Table 2 – Confusion Matrix

Metric used	***k*-NN** **1**	**M_*k*-NN** **2**	***k*-NN+CB** **3**	**M_*k*-NN+CB** **4**	***k*-NN+CB+CS** **5**	**M_*k*-NN + CB+ CS** **(Proposed Model)** **6**
Confusion Matrix	a b 435 9 a= 2 26 213 b= 4 a- benign b- malignant	a b 435 9 a= 2 25 214 b= 4 a- benign b- malignant	a b 330 11 a= 2 6 335 b= 4 a- benign b- malignant	a b 338 3 a= 2 4 337 b= 4 a- benign b- malignant	a b 328 13 a= 2 6 335 b= 4 a- benign b- malignant	a b 338 3 a= 2 1 340 b= 4 a- benign b- malignant

The curve is seen at the upper leftmost corner near the y axis adjacent to one indicating a high ROC value and the discriminating power of the proposed model. The proposed model has a better value for MCC with a value of 0.988, FPR rate with 0.006, and recall of 0.994 when compared with other classifiers. A comparison of the FPR of the classifiers is illustrated (Fig 4). The least efficient result was shown by the Naïve Bayes classifier. The FPR indicates how much the model incorrectly predicts the positive class. Hence a lower value is preferred. The least FPR is shown by the proposed model. It indicates that the number of correctly classified

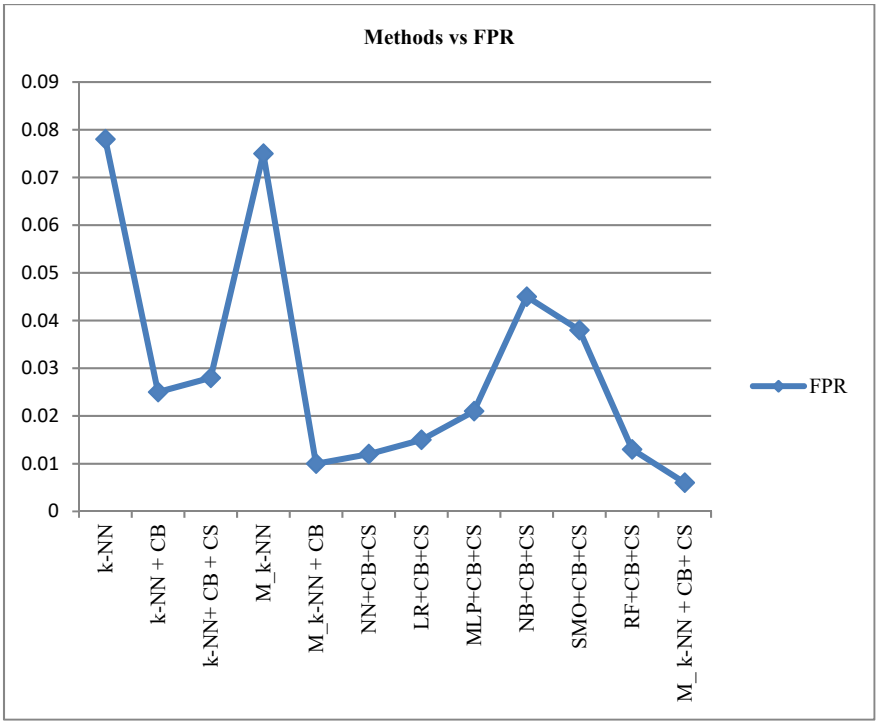

Figure 4 Methods vs FPR

positive instances are high and that with least misclassification was done by the proposed model. Figure 5 illustrates the MCC values plotted against the proposed model and various *k*-NN models. MCC gives a high score only if good prediction results are achieved in all four categories of the confusion matrix - (TP, TN, FP, FN). MCC is seen to be effective if the classes are balanced and deteriorate if they are unbalanced since it gets unevenly distributed [Zhu, (2020)]. Table 3 gives the comparison with other data mining methods.

