A MODIFIED- WEIGHTED- K - NEAREST NEIGHBOUR AND CUCKOO SEARCH HYBRID MODEL FOR BREAST CANCER CLASSIFICATION

Tina Elizabeth Mathew
Research Scholar, Technology Management, Department of Future Studies,
University of Kerala, Thiruvananthapuram, Kerala, India
Email:tinamathew04@gmail.com

K S Anil Kumar
Research Guide, Technology Management, Department of Future Studies,
University of Kerala, Thiruvananthapuram, Kerala, India
Email:ksanilksitm@gmail.com

Abstract

One of the leading death-causing cancers in women is Breast Cancer. Accurate, precise, and early diagnosis is a crucial solution to survival. Data mining techniques have proved to produce good results in disease diagnosis. Feature search techniques are useful in identifying the relevant features for classification thus reducing time and effort. Class inequality is a significant challenge and one of the methods to overcome it is class balancing. In certain cases, the negative class is the majority class. To be specific; the negative class has a more number of instances than the positive class, so the overall classifier performance may be high; consequently, the classifier performance in accurately identifying positive instances gets overlooked. In this paper, a combination of two class balancing approaches is applied. It is used to balance the number of instances in each of the target classes. *k*-Nearest Neighbour classifier is a simple, easy to implement, and robust classifier with few parameters needed to be tuned. In this paper, we propose a *k*- Nearest Neighbour Classifier model implemented with feature search using Cuckoo search and Class balancing to classify Breast Cancer. The proposed model produced an accuracy of 99.41 %., ROC of 0.999, and MCC of 0.988.

Keywords: *k*-Nearest neighbors (*k*-NN); Cuckoo Search (CS); Class Balancing (CB); Breast Cancer (BC), Metaheuristic Search

1. Introduction

Breast cancer is a leading cause of death in women worldwide [Paul et al. 2015, Sharma et al., (2010)]. Many methods are available for diagnosing Breast cancer, yet the disease is still on the rise and claiming the lives of thousands of women [Momenimovahed & Salehiniya, (2019)]. A key to survival is the early diagnosis of cancer. Medical diagnosis at times can be inconclusive, stressful, and painful to the patients. Applying data mining techniques for the diagnosis of the disease can provide added support to healthcare personnel in disease diagnosis. Many classification models and case studies are available with medical images [Enireddy & Kumar, (2015), Chakravarthy & Rajguru, (2019), Rajathi, (2020)]. Disease classification can be done as well using cytological feature analysis besides, image analysis [Arya & Tiwari, (2016), Daoudy & Maalmi, (2020), Mathew, (2019)]. Simple as well as ensemble machine learning models [Mathew, (2019)] are being to address the breast cancer classification problem. But developing models that are precise as well as accurate is a challenging task. Misclassification of the disease is one matter to be taken into account. In most cases, the positive class which indicates the existence of the disease is a minority and the negative class which indicates no disease is the majority class. So usually an overall moderate accuracy is received in classification when the various machine learning techniques are used and a major constraint is that the classification accuracy of the minority class is overlooked [Pelayo, (2012), Park & Park, (2020)]. Misdiagnosis of a positive class is a relatively serious issue than vice versa. So to provide equal importance to both the classes in the two-class problem of Breast cancer classification class balancing is used. Medical datasets usually have high dimensional datasets and hence suffer from the curse of dimensionality [Nabila, Boukadoum & Proulx, (2020)]. Feature selection has become a necessity in many applications [Saeys, Inza & Larrananga, (2007), Mathew, (2019)] Feature selection helps in reducing irrelevant features [Perelta et al., (2015)]. Many feature search algorithms are available in the literature. Nature-inspired, Metaheuristic, Swarm Intelligence algorithms have been seen to show better performance than conventional search methods [(Fong, Aghai, Milhalm, (2018), Mathew, AnilKumar, (2020)]. These algorithms have been used for various applications from disease diagnosis [Nagthane & Rajurkar, (2017)], drug design [Houssein et al., (2020)], Traffic engineering in Networks (Ammal, Sajimon & Vinodchandra, (2020)] to Big data cybersecurity problems [Mylavathi & Sreenivasan, (2019)] and many more. The combination of feature search algorithms can also be used to improve searching [Houssein et al, (2020)]. Cuckoo search has the advantage that it is easy to implement and it has been found capable of solving many combinatorial optimization problems. A significant number of studies have used Cuckoo search to analyze images for breast cancer classification [Michahial, (2019). Adam based Cuckoo search algorithm proposed by [Mohsin, Li & Abdalla, (2020)] used Deep Belief networks to classify various datasets and the method was found to improve performance. Cuckoo search optimization and SVM were used by [Prabhukumar, Agilandeeshwari & Ganesan, (2019)] to classify lung cancer and achieved an accuracy of 98.51%. In their proposed work [Jaddi, Abdullah & Malik, (2017)] used a modified cuckoo search with ANN to predict water quality prediction. They modified the CS algorithm so that the parameter P_a .takes a maximum value initially instead of the default value of one and reduced it during the search. [Peng et al., (2020)] proposed a composite firefly algorithm with *k*-NN and applied it on Breast cancer datasets. It was seen to improve the performance and an accuracy of 98% was obtained on the WBCD dataset. An enhanced cuckoo search with *k*-NN was proposed by [Sudha, & Selvarajan, (2016)]. The model achieved an accuracy of 98.75%. A new cuckoo search based extreme learning-based model was proposed by [Mohapatra, (2015)] and it was seen to outperform other compared models. The improved cuckoo search was used to pre-train the machine.

