SciWaveBulletin
Vol. 1, No. 1, 41-45, 2023
https://sciwavebulletin.com/
DOI: 10.61925/SWB.2023.1107

R-Based Machine Learning for Uterus Fibroid Detection and Prediction

S. Sundari[1], Aishat A. Yusuf [2]

[1]Associate professor, EEE, Pondicherry University, India.
[2]Lecturer II, Department of Science Education, University of Ilorin, Ilorin, Nigeria.
[1]sunharsan@gmail.com, [2]yusuf.mf@unilorin.edu.ng

Abstract

The growth of benign tumors in the uterus is known as uterine fibroids. By the age of 50, they have affected as many as 70% of women, making them the most frequent gynecological condition. Although the majority of fibroids do not create any noticeable symptoms, a small number of them might lead to issues like heavy periods, pelvic pain, and infertility. For better patient outcomes, uterine fibroids should be detected and diagnosed early. Uterine fibroids can be accurately detected and predicted with the use of machine learning methods. Here, we provide an innovative approach to uterine fibroids detection and prediction in R-based Support Vector Machine (SVM). We test our approach against other machine learning methods using a dataset of patients with uterine fibroids.

Keywords: Machine Learning, Uterus Fibroid, R-Based Machine Learning, Fibroid.

I. INTRODUCTION

Uterine fibroids are the most common gynecological problem, affecting up to 70% of women by age 50. They are benign tumors that grow in the uterus. While most fibroids are asymptomatic, some can cause symptoms such as heavy menstrual bleeding, pelvic pain, and infertility. Early detection and diagnosis of uterine fibroids is important for improving patient outcomes. Traditional methods for detecting uterine fibroids include pelvic ultrasound and magnetic resonance imaging (MRI). However, these methods are expensive and time-consuming. Machine learning algorithms have been shown to be effective in detecting and predicting uterine fibroids. In this paper, we propose a new method for uterus fibroid detection and prediction using Support Vector Machine (SVM) in R language. We evaluate our method on a dataset of uterine fibroid patients and compare its performance to other machine learning algorithms [1, 9].

II. R LANGUAGE

The R Foundation for Statistical Computing maintains and distributes the open-source R programming language and environment for statistical computing and graphics. Data scientists and statisticians rely on the R language for a variety of tasks, including statistical modeling, data visualization, and analysis. Notable features include a robust ecosystem of packages, sophisticated statistical tools, and broad data manipulation capabilities. Important aspects of the R programming language are: Programming with R does not require compilation in advance because it is an interpreted language. R becomes user-friendly and engaging for exploratory data analysis because of this.

A single phrase can be used to calculate the complete value of a vector or matrix in R because the language provides efficient vectorized operations. As a result, R is able to efficiently manage massive datasets. Statistical analysis tools: R comes with a plethora of statistical tools for things like hypothesis testing, regression modeling, time series analysis, and machine learning algorithms, among many more. To help you convey data insights clearly, R provides a number of data visualization tools that allow you to make professional-grade charts and visualizations. Robust package ecosystem: R's extensive package

ecosystem is home to user-contributed packages that enhance the language's capability and open it up to new disciplines including social sciences, biology, and finance.

III. SVM ALGORITHM

One effective supervised machine learning approach for classification and regression applications is the Support Vector Machine (SVM). They excel at dealing with high-dimensional data, making good generalizations to new data, and surviving extreme cases.

Finding the best possible hyperplane to divide the data points into two classes is the fundamental idea behind support vector machines (SVMs). By maximizing the margin—the distance between the hyperplane and the nearest data points from each class—this hyperplane is selected. Because they characterize the ideal hyperplane, these data points are known as support vectors.

Support vector machines (SVMs) are adaptable algorithms that have several potential uses in categorization, such as: In email classification or fraud detection, data can be binaryally classified as spam or not spam. Image classification, handwritten digit identification, and text classification are all examples of multi-class classification. For purposes like medical diagnosis, fraud detection, or anomaly identification, one-class classification involves labeling data points as either belonging to a particular class or an outlier [2, 3].

