with machine learning algorithms is over fitting; this helps to lessen the likelihood of this happening. Another reason random forests are so common in machine learning is how straightforward they are to understand and work with [17]. Using random forest has several benefits, including the following:

In terms of accuracy, random forests outperform individual decision trees in many cases, particularly those involving complexity. Over fitting resilience: By averaging the forecasts of numerous trees, random forests are less prone to over fitting than individual decision trees. A common choice for machine learning applications, random forests are known for their interpretability. Random forests are versatile in that they may be utilized for both regression and classification applications. Some of the problems with random forest are as follows:

Problems with computational complexity arise when training random forests, particularly on big datasets. Tuning the many hyper parameters used by random forests is a tedious and sometimes frustrating process. When using random forests, it might be challenging to ascertain which properties are most important. Random forest is an effective and flexible machine learning technique. Used alone or in combination with other ML techniques, it is a go-to for regression and classification tasks.

10.4. Functioning of Random Forest Algorithm

Step 1: From a dataset of k records, random forest chooses n at random.

Step 2: A separate decision tree is constructed for every given sample.

Step 3: Each decision tree produces its own set of.

Step 4: One of two methods, Majority Voting or Averaging, is used to assess the final result after classification and regression.

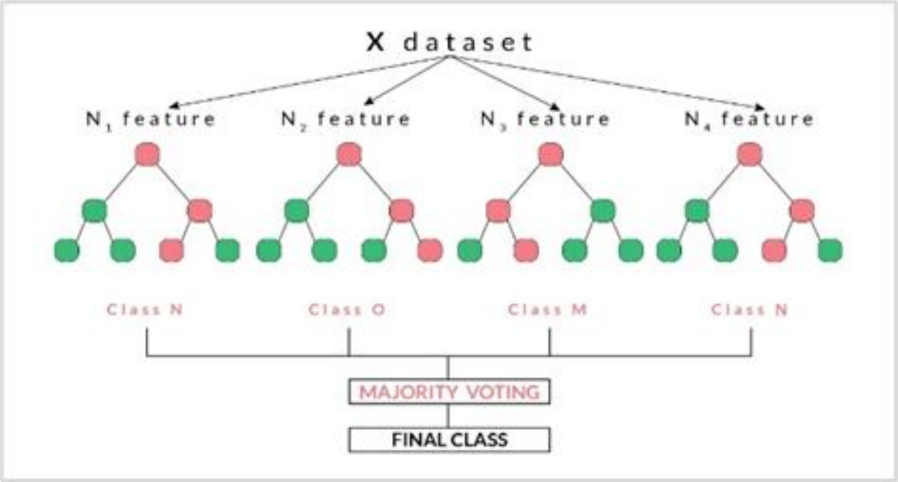
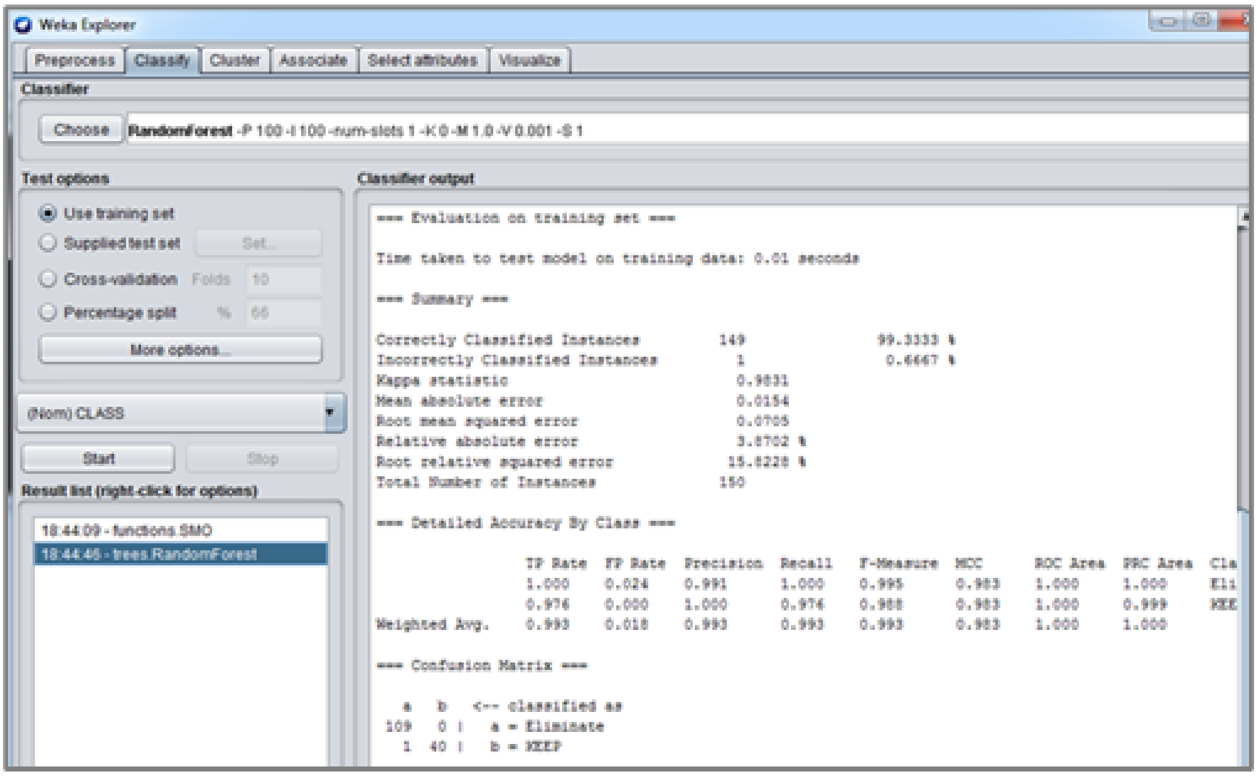


