3.1. Data Mining Techniques Used

3.1.1 Logistic Regression

Logistic Regression is a supervised classifier that models data using a sigmoid function. The advantage of using LR is that it maximizes the quality of output on a training set and makes no assumptions on the distribution of classes in the feature space. A major problem with the classifier is that it is insensitive to imbalanced data and outliers. The model provides the relationship between one dependent binary variable and the many independent variables. The probability that a data point belongs to a specific class is given by the Logistic model. For a dataset with n features and p instances the feature matric X = $\begin{bmatrix} 1\, x11\, x12 & \cdots & x1n \\ \vdots & \ddots & \vdots \\ 1\, xp1\, xp2 & \cdots & xpn \end{bmatrix}$, where x_{ij} represent the j^{th} feature of the i^{th} instance.

The sigmoid function is used as objective function and the aim is to minimize it. The logistic function "Eq.(1)" is given as

$$Ø(z)= \frac{1}{1-e^{-z}} \tag{1}$$

Ø(z) is always bounded within (0,1). The net input function, z, is a dot product of the input features and the respective regression coefficients and is denoted "Eq. (2)" as

$$z = x_0w_o+x_1w_1+\ldots..x_nw_n = \sum_{j=1}^{n} \mathrm{wjxj} \quad = \mathrm{wTx} \tag{2}$$

z can also be represented as logit(p(y=1|x), (3)

Where p(y=1|x) is the conditional probability that data point belongs to class 1 given its features x. This is inverse to the logistic function and once model fitting is done, the conditional probability p(y=1|x) is transformed to a binary class label thru g(z), a threshold function "Eq.(4)" and

$$g(z)= 1 \text{ if } Ø(z) \geq 0.5 \tag{4}$$
$$= 0 \text{ else}$$

To minimize the logistic function the Maximum likelihood function is used. The log likelihood function is maximized or alternatively a cost function can be defined to be minimized. This cost function "Eq. (5)" can be defined as

$$H(w)= - \log(Ø(z)) \text{ if } y = 1$$
$$= -\log(1- Ø(z)) \text{ if } y = 0 \tag{5}$$

To prevent overfitting L2 parameter regularization is done. Large weight values are penalized to reduce the model complexity. The regularization term is added to the cost function. The L2 parameter "Eq. (6)" is given as

$$L2 = \frac{\lambda}{2} \sum w_j2,\ j = 1\ldots.n \tag{6}$$

3.2 Class Balancing Techniques

3.2.1 Oversampling using SMOTE (OS)

Synthetic minority oversampling technique (SMOTE) is an oversampling method used for class imbalance problems. In SMOTE minority class examples are randomly increased by replicating them. To be precise, new minority instances are synthesized between the existing minority instances. The synthetic records are generated by the random selection of k nearest neighbours of the instances in the minority class.

3.2.2 Undersampling using Random Subsampling (US)

Random Undersampling involves randomly selecting examples from the majority class to be removed from the training dataset. To be specific, a sample down procedure is done on the majority class data until it occurs with the same frequency as the minority class The major limitation of Undersampling is that instances from the majority class that are deleted may be useful important information or even perhaps critical to fitting a robust decision boundary and this can influence the performance of the model.

3.2.3 Combination of Oversampling and Undersampling (OS+US)

A moderate increase in minority class instances and moderate reduction in majority class instances help in improving and reducing the bias involved in the two situations. First the difference between the majority and minority class samples are calculated. Then the number of samples to be removed from the majority class and number of samples to be increased for the minority class are determined. Then, the majority class samples are reduced and the minority class samples are increased accordingly.