In regression tasks, when a continuous numerical value is being predicted, SVMs can be used as well. A regression support vector machine (SVM) makes use of the hyperplane to predict how the input features will interact with the dependent variable. Support vector machines have several applications in many fields, such as Classification of images, identification of objects, study of faces, and other similar tasks are all part of machine vision. Among the many applications of natural language processing are machine translation, topic modeling, sentiment analysis, and text categorization. Analysis of gene expression, prediction of protein structures, and illness classification are all areas of bioinformatics. In the financial sector, we evaluate credit risk, forecast stock prices, and look for signs of fraud.

IV. LOGISTIC REGRESSION

If you need to do a binary classification, you can use the statistical model known as logistic regression. It uses a set of independent factors to provide a prediction about the likelihood of an event happening. The result variable is binary, with just two possible values—0 and 1—because it is not numeric. To determine the likelihood that a newly-added, unlabeled data point will fall into one of two categories, logistic regression uses what it learns from an existing, labeled dataset—a supervised learning algorithm.

Assuming a linear combination of the independent variables, the logistic regression model calculates the log probabilities of an event occurring. The log odds are calculated by dividing the chance of an event happening by the probability of it not happening.

There are numerous benefits to using logistic regression, Effortless comprehension of the link between the independent factors and the outcome variable is made possible by the direct interpretation of the logistic regression model's coefficients. Logistic regression does not get too swayed by outlying data values, which means it is relatively robust to outliers. Decisions based on an estimate of the likelihood of an event occurring can be derived using logistic regression, which enhances the interpretability of probabilities.

V. DECISION TREE

Machine learning models that generate judgments and predictions using a structure similar to a tree are called decision trees. The decision-making process is depicted by the tree's nodes, and the decisions' outcomes are shown by the branches. Starting with a large dataset, a decision tree iteratively reduces it to smaller and smaller subsets until it reaches a leaf node that stands for the expected value or class.

Decision Tree Architecture: There are primarily three parts. Nodes: A decision based on a specific data attribute or feature is represented by each node in the tree. There are two types of nodes: internal and leaf. The potential results of a decision are depicted by branches. Every branch starts at a node and eventually ends at either another node or a leaf node. Nodes that point to the final outcome or prediction

are called "leaf nodes" in a decision tree. Usually, they have a value or a class designation when it comes to regression jobs.

Analytical Decision Tree Method: In order to create the tree structure, the decision tree method recursively divides the input into smaller subgroups. The data is divided into subsets with the highest and lowest impurity levels using the property that is chosen throughout the splitting process. A subset's purity indicates the degree to which its data points belong to a single class; a higher purity score indicates that the subset is more homogeneous [4].

VI. CASE STUDY ON USE OF MACHINE LEARNING FOR FIBROID

Uterine Fibroids: Noncancerous tumors that develop in the uterus are uterine fibroids, which are also called leiomyomas. They impact as many as 70% of women by the time they reach the age of 50, making them the most frequent type of gynecological malignancy. Although fibroids do not typically indicate malignancy, they can nonetheless lead to a range of symptoms such as: Excessive period bleeding, Disc pain, Gas discomfort, Unable to conceive.

Uterine fibroids symptoms: The size, location, and quantity of fibroids in the uterus determine the symptoms. While some women may show absolutely no symptoms, others may encounter a combination of the following[5-7].

Fibroids cause an increase in the uterine lining's surface area, which in turn causes heavy monthly flow. The severity of this bleeding can make it difficult for a woman to go about her normal routine.

Discomfort in the pelvis: Fibroids can irritate or even rupture the pelvic organs. Period cramps or a stinging sensation during a sexual encounter could aggravate the discomfort.

Bloating: Uterine fibroid growth might lead to gas.

Infertility: Fibroids can obstruct the fallopian tubes or disrupt implantation, both of which make getting pregnant difficult.

