Resampling and Hyperparameter Tuning for Optimizing Breast Cancer Prediction Using Light Gradient Boosting

Kartika Handayani[1], Erni[1], Rangga Pebrianto[1], Ari Abdilah[1], Rifky Permana[1] and Eni Pudjiarti[2]

[1] *Universitas Bina Sarana Informatika, Jakarta, Indonesia*

[2] *Universitas Nusa Mandiri, Jakarta,Indonesia*

Keywords: Breast Cancer, Light Gradient Boosting, Resampling, Hyperparameter Tuning.

Abstract: Breast cancer is the most frequently diagnosed cancer and the leading cause of death. The main cause of breast cancer is mainly related to patients who inherit genetic mutations in genes. Early diagnosis of breast cancer patients is very important to prevent the rapid development of breast cancer apart from the evolution of preventive procedures. A machine learning (ML) approach can be used for early diagnosis of breast cancer. In this study, testing was performed using the Wisconsin Diagnostic Breast Cancer Dataset, also known as WDBC (Diagnostics) which consists of 569 instances with no missing values and has one target class attribute, either benign (B) or malignant (M). Tests were carried out using the ROS, RUS, SMOTE, and SMOTE-Tomek resampling techniques to see the effect of overcoming unbalanced data. Then tested with Light Gradient Boosting and optimized to get the best results using hyperparameter tuning. The best results are obtained after tuning the hyperparameter with accuracy 99.12%, recall 99.12%, precisions 99.13%, f1-score 99.13% and AUC 0.988.

1 INTRODUCTION

Breast cancer is the most frequently diagnosed cancer and the leading cause of death from cancer with a percentage of 11.6%. In 2018 there were around 2.1 million cases of breast cancer in the world. It is the most frequently diagnosed cancer in most countries (154 out of 185) and is also the leading cause of death from cancer in more than 100 countries (Bray et al., 2018). The main cause of breast cancer is mainly related to patients inheriting genetic mutations in their genes (Majeed, 2014). Breast cancer has two different classes, namely benign and malignant. Benign tumors, usually known as non-cancerous, while malignant tumors or known as cancer can damage the surrounding tissue and if the patient is diagnosed as malignant, the doctor will perform a biopsy to determine the aggressiveness or severity of the tumor(Khan et al., 2017).

Early diagnosis of breast cancer patients is very important to prevent the rapid development of breast cancer apart from the evolution of preventive procedures(Sun, 2017) and aids in the speedy recovery and reduces the chance of death (Badr et al., 2019). Diagnosis of breast cancer can be done manually by a doctor, but it takes a longer time and must be very complicated for doctors to apply this classification (Khuriwal and Mishra, 2018). The incompleteness of relevant data can also lead to human error in diagnosis (Zaitseva et al., 2020). Although advances in the treatment of breast cancer have led to a reduction in mortality from breast cancer across all age groups, young age remains a high-risk factor and has a low survival rate (Lee and Han, 2014). Due to the severity of the disease, a computer-assisted detection (CAD) system using a machine learning (ML) approach is required for the diagnosis of breast cancer(Omondiagbe et al., 2019). Therefore improving the accuracy of identifying breast cancer is a very important task (Badr et al., 2020). Various methods can be applied to the classification of breast cancer (Idris and Ismail, 2021) to distinguish between two types of breast tumors namely Benign and Malignant (Khan et al., 2017).

Classification is a data mining process that aims to divide data into classes to facilitate decision-making because that is an important task in the medical field. Many researchers have done research to predict breast cancer.Research done by proposing Support Vector Machine (SVM) on the WBCD dataset produces an accuracy of 86.10% (Kumari, 2018). Research that

Handayani, K., Erni, E., Pebrianto, R., Abdilah, A., Permana, R. and Pudjiarti, E.
Resampling and Hyperparameter Tuning for Optimizing Breast Cancer Prediction Using Light Gradient Boosting.
DOI: 10.5220/0012446100003848

In *Proceedings of the 3rd International Conference on Advanced Information Scientific Development (ICAISD 2023)*, pages 177-180
ISBN: 978-989-758-678-1

proposes Random Forest using the WBCD dataset obtains 91.66% accuracy (Pyingkodi et al., 2020). Research with K- Nearest Neighbors(KNN) using the WBCD dataset obtained an accuracy of 92.57%.