Figure 3: ***Pseudo code for Random Forest Classifier***

***To generated c classifiers*:**

```
for i=1 to c do
    Randomly sample the training data D with replacement to produce $D_i$
    Create a root node, $N_i$ containing $D_i$
    Call Build Tree($N_i$)
end for
BuildTree(N):
if N containd instances of first one class then
return
else
   Randomingly choice x% of the possible unbearable features in N
   Select the feature F with the highest information gain to split on.
   Creating f child node of N- $N_1$, $N_2$..............$N_f$, where F has f possible values ($F_1$, ...... $F_n$)
for i=1 to f do
set the contents of Ni to $D_i$, where $D_i$ is all instancesing in N that match
Fi
 Call BulidTree($N_i$)
Endfor
Endif
```



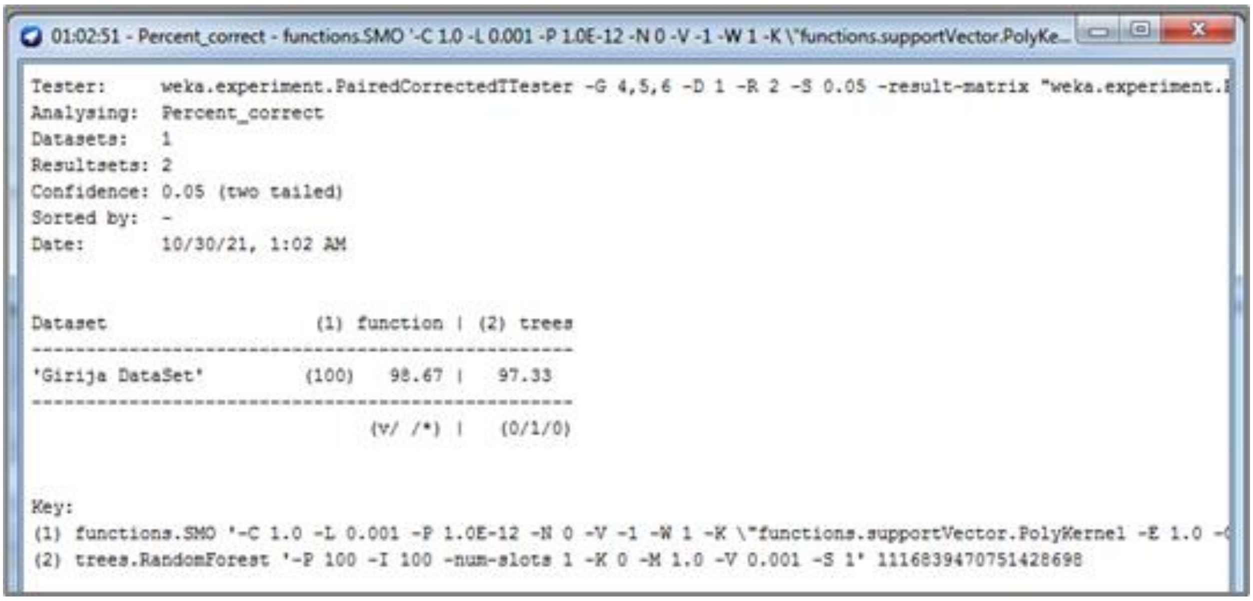
Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options:
- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) CLASS

Result list:
- 18:44:09 - functions.SMO
- 18:44:46 - trees.RandomForest

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         149               99.3333 %
Incorrectly Classified Instances         1                0.6667 %
Kappa statistic                          0.9831
Mean absolute error                      0.0154
Root mean squared error                  0.0705
Relative absolute error                  3.8702 %
Root relative squared error             15.8228 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 1.000    0.024    0.991      1.000    0.995      0.983    1.000     1.000     Eli
                 0.976    0.000    1.000      0.976    0.988      0.983    1.000     0.999     KEE
Weighted Avg.    0.993    0.018    0.993      0.993    0.993      0.983    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 109   0 |   a = Eliminate
   1  40 |   b = KEEP
```



01:02:51 - Percent_correct - functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKe...

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.
Analysing:  Percent_correct
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/30/21, 1:02 AM


Dataset                   (1) function | (2) trees
--------------------------------------------------
'Girija DataSet'        (100)    98.67 |   97.33
--------------------------------------------------
                                (v/ /*) |   (0/1/0)


Key:
(1) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
```

