



## Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara

\*Nur Ghaniaviyanto Ramadhan<sup>1</sup>, Faisal Dharma Adhinata<sup>2</sup>

<sup>1,2</sup>Fakultas Informatika, Program Studi Rekayasa Perangkat Lunak, Institut Teknologi Telkom Purwokerto, Indonesia

<sup>1</sup>ghani@ittelkom-pwt.ac.id, <sup>2</sup>faisal@ittelkom-pwt.ac.id

### Abstrak: Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara.

Kanker payudara merupakan keganasan pada jaringan payudara yang dapat berasal dari epitel duktus maupun lobulusnya. WHO menyebutkan 30 % - 50 % kasus kanker dapat dicegah. Pencegahan kanker payudara dapat dilakukan dengan cara skrining atau diagnosis awal. Tujuan dari diagnosis awal tersebut agar jika muncul benjolan dapat dilakukan prediksi apakah tergolong kategori ganas atau jinak. Prediksi kanker payudara dapat dilakukan dengan menggunakan kumpulan dataset yang berisi parameter terkait kanker. Akan tetapi terkadang dataset yang digunakan juga memiliki permasalahan seperti jumlah data tidak seimbang dan penggunaan fitur yang tidak relevan. Pada penelitian ini bertujuan untuk meningkatkan hasil prediksi kanker payudara dengan menyeimbangkan jumlah kelas data dan menggunakan rank fitur. Metode yang digunakan yaitu SMOTE untuk data tidak seimbang dan gini score untuk peringkat fitur. Model klasifikasi yang digunakan yaitu random forest dan naïve bayes. Hasil yang didapatkan model klasifikasi random forest lebih unggul dibandingkan naïve bayes.

**Kata kunci:** Kanker Payudara; Imbalanced Data; Feature Selection; Random Forest, Naïve Bayes

### Abstract: SMOTE Technique and Gini Score in Breast Cancer Classification.

Breast cancer is a malignancy in breast tissue that can originate from the epithelium of the ducts and lobules. WHO says 30% - 50% of cancer cases can be prevented. Breast cancer prevention can be done utilizing screening or early diagnosis. The purpose of the initial diagnosis is that if a lump appears, predictions can be made whether it is classified as malignant or benign. Breast cancer prediction can be done using a dataset containing cancer-related parameters. However, sometimes the dataset used also has problems such as the amount of data is not balanced and the use of irrelevant features. This study aims to improve breast cancer prediction results by balancing the number of data classes and using the rank feature. The method used is SMOTE for imbalanced data and Gini score for rank features. The classification model used is random forest and naïve Bayes. The results obtained by the random forest classification model are superior to Naïve Bayes.

**Keywords:** Breast Cancer; Imbalanced Data; Feature Selection; Random Forest, Naïve Bayes

---

#### History & License of Article Publication:

Received: 17/10/2021 Revision: 13/11/2021 Published: 18/12/2021

---

DOI: <https://doi.org/10.37971/radial.v9i2.229>

---



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

---

## PENDAHULUAN

Kanker payudara adalah keganasan pada jaringan payudara yang dapat berasal dari *epitel duktus* dan *lobulus* (Indonesia, 2021). Faktor risiko yang berkaitan erat dengan peningkatan kejadian kanker payudara antara lain jenis kelamin perempuan, usia > 50 tahun, riwayat keluarga dan genetika (pembawa mutasi gen BRCA1, BRCA2, ATM atau TP53 (p53)), riwayat penyakit payudara sebelumnya (DCIS pada payudara serupa, LCIS, kepadatan tinggi pada mamografi), riwayat menstruasi dini (<12 tahun) atau menstruasi lanjut (> 55 tahun), riwayat reproduksi (tidak ada anak dan tidak menyusui), hormon, obesitas, konsumsi alkohol, riwayat radiasi dinding dada, dan faktor lingkungan (Indonesia, 2021). Pencegahan kanker payudara dapat dilakukan dengan salah satu cara yaitu skrining atau diagnosa dini. Menurut WHO, antara 30% - 50% kasus kanker dapat dicegah (Organization, 2019). Diagnosis dini kanker payudara seringkali berupa keluhan adanya benjolan di area payudara dan benjolan di area ketiak (Indonesia, 2021). Tujuan diagnosis awal adalah jika benjolan muncul, dapat diprediksi apakah benjolan itu tergolong ganas atau jinak.