The objectives of the paper are

- To investigate whether class balancing and the combination with feature selection improves the classification of breast cancer.
- To investigate whether the modification to the *k*-NN algorithm improves breast cancer classification.
- To develop an effective data mining approach that helps the classification of Breast cancer as malignant or benign with minimal misclassification of the positive class.

The major contribution of the work is

- A hybrid model for the classification of breast cancer based on a modified weighted *k*-Nearest Neighbour using Class balancing and Cuckoo Search with minimal misclassification of the positive class.

The rest of this paper is structured as follows. The next section describes the materials and methods used. This is followed by the results obtained. The subsequent section is a discussion of the results, and finally, a conclusion of the study is provided.

2. Methodology and Techniques

2.1. *Dataset*

The Wisconsin Breast Cancer original dataset publicly available in the University of California, Irvine Machine Learning Repository created by Dr. William H Wolberg is used. The dataset has 699 instances, 11 attributes with 458 benign (65.5%) and 241 (34.5%) malignant cases. Since 13 instances have missing attribute values only 683 instances are used and the rest is discarded. The first attribute Id number is of no relevance in classification so it is also removed from the dataset. All the attribute values are in the range 1-10. And the class has two labels 2 for benign and 4 for malignant.

2.2. *Cuckoo Search (CS)*

Cuckoo search (CS) is a nature-inspired metaheuristic algorithm belonging to the family of swarm intelligence [.Meng et al. (2018)]. The propounders are Xin-She Yang and Suash Deb [Yang & Deb, (2009)]. Cuckoo Search is considered to be easier in tuning as it has a lesser number of parameters than other metaheuristic techniques. Cuckoo search is based on the principles of the brood parasitizing mechanism of some species of cuckoo birds and Levy Flight search. Some Cuckoo species lay their eggs in the nest of other birds. To increase the hatching probability of its eggs the cuckoos at times remove the host eggs and the host bird nurtures the eggs. Three types of brood parasitism can be adopted in the Cuckoo search- intraspecific brood parasitism, nest takeover, and cooperative breeding [Shehab, Khader & Al-Betar, (2017)]. If the host bird discovers that the eggs in the nest are not its own, it throws out unknown eggs or leaves the nest and builds a new nest elsewhere. Certain species of cuckoos are capable of mimicking egg colors and patterns of the host eggs, thus preventing their eggs from getting abandoned. Levy flight, a term coined by Benoit Mandelbrot, is a random walk with step size having a levy tailed probability distribution. Many species of birds and insects follow the Levy flight properties. Here steps are defined in terms of step length with a definite probability distribution and isotropic and random direction. In the cuckoo search algorithm, each egg in the nest represents a solution. The cuckoo eggs denote new solutions. The cuckoo search aims in replacing the solution in the nests (host eggs) with better solutions (cuckoo eggs). Three rules followed in cuckoo search are [Yang & Deb, (2009)]:

- A cuckoo places an egg one at a time in an arbitrarily selected nest.
- The next generation includes the nest with the best fitness namely, eggs
- The number of host nests is fixed and, the detection of the cuckoo egg by the host bird has a probability index ϵ (0, 1) [Kuldeep et al., (2014)].