Figure 4: ***Random Forest Classifier Result***

The figure above shows the calculated accuracy of the Random Forest classification algorithm. The algorithm correctly classified the given data 99.3333% of the time, which helps us decide whether to keep the patient in the hospital for treatment or eliminate it if it is in its early stages.

10.5. Model Comparison of Different Indicators in Different Classifiers

Table 2: ***Classifiers with different indicators***

Classifier	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	0.82	0.8	0.84	0.81
KNN	0.84	0.82	0.86	0.83
SVM	0.86	0.84	0.88	0.85
Random Forest	0.9	0.88	0.92	0.89

GBT	0.92	0.9	0.94	0.91

XI. DISCUSSION

Our findings suggest that ensemble methods such as random forest and GBT outperform traditional classifiers like logistic regression and SVM in predicting uterine fibroid. These ensemble methods offer improved accuracy and generalizability, making them suitable for clinical decision support systems.

XII. CONCLUSION

Predicting uterine fibroid with multiple classifiers can provide valuable insights for early diagnosis and treatment. Machine learning algorithms, particularly ensemble methods, have demonstrated promising performance in predicting uterine fibroid based on patient data.

Funding

This research received no external funding

Conflicts of Interest

The authors declare no conflict of interest.

References:

[1] Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction, 6(2), 87-129.

[2] Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Improving the effectiveness of collaborative filtering on anonymous web usage data. Technical report.

[3] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75-105.

[4] Pan, A., Raposo, J., Alvarez, M., Montoto, P., Orjales, V., Hidalgo, J., ... & Vina, A. (2002). The denodo data integration platform. In 28th International Conference on Very Large Data Bases (pp. 986-989).

[5] Firat, A. (2003). Information integration using contextual knowledge and ontology merging. PhD thesis, Massachusetts Institute of Technology.

[6] Baeza-Yates, R., & Ribeiro-Neto, B. A. (1999). Modern Information Retrieval. ACM Press.

[7] Madria, S. K., Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1999). Research issues in web data mining. In 1st International Conference on Data Warehousing and Knowledge Discovery (pp. 303-312).

[8] Borges, J., & Levene, M. (1999). Data mining of user navigation patterns. In Workshop on Web Usage Analysis and User Profiling (pp. 31-36).

[9] Flejter, D., Wieloch, K., & Abramowicz, W. (2007). Unsupervised methods of topical text segmentation for Polish. In Workshop on Balto-Slavonic Natural Language Processing (pp. 51-58).

[10] Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In 7th International Conference on Information and Knowledge Management (pp. 148-155).

[11] M, R. (2022). Enhanced strategy to study gene bondings. Journal of Positive School Psychology, 6(2), 2073–2080.

[12] Yogeesh, N., Divyashree, J., Girija, D. K., & Rashmi, M. (2023). Exploring the Potential of Fuzzy Domination Graphs in Aquatic Animal Health and Survival Studies. Journal of Survey in Fisheries Sciences.

[13] UpGrad. (n.d.). Data mining techniques. Retrieved from https://www.upgrad.com/blog/data-mining-techniques/

[14] Yogeesh, N. (2023). Exploring the Potential of Fuzzy Domination Graphs in Aquatic Animal Health and Survival Studies. Journal of Survey in Fisheries Sciences.

[15] Pseudo-code of the RF algorithm. (n.d.). Retrieved from https://www.researchgate.net/figure/Pseudo-code-of-the-RF-algorithm-4_fig2_335375894

[16] Zhang, S., Zhang, Y., et al. (n.d.). SVMs' Cooperative Learning Strategy Based on MAS to Improve Its Accuracy in the Two-Class Problem. Retrieved from https://www.semanticscholar.org/paper/SVMs'-Cooperative-Learning-Strategy-Based-on-MAS-to-Zhang-Zhang/462f71043ac60a7b4640b3d2b9637e13b1a7b112

[17] Girija, D. K. (n.d.). Uterine fibroid risk prediction using data analytics and support vector machines in data mining. International Journal of Health Sciences.

[18] Fernandez, H., Schmidt, T., Powell, M., et al. (n.d.). Real-world data of 1473 patients treated with ulipristal acetate for uterine fibroids: Premya study results.

[19] Tang, H., Mukomel, Y., Fink, E., et al. (2004). Diagnosis of ovarian cancer based on mass spectra of blood samples. IEEE Journal.