Prediksi kanker payudara dapat dilakukan dengan menggunakan dataset yang berisi parameter terkait kanker. Namun terkadang dataset yang digunakan juga mengalami kendala seperti jumlah data yang tidak seimbang dan penggunaan fitur yang tidak relevan. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan hasil prediksi kanker payudara dengan cara menyeimbangkan jumlah kelas data dan menggunakan fitur rank.

Dalam studi sebelumnya, ada banyak prediksi kanker payudara menggunakan dataset terbuka dari repositori UCI, seperti pada pembahasan makalah-makalah referensi penelitian ini. Makalah (Mishra, Arnab Kumar, Pinki Roy, 2021) menyajikan pendekatan pemilihan fitur berdasarkan *Particle Swarm Optimization* (BPSO) yang dapat digunakan untuk meningkatkan kinerja CDSS dalam prediksi kanker payudara otomatis. Makalah (Jabbar, 2021) membahas pembangunan sistem pendukung keputusan menggunakan model ensemble yang dibangun dengan *Bayesian Tissue* dan *Radial Basis Functions* untuk masalah kanker payudara. Makalah (Grover, Ankit, Nitesh Pradhan, 2021) menyajikan studi perbandingan untuk memprediksi kanker payudara pada dataset Kanker Payudara *Wisconsin Diagnostic* dengan menerapkan enam algoritma pembelajaran mesin yang berbeda seperti CART, *Logistic Regression*, *Support Vector Classifier*, *Hard Voting Classifier*, *Extreme Gradient Boosting*, dan *Artificial Neural Network*.

Penelitian (Dubey, Ashutosh Kumar, Umesh Gupta, 2016) bertujuan untuk mengetahui pengaruh algoritma pengelompokan *k-means* dengan ukuran komputasi yang berbeda seperti sentroid, jarak, metode split, epoch, atribut, dan iterasi serta untuk mempertimbangkan dan mengidentifikasi kombinasi pengukuran yang berpotensi memiliki akurasi sangat tinggi. akurasi pengelompokan. Dalam artikel (Chaurasia, Vikas, 2021) mengembangkan teknik ensemble berbasis tumpukan dan metode pemilihan fitur untuk kinerja algoritma yang komprehensif dan analisis komparatif dari kumpulan data kanker payudara dengan atribut yang dikurangi dan semua atribut. Penelitian (Laghmati, Sara, 2020) menerapkan dan membandingkan empat algoritma pembelajaran mesin (KNN,

pohon keputusan, SVM Biner, dan *Adaboost*) untuk memprediksi apakah pasien menderita tumor payudara ganas atau jinak.

Makalah (Kumar, Vivek, 2020) berfokus pada penerapan teknik klasifikasi yang berbeda untuk data mining dalam memprediksi kanker payudara ganas dan jinak. Dataset *Breast Cancer Wisconsin* dari repositori UCI telah digunakan sebagai dataset eksperimental sedangkan ketebalan rumpun atribut digunakan sebagai kelas evaluasi. Studi (Mushtaq, Zohaib, 2020) melibatkan eksplorasi kinerja KNN menggunakan berbagai fungsi jarak dan nilai K untuk menemukan KNN yang efektif. Dataset *Wisconsin Breast Cancer* (WBC) dan *Wisconsin Breast Cancer Diagnostic* (WDBC) dari repositori pembelajaran mesin UC Irvine digunakan sebagai sumber data utama. Sebuah studi (Prastyo, Pulung Hendro, 2020) membandingkan delapan algoritma pembelajaran mesin: *Gaussian Naïve Bayes* (GNB), *K-Nearest Neighbours* (K-NN), *Support Vector Machine* (SVM), *Random Forest* (RF), *AdaBoost*, *Gradient Boosting* (GB), *XGBoost*, dan *Multi-Layer Perceptron* (MLP). Eksperimen dilakukan dengan menggunakan dataset *Breast Cancer Wisconsin*. Makalah (Mohammed, Siham A., 2020) mengusulkan pendekatan yang meningkatkan akurasi dan kinerja tiga pengklasifikasi yang berbeda: *Decision Tree* (J48), *Naïve Bayes* (NB), dan *Sequential Minimal Optimization* (SMO). Masalah yang diangkat di penelitian (Mohammed, Siham A., 2020) adalah data yang tidak seimbang.