Table 3- Comparison with other classifiers

Methods	Accuracy 1	ROC 2	Recall 3	MCC 4	FPR 5	Kappa 6
NN+CB+CS	98.82	0.988	0.988	0.977	0.012	0.9765
LR+CB+CS	98.53	0.987	0.985	0.971	0.015	0.9707
MLP+CB+CS	97.94	0.989	0.979	0.959	0.021	0.9589
NB+CB+CS	95.45	0.991	0.955	0.909	0.045	0.9091
SMO+CB+CS	96.18	0.962	0.962	0.924	0.038	0.9238
RF+CB+CS	98.68	0.997	0.987	0.974	0.013	0.9736
Modified_ *k*-NN + CB+ CS	99.4135	0.999	0.994	0.988	0.006	0.9883

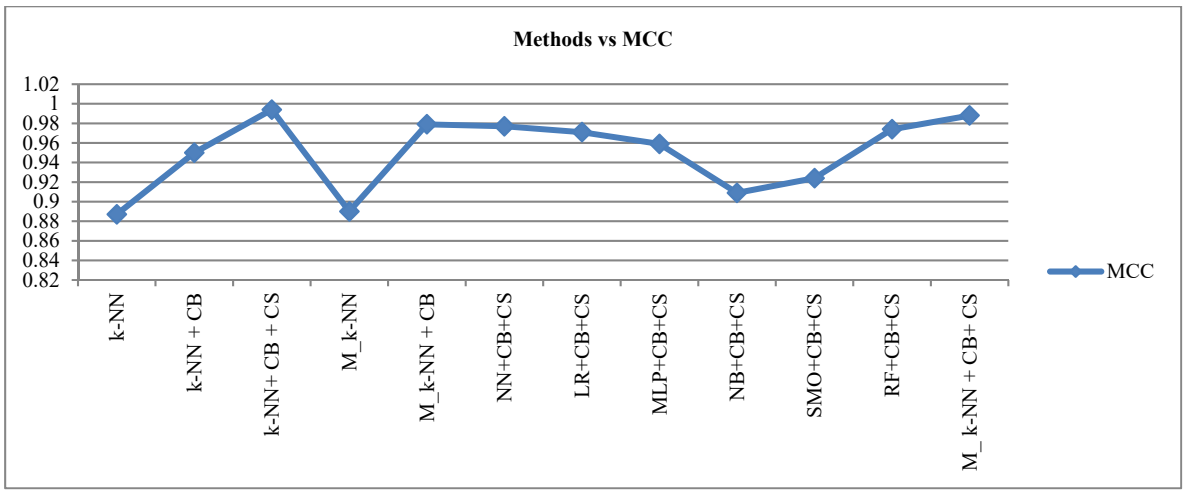

Figure 5 Methods vs MCC

The proposed model has the best MCC value among the models. The least efficient model is given by Naïve Bayes and *k*-NN classifiers. (Fig 6) illustrates the Precision-Recall curve for the positive class with the X-axis plotted with TPR values and Y-axis with Precision values. Recall is the ability of the classifier to correctly predict the positive samples. The P-R curve helps in visualizing classifier performance and threshold. Precision shows how closely the results agree with one another. The P-R curve is at the upper rightmost corner indicating the good performance of the classifier.

