International Journal of Computer Applications Technology and Research
Volume 11–Issue 12, 423-430, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1112.1007

A Comparative Analysis of Advanced Ensemble Models in Cervical Cancer Prediction

Rebecca Adhiambo Okaka
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

Abstract: There are no symptoms in the early stages of cervical cancer, detection can only be done through regular Papanicolaou (Pap) and Human papillomavirus (HPV) tests. However, most women are not aware of these tests, if not, shy away from taking these tests, this has led to late cervical cancer diagnosis, and now, cervical cancer is one of the most common causes of cancer deaths among women. Successful cervical cancer treatment can be improved by early diagnosis, which can be achieved by analysing potential risk factors. This paper presents the performance of two advanced ensemble models; Bagging Classifier and Adaptive Boosting (AdaBoost) Classifier in predicting cervical cancer diagnosis based on documented cancer risk factors and target variables. The models were evaluated using accuracy, sensitivity and specificity metrics. Experiments done using the Cervical Cancer Risk Factors dataset found in the University of California at Irvine (UCI) repository shows that both models achieved good accuracy levels and can thus be used in early cervical cancer detection to avoid late diagnosis that has led to massive loss of lives.

Keywords — adaboost classifier; bagging classifier; biopsy; cervical cancer; cytology; hinselmann; schiller

1. INTRODUCTION

Recently, cancer was declared a national disaster in Kenya. Cervical cancer is the second cause of cancer deaths in women after the breast cancer in Kenya, and the fourth most frequent cancer in women in the whole world [1]. Cervical cancer arises from abnormal growth of cervical cells, the cancer can spread from the cervix to other parts of the body like the lungs, liver and bladder. Cervical cancer grows slowly, and has no symptoms in the early stages, even though regular Pap test and HPV test can help detect cervical cancer early, many women feel ashamed of going for the tests and seeking early treatment. Its symptoms such as pelvic pain, abnormal vaginal bleeding and discharge and kidney failure appear in late stages. HPV, a common sexually transmitted infection (STI), is the leading cause of cervical cancer [2], other factors that may lead to cervical cancer include; prolonged use of contraceptives, cigarette smoking and multiple pregnancies.

There are opportunities to improve cervical cancer diagnosis, using Artificial Intelligence (AI) and machine learning approaches. Unassisted medical practitioner is likely to make wrong diagnosis, because they are exposed to imperfect human memory, and varying disease presentation [3], besides, machine learning models can be used to assist medical practitioners in disease diagnosis [4]. At the moment, most computer aided medical diagnosis systems use medical images and frequency signals to assist doctors interpret disease diagnosis, there is need to create systems that could also predict disease diagnosis based on its documented risk factors, to enable early diagnosis and treatment.

Ensemble models, also called multiple classifier models combine several machine learning algorithms to improve their predictive power, they have proven to be very effective and extremely versatile; can be used in a wide variety of problem domains and real world applications [5]. They were

originally developed to reduce variance, bias and improve accuracy in automated systems, but today they have become very successful in addressing a variety of machine learning problems. There are two categories of ensemble models; simple ensemble models and advanced ensemble models. Simple ensemble models use, max voting, averaging and weighted average techniques, advanced ensemble models use, stacking, blending, bagging and boosting techniques. In our study, we use two advanced ensemble algorithms; bagging algorithms and boosting algorithms to predict cervical cancer diagnosis, based on documented risk factors and four cervical cancer indicators tests.

The rest of the paper is organized as follows; related work is discussed in section II, methodology is discussed in section III. Experiments, which includes the dataset used, experimental setup, metrics, results and discussions in section IV, then finally, conclusion and future work in section V and VI.

