RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 07, Issue 04 April

Check for updates

http://dx.doi.org/10.47392/IRJASH.2025.039

Gynaecological Disease Detection: A Machine Learning Algorithm and Natural Language Processing Approach

Naga Swaroopa Chitveli[1], Hima chandana Nandaluri[2], Bharath kumar Kamatham[3], Harathi Badugu[4], Jaswanth Battu Siddapuram[5]

[1](M.tech, Ph.D),Assistant professor, Dept. Of CSE, Annamacharya University, Rajampet, India.

[2,3,4,5]UG Scholar, Dept. Of CSE, Annamacharya Institute of Technology and Sciences, Rajampet, India.

Emails: *swarupabaalu@gmail.com[1], nandalurihimachandana@gmail.com[2], bharathkamatham143@gmail.com[3], harathi.badugu@gmail.com[4], jaswanthbattusiddapuram@gmail.com[5]*

Article history

Received: 06 March 2025
Accepted: 21 March 2025
Published: 18 April 2025

Keywords:
Preliminary Diagnosis of Gynecological diseases, Supervised learning, Decision trees, Logistic models, gradient boosting.

Abstract

As demonstrated in this project, the GDDES is improved through the application of machine learning as well as natural language processing to diagnose the common gynecological diseases that include Urinary Tract Infection (UTI) as well as Polycystic Ovary Syndrome (PCOS). The constructiveness of the current structure employs traditional probabilistic schemes including the Decision Tree, Random Forest Classifier, SVC, Naïve Bayes, and the K-Nearest Neighbor to classify. The performance and diagnostic capability increases with the help of algorithms such as Logistic Regression and Gradient Boosting Models in the proposed system. As such, it leans on the power of the NLP algorithm to scour through the patient records and symptoms for a completely automated diagnosis. The above approach is meant to enhance precision in detection or diagnosis processes as well as the amount of time needed for such diagnosis so as to develop a tool that can be regarded as reliable in the profession.

1. Introduction

The Gynecological Disease Diagnosis Expert System (GDDES) is new and unique in Women's health field which uses machine learning-NLP. UTIs and PCOS present difficulties in their diagnosis making process, and this results to delayed prognosis leading to longer suffering of the affected patient. The proposed GDDES endeavor is to optimize the diagnostic protocols by blending the traditional probabilistic approaches. Using NLP to analyze patient records and reported symptoms the whole process of diagnostics turns into fully automated process that not only increases its efficiency but also its accuracy.

1.1. Objective of the Study

The It is for this reason that the basic goal of this study is to propose an improved Gynecological Disease Diagnosis Expert System (GDDES) that is based on ML and NLP mechanisms, in a bid to provide improved diagnosis on the commonly prevalent gynecological diseases .The system, using Decision Trees, Random Forest Classifier, Support Vector Classifier, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, and Gradient Boosting Models in setting of different classifiers, is designed to increase the diagnostic accuracy and also work faster for first-round preliminary diagnosis. Moreover, thanks to natural language processing it will be possible to perform the analysis

of text content in the patient records, as well as in the description of the symptoms that, in turn, will help develop the completely automated and efficient diagnostic tool for the healthcare professionals.

1.2. Scope of the Study

This research aims at construing a Gynecological Disease Diagnosis Expert System (GDDES) that integrates machine learning .which include Logistic Regression and Gradient Boosting, will enhance the effectiveness of diagnostics in the system. This research aims at helping healthcare professionals get an easy and efficient method of diagnosing gynecological diseases in order to increase the quality of patient care and definitely increase the efficiency of the health care delivery systems in gynecological practices.

1.3. Problem Statement

The Gynecological Disease Diagnosis Expert System (GDDES) shall operate to try to solve the existing specific issues of amorphous diagnosis of the diseases, including UTI and PCOS. Present day diagnosis techniques are slow and inconclusive causing conventional human driven processes to perform poorly and ineffectively. When it comes to addressing critical patient records the intent is to optimize the level of diagnostic accuracy and minimize the time required for a model to reach the decision using techniques.