3.3 Feature Search Techniques

3.3.1 Ant Search (AntSrch/AS)

Ant Search is based on the Ant Colony optimization technique proposed by Marco Dorigo and colleagues in 1990s. It is a population based metaheuristic technique. It is inspired by the foraging behavior of ants seeking an optimal path from the food source to their colony. Ants live and work in colonies and as a group exhibit highly organized capabilities. They travel the shortest path between their food sources and nest. They communicate with each other through pheromones as they have low visibility. Pheromones are chemical substances released while an ant travels on the ground. These mark trails on the ground and other ants follow this path. The collective behavior of ants is used as an optimization tool. Initially ants move in random searching for food. Hence multiple paths are created. A portion of food is carried back to the nest if the quantity and quality are right, and they leave pheromone trails on the way back. This acts as a guide to other ants. Pheromone evaporation is also to be taken into account. On the less travelled trail the pheromones evaporate and hence the most frequently travelled path will have a high intensity of pheromones. The intensity of pheromones on the travelled path increases as each ant traversing it deposits pheromones on it. There are various variants to the original ACO algorithm. Elitist ant systems, Ant colony system, Max- Min ant system, rank based ant systems, and continuous orthogonal ant systems,

The algorithm for ant search is given as

```
Procedure AntSearch()
Initialize pheromone trails and parameters.- population size n, maximum iterations, pheromone value, fitness value, pheromone evaporation rate T,
While (not terminated)
        Generate ant population
        Calculate fitness value for each ant
        Find best solution through roulette wheel selection criteria
        Update pheromone trail.
End while
Display best ant(solution), best fitness value
End
```

3.4. Data Used

The benchmark Breast Cancer datasets of the Wisconsin Hospitals from the UCI repository is being used. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The Breast Cancer Wisconsin Original data Set has 699 instances, 11 attributes and 13 missing instances with 458 benign (65.5%) and 241 (34.5%) malignant cases. The data features are computed from the digitized image of a fine needle aspirate (FNA) of a breast mass and describe characteristics of the cell nuclei present in the image. The attributes are shown in Table 1. Id Number, since it has no relevance in the classification process is discarded from the set of attributes.

TABLE 1 Attributes

Number	Attribute Name	Values	Comparison of malignant and benign cells	
			Malignant	Benign
1	Clump_thickness	1-10	Seen in Multilayers	Seen in monolayers
2	Size_uniformity	1-10	Size differs	Unifrom size
3	Shape_uniformity	1-10	Shape differs	Unifrom Shape
4	Marginal_adhesion	1-10	Cells don not stick together	Cells stick together
5	Epithelial_size	1-10	Enlarged	Small
6	Bare_nucleoli	1-10	Have bare Nucleoli	No Bare Nucleoli
7	Bland_chromatin	1-10	Coarse in texture	Uniform texture
8	Normal_nucleoli	1-10	Nucleus is bigger	Nucleus is small
9	Mitosis	1-10	More Mitosis	Not so
10	Class	2-Benign 4-Malignant		

3.5 Evaluation Metrics used

Various metrics are available for evaluation of models. The metrics used for evaluation of the model in this study are accuracy, ROC, Mathews Correlation Coefficient (MCC), Kappa Statistic, Precision and Recall.

3.5.1 Accuracy

Accuracy is the number of correct classifications made by the model. It is evaluated as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100 \qquad (7)$$

TP, TN, FP, FN being the True Positives, True Negative, False Positives and False Negatives obtained from the confusion matrix.

3.5.2 Mathews Correlation Coefficient

$$\text{MCC=} \quad \frac{=TPxTN-FPxFN}{\sqrt{}(TP+FP)(TP+FN)(TN+FP)(TN+FN)} \qquad (8)$$

3.5.3 F Score

F score is the harmonic mean of Precision and Recall

$$\text{F score} = \frac{2TP}{2TP+FP+FN} \qquad (9)$$

3.5.4 Kappa Statistic

It compares the expected and observed outcome and is given by

$$\text{Kappa} = \frac{total\ accuracy-random\ accuracy}{1-random\ accuracy} \qquad (10)$$

And random accuracy = $\frac{(TN+FP)(TN+FN)+(FN+TP)(FP+TP)}{total\ x\ total}$

And total accuracy = Accuracy

3.5.5 Recall

Recall is also known as sensitivity gives the number of correctly classified true positives.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (11)$$