The new solution (cuckoo egg) $x_i^{(t+i)}$ is generated by the Levy flight principle and is given as

$$x_i^{(t+i)} = x_i^{(t)} + a \oplus \text{Levy}(\lambda), \tag{1}$$

Where step size, a= 1, & $x_i^{(t)}$ is the current position.

$$\text{Levy} \sim u = t^{-\lambda}, (1 < \lambda <= 3) \tag{2}$$

ƛ is the infinite variance with infinite mean

The best fitness is denoted by x_{best} and the control parameters used are scale factor (β= 1.5) and probability index (p_a). Evolution of x_i is defined by v= x_i and

$$\text{Stepsize} = 0.01(u_i/v_i)^{1/\beta} . (v - x_{best}) \tag{3}$$

Many variants of the Cuckoo search are available in the literature [Yang, (2014)]. A few are

a) Gradient free Cuckoo Search which improves the convergence rate (Walton et al., 2011). Two modifications were made. The original CS algorithm uses a step size of value 1. In the gradient method, the value of step size, (a) is varied at each generation as (a / √G), where G is the generation number, instead of being assigned a constant value of 1. The second modification was that to speed convergence information is exchanged between eggs.

b) Improved Cuckoo search [Valian, Mohana &Tavakoli, (2011)] proposed a technique to vary the parameters p_a and a, which was assumed as constant in the original version. They are initially kept high and are decreased in the final generations for fine-tuning values.

c) Binary Cuckoo search [Rodrigues et al., (2013)] is another variant where the search space is represented as n-cube, n being the number of features. A set of binary coordinates are assigned to each nest. This is an indication of whether a feature belongs to the final set of features or not. The accuracy of the classifier is the objective function used. This is to be maximized [Rodrigues et al., (2013)].

2.3. *Class balancing (CB)*

Mostly class balancing is done either by reducing the majority class or increasing the minority class. Usually increasing minority classes is seen as more effective with classifiers than vice versa, since relevant information can get lost in the process of reducing instances [Sreejith, Nehemiah & Kannan, (2020)]. Class balancing of the datasets is seen to improve classification performance [Vasquez, (2020)]. A combination of both techniques has been seen to be effective [Chawla et al., (2002), Pradoa et al., (2020)]. Resampling is being used to implement the combination method.

2.4. *k -Nearest Neighbour (k-NN)*

k-NN, also known as a lazy learner is a non-parametric supervised method developed by Thomas Cover. It assumes the similarity between the new cases and existing cases. The similarity is measured using distance metrics. The conventional *k*-NN algorithm uses the Euclidian distance. The advantage of *k*-NN is that it is simple to implement. The working of *k*-NN is as below

- *Define the objective function*
- *Select the value for k*
- *For a new point, estimate the Euclidean distance of k number of neighbors from the new point*
- *Take the k nearest neighbors for the calculated Euclidean distance.*
- *Among these k neighbors, count the labels in each category.*
- *Assign the new label to that category for which the number of labels of the neighbors is the highest.*

2.5. *Proposed model Modified -weighted k-NN (M_W-k-NN)*

The distance metric used in the conventional *k*-NN is Euclidian distance. Many other distance metrics are available. The proposed algorithm uses the Manhattan distance which is calculated as the sum of the absolute differences of the Cartesian coordinates of two points. Equation 4 gives the formula for Manhattan distance. For two given points (x_1, y_1) and (x_2, y_2)

$$\text{Manhattan distance} = |x_1 - x_2| + |y_1 - y_2|. \tag{4}$$

The *k*-NN classifier assigns the *k* nearest neighbors a weight (1/k) and all others a weight 0. The proposed model uses a weighting factor for the k nearest neighbors as shown in equation 5.