Based on the research that has been done, this study proposes a comparison of sampling methods such as Random Under Sampling (RUS), Random Over Sampling (ROS), and Synthetic Minority Over Sampling Technique (SMOTE) (Ernawan et al., 2022) to see the effect on the problem of unbalanced data. Then do the optimization of the best model with hyperpa- rameter tuning. The purpose of this study is to get the best predictive results from the proposed classification model. This research is expected to produce better accuracy, recall, precision, f-score, and AUC than previous research and add to research contributions related to the data used.

2 MATERIALS AND METHODS

2.1 Datasets

The Wisconsin Diagnostic Breast Cancer Dataset, also known as WDBC (Diagnostics). The source of this dataset is the University of Wisconsin. This data set consists of 569 instances with no missing values and has one target class attribute, either benign (B) or malignant (M). The predictive attributes of the data set consisted of ten real-valued features calculated for each core, such as radius, texture, circumference, area, smoothness, compactness, concavity, symmetry, and fractal dimension. Mean, standard error, and radius (average of the three largest value readings) were calculated for each lead leading data set that has more than 32 attributes, including the non-predictive attribute i.e. patient ID.

2.2 Preprocessing

Data preparation stages are carried out to ensure that the dataset used for training and testing is quality data. If the dataset used still has noise, the resulting model will also be of low quality and will have bias.

2.2.1 Data Cleaning

Data pre-processing begins with data cleaning which consists of deleting the ”id” column in the original dataset (WBCD) and diagnostic dataset (WDBC). Then change the feature class name (as the target class) to a feature with the name ”diagnosis” for the three datasets, this is done because if you don’t change the feature class name it will be read as a function.

2.2.2 Label Encodig

Data pre-processing begins with data cleaning which consists of deleting the ”id” column in the original dataset (WBCD) and diagnostic dataset (WDBC). Then change the feature class name (as the target class) to a feature with the name ”diagnosis” for the three datasets, this is done because if you don’t change the feature class name it will be read as a function.

2.2.3 Split Data

Split data: 80% data training, 20% data testing, training data is used to build and train the model, and data testing will be used for model evaluation and model testing.

2.2.4 Resampling

The WDBC dataset with the target class ”benign” has a total of 357 data and the target class ”malignant” has a total of 212 data with a total of 569 instances. The following is a visualization of the data distribution of the WDBC dataset target class.

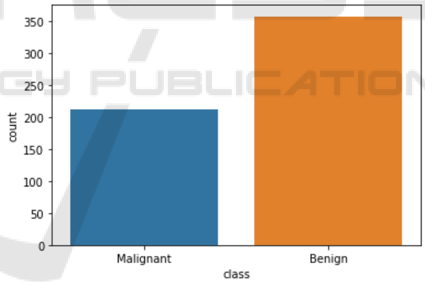

Figure 1: Class Distribution.

For the Random Under Sampling (RUS), most of the category examples are discarded until an even distribution of information is achieved (Dorn, 2021). The Random Over Sampling (ROS) algorithm randomly replicates samples from the minority class(Rendón et al., 2020). Oversampling can be done by increasing the number of instances or minority class samples based on the production of new samples or repeated samples(Mohammed et al., 2020). The Synthetic Minority Over Sampling Technique SMOTE generates an artificial sample of the minority class by interpolating existing instances that are very close to each other (Rendón et al., 2020). For the minority category in the information set, SMOTE initially selects the minority class data instance at random. Then, Ma’s k nearest neighbors are related to the minority class, ID (Dorn, 2021). SMOTE with Tomek Link balances knowledge and instant builds of well-separated additional categories. during this approach, any data instance that generates a link that Tomek discards are of either the minority or the majority class (Dorn, 2021).