3.5.6 Precision

Precision gives the number of true positives against the number of positives identified.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (12)$$

4. Results and Discussion

The performance of the hybrid model obtained is shown in Table 2. Accuracy “Eq. (7)” of 99.4 % was attained by the proposed model. MCC “Eq. (8) is a reliable score which produces a good result only if a good prediction score is arrived for the four categories of the confusion matrix. It shows how well the classifier performs and the proposed model displayed it at 0.988. Table 2 gives the comparison of the various cases of accuracy of the classifier. ROC, Fig 2 depicts the tradeoff between TPR and FPR. The ROC value of 0.998 was obtained and in figure 2 it can be seen along the y axis at leftmost edge of the graph. The Confusion matrix presents the correctly classified and misclassified instances of the two classes. The hybrid model was seen to classify the positive classes correctly except for one instance and 3 wrongly classified instance for the negative class. The kappa statistic “Eq. (10)” measures the interrater reliability viz. expected and observed outcomes. A value of 1 shows perfect agreement. The proposed model achieved a good value of 0.9883. The F measure “Eq.(9)” is obtained as the harmonic mean of precision and recall. Recall “Eq. (11)” gives the ratio of correctly predicted positive observations to the all observations in actual class. The proposed model achieved a recall, precision “Eq. (12)” and F measure of 0.994 each.

The proposed model is compared with the conventional Logistic Regression classifier, the LR classifier with Oversampling alone performed, LR classifier with Undersampling alone done and with feature search performed with the LR classifier. The proposed model outperformed them in all the cases. Oversampling improved the accuracy measure considerably when compared with the technique with no class balancing. This is due to the increase in samples of the minority class. The performance of Undersampling was reduced and seen as the least efficient among the class balancing methods applied in the models. This is due to the loss of useful information when samples are reduced from the majority class. The combination of Undersampling and Oversampling produced the best results with the Logistic Regression classifier. This was similar with the four other classifiers used for comparison, (Table 3). Table 2 shows the various performance metrics used. In every case the proposed hybrid model performed comparatively better than them all.

Table 2 Performance Metrics

Classifiers	Accuracy	Kappa	ROC	FPR	F-Measure	MCC	Recall	Precision	Time taken to build model (secs)	Confusion Matrix
LR	92.53	0.8321	0.928	0.112	0.924	0.835	0.925	0.926	0.06	430 14 37 202
LR+OS	96.42	0.9284	0.988	0.035	0.964	0.929	0.964	0.965	0.28	433 11 22 456
LR+US	92.25	0.8452	0.958	0.077	0.923	0.846	0.923	0.924	0.16	
LR+(OS+US)	99.1	0.9824	0.999	0.009	0.991	0.982	0.991	0.991	0.05	339 2 4 337
LR+Ant Search	91.9	0.8311	0.920	0.116	0.921	0.832	0.921	0.921	0.06	418 26 29 210
Proposed Hybrid Model-LR+OS+US+AntSrch	99.4	0.9883	0.998	0.006	0.994	0.988	0.994	0.994	0.03	338 3 1 340