Tujuan dari penelitian (Ojha, 2017) adalah untuk mengetahui seberapa tepat algoritma data mining ini dapat memprediksi kemungkinan kekambuhan penyakit di antara pasien berdasarkan parameter yang dianggap penting. Makalah (Silva, Jesús, 2019) membahas integrasi teknik klasifikasi data mining dan pembelajaran *ensemble* untuk memprediksi jenis kekambuhan kanker payudara. Makalah (Sharma, Alpna, 2021) telah melakukan upaya untuk mendeteksi kanker payudara menggunakan jaringan saraf dalam menggunakan dataset kanker payudara *Wisconsin*. Studi (Nguyen, Cuong, Yong Wang, 2013) bertujuan untuk mendiagnosis dan prognostik kanker payudara menggunakan metode pembelajaran mesin berdasarkan pengklasifikasi hutan acak dan teknik pemilihan fitur.

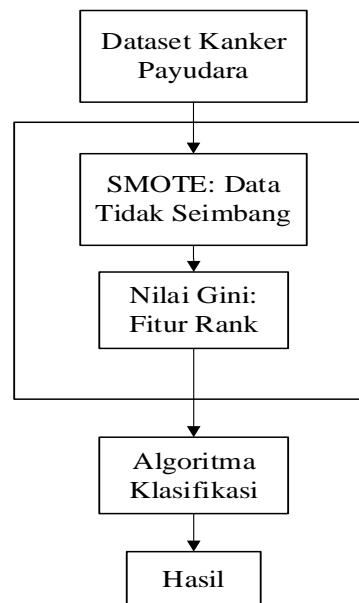
Pada penelitian ini bertujuan untuk klasifikasi dataset kanker payudara dengan menyelesaikan permasalahan pada proses *pre-processing* (data tidak seimbang dan pemilihan fitur). Tabel 1 merupakan perbandingan kontribusi penelitian ini dengan beberapa penelitian yang telah dilakukan.

**Tabel 1.** Perbandingan Kontribusi

Penulis	Data Tidak Seimbang	Pemilihan Fitur
(Mishra, Arnab Kumar, Pinki Roy, 2021), (Chaurasia, Vikas, 2021), (Ojha, 2017), (Nguyen, Cuong, Yong Wang, 2013)	Tidak	Iya
(Mohammed, Siham A., 2020)	Iya	Tidak
Penelitian ini	Iya	Iya

## METODE

Penelitian ini akan menggunakan metode *SMOTE oversampling* untuk masalah data tidak seimbang dan skor *Gini* untuk fitur peringkat. Pada gambar 1 adalah diagram alir yang digunakan pada penelitian ini.



Gambar 1. Skema Usulan

## Dataset

Penelitian ini menggunakan dataset terbuka dari UCI *Repostiroy* yaitu *Breast Cancer Wisconsin Diagnostic* (BCWD) (Indrawati, 2021). Dataset tersebut berisi 569 baris data, 31 fitur yang berkaitan dengan kanker payudara, dan dua jenis golongan, yaitu ganas (*malignant*) dan jinak (*benign*).

Tabel 2. Deskripsi Dataset

No	Fitur	Tipe
1	<i>radius_mean</i>	Float
2	<i>texture_mean</i>	Float
3	<i>perimeter_mean</i>	Float
4	<i>area_mean</i>	Float
5	<i>smoothness_mean</i>	Float
6	<i>compactness_mean</i>	Float
7	<i>concavity_mean</i>	Float
8	<i>concave points_mean</i>	Float
9	<i>symmetry_mean</i>	Float
10	<i>fractal_dimension_mean</i>	Float
11	<i>radius_se</i>	Float
12	<i>texture_se</i>	Float
13	<i>perimeter_se</i>	Float
14	<i>area_se</i>	Float

Teknik *SMOTE* dan *Gini Score* dalam Klasifikasi Kanker Payudara (Ramadhan)

