Classification for Imbalanced Breast Cancer Dataset Using Resampling Methods

Hana Babiker Nassar
hananassar2011@gmail.com
Department of Information Technology, University of Science and Technology, Omdurman, Sudan

Abstract
Analyzing breast cancer patient files is becoming an exciting area of medical information analysis, especially with the increasing number of patient files. In this paper, breast cancer data is collected from Khartoum state hospital, and the dataset is classified into recurrence and no recurrence. The data is imbalanced, meaning that one of the two classes have more sample than the other. Many pre-processing techniques are applied to classify this imbalanced data, resampling, attribute selection, and handling missing values, and then different classifiers models are built. In the first experiment, five classifiers (ANN, REP TREE, SVM, and J48) are used, and in the second experiment, meta-learning algorithms (Bagging, Boosting, and Random subspace). Finally, the ensemble model is used. The best result was obtained from the ensemble model (Boosting with J48) with the highest accuracy 95.2797% among all the algorithms, followed by Bagging with J48(90.559%) and random subspace with J48(84.2657%). The breast cancer imbalanced dataset was classified into recurrence, and no recurrence with different classified algorithms and the best result was obtained from the ensemble model.

Keywords:
Breast Cancer, Imbalance dataset, attribute selection, Resampling method, Ensemble model.

1. Introduction

The initial stage of breast cancer begins when cells in the breast begin to grow out of control. A tumor is formed by these cells that can be seen on X-ray or be felt as a lump. The common symptom of breast cancer is a new swelling or accumulation. A hard mass with irregular edges without pain is more likely to be cancer, but breast cancers can be tender, soft, or rounded. They can even be painful[1], [2]. Breast cancer is a malignant or benign tumor, inside the east, wherein cells divide and grow without control. As a few risk factors increase the likelihood of a woman developing breast cancer, the researchers tried to find the exact reason behind breast cancer. The cancer stage is one of the most essential factors in selecting treatment options, and it uses the Tumor, Nodes, and Metastasis (TNM) system. This is indicated as a form of tumor from Stage 0 (the least advanced stage) to the most advanced stage[1]. Imbalanced data classification often arises in many practical applications. Many classification approaches are developed by assuming the underlying training set is evenly distributed. However, those approaches face a severe bias problem when the training set is highly imbalanced. There are many real-world[3] problems faced with severe learning problems for imbalanced classes[4]. Most of the data in the real word are imbalanced in nature. This situation occurs when the distribution of the target class is not uniform among the different class levels. Classification of this type of data is one of the most challenging problems in the field of machine learning and has recently gained a great deal of interest[4].

In an imbalanced data set, the class; having more instances is called a significant mas, while the one having a relative and several instances are recalled as a minor class [4].

Several methods were developed to solve this imbalance data problem these methods, include methods based on sampling techniques, cost-sensitive learning, ensemble learning, feature selection, and algorithmic modification.

2. Related Work

Wang [5] used a Bayesian classifier as a probabilistic classification technique. To help with interpreting classification results, a hybrid network that combined both ANN and Bayesian networks was studied. The SEER dataset was used to compare the performance of Bayesian Networks, ANN, and Hybrid Networks. Their results showed that the hybrid model can produce better accuracy and can make the decision; easier.

William et al. [6]worked on the application of data mining techniques to model breast cancer using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. The input contained sample code number, clump thickness,

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.12

cell size and shape uniformity, cell growth, and other physical examination results The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and the error rate of 0 while CART had the lowest accuracy with a value of 92.99%, but naïve Bayes 'had an accuracy of 97.42% with an error rate of 0.0258. Delen et al. [3]used an evolutionary ANN (EANN) for breast cancer Diagnosis the EANN, was able to achieve an average test accuracy of 0.981 with a standard deviation of 0.005. it is also used, in an ANN ensemble to predict cardiorespiratory morbidity. The ANN ensemble performed very well and achieved an area under the ROC curve value of 0.98.

G.I. Salama et al.[7] used Decision tree J48 to implement Quinlan 's C4.5 algorithm for generating a pruned, unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to decide by splitting the data into smaller subsets.

J48 examines the normalized information gain that results from choosing an attribute for spilling the data. To

The attribute with the highest normalized information gain is used to make the decision algorithm recurs on

the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. However, it can also happen that none of the features give any information gain. In this case, J48 creates a decision node higher up in the tree using the expected value of the classes[8].