2. Related Work

Intelligence of ML and NLP are new development in diagnosis of gynecological disease. Current research has focused[1] in using various ML algorism in enhancing diagnosis for common illnesses such as UTIs and PCOS. For example[2]employed clinical data to implement Random Forest Classifier and Decision Tree[3] that revealed higher diagnostic performance into comparison with original methods. In addition, the role of NLP in the strategy of medical diagnosis has been already described to some extent. [4]noted that some NLP algorithms[5] could assist clinicians to build patterns from the notes to make right diagnoses at the right time. Thus, these algorithms assist clinicians transform a patient's complaints[6] and medical history into a format that can support a decision on treatment. Furthermore, the usage of the Logistic Regression is appreciated as providing the opportunities for finding the predictors concerning such gynecological disorders[8]. In studying the pattern of distribution in patient-level analysis by [7]logistic models of clinically and demographically significant differences were used to profile essential characteristics. Gradient Boosting techniques have[9] been advanced as some of the most effective methods of boosting accuracy on any kind of classification models. In a 2022 paper by[10]. it was possible to demonstrate that the accuracy of these models with patient data for prediction of the PCOS results[11] in better precision and recall rate than the classifiers. Thus, it is reasonable to recommend that stimulating the development of integration of ML and NLP in gynecology[12], the rate and accuracy of detection of diseases has increased. This integrated strategy[13] is hoped to help reduce the degree of cross-professional working needed in order to take a diagnosis[14] .

2.1. Problem Defination

Determine what gynecological diseases are to be included in diagnosis – namely UTI and infertility due to PCOS. There are two distinct objectives of the system which are as follows; reducing diagnostic errors and minimum response time.

2.2. Data Collection

Retrieve data sets which include patient records, signs, and diagnoses. The datasets that are to be formulated should be different with limited volumes for augmenting model generalizability.

2.3. Data Preprocessing

Clean it by eliminating redundant records on the set, add more columns, and standardize some of the attributes in a data set.

2.4. Normalization

It is also important to also make the features normal and/or to make all features standard in relation to all the others. A discretize is applied on categorical variables to convert them into more workable formats with encodings.

2.5. Natural Language Process

Resizing of the naïve and indistinct symptomatology and histories from the text data into the ‘‘loose description’’ form for the purpose of formation of the NLP for the construction of its database. When applying treatment to textual data, the following processes are used; tokenization, stemming and lemmatization. Text pre-processing: For turning the text into number, one must make use of Vectorization techniques such as TF-IDF or the Word Embedding's.

2.6. Feature Selection

By means of visualization, reduce the number of

features necessary for the diagnosis based on some criteria, such as Recursive Feature Elimination (RFE) or feature importance through tree-based models.

Implement classical machine learning algorithms:

- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier (SVC).
- Naïve Bayes Classifier
- K-Nearest Neighbor

2.7. Gradient Boosting Models

They should then train the models on the features selected above alongside perform hyperparameter tuning in order to obtain the model with good performance.

2.8. Model Evaluation

Then, split the given set of data which is very important to analyses the performance of the model. The literals should be chosen correctly about the metrics of evaluation such as accuracy, precision, recall value, F1 score, and ROC-AUC in order to measure the performances of the models.

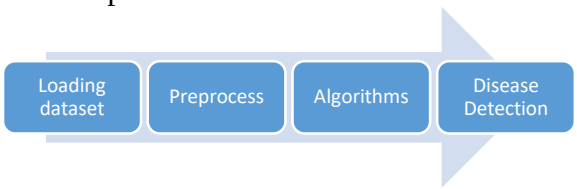

Figure 1 System Architecture of Analysis

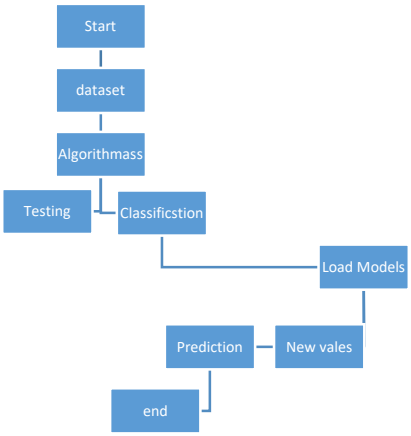

Figure 2 Testing Flow chart of Analysis

3. Methodology

3.1. Random Forest Classifier

Is method of ensemble learning technique where while training it generates multiple decision trees and in turn provides the mode of categories (classification) or mean of prediction by trees (regression). It helps in increasing accuracy, and then reducing the level of overfitting next.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Where TTT is the number of trees and ht(x)h_t(x)ht(x) is the prediction of tree for input.

3.2. Decision Tree

A tree is like many things in real life and it novely turns out that it has to do with a large swath of the field of machine learning that includes both classification and regression trees. In decision analysis, decision trees may also be effectively defined as graphical and unambiguous representations of decisions. As the name suggests they are a type of model that is in the form of a tree of decisions in an organization.