<https://stitek-binataruna.e-journal.id/radial/index>

No	Fitur	Tipe
15	<i>smoothness_se</i>	Float
16	<i>compactness_se</i>	Float
17	<i>concavity_se</i>	Float
18	<i>concave points_se</i>	Float
19	<i>symmetry_se</i>	Float
20	<i>fractal_dimension_se</i>	Float
21	<i>radius_worst</i>	Float
22	<i>texture_worst</i>	Float
23	<i>perimeter_worst</i>	Float
24	<i>area_worst</i>	Float
25	<i>smoothness_worst</i>	Float
26	<i>compactness_worst</i>	Float
27	<i>concavity_worst</i>	Float
28	<i>concave points_worst</i>	Float
29	<i>symmetry_worst</i>	Float
30	<i>fractal_dimension_worst</i>	Float
31	<i>diagnosis</i>	Binary

Sumber: Kaggle.com

### **SMOTE**

*Synthetic Minority Oversampling Technique (SMOTE)* merupakan algoritma preprocessing yang sering digunakan untuk masalah ketidakseimbangan data (Fernández, 2018). Dataset yang digunakan tidak seimbang di kelasnya dengan 212 baris ganas dan 357 baris jinak.

**Tabel 3. SMOTE**

Sebelum Penerapan SMOTE	Setelah Penerapan SMOTE
Ganas: 212 baris	Ganas: 312 baris
Jinak: 357 baris	Jinak: 357 baris

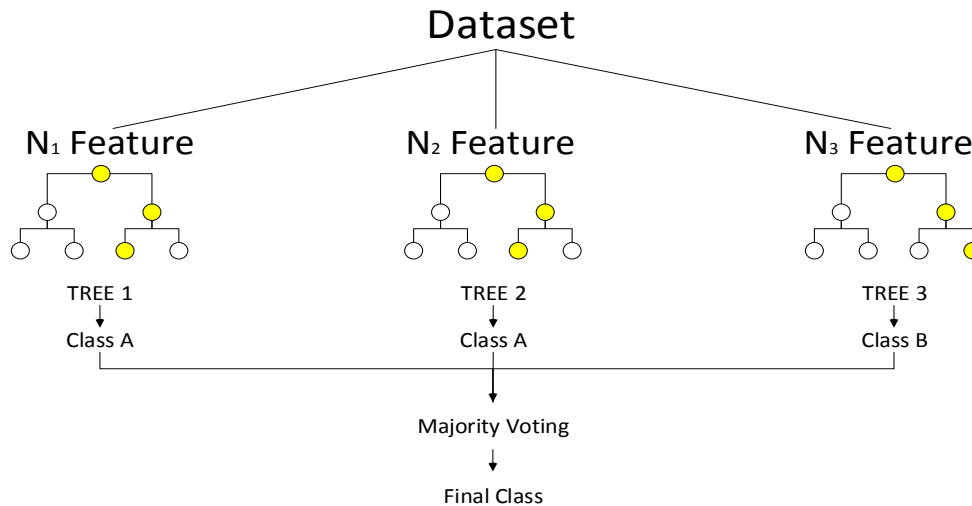
Tabel 3 menunjukkan hasil penerapan teknik *SMOTE* yang berhasil meningkatkan jumlah kelas ganas yang sebelumnya memiliki persentase yang sangat jauh dibandingkan dengan kelas jinak.

### **Gini Score**

Proses ini mencari nilai *Gini* ini untuk mendapatkan peringkat fitur. Pemeringkatan fitur merupakan suatu teknik dalam pemilihan fitur yang mencari nilai hubungan antara fitur dan kelas (Khalid, 2014). Hasil yang diperoleh peringkat fitur setelah menerapkan algoritma indeks Gini adalah fitur 7, 27, 22, 6, 20, 23, 26, 2, 13, 3, 10, 0, 1, 25, 12, 21, 24, 29, 28, 5, 11, 19, 17, 8, 18, 4, 9, 15, 14, 16 dengan mengacu pada urutan tabel fitur sebelumnya yaitu Tabel 2.

### **Random Forest**

*Random forest* adalah metode pembelajaran *ensemble* untuk klasifikasi atau regresi yang beroperasi dengan membangun banyak pohon keputusan selama proses pelatihan dan memberikan output berupa mode kelas (klasifikasi) atau prediksi rata-rata (regresi) pohon individu (Saru, 2019).