Weighting factor= (1/distance). (5)

The Proposed M_W-*k*-NN Algorithm

```
•   Let m be the number of data samples. Let p be an unknown point.
•   Read k
•   Store the samples in an array of data points arr[]
•   For j= 0 to k
        o   Calculate  d= Manhattandistance(arr[j], p)
        o   Add d to set S = (k distances obtained from p)
•   Return the majority label of S
•   Manhattandistance ($A(x_1,y_1)$, $B(x_2,y_2)$)
        o   $d= |x_1 - x_2| + |y_1 - y_2|$
        o   Return d
        o
```

Finding the value of k

```
•   The optimal value of k is obtained by using k fold validation
•   Initialize neighbors N with values[0 or 1,50, 2]
•   Initialize L_CV= empty list of cv scores
•   Perform cv(for cv= 10)
•   For each k in N
        o   Find the cv_scores using accuracy,
        o   Add  mean of cv_scores to L_CV
•   Plot accuracy vs.  k
```

2.6. *Methodology*

The working of the proposed model is in four stages – preprocessing, class balancing, feature selection, and classification.

- *Data preprocessing by removing instances with missing values*
- *Applying a Combination of Oversampling and Undersampling*
- *Cuckoo Search with Chaotic logistic map is applied*
- *Relevant features are identified*
- *Modified k-NN classifier with, k= 5, Manhattan distance metric and a weighting factor of (1/distance) is applied on the new set of feature vectors*
- *10 fold cross-validation to avoid overfitting.*
- *Evaluation of the classifier based on performance metrics –Accuracy, Receiver Operating Characteristics (ROC), False Positive Rate (FPR), Kappa Statistic, Matthews correlation coefficient (MCC), and Recall. To identify the performance of the classifier in identifying positive classes other performance metrics such as ROC, FPR, and MCC are used.*

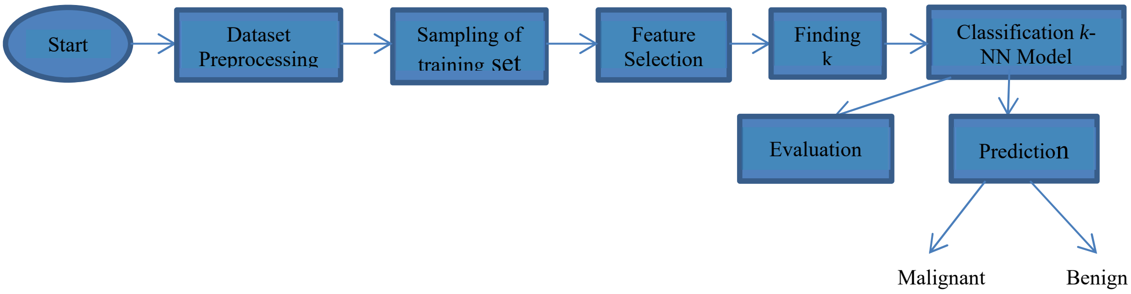

Figure 1 Working of Proposed Model

3. Results and Discussion

Table 1 summarizes the results obtained by the various models. The comparison of the performance of the classifiers is done using the standard metrics Accuracy, ROC, Recall, MCC, Kappa statistic, and FPR. The proposed method presented an accuracy of 99.41% with a ROC value of 0.999, recall of 0.996, MCC value of 0.988, FPR value of 0.006, and Kappa statistic value of 0.9883. The conventional *k*-NN method without class balancing obtained an accuracy of 94.87%. With class balancing alone the conventional method achieved an accuracy of 97.50%. With feature selection using cuckoo search the accuracy slightly decreased to 97.21%. But with the proposed modification, the model with class balancing alone achieved an accuracy of 98.97% and 95.02% without class balancing. Accuracy improved to 99.4 with feature selection using cuckoo search.