J. Joshi, et al. [7]constructed the work to discover the effectiveness of pre-processing algorithms on datasets that are used to achieve more accurate results.

The cost-sensitive algorithm is used along with base classifiers on naïve Bayes, minimal sequential optimization (SMO), K-Nearest Neighbors (KNN)[9], [10], Support Vector Machine (SVM), and C4.5 were closed work with SMO algorithm have very high sensitivity (97.22%) and accuracy (92.09%) rates. As a final point, they said that the proposed cost-sensitive algorithm could be used on other diseases such as cancer.

J. Joshi, et al. [10] [11]analyzed Brest cancer using data mining technique classification. They use machine learning.

Techniques like Decision Tree (C4.5), Artificial Neural Networks, RK, and support vector machines for predicting breast cancer. This work targeted analyzing the performance to achieve higher accuracy specificity, and the sensitivity result by the SVM technique indicated a promising accuracy level of 95.7%, 97. levels 94.5%. of 95.7%, 97.1%, and 94.5%.

Cha[12]rasia [13]compared the performance criterion of supervised learning classifiers, such as Naïve Bayes, SVM-RBF Kernel, RBF neural network, Decision classifier classifiers, such as Naive Bayes, SVM-RBF kernel, RBF neural networks, and Decision.

Tree (DT) (J48), and simple classification and regression tree (CART)[14], to find the best classifier in breast Cancer datasets. The experimental result shows that the SVM-RBF kernel is more accurate than other classifiers; it, Score at the accuracy level of 96.84% in the Wisconsin breast cancer (original) datasets.

V. Chaurasia offered three popular data mining algorithms: CART, ID3 (iterative dichotomized 3), and

DT for diagnosing heart diseases, and the results presented demonstrated that CART obtained higher accuracy Within less time.

Choi et al.[2] compared three data mining methods: ANN, decision trees, and logistic regression they

used 202,932 records obtained from the SEER (1973-2000) data set they selected 20 variables (race, marital status, primary site, histology, behavior, grade, extension of disease (includes five subfields), lymph node involvement, radiation, stage of cancer, site-specific surgery, age, tumor size, number of positive nodes, number of nodes, number of primaries). The accuracies of the models were high (93.6%, 91.2%, and 89.2%, respectively). However, in the tenfold cross-validation, 90% of the data set was used to train the model, and only 10%is used to test the model.

Vikas Chourasia et al. [15] used RepTree, RBF Network, and Simple Logistics to predict the survivability of breast cancer patients. They consider the effect of an ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set compared to previous results. Liu-Qin experimented with breast cancer data using the C5 algorithm with bagging to predict breast cancer survivability.

Ravi Avula et al. [16]studied the performance of the synthetic minority oversampling technique (SMOTE) to

deal with imbalanced breast cancer data. Studied classification algorithms like Random Tree, J48, Naive Bayes, One R, and Zero R. J48 showed better performance.

3. Methods

The methods start with steps that include dataset description, tool selection, pre-processing, resampling, attribute selection, classification algorithms, ensemble model, and evaluation results.

The procedure followed consists of the steps starting from literature review and creation of the data set, data

Transformation, selecting tools, and preprocessing in this step preprocessing, we used, missing values, and the attribute selection resamples method to solve an imbalanced dataset. The algorithms selected in this research were base classifier, meta classifier, ensemble model finally, and evaluation results and model.

3.1 Data set Description

A data set is a set of data collected. for a specific purpose, there are, many ways in which data can be collected.

For example, extraction, surveys, interviews, observations, and so on.

The data set is obtained from files and records of Khartoum State Hospital, the total data sampled included 1144 Patients from the hospital.

The data include (Age), (T, Tumor Size), (Node – Caps), (deg – malign), (Metastasis), (L, left breast R, right breast) and irradiate, and Class.

This dataset is Imbalanced, including 336 recurrences and 808 no recurrences.

In an imbalance data set, the class having more instances is called a major class, while the one having relatively a smaller number of instances are called mina or classes.

Table: 1 dataset description.