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2$$

3.3. Support Vector Classifier(SVC)

Is a supervised machine learning algorithm that focuses on maximizing the margin from the target functional in a feature space. This it does by seeking to make the widest margin between the closest points of different classes that is the support vectors. SVC is efficient in the high dimensional space and there also different kernel types (linear, polynomial, RBF).

$$\hat{y} = \text{mode}(y_i) \text{ for } i \in \{1, \ldots, k\}$$

3.4. K-Nearest Neighbour(KNN)

Is a simple and easily understandable classification and regression algorithm that does not place restrictions on the data. They do so by identifying the ‘k’ nearest data points or neighbours relative to the input by using a specific distance measurement. In the same manner in classification the class or the value of the input is the most commonly predicted by the neighbors in regard to regression it is the average of the neighbor values.

$$\hat{y} = \text{mode}(y_i) \text{ for } i \in \{1, \ldots, k\}$$

3.5. Logistics Regression

This is well suited for Logistic regression because the model is chosen for binary classification based problems where most of the target values in binary and the aim of the model is to predict the probability of an observation originating from the positive class for one or more independent variables.

Equation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}$$

3.6. Gradient Boost Models

Therefore, Gradient Boosting Models (GBMs) are machine learning methods that aimed at increasing predictive power of a model by using a number of weak predictors to make final decision that is actually a more accurate decision. The technique works in the fashion that it gradually introduces trees to reduce the error margins of the preceding models. Each of the new models fixes the errors it has found in the previous ones which is done by paying more attention to the cases that give higher errors.

3.7. Naïve Bayes

Naive Bayes is a group of probabilistic methods developed based on the Bayes theorem with an assumption about identical and independence predictors. It categorizes data by determining the likelihood of a class given the features which makes it ideal for use on big data sets.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

4. Result

4.1. Decision Tree Results

These results depict the evaluation measures of a model for a Decision Tree classifier that has been provided. The measurement of accuracy is almost 99.71%, the F1 score is 99.72%, the figuring of precision is 99.71%, and the figuring of recall is 99.73%. (Figure 3)

```
Accuracy Score =  0.9971428571428571
==================================================
f1 score score fo Decison tree = 0.997160081
==================================================
precision_score of Decision tree is = 0.9970
==================================================
recall_score of Decision tree is = 0.9972508
==================================================
coonfusion matrxi of decision tree =
 [[429   0   0]
 [  0 486   0]
 [  2   2 481]]
```

Figure 3 Metrics-Accuracy, F1 score etc

It proves the quantified results of the true positive, true negative and the misclassification of three classes. (Figure 4)

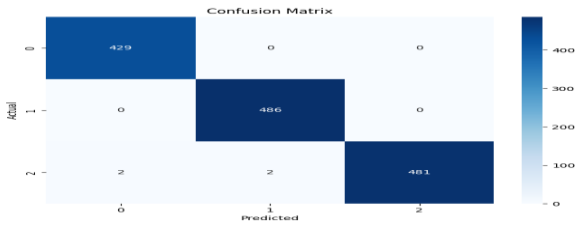

Figure 4 Metrics-Confusion Metrix

4.2. Random Forest Classifier Results

```
Accuracy Score of RandomForest=  0.9971428571428571
==================================================
f1 score score fo RandomForest = 0.9971629505314473
==================================================
precision_score of RandomForest is= 0.99707863914336
==================================================
recall_score of RandomForest is = 0.9972536874407818
==================================================
coonfusion matrxi of RandomForest =
 [[429   0   0]
 [  0 484   2]
 [  2   0 483]]
```

Figure 5 Metrics-Accuracy, F1 Score etc Random Forest

This confusion matrix illustrates the performance of the Decision Tree model across three classes. It shows 429 true positives for class 0, 484 true positives for class 1, and 483 for class 2. The model misclassified only a few instances, with two false positives and two false negatives, indicating high accuracy. (Figure 6)