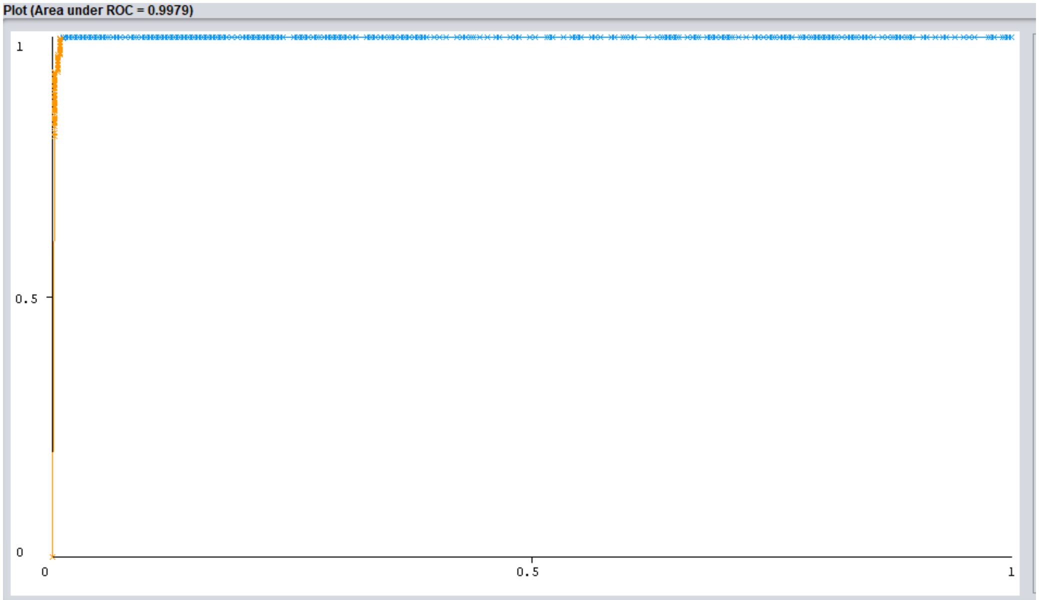

Figure 2 ROC

The performance of the proposed model is evaluated other four other classifiers- Support Vector Machines, Neural Networks and Naïve Bayes is shown in Table 3. The accuracy obtained for each classifier is displayed. SVM obtained an accuracy of 98.87%, Neural Network 98.97%, Naïve Bayes 98.3% and Random Forest 98.82%. The results show that the proposed model outperformed them all with an accuracy of 99.4%.

Table 3 Comparison with other classifiers

Classifier	Accuracy%	Confusion Matrix
SVM	98.97	338 3 4 337
Neural Networks	98.97	338 3 4 337
Naïve Bayes	98.3	332 9 2 339
Random Forest	98.82	337 4 4 337
Proposed Model	99.4	338 3 1 340