Item	Describe	Attribute Type
(T, Tumor)	Patient's tumor in the breasts	Numeric
Age	Patient's Age	Numeric
(N, Nodes)	Node is present or not in the cap of the breast	Nominal
(M, Metastasis)	Tumors spread to other parts of the body	Nominal
Deg-Malig	Stage of breast cancer	Numeric
L/R	Breast left and right.	Nominal
Irradiate	Present or not	Nominal
Class	No recurrence-events, recurrence-events (reduce the risk of breast cancer)	Nominal
Item	Describe	Attribute Type

3.2 A selected Tool

The software framework of this work has been developed with the WEKA tool. WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, and association rules, including visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open-source software issued under a general public license[17].

3.3 Pre-processing

Data pre-processing[18] is an essential step in the data mining process. Data pre-processing describes any type of processing, the data was inserted in an Excel sheet and saved by CSV[19], missing, values were processed in the data all, data are numeric and nominal[20].

3.4 Resampling

Sampling techniques used to solve imbalance data problems with the distribution of a dataset, sampling techniques involve artificially resampling the data set, which is also known then as the data pre-processing data[21].

3.4.1 Under-sampling

The most essential method in under-sampling is a random under-sampling method that tries to balance the class distribution by randomly removing the majority class sample[10], [22]. Figure 1 shows the random under-sampling method. The problem with this method is the loss of valuable information[23].

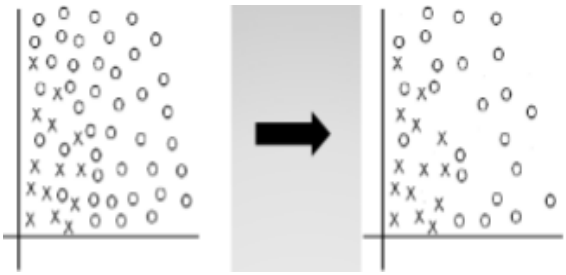

Fig.1 Randomly removes the majority sample. [4]

3.4.2 Oversampling

Random oversampling methods also help achieve balance class distribution by replicating minatory class samples. Figure 2 shows random oversampling[24].

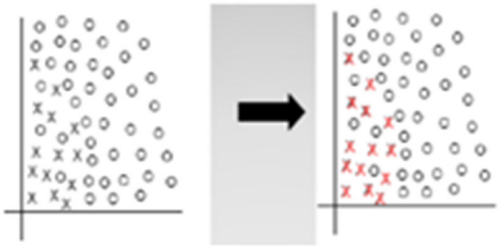

Fig:2 Randomly removes the majority sample. [4]

3.5 Classification of algorithms selected

This paper used four base classifier algorithms (j48, SVM, rep tree, neural network; and three Meta classifiers.

(Bagging, boosting, random subspace) to build the classifier model.

3.6 Ensemble learning

In this paper, I used an ensemble model to improve classification accuracy. Ensemble for classification composite model comprising a combination of classifiers base and Meta classifiers. Ensemble methods can be used to increase overall accuracy by learning and combining a series base classifier model. Bagging, boosting, and random subspace are popular ensemble methods[21]. The ensemble Model combined different types of classifiers to find the optimal classification performance from the combi sub-model. It contains two layers, the first layer consists of base classifiers, and the second layer is a Meta classifier which receives the prediction of the base classifiers as an input and then generates the final prediction.

3.7 Experiment and result

The paper consists of three experiments ensemble model combination meta classifier with base classifier with resampling and attribute selection.

3.8 First experiment boosting with attribute selection and resampling combined j4 [16], [25] 8,

The experiment applies to boost the ensemble learning algorithm with the J48 tree classification algorithm after The resampling technique is used to improve the performance of ensemble learning on imbalanced data. Boosting algorithm is tried with many classification algorithms (J48, REP, Random Forest tree, SVM, Neural Network), the best-boosting performance after resampling was obtained with the J48 decision tree.

The accuracy is 95.2797 %, the time taken to build the model was 0.14 seconds, correctly classified instances 1090, incorrectly classified instances 54, kappa statistic 0.8852 mean absolute error 0.0495, root mean squared error 0.2134, relative absolute error 11.9292 %, root relative squared error 46.8643 %.