2.3 Model

2.3.1 Light Gradient Boosting

LightGBM uses gradient enhancement in its construction, but light GBM does not split the eigenvalues individually, so it is necessary to calculate the splitting benefit of each eigenvalue. LightGBM algorithm on the model to improve forecasting accuracy and robustness (Ju et al., 2019). Can find the optimal split value (Su, 2020).

2.3.2 Hyperparameter Tuning

Hyperparameter tuning can be complete when the data is large (Joy et al., 2016), showing more hyperparameters to tune, and models with carefully designed structures imply that hyperparameters must be set too tight ranges to reproduce the precision (Yu and Zhu, 2020).

2.3.3 GridSearch

To increase the optimal hyperparameter, the GridSearchCV model taken from Scikit learn is used [22]to get the best parameters. GridSearchCV implements the fit and score methods. It also implements score_samples, predict, predict_proba, decision_function, transform and inverse_transform if implemented in the estimator used (Pedregosa, 2011). The parameter estimator used to implement this method is optimized by cross-validation through grid parameters.

3 RESULT AND DISCUSSION

The following tests are carried out using light gradient boosting without using hyperparameter tuning.

Table 1: Testing Without Hyperparameter Tuning.

Resampling	Accuracy	Recall	Precision	F1-Score	AUC
Without Re-sampling	95.61%	90.48%	97.44%	93.83%	0.945
ROS	95.61%	90.48%	97.44%	93.83%	0.945
SMOTE	96.49%	92.86%	97.50%	95.12%	0.957
RUS	97.37%	95.24%	97.56%	96.36%	0.969
SMOTE Tomek	94.73%	88.09%	97.36%	92.50%	0.934

The best results before testing without hyperparameter tuning were obtained by resampling RUS with accuracy results 97.37%, recall 95.24%, precision 97.56%,f1-score 96.36%a nd AUC 0.934. Then a search for hyperparameter tuning is performed with a grid search based on the following table 2.

Table 2: Parameters Tested in the Hyperparameter tuning Grid Search.

n_estimators	100, 400, 10
min_child_weight	3, 20, 2
colsample_bytree	0.4, 1.0
max_deph	5, 15, 2
num_leaves	8, 40
min_child_weight	10.30
learning_rate	0.01,1

After searching for hyperparameter tuning, the best parameters are obtained as follows

Table 3: Best Parameters.

’boosting_type’	’gbdt’,
’class_weight’	none,
’colsample_bytree’	1.0,
’importance_type’	’split’,
’learning_rate’	0.1,
’max_depth’	15,
’min_child_samples’	10,
’min_child_weight’	20,
’min_split_gain’	0.0,
’n_estimators’	400,
’n_jobs’	-1,
’num_leaves’	31,
’objective’	none,
’random_state’	none,
’reg_alpha’	0.0,
’reg_lambda’	0.0,
’silent’	true,
’subsamples’	1.0,
’subsample_for_bin’	200000,
’subsample_freq’	0

After the best parameter results were obtained, the test was carried out again with the results in the following table

Table 4: Testing With Hyperparameter Tuning.

Resampling	accuracy	recall	Precision	F1-Score	AUC
Without Re-sampling	99.12%	99.12%	99.13%	99.13%	0.988
ROS	99.12%	99.12%	99.13%	99.13%	0.988
SMOTE	99.12%	99.12%	99.13%	99.13%	0.988
RUS	98.24%	98.24%	98.24%	98.24%	0.981
SMOTE Tomek	96.49%	96.49%	96.68%	96.68%	0.952

The results using hyperparameter tuning have increased in all tests. With the best results with the same value on the test without resampling, ROS, and SMOTE. The best results are obtained with the accuracy 99.12%, recall 99.12%, precisions 99.13%, f1-score 99.13% and AUC 0.988.