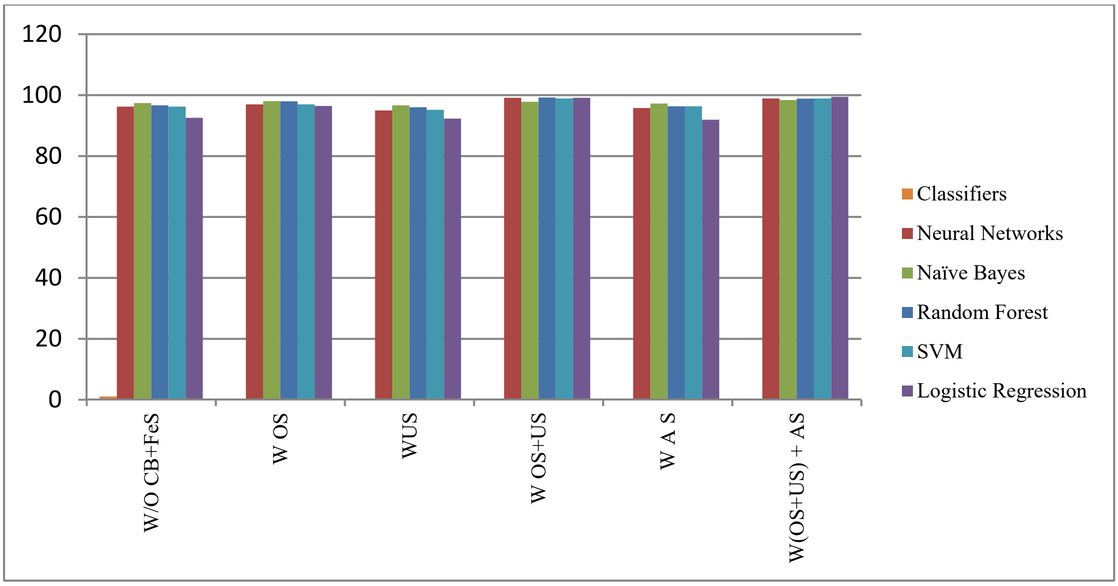

Fig 3 Comparison of Models against accuracy

In Fig 3, the performance of Logistic Regression models is presented alongwith that of the other four classifiers. The models using the classifiers alone performed comparatively lesser. The best accuracy among them was given by Naïve Bayes. With Oversampling alone Random Forest and Naïve Bayes models produced the best result. With Undersampling alone Naïve Bayes was the best classifier. With the combination of Oversampling and Undersampling performance of Logistic Regression improved considerably. Similar was the case with Random Forest, NN, and SVM, while Naïve Bayes showed a moderate performance. The best performance was by the proposed Logistic Regression and Ant search model. With Class balancing the results of all the classifiers increased considerably. With Ant search alone the best result was given by Naïve Bayes and the combination of OS+US increased the performance of all the classifiers significantly. Logistic regression improved with the combination class balancing methods and ant search techniques. Logistic regression classifier is sensitive to class imbalance. Hence applying a combination of oversampling and undersampling brought a significant accuracy increase. Combining ant search with this further increased the performance of the model. The proposed models are compared with a few other works that used swarm intelligence methods with classification techniques in Table 4

Table 4 Comparisons with other Related Works using various Search Techniques

Literature	Search Technique	Data Mining technique	Accuracy %
Saoud *et al.* (2019) [25]	Best First	LR	96.7096
Dhahri *et al.* (2020) [26]	Tabu Search	LR	98
Mathew (2019)[27]	RFE	LR	95.98
Proposed work	Ant search	LR	99.4

The proposed model is compared against related works in Logistic Regression with various feature selection methods. The proposed model outperformed all in terms of accuracy.

5. Conclusion

The paper produced a hybrid model for breast Cancer classification using Logistic Regression on the WBCD dataset using ant search and class balancing techniques. The model was compared against four other Meta heuristic methods for attribute selection and reduction and four other classifiers alongwith three class balancing methods. Among the balancing methods used combination of Undersampling and Oversampling was seen most effective in Logistic Regression. Feature selection using Ant Search, when applied on the model made improvement to the various performance measures. The best accuracy measure was obtained by the hybrid Logistic Regression model using the combination of class balancing and Ant Search methods with a value of 99.4%. Ensemble methods can be devised so as to improve the performance of other feature selection methods with logistic Regression for classification in two class problems. Modification of the cost function and application of optimization techniques for improving the Logistic regression model can also be explored

Acknowledgments

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

References

[1] Arafat, H, Barakat, S, Goweda, A F, Using Intelligent Techniques for Breast Cancer Classification, International Journal of Emerging Trends & Technology in Computer Science Volume I, Issue 3, September-October 2012

[2] Chawla, N V, Bowyer, K W , Hall, L, O, Kegelmeyer, W P, SMOTE: Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research 16 (2002),321-357

[3] Dhahri, H, Mahmood,R A, Maghayreh, E A, and Elkilani·,W Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions, BioMed research International, Volume 2020, Article ID 4671349

[4] Deb, S, Fong,S, Tian, Z, Elephant Search Algorithm for Optimization Problems, The Tenth International Conference on Digital Information Management (ICDIM 2015)

[5] Durga, T S, Assiri, A S, Nazir,S, Velastin,S A, breast Tumor Classification using an Ensemble Machine Learning Method, J Imaging, 2020,6,39

[6] Emami, N A. Pakzad, A New Knowledge-based System for Diagnosis of Breast Cancer by a combination of Affinity Propagation Clustering and Firefly Algorithm, Journal of AI and Data Mining Vol 7, No 1, 2019, 59-68

[7] Fang, M, Lei, X, Cheng, S, Shi, Y, Wu, F X, Feature Selection via Swarm Intelligence for Determining Protein Essentiality, Molecules 2018, 23, 1569

[8] Fong, S Biuk-Aghai, R P, RicMillham ,R C, 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

[9] Jabbar, S F, A classification model on tumor cancer disease based mutual information and firefly algorithm, Periodicals of Engineering and Natural Sciences ISSN 2303-4521 Vol. 7, No. 3, September 2019, pp.1152-1162

[10] Kotsiantis, S B,"Supervised machine learning: a review of classification techniques," Informatica, vol. 31, no. 3, pp. 249–268, 2007.