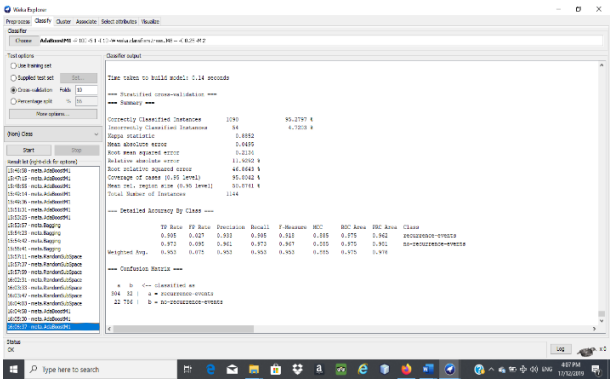

Fig 3: Result of boosting with attribute selection combined (J48)

3.9 Second experiment Bagging with attribute selection and resampling combined (J48)

The experiment applies the bagging ensemble learning algorithm with the J48 tree classification algorithm after The resampling technique is used to improve the performance of ensemble learning on imbalanced[26].

The algorithm is tried with many classification algorithms (J48, rep tree, SVM, and neural network), and the best Bagging performance after resampling was obtained with the J48 decision tree. The accuracy was 90.5594 %, and the time taken to build the model is 0secondscond correctly, classified instances 1036, incorrectly classified instances 108, kappa statistic 0.7621, mean absolute error 0.1719, root mean squared error 0.2717, relative absolute error 41.4193%, root relative squared error 59.6497%.

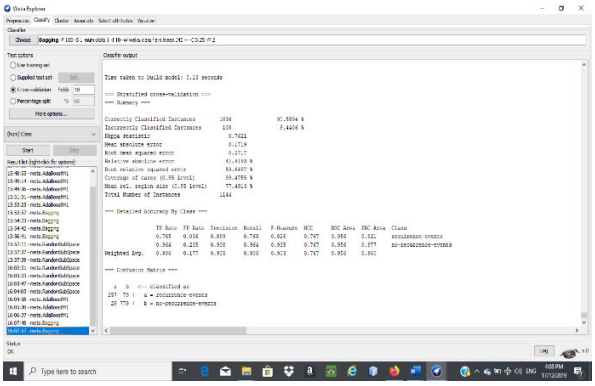

Figure 4: Result of bagging with attribute selection and resampling combined (j48)

3.10 Third Experiment random subspace with attribute selection and resampling with j48

The experiment applies a random subspace ensemble learning algorithm with the j48 tree classification algorithm after the resampling techniques are used to improve the performance of ensemble learning on imbalanced data. Random subspace algorithm is tried with many classification algorithms (j48, rep tree, SVM, and neural Network). The best performance of random subspace after resampling is obtained with the j48 decision tree.

The accuracy is 84.2657%, the time taken to build the model is 0.09 seconds correctly classified instances 964, incorrectly classified instances 180, kappa statistics 0.5639, mean absolute error 0.28, 77 root mean squared error 0.5436, relative absolute error 69.3097%, root relative squared error 75.4349%.

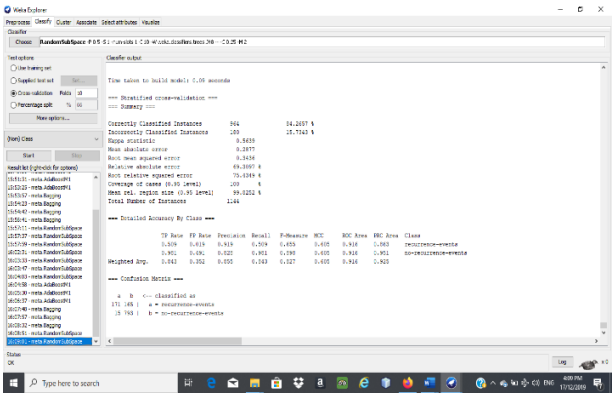

Figure 5: Result of Random with attribute selection and resampling combined (J84)

4. Evaluation of classification result

Bagging, boosting, and random subspace algorithm is tried with many classification algorithms (J48, REP, Random Forest tree, SVM, and Neural Network), and the best performance of each ensemble classifier before and after resampling is obtained. Depending on the final result of the model that is constructed, classification model efficiency is evaluated based on correct/incorrect instances, accuracy regarding correct and incorrect instances generated with a confusion matrix, Precision, Recall, F-Measure, and time taken to build the model. The results of all base and meta classifiers with attribute selection and resampling and all ensemble classification experiment before using the resampling technique and after using it are shown in the figure (6).