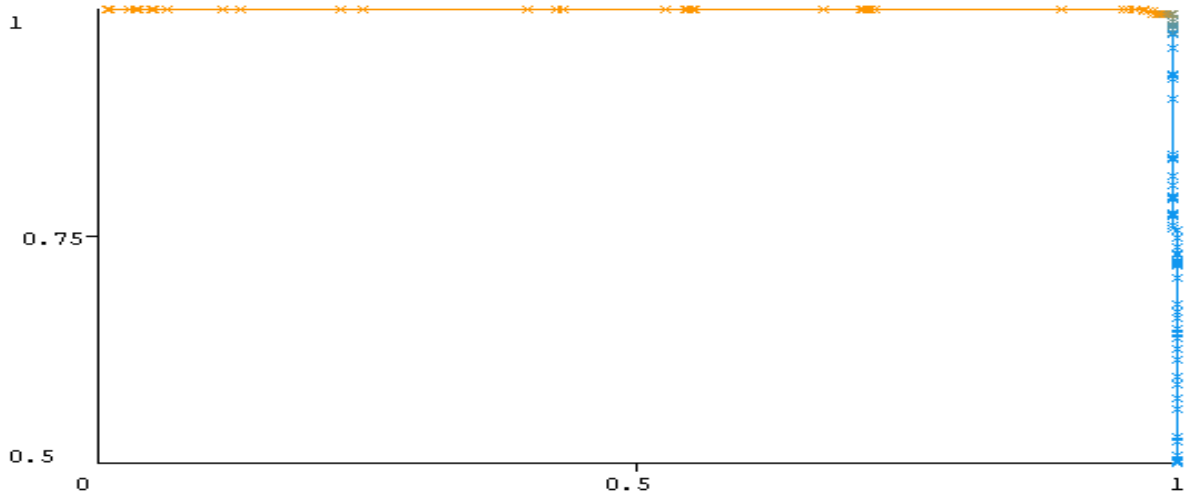

Figure 6: Precision-Recall Curve

K is chosen a value 5(Fig 7).A set of values from 1 to 50 were used for k, and k=5 was taken based on the cross validation scores obtained. Higher k leads to high bias and too low value for k leads to high variance. Hence to select a suitable k value the cv scores for different values of k is compared. A large k is also computationally costly, besides to avoid ties between classes that are chosen an odd value of k will be better.

Comparison of Modified- Weighted -*k*-NN+CS with a few other nature-inspired feature search methods: Firefly search (FFS), Bee search (BS), Flower search (FS), Elephant search (ES) is also shown (Table 4). The proposed model shows better results against *k*-NN with other search methods. In all the methods class balancing is used.

Table 4 Comparison with other search methods

Methods	**Accuracy** **1**	**ROC** **2**	**Recall** **3**	**MCC** **4**	**FPR** **5**	**Kappa** **6**
k-NN + CB+ BS	98.97	1	0.990	0.980	0.010	0.9795
k-NN + CB+ ES	98.8	1	0.988	0.977	0.012	0.9765
k-NN + CB+ FFS	98.68	0.999	0.987	0.974	0.013	0.9736
k-NN + CB+ FS	99.1	1	0.991	0.982	0.009	0.9824
M_*k*-NN + CB+ CS	99.4135	0.999	0.994	0.988	0.006	0.9883

Comparison with related existing works in literature with the proposed model in terms of accuracy is summarized (Table 5). The proposed model exhibited an improved performance when compared to other models. To observe the stability and performance of the model. It was evaluated on three other datasets from the UCI machine learning repository- the Cleveland heart dataset, Hepatitis Dataset, and Kidney dataset. (Table 6) presents the results obtained. The model was seen to produce better performance in all three cases.

Table 5 Comparison with literature

Previous Literature	**Classifier used 1**	**Dataset used 2**	**Accuracy obtained 3**
(Sudha & Selvarajan 2016)	*k*-NN+ECS	DDSM	99.13
(Prabhukumar, Agilandeeswari & Sangaiah, 2017)	SVM +CS	MIAS	96.72
(Chakravarthy & Rajguru, 2019)	ECS	MIAS	97.5
(Peng et al., 2020)	k-NN+ CoFF	WBCD	98
Proposed Model	M_*k*-NN + CS +CB	WBCD	99.41

Table 6 Comparison of the model with other datasets

Dataset used	**Accuracy of the *k*-NN method 1**	**Accuracy of the proposed Model 2**
Cleveland Heart database	88.4106	98.0132
Hepatitis dataset	60.3196	82.4675
Kidney dataset	79.75	87.25