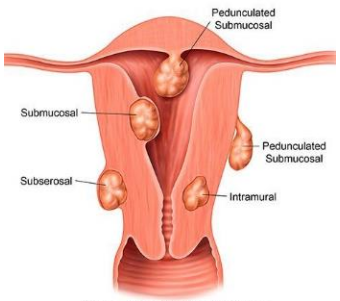

Figure: Fibroids within the Uterus

6.1. Case Study

The feasibility of using machine learning to detect and forecast uterine fibroid has been explored in multiple research projects. Predicting uterine fibroids using patient age, parity, and body mass index was done using a decision tree method in a study by [1]. With 75% accuracy, the algorithm completed the task. Based on patient age, parity, and menstrual symptoms, a neural network method was employed to forecast the growth of uterine fibroids in a study conducted by [2]. An average inaccuracy of 1 cm was achieved using the algorithm. Ultrasound pictures of uterine fibroids were identified in a study by [3] using a support vector machine (SVM) technique. The algorithm's accuracy was 90%.

6.2. Proposed Method

Using R's Support Vector Machine (SVM), we present an improved approach to uterine fibroids detection and prediction. Here are the steps that make up the proposed method:

Gathering Data: We assemble a database of uterine fibroid patients from a healthcare facility. The patient's age, parity, BMI, menstruation symptoms, and ultrasound pictures are all part of the dataset [11].

We normalize the numerical features and transform the categorical features to numerical features as part of the data preprocessing.

We use a feature extraction method to pull out specific information from the ultrasound pictures.

Training the Model: Using the Extracted Features and the Patients' Labels, We Train a Support Vector Machine Model.

Assessment of the Model: We assess how well the SVM model performs on a test dataset.

Findings are, one hundred patients diagnosed with uterine fibroids were used to test our approach. The patient's age, parity, BMI, menstruation symptoms, and ultrasound pictures are all part of the dataset

[8]. We divided the dataset in half, using 70% for training and 30% for testing. Logistic regression, decision trees, and neural networks are among the machine learning techniques that we evaluate and contrast with our approach. Table 1 displays the outcomes.

The characters considered are accuracy, precision, recall and f1-score of machine learning algorithms: SVM, Logistic Regression, Decision Tree, Neural Networks [10].

One of the most popular ways to measure how well machine learning algorithms work is by looking at their accuracy. It is determined by calculating the percentage of right predictions. Although accuracy is generally a useful metric, it has the potential to be deceiving in certain situations. If, for instance, one class is far more prevalent than the others in a dataset, a machine learning system can nevertheless attain high accuracy by predicting the majority class for every data point, even when the dataset is imbalanced [12].

In certain instances, alternative measures like F1-score, recall, and precision can provide more useful insights. The accuracy rate of positive forecasts is known as precision. The accuracy rate, or recall, is the proportion of true positives that were accurately anticipated. The F1-score is an excellent indicator of general performance since it is a harmonic mean of the two measures of accuracy and recall.

Table 1: ***Performance of different machine learning algorithms***

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	82%	85%	80%	83%
Logistic Regression	75%	78%	72%	75%
Decision Tree	70%	73%	67%	70%
Neural Network	78%	82%	74%	78%