Gambar 2. Proses Random Forest  
Sumber: (Saru, 2019)

### Naïve Bayes

*Naïve bayes* merupakan metode *machine learning* yang menggunakan pendekatan *bayes* (Yulita, 2021). Penggunaan teorema *bayes* dapat menggunakan formula (1).

$$P(H|X) = \frac{P(H) \cdot P(X|H)}{P(X)} \quad (1)$$

### HASIL DAN PEMBAHASAN

Proses selanjutnya akan dilakukan dengan melakukan eksperimen dengan menerapkan skema yang diusulkan dan asli yang kemudian diklasifikasikan menggunakan model pembelajaran mesin *random forest* untuk melihat hasil presisi dan *recall* yang diperoleh di masing-masing kelas. Rumus presisi dan *recall* dapat dilihat pada rumus (2) dan (3) (Ramadhan, 2019).

$$\text{Recall/Sensitivity} = \left( \frac{TP}{TP + FN} \right) \quad (2)$$

$$\text{Precision/PPV} = \left( \frac{TP}{TP + FP} \right) \quad (3)$$

Tabel 4. Confusion Matrix

	<i>Actual</i>	
<i>Predicted</i>	TP	FP
	FN	TN

Sumber: (Ramadhan, 2019)

Tabel 4 merupakan bentuk matriks konfusi yang memiliki arti *True Positive* (TP) yaitu prediksi kanker dan aslinya kanker. *True Negative* (TN) adalah prediksi non-kanker dan sebenarnya bukan kanker. *False Positive* (FP) adalah prediksi kanker dan bukan kanker. *False Negative* (FN) adalah prediksi non-kanker dan aslinya merupakan kanker.

**Tabel 5.** *Confusion Matrix Original*

	<i>Actual</i>	
<i>Predicted</i>	104	4
	2	61

**Tabel 6.** Hasil *Original*

<i>Class</i>	<i>Precision</i>	<i>Recall</i>
Ganas	94 %	97 %
Jinak	98 %	96 %

Tabel 5 adalah nilai matriks konfusi yang diperoleh. Tabel 6 adalah hasil yang diperoleh sebelum metode yang diusulkan diterapkan.

**Tabel 7.** *Confusion Matrix Skema Usulan*

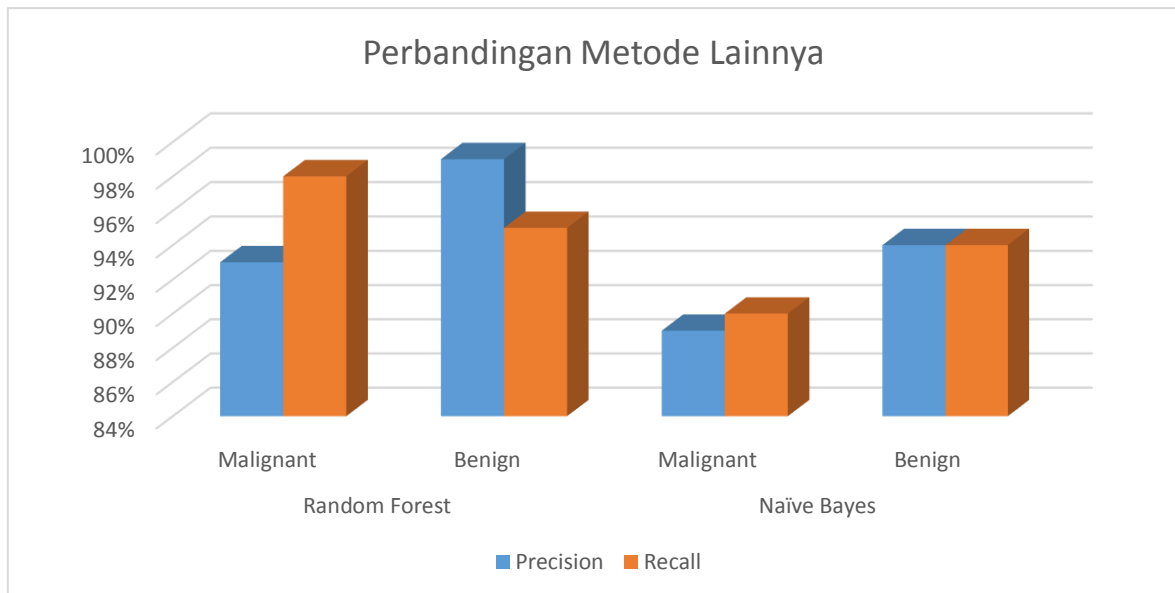
	<i>Actual</i>	
<i>Predicted</i>	103	5
	1	62