4 CONCLUSIONS

This study conducted a test for breast cancer prediction. The use of resampling techniques such as ROS, SMOTE, RUS, and SMOTE-Tomek is done to overcome unbalanced data. The use of hyperparameter tuning using a grid search with light gradient boosting results in an increase and optimization of results. Obtain the best results with accuracy 99.12%, recall 99.12%, precisions 99.13%, f1-score 99.13% and AUC 0.988. In further research, testing can be done with other breast cancer datasets or with other methods.

REFERENCES

Badr, E., Abdulsalam, M., and Ahmed, H. (2020). The impact of scaling on support vector machine in breast cancer diagnosis. *Int. J. Comput. Appl*, 175(19):15–19,.

Badr, E., Salam, M., and Ahmed, H. (2019). Optimizing support vector machine using gray wolf optimizer algorithm for breast cancer detection.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin*, 68(6):394–424,.

Dorn, M. (2021). Comparison of machine learning techniques to handle imbalanced covid-19 cbc datasets. *PeerJ Comput. Sci*, 7:670,.

Ernawan, F., Handayani, K., Fakhreldin, M., and Abbker, Y. (2022). Light gradient boosting with hyper parameter tuning optimization for covid-19 prediction. *Int. J. Adv. Comput. Sci. Appl*, 13(8):514–523,.

Idris, N. and Ismail, M. (2021). Breast cancer disease classification using fuzzy-id3 algorithm with fuzzydbd method: Automatic fuzzy database definition. *PeerJ Comput. Sci*, 7:1–22,.

Joy, T., Rana, S., Gupta, S., and Venkatesh, S. (2016). Hyperparameter tuning for big data using bayesian optimisation. *Proc. - Int. Conf. Pattern Recognit*, 0:2574–2579,.

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H., and Rehman, M. (2019). A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*, 7(c):28309–28318,.

Khan, R., Suleman, T., Farooq, M., Rafiq, M., and Tariq, M. (2017). Data mining algorithms for classification of diagnostic cancer using genetic optimization algorithms. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, 12(March):207–211,.

Khuriwal, N. and Mishra, N. (2018). Breast cancer diagnosis using deep learning algorithm. In *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, pages 98–103,.

Kumari, M. (2018). The impact of news information on the stock recommendation system: A survey.

Lee, H. and Han, W. (2014). Unique features of young age breast cancer and its management. *J. Breast Cancer*, 17(4):301–307,.

Majeed, W. (2014). Breast cancer: Major risk factors and recent developments in treatment. *Asian Pacific J. Cancer Prev*, 15(8):3353–3358,.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, (May):243–248,.

Omondiagbe, D., Veeramani, S., and Sidhu, A. (2019). Machine learning classification techniques for breast cancer diagnosis. *IOP Conf. Ser. Mater. Sci. Eng*, 495(1).

Pedregosa, F. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res*, 12:2825–2830,.

Pyingkodi, M., M., M., Shanthi, S., Saravanan, T., Thenmozhi, K., Nanthini, K., Hemalatha, D., and Dhivya (2020). Performance study of classification algorithms using the breast cancer dataset. *Int. J. Futur. Gener. Commun. Netw*.

Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F., and Granda-Gutiérrez, E. (2020). Data sampling methods to dealwith the big data multi-class imbalance problem. *Appl. Sci*, 10(4).

Su, Y. (2020). Prediction of air quality based on gradient boosting machine method. In *Proc. - 2020 Int. Conf. Big Data Informatiz. Educ. ICBDIE 2020*, pages 395–397,.

Sun, Y. (2017). Risk factors and preventions of breast cancer. *Int. J. Biol. Sci*, 13(11):1387–1397,.

Yu, T. and Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications.

Zaitseva, E., Levashenko, V., Rabcan, J., and Krsak, E. (2020). Application of the structure function in the evaluation of the human factor in healthcare. *Symmetry*, 12(1):93,.