2. RELATED WORK

Machine learning algorithms provide several tools for smart data analysis [6], with the recent digital revolution, many modern hospitals are now equipped with means for data capture, storage and sharing. Decision trees [7] have been used diagnosing cervical cancer, from experiments, a decision tree achieved accuracies of 92.54%, 92.80%, 94.41% and 90.44% for Biopsy, Cytology, Hinselmann, and Schiller tests respectively. Multilayer Perceptron (MLP), Bayes Net and k-Nearest Neighbour have also been used [8] to correctly classify cervical cancer instances, experiments showed that, Bayes Net achieved the highest classification accuracy, by classifying 97.26% instances correctly, followed by both k-Nearest Neighbour and MLP at 95.89%. The effectiveness of Iterative Dichotomous (ID)3, C4.5 and Naïve Bayes in predicting cervical cancer were analysed [9], the results from the test set of each model was averaged, Naïve Bayes got the highest accuracy score of 81%, followed by C4.5 at 72%, then ID3 at 69%. Medical diagnosis is sensitive, therefore apart from accuracy analysis, it is also important to get from a model how often it predicts a disease when the patient actually has the disease, and how often it predicts no disease when a person actually does not have the disease. From existing literature, many models including the ones discussed in this section, only present their accuracy levels or scores but fail to present their sensitivity and specificity levels.

Many other studies have explored different methods to predict cervical cancer, data based approaches such as support vector machines (SVM), linear regression (LR), principal component analysis (PCA), particle swarm optimization (PSO), artificial neural networks (ANN) and clustering algorithms [10 - 15] have been used.

This paper makes use of two advanced ensemble algorithms; for Bagging ensemble algorithms, we use the Bagging Classifier model and for Boosting ensemble algorithms, we use the AdaBoost Classifier model. The two models are separately used to predict cervical cancer diagnosis based on 32 documented risk factors and four target variables; Biopsy, Cytology, Hinselmann and Schiller.

3. METHODOLOGY

In this section we describe the models used in our study.

3.1 Bagging Classifier

A bagging classifier is an advanced ensemble technique that combines predictions from several base models to get the final prediction, it does this by fitting each base model on random subsets of the original dataset, then, aggregating their individual predictions either by voting or averaging to get the final prediction. Bagging classifier takes different dimensions; it is known as Pasting, when random subsets of the dataset are drawn as random subsets of the sample [16], Bagging, when samples are drawn with replacement [17], Random Subspaces, when random subsets of the dataset are drawn as random subsets of features [18], and, as Random Patches, when the base models are built on subsets of both samples and features [19].

International Journal of Computer Applications Technology and Research
Volume 11–Issue 12, 423-430, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1112.1007

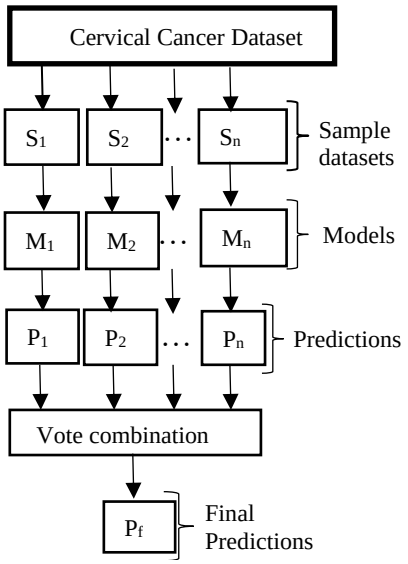

Figure 1. Structure of bagging classifier algorithm

Bagging Classifier Algorithm;

Inputs: Training data **S**; supervised learning algorithm, Base Classifier, integer **T** specifying ensemble size, percent **R** to create bootstrapped training data.

For t = 1,…,T Do

i. Take a bootstrapped replica $\mathbf{S_t}$ by randomly drawing **R%** of **S**
ii. Call Base Classifier with $\mathbf{S_t}$ and receive the hypothesis (classifier) $\mathbf{h_t}$
iii. Add $\mathbf{h_t}$ to the ensemble, $\mathbf{\mathcal{E} \leftarrow \mathcal{E} \cap h_t}$

End For

Simple Majority Voting; given unlabelled instance **x**

i. Evaluate the ensemble $\mathbf{\mathcal{E} = \{h_1,\ldots,h_T\}}$ on **x**
ii. Let $\mathbf{V_{t,c} = 1}$ if $\mathbf{h_t}$ chooses class $\boldsymbol{\omega_c}$, and **0**, otherwise
iii. Obtain total vote received by each class

$$\mathbf{V_c} = \sum_{t=1}^{T} \mathbf{v_{t,c}},\ \mathbf{c = 1,\ldots,C} \qquad (1)$$

Output: Class with the highest $\mathbf{V_c}$

3.2 AdaBoost Classifier

AdaBoost is an advanced ensemble technique that makes use of multiple models in a sequential process, where each of the subsequent model attempts to correct the errors of the previous model, it does this by assigning weights to observations which are incorrectly predicted, so that the subsequent model can work to predict these values correctly, it also chooses the training set for each new classifier based on the result of the previous classifier. The succeeding models are dependent on the previous model. AdaBoost can be viewed as a technique that builds on top of other classifiers as opposed to being a classifier itself; it combines multiple weak classifiers into a strong classifier [16,17].