Table 1 Results

Methods	**Accuracy** **1**	**ROC** **2**	**Recall** **3**	**MCC** **4**	**FPR** **5**	**Kappa** **6**
k-NN	94.8755	0.989	0.949	0.887	0.078	0.8855
k-NN + CB	97.5073	0.996	0.975	0.950	0.025	0.9501
k-NN+ CB + CS	97.2141	0.996	0.972	0.994	0.028	0.9443
M_W-*k*-NN	95.022	0.994	0.950	0.890	0.075	0.8889
M_W-*k*-NN + CB	98.97	0.999	0.990	0.979	0.010	0.9795
M_W-*k*-NN + CB+ CS	99.4135	0.999	0.994	0.988	0.006	0.9883

In the current study, we focused on improving breast cancer classification using the k-NN classifier and implemented class balancing to avoid the issues related to class imbalance along with feature selection using cuckoo search to reduce irrelevant features. The various metrics applied to the proposed model demonstrate better performance than the original *k-NN* model with class balancing and Cuckoo search. The proposed model is better in terms of Accuracy, Recall, MCC, FPR, ROC, and Kappa statistics. The confusion matrix (Table 2) presents the classification done by the model. The proposed model misclassified one instance of the positive class and three instances of the negative class while the conventional model, having a classification accuracy of 97.21%, even with class balancing and feature selection has thirteen and 6 misclassified cases in each class. When compared against the various models' classification of the positive classes as well as the negative class in the proposed model has improved significantly. A significant improvement was seen in classifying the positive class. Thus, proving the fact that classifier accuracy alone does not imply the best classification was done [Araya & Cipriano, (2007)]. The variance of performance between the simple *k*-NN and *k*-NN with class balancing shows the sensitiveness of the classifier to an imbalanced dataset. The conventional algorithm of *k*-NN uses majority voting of the labels of the nearest neighbors to classify a new instance. In cases when the probability distribution of the data is skewed the predictions of the majority class dominate and affect the accuracy. Providing weights to the distance measured from the new point to its neighbors by multiplying it with the weighting factor (1/distance) helps in overcoming this problem. Moreover, class balancing improves the performance significantly, by keeping a balance among class labels. Feature selection using Cuckoo search further enhanced the performance. By using Manhattan distance in the proposed model, it gives the advantage that the closest approximation of the real distance is taken, whereas Euclidean distance gives the shortest distance. The chaotic map variable used in the CS algorithm is seen to improve performance as it increases the speed of search and avoids local optima. Similar performance was seen with whale optimization methods [Houssein et al, (2020)] The *k*-NN classifier without CS and CB has low performance when compared to the proposed model. FPR rate is a mere value of 0.078 compared to 0.006 achieved by the proposed model.

The proposed model is compared with the Nearest Neighbour (NN) model which uses normalized Euclidean distance as the distance metric, Logistic Regression, and Multilayer Perceptron, Random Forest, Naïve Bayes, and SMO Models (Table III). The proposed method shows better performance among the 7 classifiers. The second-best model is the NN model with an accuracy of 98.82%. The Logistic Regression and Random Forest have an accuracy of 98.53% and 98.68% respectively. The lowest performance was demonstrated by the Naïve Bayes classifier. Besides accuracy, the kappa value of the proposed model is much better than that of other classifiers. Kappa statistics compare the observed accuracy with the expected accuracy. Accuracy, inaccuracy is plotted against the proposed model and various k-NN models (Fig 2).