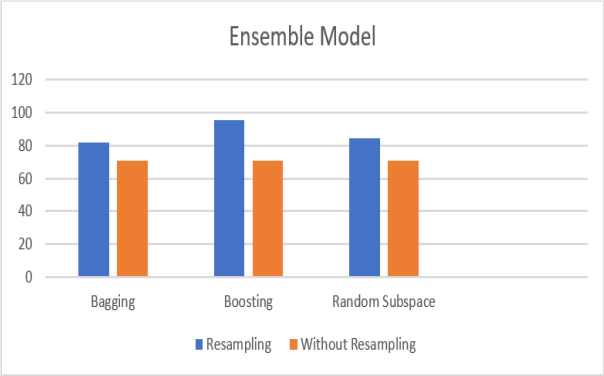

Figure 6: Shows comparative graph of different base classifiers with different evaluation accuracy of ensemble model

4. Conclusion

This research aims to build a classification model to classify the imbalanced breast cancer data available in the IT departments in Sudanese hospitals. This will help us to predict cancer recurrence or no recurrence events. The research concluded that boosting the ensemble learning algorithm with a single misclassified J48 is the best model of classification that can be used in breast cancer data. In this research, the accuracy of classification techniques is evaluated based on a selected single classifier with a combination ensemble Meta algorithm with used three popular Meta-learning algorithms (bagging, boosting, random subspace). Also, the accuracy of classification techniques is evaluated based on the resampling method.

The research also shows the most essential attributes selection for breast cancer survival by using methods: gain ratio and Ranker. AdaBoost Meta Learning combination with single classifier J48 was suggested for the classification of breast cancer-based classification to get the best results with an accuracy of 95.2797 % with a low error rate and performance.

References

[1] M. M. El-Lamey, M. M. Eid, M. Gamal, N. E. M. Bishady, and A. W. Mohamed, "Using machine learning algorithms for breast cancer diagnosis," *International Journal of Applied Metaheuristic Computing*, vol. 12, no. 4, pp. 117–137, 2021, doi: 10.4018/IJAMC.2021100107.

[2] J. P. Choi, T. H. Han, and R. W. Park, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," *Journal of Korean Society of Medical Informatics*, vol. 15, no. 1. pp. 49–57, 2009. [Online]. Available: www.seer.cancer.gov

[3] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif Intell Med*, vol. 34, no. 2, pp. 113–127, Jan. 2005, doi: 10.1016/j.artmed.2004.07.002.

[4] F. Ibrahim and N. A. Osman, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," vol. 15. pp. 520–523, 2007. [Online]. Available: www.springerlink.com

[5] H. Wang, "Breast Cancer Prediction Using Data Mining Method Machine Learning and Data Mining Techniques View project Optimization Techniques View project." 2015. [Online]. Available: https://www.researchgate.net/publication/319688741

[6] J. T. McDonald, N. Herron, W. B. Glisson, and R. K. Benton, "Machine learning-based android malware detection using manifest permissions," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021, vol. 2020-January, pp. 6976–6985. doi: 10.24251/hicss.2021.839.

[7] M. R. Longadge, M. Snehlata, S. Dongre, and D. L. Malik, "Class Imbalance Problem in Data Mining: Review," *International Journal of Computer Science and Network*, vol. 2, no. 1. 2013. [Online]. Available: www.ijcsn.org

[8] A. E. Karrar, "Adopting Graph-Based Machine Learning Algorithms to Classify Android Malware," *IJCSNS International Journal of Computer Science and Network Security*, vol. 22, no. 9, p. 840, 2022, doi: 10.22937/IJCSNS.2022.22.9.109.

[9] A. E. Karrar, "A Novel Approach for Semi Supervised Clustering Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 6, no. 1, pp. 1–7, 2017, [Online]. Available: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01612017.pdf

[10] A. E. Karrar, "A Proposed Model for Improving the Performance of Knowledge Bases in Real-World Applications by Extracting Semantic Information," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, 2022, doi: 10.14569/IJACSA.2022.0130214.