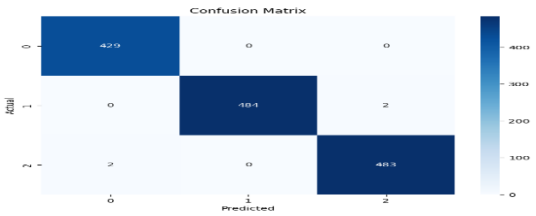

Figure 6 Metrics-Confusion Metrix for Random Forest

4.3. Support Vector Machine(SVM)

The results indicate a Support Vector Classifier (SVC) accuracy of approximately 74.07%. The F1 score is about 73.98%, with a precision of 74.11% and recall of 74.07%. The confusion matrix reveals several misclassifications, with notable overlaps among classes. (Figure 7)

```
Accuracy Score of SVC=  0.7407142857142858
==================================================
f1 score score fo SVC = 0.7397773489087821
==================================================
precision_score of SVC is= 0.7410999370345377
==================================================
recall_score of SVC is = 0.7407055250125124
==================================================
coonfusion matrxi of SVC =
 [[318  59  52]
 [ 61 381  44]
 [ 76  71 338]]
```

Figure 7 Metrics-Accuracy, F1 Score etc SVM

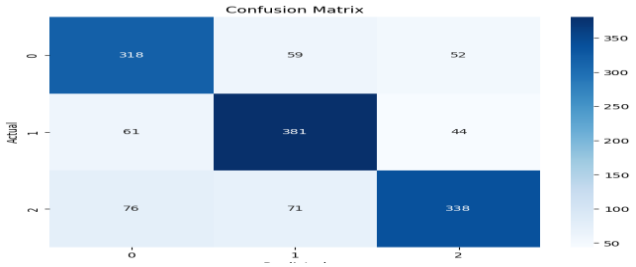

Figure 8 Metrics-Confusion Metrix for SVC

4.4. Logistic Regression

The Logistic Regression model achieves an accuracy of approximately 66.79%. Its F1 score is about 50.90%, with a precision of 51.46% and recall of 50.99%. The confusion matrix indicates notable misclassifications,

```
Accuracy Score of Logistic -  0.667857142857142
--------------------------------------------------
f1 score score fo Logistic - 0.5090145180482971
--------------------------------------------------
precision_score of Logistic is- 0.5145760153988
--------------------------------------------------
recall_score of Logistic is - 0.5099016817611
--------------------------------------------------
coonfusion matrxi of Logistic -
 [[172 138 119]
 [ 90 265 131]
 [ 71 131 283]]
```

Figure 10 Metrics-Accuracy, F1 score etc Logistic

The confusion matrix for the Logistic Regression model shows the distribution of actual versus predicted class labels. (Figure 11)

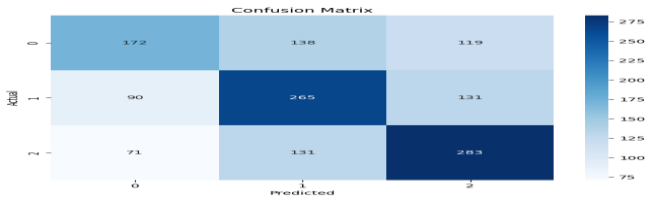

Figure 11 Metrics-Confusion Metrix for Logistic

4.5. Gradient Boosting Classifier Results

The model obtains an accuracy of about 67.57% with the Gradient Boosting Classifier. The F1 score of a model is approximately 50.90 % and its precision and recall ratios are 67.67 % and 67.66 % respectively.

```
Accuracy Score of GradientBoostingClassifier -  0.6757142857142
--------------------------------------------------
f1 score score fo GradientBoostingClassifier - 0.5090145180482971
--------------------------------------------------
precision_score of GradientBoostingClassifier is- 0.6767697915
--------------------------------------------------
recall_score of GradientBoostingClassifier is - 0.6766254689126
--------------------------------------------------
coonfusion matrxi of GradientBoostingClassifier -
 [[300  68  61]
 [ 93 326  67]
 [ 89  76 320]]
--------------------------------------------------
```

Figure 12 Metrics-Accuracy, F1 Score etc Gradient Models

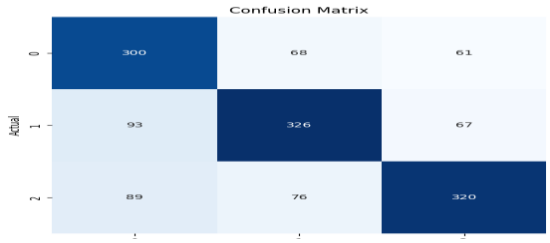

Figure 13 Metrics-Confusion Metrix for Gradient Boost

4.6. K-Nearest Neighbours(KNN) Results

The Naive Bayes classifier gives an accuracy of about 66.79 %. It's somehow efficient, given an F1 score of approximately 66.83%, precision of around 69.02% and recall of 51.57 %.

```
Accuracy Score of navi-  0.6678571428571428
--------------------------------------------------
f1 score score fo navi - 0.6683544394059764
--------------------------------------------------
precision_score of navi is- 0.6902272515223541
--------------------------------------------------
recall_score of navi is - 0.5150701263721501
--------------------------------------------------
coonfusion matrxi of navi -
 [[338  66  25]
 [130 318  38]
 [127  79 279]]
--------------------------------------------------
```