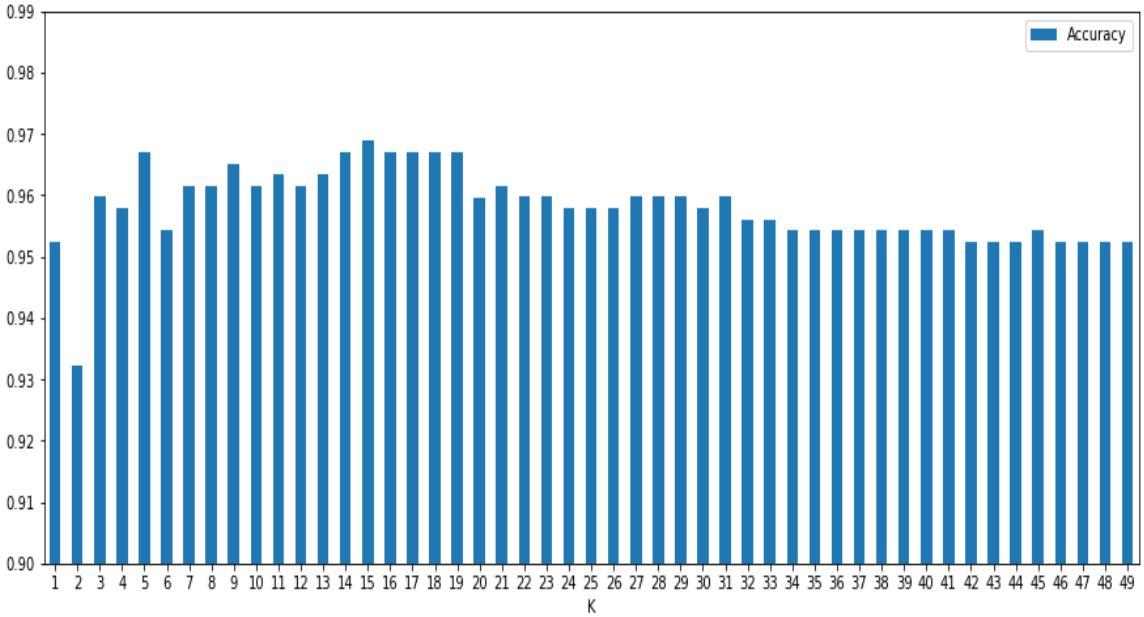

Figure 7: K vs CV Scores

4. Conclusion

The study proposed a model with the *k*-Nearest Neighbour classifier used in combination with the feature selection method of Cuckoo search and class balancing. It demonstrated good performance with an accuracy of 99.41% and MCC of 0.988 for Breast Cancer classification into Malignant or Benign Class. The proposed model was used on small datasets and its performance on large and high dimensional datasets is to be evaluated. Further work can be done to evaluate the performance and build ensemble classifiers with metaheuristic search techniques and also to use the combination of different feature search optimization methods to improve feature search thus aiding disease diagnosis. Moreover, Deep learning models can be implemented to provide better models for diagnosis and classification.

Acknowledgments

We are indebted to Dr. William H. Wolberg of the University of Wisconsin Hospitals, Madison for the Breast cancer dataset made available