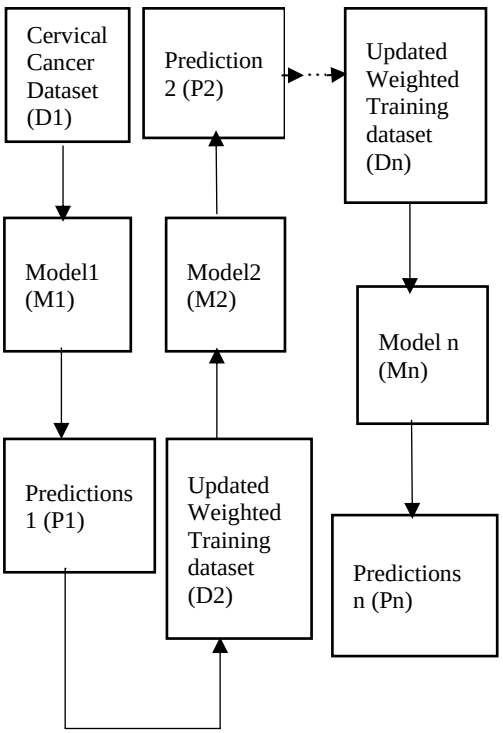

Figure 2. Structure of adaboost classifier algorithm

International Journal of Computer Applications Technology and Research
Volume 11–Issue 12, 423-430, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1112.1007

AdaBoost Classifier Algorithm:

Inputs: Given training data = (x_i, y_i), $i = 1,\ldots,N$ $y_i \in \{\omega_1,\ldots,\omega_C\}$, supervised learner, Base Classifier; ensemble size T

Initialize the distribution $D_1(i) = \frac{1}{N}$

For $t = 1,\ldots,T$ **DO**

i. Draw training subset S_t from the distribution D_t

ii. Train Base Classifier on S_t, receive hypothesis h_t: $X \rightarrow Y$

iii. Calculate the error of h_t

$$\varepsilon_t = \sum_i I[\![h_t(x_i) \neq y_i)]\!] D_t(x_i) \quad (2)$$

If $\mathcal{E}t > 0.5$ *abort*

iv. Set $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$

v. Update sampling distribution

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot \{\beta_t, \text{ if } h_t(x_i) = y_i \text{ or } 1, \text{otherwise} \quad (3)$$

where $Z_t = \sum_i D_t(i)$ is a normalization factor chosen so that D_{t+1} is a proper distribution function

End For

Weighted Majority Voting; given unlabelled instance z,

i. Obtain total vote received by each class

$$V_c = \sum_{t:h_t(z)=\omega_c} \log\left(\frac{1}{\beta_t}\right), c = 1,\ldots,C \quad (4)$$

Output: Class with the highest V_c

4. EXPERIMENTS

4.1 Dataset

The dataset used is a Cervical Cancer Risk Factors dataset that was donated by the Hospital Universitario de Caracas in Caracas, Venezuela, on 3rd March 2017, it is found in the University of California at Irvine (UCI) repository. The dataset contains historical records of 858 patients each containing 36 variables; 32 risk factors and four target variables: Hinselmann, Schiller, Biopsy and Cytology. The dataset features contain, patients' historical medical records, habits and demographic information. The original dataset target variables outcome, yes for '1' and 'no' for '0' distribution is shown in Figure 3.