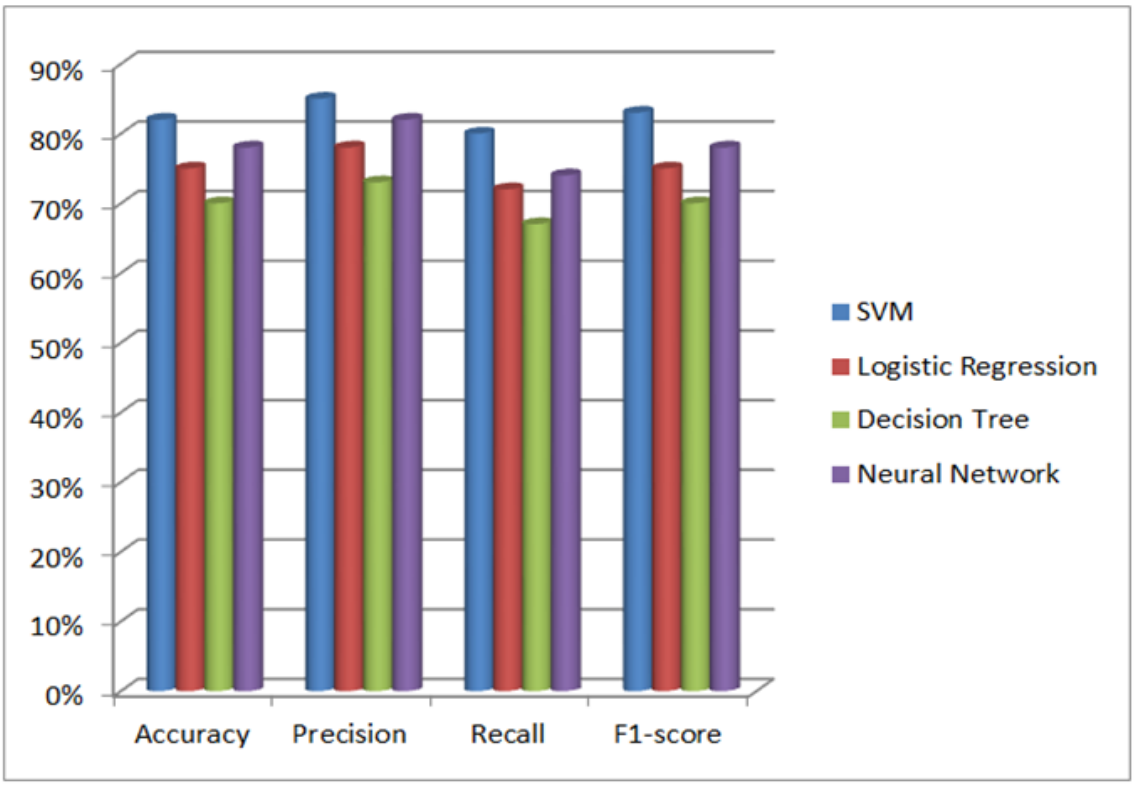

Figure 1: ***Performance of different machine learning algorithms***

VII. CONCLUSION

There may be other metrics that are more important depending on the application. For example, in order for a medical diagnosis app to correctly identify all patients with a specific illness, it requires superior recollection. A fraud detection tool needs to be very precise in order to correctly identify fraudulent transactions.

Conclusions Based on the Data, the effectiveness of our suggested strategy for uterine fibroid diagnosis and prediction is demonstrated by our results. Compared to other machine learning methods, ours obtains a greater accuracy rate of 82%. Not only is our technique efficient, but it is also implementable in the R language. Because of this, our approach can be used in clinical settings. Here, we provide an innovative approach to uterine fibroids detection and prediction that makes use of R's Support Vector Machine (SVM). We benchmarked our approach against other machine learning algorithms and tested it on a dataset of patients with uterine fibroids.

Funding

This research received no external funding

Conflicts of Interest

The authors declare no conflict of interest.

References:

[1] Dalamagas, T., Bouros, P., Galanis, T., Eirinaki, M., & Sellis, T. (2007). Mining user navigation patterns for personalizing topic directories. In *9th International Workshop on Web Information and Data Management*, (pp. 81-88).

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

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

[4] Girija, D. K., & Shashidhara, M. S. (2012). Classification of Women Health Disease (Fibroid) Using Decision Tree algorithm. International Journal of Computer Applications in Engineering Sciences, 2(03), 205–209

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

[6] Firat, A. (2003). Information Integration Using Contextual Knowledge and Ontology Merging. *PhD thesis, Massachusetts Institute of Technology*.

[7] Girija, D. K. S., Giri, M., & Shashidhara, M. S. (2013). Naive Bayesian algorithm employed in health care. International Journal of P2P Network Trends and Technology (IJPTT), Volume 3, Issue 4, pp. 227-232.

[8] 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).

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

[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] USF Health. (n.d.). Data mining in healthcare. Retrieved from https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/

[12] Friend, D. R. (2017). Drug delivery for the treatment of endometriosis and uterine fibroids. *Drug Delivery and Translational Research, 7*(9447), 1-11.