[11] A. Puri and M. Kumar Gupta, "Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data," *Comput J*, Nov. 2021, doi: 10.1093/comjnl/bxab039.

[12] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.

[13] V. Chaurasia and S. Pal, "9 A Novel Related papers A Novel Approach on Ensemble Classifiers wit h Fast Rot at ion Forest Algorit hm Azhagu Sundari A Pragmat ic Approach of Preprocessing t he Dat a Set for Heart Disease Predict ion sivagowry sabanat han, Eugene Bern PERFORMANCE ANALYSIS OF DATA MINING ALGORIT HMS FOR DIAGNOSIS AND PREDICT ION OF HEART … A Novel Approach for Breast Cancer

Detection using Data Mining Techniques," *International Journal of Innovative Research in Computer and Communication Engineering (An ISO*, vol. 3297, no. 1, 2007, [Online]. Available: www.ijircce.com

[14] N. v Chawla, "Data Mining for Imbalanced Datasets: An Overview," *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, pp. 853–867, Jan. 2006. doi: 10.1007/0-387-25465-x_40.

[15] V. Chaurasia and S. Pal, "9 A Novel Related papers A Novel Approach on Ensemble Classifiers wit h Fast Rot at ion Forest Algorit hm Azhagu Sundari A Pragmat ic Approach of Preprocessing t he Dat a Set for Heart Disease Predict ion sivagowry sabanat han, Eugene Bern PERFORMANCE ANALYSIS OF DATA MINING ALGORIT HMS FOR DIAGNOSIS AND PREDICT ION OF HEART … A Novel Approach for Breast Cancer Detection using Data Mining Techniques," *International Journal of Innovative Research in Computer and Communication Engineering (An ISO*, vol. 3297, no. 1, 2007, [Online]. Available: www.ijircce.com

[16] R. Aavula and R. Bhramaramba, "XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability Data mining and machine learning Techniques View project Lung cancer related genes identification View project XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability," *International Journal of Engineering and Advanced Technology (IJEAT)*, no. 5. pp. 2249–8958, 2019. [Online]. Available: https://www.researchgate.net/publication/337077283

[17] A. Elsharif Karrar, "The Use of Case-based Reasoning in a Knowledge-based (Learning) Software Development Organizations," *International Journal of Innovative Research in Science, Engineering and Technology (An ISO*, vol. 3297, no. 5, 2007, doi: 10.15680/IJIRSET.2016.0505331.

[18] T. Chakraborty and A. K. Chakraborty, "Superensemble classifier for improving predictions in imbalanced datasets," *Commun Stat Case Stud Data Anal Appl*, pp. 1–19, Nov. 2020, doi: 10.1080/23737484.2020.1740065.

[19] "A Review on Data Mining Techniques for Treatment of Cancer in Ayurveda Therapy." [Online]. Available: www.ijcset.net

[20] A. E. Karrar, "The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 10, no. 2, Apr. 2022, doi: 10.52549/ijeei.v10i2.3730.

[21] A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130122.

[22] M. Umair *et al.*, "Main path analysis to filter unbiased literature," *Intelligent Automation and Soft Computing*, vol. 32, no. 2, 2022, doi: 10.32604/iasc.2022.018952.

[23] B. Mirzaei, B. Nikpour, and H. Nezamabadi-Pour, "An under-sampling technique for imbalanced data classification based on DBSCAN algorithm," *2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS)*, Nov. 2020, doi: 10.1109/cfis49607.2020.9238718.

[24] M. F. Ijaz, M. Attique, and Y. Son, "Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods," *Sensors*, vol. 20, p. 2809, Nov. 2020, doi: 10.3390/s20102809.

[25] A. M. Morey, F. Noo, and D. J. Kadrmas, "Effect of Using 2 mm Voxels on Observer Performance for PET Lesion Detection," *IEEE Trans Nucl Sci*, vol. 63, pp. 1359–1366, Nov. 2016, doi: 10.1109/tns.2016.2518177.

[26] B. Elhussein, M. Khalifa, A. E. Karrar, and M. M. Alsharani, "A Client-Side App Model for Classifying and Storing Documents," *IJCSNS International Journal of Computer Science and Network Security*, vol. 22, no. 5, p. 225, 2022, doi: 10.22937/IJCSNS.2022.22.5.32.