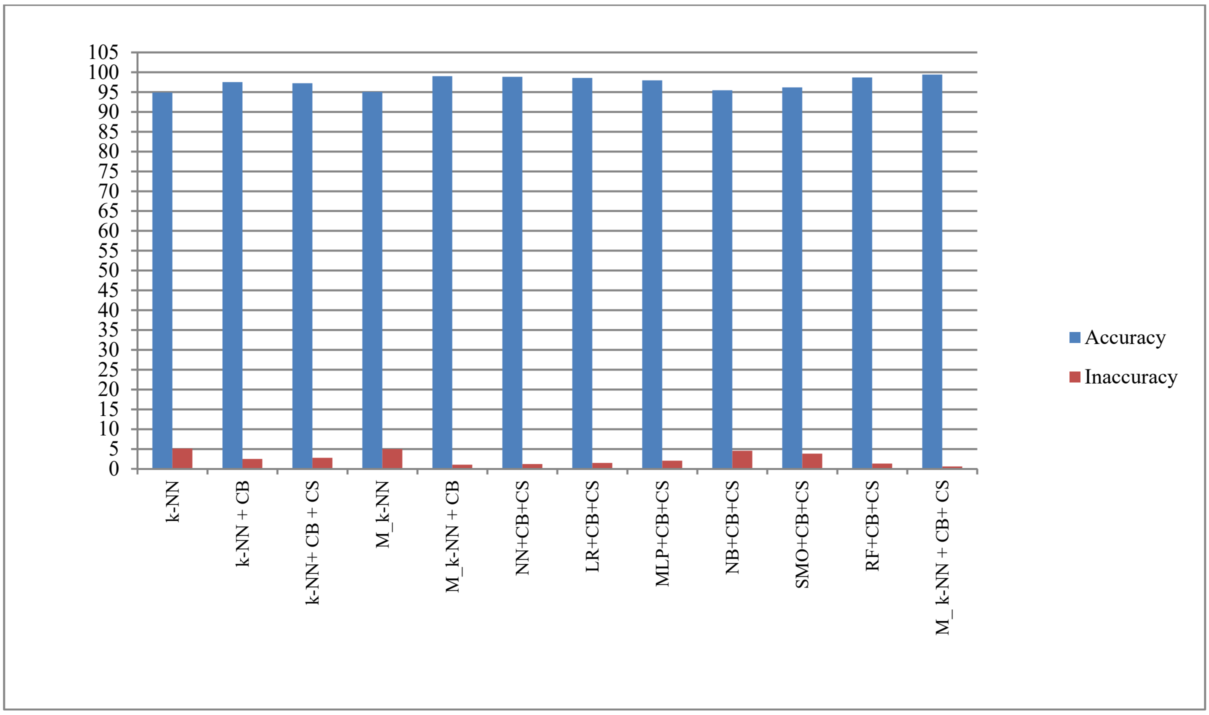

Figure 2: Methods vs Accuracy vs Inaccuracy

The ROC of the positive class, (Fig 3), with FPR, plotted on the x-axis and TPR on the y axis is shown. The ROC metric helps in discriminating between classes and is found effective for medical diagnostic evaluation (Swets, 1986, Tilaki, 2013).

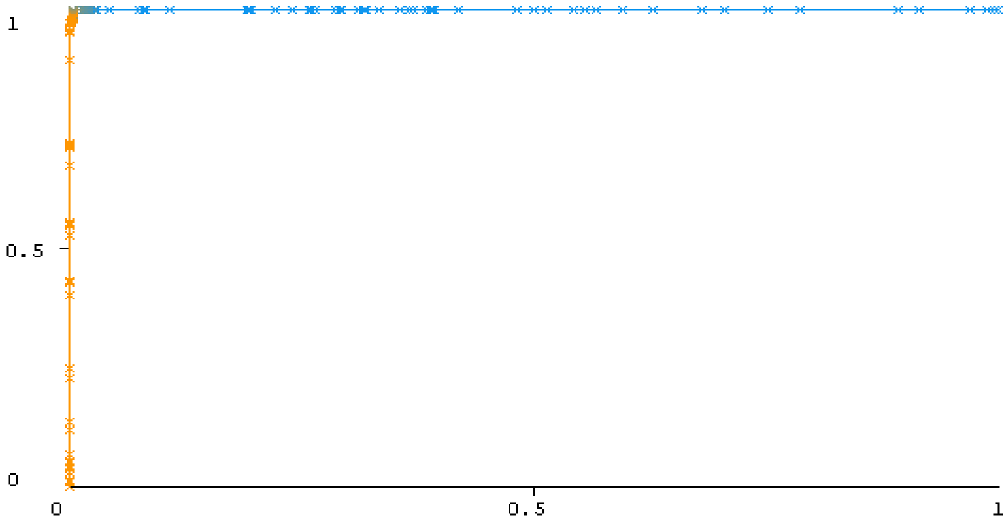

Figure 3 ROC curve:

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 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.