Figure 13 Metrics-Accuracy, F1 Score etc KNN

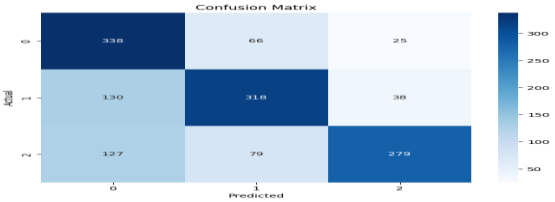

Figure 14 Metrics-Confusion Metrix for KNN

4.7. Naïve Bayes Results

I conclude that the Naive Bayes classifier gives an accuracy of around 51.93%. The F1 score of the proposed model is approximately 51.42% having precision value of 51.81% and recall of 51.57%. (Figure 15)

```
Accuracy Score of navi-  0.5192857142857142
--------------------------------------------------
f1 score score fo navi - 0.5141901139242296
--------------------------------------------------
precision_score of navi is- 0.5180488898523911
--------------------------------------------------
recall_score of navi is - 0.5150701263721501
--------------------------------------------------
coonfusion matrxi of navi -
 [[176 137 116]
 [ 93 266 127]
 [ 77 123 285]]
```

Figure 15 Metrics-Accuracy, F1 Score etc naive Bayes

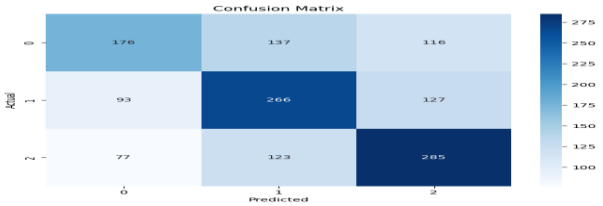

Figure 16 Metrics-Confusion Metrix for KNN

4.8. Comparative Analysis

Model Performance Comparison

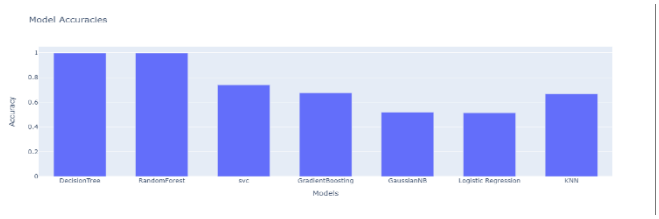

Figure 17 Comparative Plot for All Model

Table 1 Model Performance Comparison

Model	Accuracy
Decision Tree	0.997
Random Forest	0.997
SVC	0.741
Gradient Boosting	0.676
Gaussian Naive Bayes	0.519
Logistic Regression	0.514
KNN	0.668

The accuracy table shows the various models from the machine learning technique used in a classification problem. The results showed that, for both Decision Tree and Random Forest models, the overall accuracy of predictions of the target variable was of 99.7%. Random Forest, one of these ensemble methods comprising numerous decision trees, hails from decision tree learning as they help improve the lower susceptibility to overfitting.

4.9. Key Observations

The insights into classification effectiveness are related to the key observations derived from the structure of models and their performance on training and test datasets as well as the confusion matrices which belong to the models. The models consist of the Decision Tree and Random Forest; the accuracies for ironwomen presented high accuracy levels of 99.7%, and the confusion matrix laid down a foundation for detecting misclassification across classes. Overall, Gradient Boosting achieved 67.6% of accuracy and the following confusion matrix shows more core misclassification.

Conclusion

Consequently, the analysis of the different models of machine learning shows that there is a significant difference in the classification results. The Decision Tree and Random Forest models topped the performance list with accuracies of 99.7% proving that the models can learn intricate patterns and relationship set in the data set. Specifically, it tested KNN's performance and while showing improved outcomes as compared with the latter two models, pointed at the need for fine-tuning. As from this analysis, it is evident that choosing of right algorithms have to do with the nature of the problem at hand.