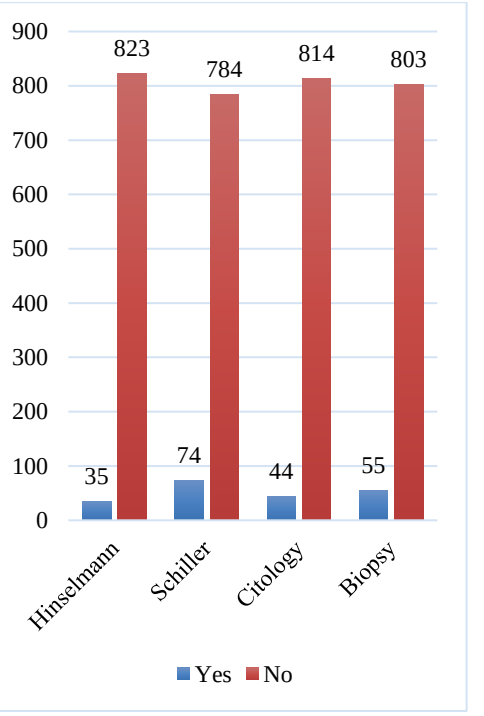

Figure 3. Original dataset target variables outcome distribution

Due to privacy issues, many patients opted not to open up on certain sensitive questions resulting to many missing values in the dataset, hence, so many risk factors and records were removed. After thorough data analysis, several trade-offs were made.

In this study we used all the four target variables and 15 risk factors out of the possible 32 as shown in Table 1, to carter for the huge variation between the Yes and No target variable outcome, each target variable used a specific number of records; for biopsy a total of 90 records were used, for Cytology, 78 records were used, for Hinselmann, 60 records were used, and for Schiller, 126 records were used, each record had the 15 variables. In the end we had four sub datasets, each for a specific target variable.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 12, 423-430, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1112.1007

Table 1. Variables used

No.	Variable	Type
1	Age	int
2	Number of sexual partners	int
3	Number of pregnancies	int
4	Smokes	bool
5	Hormonal Contraceptives	bool
6	IUD	bool
7	STDs	bool
8	STDs: Condylomatosis	bool
9	STDs:Vulvo-perineal Condylomatosis	bool
10	STDs: Syphilis	bool
11	STDs: HIV	bool
12	STDs: HPV	bool
13	STDs: Number of diagnosis	int
14	Dx: HPV	bool
15	Dx	bool
16	Hinselmann: Target variable	bool
17	Schiller: Target variable	bool
18	Cytology: Target variable	bool
19	Biopsy: Target variable	bool

4.2 Evaluation

Experiments were done using Python 3 for Windows, all the four sub datasets were divided into 70% training set and 30% testing set.

To understand how well our models have performed, we present our results in form of confusion matrix, from which we were able to compute our models' accuracy, sensitivity and specificity. We used the confusion matrix to help us identify how many No cases are predicted as No, and how many Yes cases are predicted as Yes, the accuracy metric helps us determine how often the classifier is correct, sensitivity also known as recall which refers to the True Positive Rate (TPR), helps us determine how often the classifier predicts a Yes, when it is actually a Yes, and specificity which refers to the True Negative Rate (TNR), helps us determine how often a classifier predicts a No, when it is actually a No.

The confusion matrix table structure used is shown in Table 2.

Table 2. Confusion matrix

		Predicted Values	
		N	**P**
Actual values	**N**	TN	FP
	P	FN	TP

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (7)$$

4.3 Results

Table 3. Confusion matrix for biopsy test

Biopsy Test		Predicted Values			
		Bagging Classifier		AdaBoost Classifier	
		N	**P**	**N**	**P**
Actual values	**N**	14	2	15	1
	P	1	10	2	9

From Table 3, the Classifiers each made a total of 27 predictions, in other words, 27 patients took biopsy test. Out of the 27 patients, the Bagging Classifier predicted that 10 patients tested positive and 14 patients tested negative, while AdaBoost Classifier predicted that 9 patients tested positive and 15 patients tested negative. In reality 11 patients in the sample were actually positive and 16 patients in the sample were actually negative.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 12, 423-430, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1112.1007

Table 4. Confusion matrix for cytology test

Cytology Test		Predicted Values			
		Bagging Classifier		AdaBoost Classifier	
		N	P	N	P
Actual values	N	15	1	16	0
	P	1	7	1	7

From Table 4, the Classifiers each made a total of 24 predictions, in other words, 24 patients took Cytology test. Out of the 24 patients, both Classifiers predicted that 7 patients tested positive, Bagging Classifier predicted 15 patients tested negative while AdaBoost Classifier predicted 16 patients tested negative. In reality 8 patients in the sample were actually positive and 16 patients in the sample were actually negative.