References

[1] Ammal R A, PC S, SSV, 2020, Termite inspired algorithm for traffic engineering in hybrid software-defined networks, PeerJ Computer Science6e283, https://doi.org/10.7717/peerj-cs 283
[2] Arya C, Tiwari R (2016) Expert system for breast cancer diagnosis: a survey. In: 2016 international conference on computer communication and informatics (ICCCI), pages 1–9. IEEE
[3] Deepika K. Nagthane, Dr. A.M.Rajurkar, 2017, Cuckoo Search: An Optimized Way For Mammogram Feature Selection, International Journal Of Current Engineering And Scientific Research, Volume-4, Issue-8, 2017
[4] Ed-daoudy, A., Maalmi, K., 2020, Breast cancer classification with reduced feature set using association rules and support vector machine. Network Modeling Analysis in Health Informatics Bioinformatics 9, 34 (2020). https://doi.org/10.1007/s13721-020-00237-8
[5] Ehsan Valian, Shahram Mohanna And Saeed Tavakoli, 2020, "Improved Cuckoo Search Algorithm For Feedforward Neural Network Training", International Journal Of Artificial Intelligence & Applications (Ijaia), Vol.2(3), pp. 36-43, July 2011
[6] Elias Martins Guerra Pradoa, Carlos Roberto de Souza Filho, Emmanuel John M. Carranzac , João Gabriel Motta, 2020, Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data, Ore Geology Reviews, 124,(2020), 1-20, https://doi.org/10.1016/j.oregeorev.2020.103611
[7] Essam H. Houssein, Mosa E. Hosney, Mohamed Elhoseny, Diego Oliva, Waleed M. Mohamed & M. Hassaballah,2020, Hybrid Harris hawks optimization with the cuckoo search for drug design and discovery in chemoinformatics, Scientific Reports, Nature, (2020) 10:14439, https://doi.org/10.1038/s41598-020-71502-z
[8] Ganesh N. Sharma, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and K. K Sharma, 2010, Various Types And Management Of Breast Cancer: An Overview, Journal of Advanced Pharmaceutical and Technology Research. 2010 Apr-Jun; 1(2): 109–126
[9] Hu Peng, Wenhua Zhu, Changshou Deng, Kun Yu, Zhijian Wu, 2020 Composite firefly algorithm for breast cancer recognition, Concurrency, and Computation Practice and Experience, Wiley, 2020; https://doi.org/10.1002/cpe.6032, 1- 12
[10] Juan Araya, Aldo Cipriano,2006, Optimal Identification of Takagi-Sugeno Fuzzy Models for Nonlinear FDI, A Proceedings Volume from the 6th IFAC Symposium, SAFEPROCESS 2006, Beijing, P.R. China, August 30–September 1, 2006, Volume 1, 2007, Pages 759-764, https://doi.org/10.1016/B978-008044485-7/50128-7
[11] Karimollah Hajian Tilaki,2013, Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation, Caspian Journal of Internal Medicine,2013, 4(2), 627-636
[12] Kuldeep B, V.K. Singh, A. Kumar, G.K. Singh,2014, Design of two-channel filter bank using nature-inspired optimization-based fractional derivative constraints. ISA Transactions (2014), http://dx.doi.org/10.1016/j.isatra.2014.06.005i
[13] Lourdes Pelayo, Evaluating Stratification Alternatives to Improve Software Defect Prediction, IEEE Transactions on Reliability, 2012, 61(2):516-525, DOI: 10.1109/TR.2012.2183912
[14] Manoharan Prabukumar, Loganathan Agilandeeswari, and Arun Kumar Sangaiah, 2017, An Optimized Breast Cancer Diagnosis System Using a Cuckoo Search Algorithm and Support Vector Machine Classifier, Hybrid Intelligence for Image Analysis and Understanding, 2017
[15] Manuel Torres-Vásquez , Oscar Chávez-Bosquez , Betania Hernández-Ocaña and José Hernández-Torruco ,2020, Classification of Guillain–Barré, Syndrome Subtypes Using Sampling Techniques with Binary Approach , Symmetry, 2020, 12, 482, 1- 27, doi:10.3390/sym12030482
[16] Mathew T.E, A comparative study of the performance of different Support Vector machine Kernels in Breast Cancer Diagnosis, International Journal of Information and Computing Science, Volume 6, Issue 6, pp. 432-441 June 2019
[17] Mathew T E, A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis, International Journal on Emerging Technologies 10(3): 55-63(2019)
[18] Mathew T E, Simple and Ensemble Decision tree Classifier based detection of Breast Cancer, International Journal of Scientific & Technology Research Volume 8, Issue 11, pp. 1628-1637, November 2019
[19] Mathew T E, Anilkumar K S, A Logistic Regression Based Hybrid Model For Breast Cancer Classification, Indian Journal of Computer Science and Engineering, Vol. 11 No. 6 Nov-Dec 2020, DOI : 10.21817/indjcse/2020/v11i6/201106201, pp 899- 906
[20] Michahial, S., Thomas, B.A.,2019, Applying cuckoo search based algorithm and hybrid-based neural classifier for breast cancer detection using ultrasound images. *Evol. Intel.* (2019). https://doi.org/10.1007/s12065-019-00268-9
[21] Mohammed Mohsin, Hong Li, And Hemn Barzan Abdalla 2020,, Optimization Driven Adam-Cuckoo Search-Based Deep Belief Network Classifier for Data Classification, IEEEAccess, Volume 8, 2020, 105542- 105560
[22] Mohammad Shehab, Ahamad TajudinKhader, Mohammed AzmiAl-Beta,2017, A survey on applications and variants of the cuckoo search algorithm, Applied Soft Computing, Volume 61, December 2017, Pages 1041-1059
[23] P. Mohapatra, et al., 2015,An improved cuckoo search based extreme learning machine for medical data classification, Swarm and Evolutionary Computation (2015), http://dx.doi.org/10.1016/j.swevo.2015.05.003i
[24] G.A.Mylavathi, B.Srinivasan, 2019, A Hyper Meta-Heuristic Cascaded Support Vector Machines for Big Data Cyber-Security, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019
[25] Nabila Nouaouria, Mounir Boukadoum a, Robert Proulx, 2013, Particle swarm classification: A survey and positioning, Pattern Recognition 2013, Volume 46, Issue 7,2028-2044
[26] Najmeh Sadat Jaddi, Salwani Abdullah, Marlinda Abdul Male, 2017, Master-Slave Cuckoo Search with parameter control for ANN optimization and its real-world application to water quality prediction, PLOS ONE,2017, https://doi.org/10.1371/journal.pone.0170372
[27] Nitesh V Chawla, Kevin W Bowyer, Lawrence O hall, W Philip Kegelmeyer,2020, SMOTE: Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research 16 (2020),321-357
[28] Park, S., Park, H.,2020, Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic. Computing (2020). https://doi.org/10.1007/s00607-020-00854-1
[29] Peralta D, del Río S, Ramírez-Gallego S, Triguero I, Benitez JM, Herrera F., 2015, Evolutionary feature selection for big data classification: a mapreduce approach. Math Probl Eng. 2015;501. Article ID 246139.
[30] Qiuming Zhu, 2020,On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset, Pattern Recognition Letters, Volume 136, 2020, pages 71-80
[31] Rajathi G M, 2020, Optimized Radial basis Neural Network for the classification of Breast cancer images, Current Medical Imaging, 2020,
[32] D. Rodrigues; L. A. M. Pereira; T. N. S. Almeida; J. P. Papa; A. N. Souza, C. C. O. Ramos; Xin-She Yang,2013, BCS: A Binary Cuckoo Search algorithm for feature selection, __IEEE International Symposium on Circuits and Systems (ISCAS), 2013, 10.1109/ISCAS.2013.6571881
[33] Sannasi Chakravarthy S R, Harikumar Rajaguru, Comparison Analysis of Linear Discriminant Analysis and Cuckoo-Search Algorithm in the Classification of Breast Cancer from Digital Mammograms, Asian Pacific Journal of Cancer Prevention, Vol 20, 2333-2337
[34] Simon Fong, Robert P. Biuk-Aghai, Richard C. Millham, 2018, Swarm Search Methods in Weka for Data Mining, ICMLC 2018: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, February 2018 Pages 122–127