References

[1]. Aelgani, V., & Vadlakonda, D. (2024). Analysis of Machine Learning Model of Gynaecological Cancer Diagnosis Using Multilayer Perceptron Network. 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), 1784–1789. https://doi.org/10.1109/ICESC60852.2024.10689772

[2]. Basij, M., Yan, Y., Alshahrani, S., Winer, I., Burmeister, J., Dominello, M., & Mehrmohammadi, M. (2018). Development of an Ultrasound and Photoacoustic Endoscopy System for Imaging of Gynecological Disorders. IEEE International Ultrasonics Symposium, IUS,2018-October. https://doi.org/10.1109/ULTSYM.2018.8579788

[3]. Chang, S. C., Lee, H. F., Ting, H. M., Pan, T. C., Liu, S. Y., Chen, C. F., Wang, T. Y., Juan, K. J., Liao, T. I., & Huang, E. Y. (2013). Effect of different treatment plans on irradiated small-bowel volume in gynecologic patients undergoing whole-pelvic irradiation. Journal of Radiation Research, 54(5), 909–918. https://doi.org/10.1093/JRR/RRT023

[4]. Chauhan, N. K., & Singh, K. (2022). Diagnosis of Cervical Cancer with Oversampled Unscaled and Scaled Data Using Machine Learning Classifiers. 2022 IEEE Delhi Section Conference, DELCON 2022. https://doi.org/10.1109/DELCON54057.2022.9753298

[5]. Chen, Y., Wang, W., Schmidt, E. J., Kwok, K. W., Viswanathan, A. N., Cormack, R., & Tse, Z. T. H. (2016). Design and Fabrication of MR-Tracked Metallic Stylet for Gynecologic Brachytherapy. IEEE/ASME Transactions on Mechatronics, 21(2), 956–962. https://doi.org/10.1109/TMECH.2015.2503427

[6]. Dang, C., Jin, T., Sun, X., & Su, W. (2024). Exploring the Formulation Laws of Blood Tonic Chinese Patent Medicines for Gynecological Diseases Based on Data

Mining. 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2024, 897–904. https://doi.org/10.1109/AINIT61980.2024.10581653

[7]. De, S., Goswami, P., Faujdar, N., & Singh, G. (2024). Gynaecological Disease Diagnosis Expert System (GDDES) Based on Machine Learning Algorithm and Natural Language Processing. IEEE Access, 12, 84204–84215. https://doi.org/10.1109/ACCESS.2024.3406162

[8]. Gunderman, A. L., Schmidt, E. J., Morcos, M., Tokuda, J., Seethamraju, R. T., Halperin, H. R., Viswanathan, A. N., & Chen, Y. (2022). MR-Tracked Deflectable Stylet for Gynecologic Brachytherapy. IEEE/ASME Transactions on Mechatronics, 27(1), 407–417. https://doi.org/10.1109/TMECH.2021.3064954

[9]. Karasawa, K., Wakatsuki, M., Kato, S., Kiyohara, H., & Kamada, T. (2014). Clinical trial of carbon ion radiotherapy for gynecological melanoma. Journal of Radiation Research, 55(2), 343–350. https://doi.org/10.1093/JRR/RRT120

[10]. Kondakova, N. A., Sokolova, T. M., Madonov, P. G., Usova, A. V., Denisov, D. S., & Verigin, M. A. (2022). Development of system for assessing risks of gynecological disorders. 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences, SIBIRCON 2022, 220–223. https://doi.org/10.1109/SIBIRCON56155.2022.10016931

[11]. Li, Y. (2023). Gynecological surgery safety monitoring system based on hybrid cuckoo algorithm. 59–63. https://doi.org/10.1049/ICP.2022.2377

[12]. Nasirihaghighi, S., Ghamsarian, N., Stefanics, D., Schoeffmann, K., & Husslein, H. (2023). Action Recognition in Video Recordings from Gynecologic Laparoscopy. Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2023-June, 29–34. https://doi.org/10.1109/CBMS58004.2023.00187

[13]. Shanmugasundaram, G., Malar Selvam, V., Saravanan, R., & Balaji, S. (2018). An Investigation of Heart Disease Prediction Techniques. 2018 IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018. https://doi.org/10.1109/ICSCAN.2018.8541165

[14]. Verma, V., & Singh, Y. (2023). Improved Analysis of Inflammatory Diseases in Women using Artificial Intelligence based Approach. 2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023. https://doi.org/10.1109/ICICACS57338.2023.10100269