Table 5. Confusion matrix for hinselmann test

Hinselmann Test		Predicted Values			
		Bagging Classifier		AdaBoost Classifier	
		N	P	N	P
Actual values	N	7	1	7	1
	P	2	8	1	9

From Table 5, the Classifiers each made a total of 18 predictions, in other words, 18 patients took Hinselmann test. Out of the 18 patients, both Classifiers predicted that 7 patients tested negative, Bagging Classifier predicted 8 patients tested positive, while AdaBoost Classifier predicted 9 patients tested positive. In reality 10 patients in the sample were actually positive and 9 patients in the sample were actually negative.

Table 6. Confusion matrix for schiller test

Schiller Test		Predicted Values			
		Bagging Classifier		AdaBoost Classifier	
		N	P	N	P
Actual values	N	17	2	18	1
	P	3	16	2	17

From Table 6, the Classifiers made a total of 38 predictions, in other words, 38 patients took Schiller test. Out of the 38 patients, Bagging Classifiers predicted that 16 patients tested positive and 17 patients tested negative. AdaBoost Classifiers predicted that 17 patients tested positive and 18 patients tested negative. In reality the sample had 19 patients who were actually negative and 19 patients who were actually positive.

Table 7. Overall results (%)

	Target Variables			
	Biopsy	Cytology	Hinselmann	Schiller
Bagging Classifier				
Accuracy	89	92	83	87
Sensitivity	91	88	80	84
Specificity	88	94	88	89
AdaBoost Classifier				
Accuracy	89	96	89	92
Sensitivity	82	88	90	89
Specificity	94	100	88	95

The overall results in percentages is shown in Table 7. Figure 4 shows the performance of the models. Accuracy, sensitivity and specificity are calculated from the confusion matrix tables. For example, the results for cytology test in AdaBoost is as follows;

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{16+7}{16+7+0+1} = \frac{23}{24} = 96\%$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{7}{8} = 88\%$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{16}{16} = 100\%$$

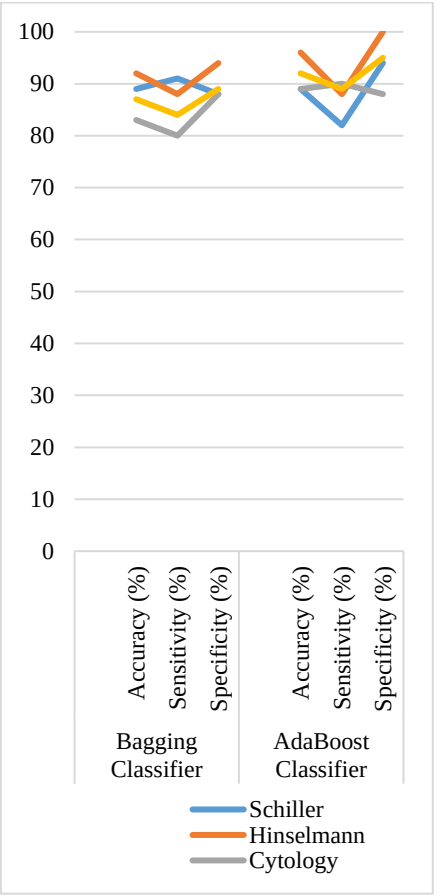

Figure 4. Advanced ensemble models overall performance

5. CONCLUSIONS

The good performance of Bagging Classifier and AdaBoost Classifier can be attributed to the fact that both models use multiple models in order to improve the accuracy of their final predictions, for instance, the Bagging Classifier combine predictions from several model to get the final prediction, either by averaging or max voting, while, the AdaBoost Classifier follows a sequential process where multiple models are used to make predictions each at a time, and the subsequent model works to correct the errors of the previous model, until the error function remains constant.

Many previous cervical cancer prediction models used the accuracy metric to evaluate how well they performed. However, medical diagnosis is a sensitive procedure that we cannot rely only on the accuracy of a model to judge how well it performs, but, we should also consider how often a model predicts a patient has a disease when the patient actually has the disease, and how often it predicts a patient does not have a disease when the patient actually does not have the disease. This study introduces the performance of two advanced ensemble algorithms; Bagging Classifier and AdaBoost Classifier in cervical cancer prediction. The models are evaluated using, accuracy, sensitivity and specificity metrics.