**Tabel 8.** Hasil Skema Usulan

<i>Class</i>	<i>Precision</i>	<i>Recall</i>
Ganas	93 %	98 %
Jinak	99 %	95 %

Tabel 7 adalah nilai matriks konfusi yang diperoleh setelah model yang diusulkan diterapkan. Tabel 8 menunjukkan hasil yang diperoleh setelah metode yang diusulkan diterapkan. Pada eksperimen tabel 5 sampai tabel 8 menggunakan model klasifikasi random forest.

Hasil yang diperoleh menunjukkan bahwa metode yang diusulkan dalam penelitian ini mampu menurunkan nilai FN dan meningkatkan nilai TN, namun metode yang diusulkan masih belum mampu meningkatkan nilai TP dan menurunkan nilai FP. Hal ini terjadi karena kurangnya pengujian sampel yang digunakan.



Gambar 3. Hasil Perbandingan Metode Lainnya

Gambar 3 merupakan perbandingan menggunakan model klasifikasi lain yaitu *Naïve Bayes*. Hasil yang diperoleh dengan model *random forest* dengan menerapkan metode yang diusulkan lebih unggul dari kelas ganas (*malignant*) dan jinak (*benign*). Berdasarkan gambar 3 membuktikan bahwa metode usulan penelitian ini mampu menghasilkan presisi yang sangat tinggi sebesar 93% dan 99%, sedangkan untuk recall sebesar 98% dan 95% dibandingkan tanpa menggunakan metode usulan penelitian. Metode klasifikasi *random forest* mampu menghasilkan *gap* presisi sebesar 6%, recall sebesar 5% dibandingkan metode klasifikasi *Naïve Bayes*.

Hasil lain menunjukkan juga bahwa penerapan *pre-processing* menggunakan teknik SMOTE mampu menghasilkan kenaikan nilai presisi kelas jinak (*benign*) dan recall kelas ganas (*malignant*) sebesar 1%, sedangkan pada nilai presisi kelas ganas (*malignant*) dan recall kelas jinak (*benign*) mengalami penurunan sebesar 1%. Hal tersebut menunjukkan ada beberapa nilai yang teknik SMOTE ini tidak dapat melakukan penyeimbangan kelas.

## KESIMPULAN

Penelitian ini bertujuan untuk klasifikasi dataset kanker payudara dengan menyelesaikan permasalahan pada proses *pre-processing* (data tidak seimbang dan pemilihan fitur). Sehingga, berdasarkan hasil yang didapatkan kesimpulan dalam penelitian ini adalah teknik *SMOTE* dan *Gini* dapat meningkatkan ketelitian hasil pada kelas jinak (*benign*) dan juga dapat meningkatkan hasil recall pada kelas ganas (*maligna*). Untuk metode klasifikasi yang dihasilkan yaitu *random forest* lebih unggul daripada *Naïve Bayes*. Untuk saran kerja kedepannya dapat dianalisa terkait masalah *missing value* pada dataset penelitian ini.

## DAFTAR PUSTAKA

- Kementerian Kesehatan Republik Indonesia, (2021). "Panduan Penatalaksanaan Kanker Payudara". [Online]. Available: <http://kanker.kemkes.go.id/guidelines/PPKPayudara.pdf>. [Accessed: Aug, 10, 2021]. [In Indonesian].
- World Health Organization, (2019). "Preventing Cancer", [Online]. Available: <https://www.who.int/activities/preventing-cancer>. [Accessed: Aug, 10, 2021].