[35] Shatabdi Paul, Prem Prakash Solanki, Uday Pratap Shahi, Saripella Srikrishna, 2015, Epidemiological Study on Breast Cancer-Associated Risk Factors and Screening Practices among Women in the Holy City of Varanasi, Uttar Pradesh, India, Asian Pacific Journal of Cancer Prevention, Vol 16, 2015, 6163-8171
[36] S. Sreejith, H. Khanna Nehemiah,, A. Kannan,2020, Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection, Computers in Biology and Medicine, 126, 2020, pages 1-14
[37] Sudha, M.N. and Selvarajan, S.,2016, Feature Selection Based on Enhanced Cuckoo Search for Breast Cancer Classification in Mammogram Image. Circuits and Systems, 7, (2016) 327-338.
[38] Swets JA. Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. Psychological Bulletin, 1986;99(1):100–17
[39] Vamsidhar Enireddy and Reddi Kiran Kumar, 2015, Improved cuckoo search with particle swarm optimization for classification of compressed images, Sadhana, Indian Academy of Sciences, Vol. 40, Part 8, December 2015, pp. 2271–2285
[40] S. Walton, O. Hassan, K. Morgan, M.R. Brown, 2011, "Modified Cuckoo Search: A new Gradient Free Optimization Algorithm", Chaos, Solutions & Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena, Elsevier, vol. 44, pp. 710-718, 2011.
[41] Xin -She Yang,2014, Cuckoo Search, Nature-Inspired Optimization Algorithms,2014, Pages 129-139
[42] Xuejiao Meng, Jianxia Chang, Xuebin Wang, Yimin Wang, 2019, Multi-objective hydropower station operation using an improved cuckoo search algorithm, Energy volume 168, February 2019, Pages 425-439
[43] Yang, X. S., and Deb, S., 2009. 'Cuckoo search via Levy flights', Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009, India), IEEE Publications, USA, pp. 210-214.
[44] Yvan Saeys, Inaki Inza, and Pedro Larranaga, 2007,A review of feature selection techniques in bioinformatics, Bioinformatics Vol. 23 no. 19 2007, pages 2507–2517 doi:10.1093/bioinformatics/btm344
[45] Zohre Momenimovahed, and Hamid Salehiniya, 2019, Epidemiological characteristics of and risk factors for breast cancer in the world, Breast Cancer Targets and Therapy, 2019; 11: 151–164.