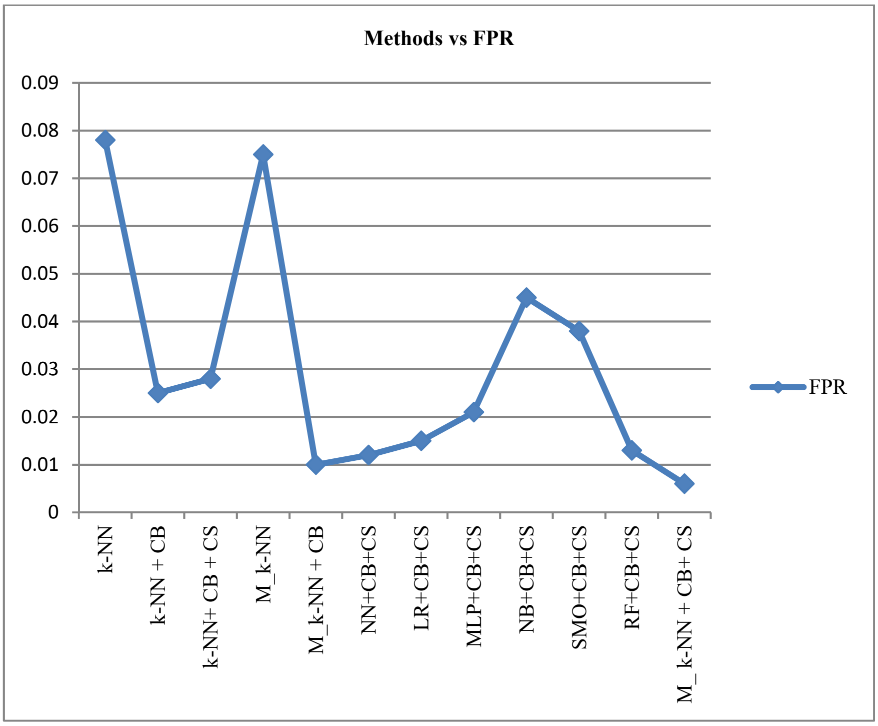

Figure 4 Methods vs FPR

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

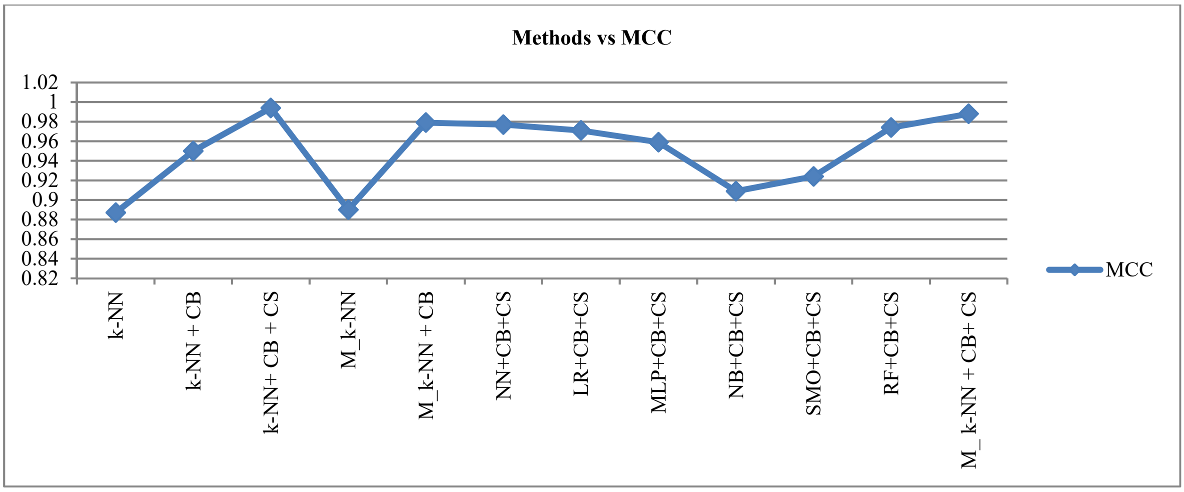

Figure 5 Methods vs MCC

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.