- Mishra, Arnab Kumar, Pinki Roy, and Sivaji Bandyopadhyay. (2021). "Binary Particle Swarm Optimization Based Feature Selection (BPSO-FS) for Improving Breast Cancer Prediction." Proceedings of International Conference on Artificial Intelligence and Applications. Springer, Singapore.
- Jabbar, Meerja Akhil. (2021). "Breast Cancer Data Classification Using Ensemble Machine Learning." Engineering and Applied Science Research 48.1. 65-72.
- Grover, Ankit, Nitesh Pradhan, and Prashant Hemrajani. (2021). "Comparison-Based Study to Predict Breast Cancer: A Survey." Innovations in Computational Intelligence and Computer Vision. Springer, Singapore. 543-550. [https://doi.org/10.1007/978-981-15-6067-5\\_61](https://doi.org/10.1007/978-981-15-6067-5_61)
- Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. (2016). "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset." International journal of computer assisted radiology and surgery 11.11. 2033-2047. <https://doi.org/10.1007/s11548-016-1437-9>
- Chaurasia, Vikas, and Saurabh Pal. (2021). "Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer." SN Computer Science 2.2. 1-13. <https://doi.org/10.1007/s42979-021-00465-3>
- Laghmati, Sara, et al. (2020). "Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques." 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet). IEEE. 10.1109/CommNet49926.2020.9199633
- Kumar, Vivek, et al. (2020). "Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications." Advances in data science and management. Springer, Singapore. 435-442. [https://doi.org/10.1007/978-981-15-0978-0\\_43](https://doi.org/10.1007/978-981-15-0978-0_43)
- Mushtaq, Zohaib, et al. (2020). "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets." Journal of the Chinese Institute of Engineers 43.1. 80-92. <https://doi.org/10.1080/02533839.2019.1676658>
- Prastyo, Pulung Hendro, et al. (2020). "Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms." Proceeding International Conference on Science and Engineering. Vol. 3. <https://doi.org/10.14421/icse.v3.545>
- Mohammed, Siham A., et al. (2020). "Analysis of breast cancer detection using different machine learning techniques." International Conference on Data Mining and Big Data. Springer, Singapore. 10.1109/CCWC51732.2021.9376007
- Ojha, Uma, and Savita Goel. (2017). "A study on prediction of breast cancer recurrence using data mining techniques." 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE. 10.1109/CONFLUENCE.2017.7943207
- Silva, Jesús, et al. (2019). "Integration of data mining classification techniques and ensemble learning for predicting the type of breast cancer recurrence." International Conference on Green, Pervasive, and Cloud Computing. Springer, Cham. [https://doi.org/10.1007/978-3-030-19223-5\\_2](https://doi.org/10.1007/978-3-030-19223-5_2)
- Sharma, Alpna, et al. (2021). "Breast Cancer Detection Using Deep Learning and Machine Learning: A Comparative Analysis." International Conference on Innovative Computing and Communications. Springer, Singapore. [https://doi.org/10.1007/978-981-15-5113-0\\_39](https://doi.org/10.1007/978-981-15-5113-0_39)
- Nguyen, Cuong, Yong Wang, and Ha Nam Nguyen. (2013). "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic." 10.4236/jbise.2013.65070

- Ariani Indrawati, (2021). "PENERAPAN TEKNIK KOMBINASI OVERSAMPLING DAN UNDERSAMPLING UNTUK MENGATASI PERMASALAHAN IMBALANCED DATASET", 38-43, vol 4(1), <http://dx.doi.org/10.33387/jiko.v4i1.2561>
- Fernández, Alberto, et al. (2018). "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." *Journal of artificial intelligence research* 61, 863-905. <https://doi.org/10.1613/jair.1.11192>
- Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. (2014). "A survey of feature selection and feature extraction techniques in machine learning." 2014 Science and Information Conference. IEEE. DOI: 10.1109/SAI.2014.6918213.
- Saru, S., and S. Subashree. (2019). "Analysis and Prediction of Diabetes Using Machine Learning." *International Journal of Emerging Technology and Innovative Engineering* 5.4.
- Yulita, W. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Data Mining dan Sistem Informasi*, 2(2), 1-9.
- Ramadhan, N. G., Putrada, A. G., & Abdurohman, M. (2019). Improving Smart Lighting with Activity Recognition Using Hierarchical Hidden Markov Model. *Indonesian Journal on Computing (Indo-JC)*, 4(2), 43-54.