[11] Li, J, Fong, S, Wong, R K, Millham , R,. Wong, K K L, Elitist Binary Wolf Search Algorithm for Heuristic Feature Selection in High-Dimensional Bioinformatics Datasets, Scientific Reports volume 7, Article number: 4354 (2017)

[12] 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

[13] 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)

[14] 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

[15] Mazen, F, AbulSeoud , R A, Gody ,A M, Genetic Algorithm and Firefly Algorithm in a Hybrid Approach for Breast Cancer Diagnosis, International Journal of Computer Trends and Technology (IJCTT) – Volume 32 Number 2 - February 2016

[16] Nadira,T , Rustam, Z, Classification of cancer data using support vector machines with features selection method based on global artificial bee colony, Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ISCPMS2017)AIP Conference Proceedings 2023, 020205 (2018),pp 1-7

[17] Rahman, M, M A, , Muniyandi, R C, An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons, Symmetry 2020, 12, 271, 2- 21

[18] Rajaguru H, Prabhakar, S K, A Study on Firefly Algorithm for Breast Cancer Classification, Lecture Notes in Computational Vision and Biomechanics Volume 30, 2018, 421-428

[19] Rajalaxmi,R R, E.Gothai, .R.Thamilselvan, P.G Bindha, P.Natesan, Naïve Bayes guided Binary Firefly Algorithm for Gene Selection in Cancer Classification, International Journal of recent technology and engineering, Volume 8, issue 4, November 2019

[20] Rajendran, K , Jayabalan, M , Thiruchelvam, V , Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020

[21] Ramasamy, R and Rani, S, Modified Binary Bat Algorithm for Feature Selection in Unsupervised Learning, The International Arab Journal of Information Technology, Vol. 15, No. 6, November 2018

[22] Rania, R R , .Ramyachitra,D, Microarray Cancer Gene Feature Selection Using Spider Monkey Optimization Algorithm and Cancer Classification using SVM, Procedia Computer Science 143 (2018) 108–116

[23] Reddy, G T, Khare, N, Hybrid Firefly-Bat Optimized Fuzzy Artificial Neural Network Based Classifier for Diabetes Diagnosis, International Journal of Intelligent Engineering and Systems, Vol.10, No.4, 2017

[24] Sadeghipour, E Sahragard, N, Sayebani,M R, Mahdizadeh,R, Breast Cancer Detection Based On A Hybrid Approach Of Firefly Algorithm And Intelligent Systems, Indian Journal of Fundamental and Applied Life Sciences , 2015, 468-472

[25] Saoud, H, Ghadi,A ,Ghailani, M, Abdelhakim, B A, Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification, © Springer Nature Switzerland AG 2019 M. Ben Ahmed et al. (Eds.): SCA 2018, LNITI, pp. 307–315, 2019.

[26] Singhal, S K, An Evolutionary Bayesian Network Learning Algorithm using Feature Subset Selection for Bayesian Network Classifiers, International Journal of Computer Applications (0975 - 8887) Volume 135 - No.13, February 2016

[27] Sivapriya,T R, A. R. Kamal, N B, Thangaiah,, P R J, Ensemble Merit Merge Feature Selection for Enhanced Multinomial Classification in Alzheimer's Dementia, Computational and Mathematical Methods in Medicine Volume 2015, Article ID 676129, 11 pages

[28] Tran, C T, Zhang, M, Andreae, P, and Xue,B, Bagging and Feature Selection for Classification with Incomplete Data, Conference: Evostar, April 2017, 471- 486

[29] Yahya, A A, Centroid particle swarm optimisation for highdimensional data classification, Journal Of Experimental & Theoretical Artificial Intelligence, 2018