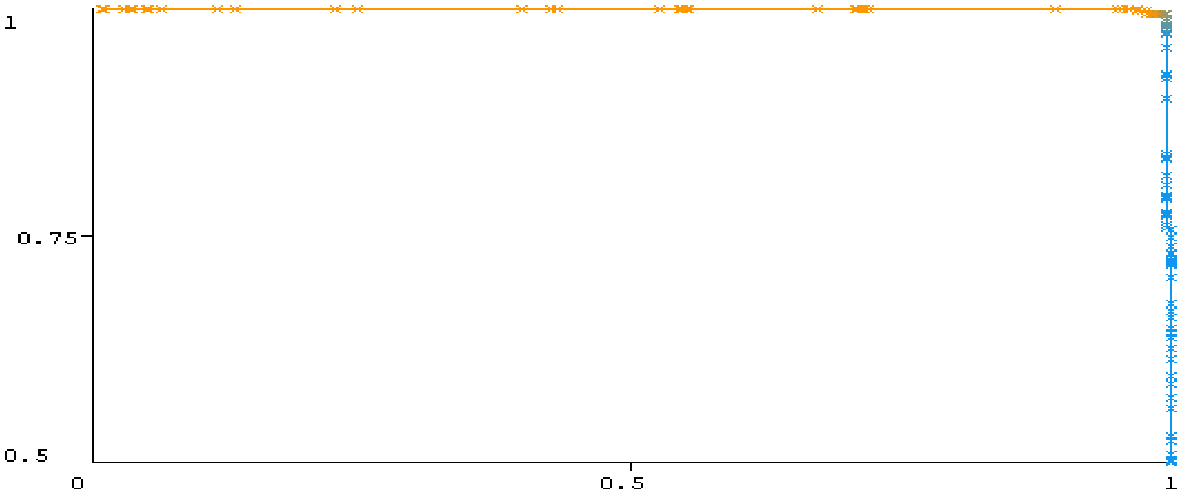

Figure 6: Precision-Recall Curve

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

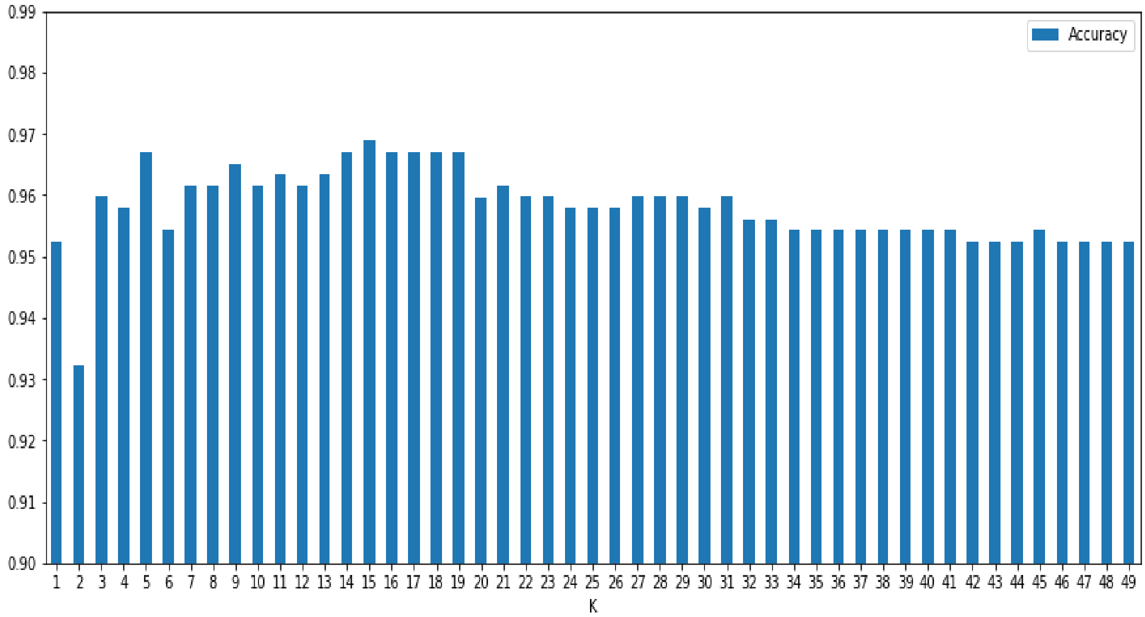

Figure 7: K vs CV Scores

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.