6. FUTURE WORK

The possible future work related to this study is to first test the efficiency of advanced ensemble models with other cancer risk factors not used in this study, as well as analyse the importance of each risk factor to the target variables, then, explore possibilities of coming up with a mobile app based on the risk factors that women can use to monitor their own cervical health status, as well as network with other women on the same platform. Lastly, is to use the same models and their modified versions in other larger cancer datasets, to see how efficient and effective they are.

REFERENCES

[1] World Cancer Report. 2014. World Health Org., Geneva, Switzerland.

[2] Gadducci, A., Barsotti, C., Cosio, S., Domenici, L., and Riccardo A. G. 2011. Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: A review of the literature. Gynecol. Endocrinal, 2011, vol. 27, no. 8, pp. 597–604.

[3] El-Kareh R., Hassan, O., and Schiff, G. 2013. Use of Health Information Technology to

reduce diagnostic error. BMJ Quality and Safety, 22(Suppl 2): ii40-ii51.

[4] Quinlann, J. R. Induction of Decision Trees. Machine Learning. 1986, 1.1:81 – 106. 1986.

[5] Zhang C., and Ma, Y. 2012. Ensemble Machine Learning: Methods and Applications. Springer Science and Business Media.

[6] Julian, P. 2004. The cervical cancer epidemic that screening has prevented in the UK. The Lancet. 364.9430:249-256.

[7] Pipti, N. P., and Kishor, H. A. "Cervical Cancer Test Identification Classifier Using Decision Tree Method", International Journal of Research in Advent Technology, 2019. Vol. 7, No. 4, E-ISSN:2321 – 9637.

[8] Mohammed, F., Kadir, U., and Muciz, S. "Determining Cervical Cancer Possibility by Using Machine Learning Methods", International Journal of Latest Research in Engineering and Technology, 2017. ISSN:2454 – 5031. Vol 03 – Issue 12, pp:65 – 71.

[9] Vidya, R., and Nasira, G. M. Predicting "Cervical Cancer Using Machine Learning Technologies – An Analysis", Global Journal of Pure and Applied Mathematics, 2016. ISSN 0973 – 1768, vol 12, no 3.

[10] Kresta, J. V., MacGregor, J. F., and MarlinT. E. 1991. Multivariate statistical monitoring of process operating performance. Can. J. Chem. Eng. vol. 69, no. 1, pp. 35–47.

[11] Salmeron, J. L., Rahimi, S. A., Navali, A. M., and Sadeghpour, A. 2017. 2017. Medical diagnosis of Rheumatoid Arthritis using data driven PSO-FCM with scarce datasets. Neuro computing, vol. 232 (April. 2017), pp. 104–112.

[12] Yin, S., and Huang, Z. 2015. Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements. IEEE/ASME Trans. Mechatronics, vol. 20, no. 5 (Oct. 2015), pp. 2613–2620.

[13] Rodrigues, P. L., Rodrigues, N. F., Fonseca, J. C., Correia-Pinto, C., and Vilaca, J. L. 2014. Automatic modeling of pectus excavatum corrective prosthesis using artificial neural networks. Med. Eng. Phys., vol. 36, no. 10, pp. 1338–1345.

[14] Yin, S., Gao, H., Qiu, j., and Kaynak, O. 2017. Descriptor reduced-order sliding mode observers design for switched systems with sensor and actuator faults. Automatica, vol. 76 (Feb. 2017), pp. 282–292.

[15] Yin, S., Yang, H., and Kaynak, O. 2017. Sliding mode observer-based FTC for Markovian jump systems with actuator and sensor faults. IEEE Trans. Autom. Control, vol. 62 (Jul. 2017), no. 7, pp. 3551–3558.

[16] Breiman, L. 1999. Pasting small votes for classification in large databases and online. Machine Learning, 36(1), 85 – 103.

[17] Breiman, L. 1996. Bagging predictors. Machine Learning, 24(2), 123 – 140.

[18] Ho, T. 1998. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, 20(8), 832 – 844.

[19] Louppe, G., and Geurts, P. 2012. Ensembles on Random Patches. Machine Learning and Knowledge Discovery in Databases, 346 – 361.