Ensemble Learning Dengan Metode *Smote Bagging* Pada Klasifikasi Data Tidak Seimbang

Rimbun Siringoringo[1,], Indra Kelana Jaya[2]

* Corresponding author : rimbun.ringo@gmail.com
[1,2] Universitas Methodist Indonesia
Jalan Hang Tuah No.8, Madras Hulu, Medan Polonia

Abstract-- Unbalanced data classification is a crucial problem in the field of machine learning and data mining. Data imbalances have a poor impact on classification results where minority classes are often misclassified as a majority class. Conventional machine learning algorithms are not equipped with the ability to work on unbalanced data, so the performance of conventional algorithms is always not optimal. In this study, ensemble learning using SMOTEBagging method was applied to classify 11 unbalanced datasets. SMOTEBagging performance is also compared with three types of conventional classification algorithms namely SVM, k-NN, and C4.5. By applying the 5 cross-validation scheme, the AUC value generated by SMOTEBagging is higher at 10 datasets. The mean values of the lowest to highest AUC were obtained by SVM, k-NN, C4.5 and SMOTEBagging algorithms with values 0.638, 0.742, 0.770 and 0.895. By applying Friedman test it was found that the performance of AUC SMOTEBagging differed significantly with the other three conventional methods SVM, k-NN and C4.5

***Keywords**: Class imbalance, ensemble learning, SMOTEBagging*

Abstrak-- *Klasifikasi data yang tidak seimbang merupakan masalah yang krusial pada bidang machine learning dan data mining. Ketidakseimbangan data memberikan dampak yang buruk pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas. Algoritma machine learning konvensional tidak dilengkapi kemampuan untuk bekerja pada data tidak seimbang, sehingga performa yang dihasilkan oleh algoritma konvensional sering tidak optimal. Pada penelitian ini, teknik ensemble metode SMOTEBagging diterapkan untuk mengklasifikasi 11 dataset tidak seimbang. Sebagai pembanding, dipilih tiga jenis algoritma klasifikasi yang konvensional yaitu SVM, k-NN dan C4.5. Dengan skema 5 cross fold validation diperoleh hasil bahwa nilai AUC SMOTEBagging lebih tinggi pada 10 dataset. Nilai rerata AUC terendah sampai tertinggi diperoleh oleh algoritma SVM, k-NN, C4.5 dan SMOTEBagging dengan nilai masing-masing 0,638, 0,742, 0,770 dan 0,895. Dengan menerapkan Friedman test diperoleh fakta bahwa performa AUC SMOTEBagging berbeda secara signifikan dengan ketiga metode konvensional yang lain yaitu SVM, k-NN dan C4.5*

***Kata kunci** : ketidakseimbangan kelas, ensemble learning, SMOTEBagging*

PENDAHULUAN

Data tidak seimbang merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data yang satu lebih sedikit (minoritas) atau lebih banyak (mayoritas) dibanding dengan jumlah kelas data lainnya [1]. Kondisi data tidak seimbang selalu memberikan

dampak buruk pada proses *machine learning* seperti klasifikasi maupun klasterisasi, pada akhirnya hasil yang diberikan selalu tidak optimal [2] dan tidak seperti yang diharapkan [3].

Hampir semua algoritma klasifikasi seperti *Naive Bayes*, *Decision Tree*, *K-Nearest Neighbor* dan yang lainnya menghasilkan kinerja yang sangat buruk jika bekerja pada data tidak seimbang [4] karena metode-metode tersebut didisain untuk bekerja pada data seimbang dan tidak dilengkapi dengan kemampuan untuk menangani masalah ketidakseimbangan kelas [5].

Ketika dihadapkan dengan data tidak seimbang, hampir semua algoritma *machine learning* akan memberikan akurasi yang jauh lebih besar pada kelas mayoritas daripada kelas minoritas [5]. Perbedaan ini merupakan suatu indikator performa klasifikasi yang buruk. Pada banyak kasus, eksistensi kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas.

Sejauh ini terdapat tiga pendekatan dalam menangani data tidak seimbang (*unbalance*), Pertama adalah pendekatan pada level data, kedua dalah level algoritma, dan pendekatan ketiga adalah pendekatan *ensemble* yaitu dengan menggabungkan tahap *pre-processing* dengan tahap klasifikasi. Pendekatan level data mencakup berbagai teknik *resampling* dan sintesis seperti SMOTE, RUS, dan ROS. Pendekatan algoritma diantaranya adalah *Genetic Fuzzy Classifier* dan *Extreme Learning*. Pada pendekatan *ensemble*, terdapat metode yang sangat populer yaitu *boosting* dan *bagging* dan *Random Forest* [6].

Banyak penelitian terkini terkait dampak penyeimbangan kelas terhadap perbaikan performa algoritma klasifikasi. Diantaraanya adalah penerapan SMOTE pada masalah klasifikasi infertilitas menggunakan metode klasifikasi *Multi Level Perceptron* (MLP), *k-Nearest Neighbor* (k-NN), Naïve Bayes, *Random Forest*, dan *Support Vector Machine* [7]. SMOTE berhasil meningkatkan performa klasifikasi dengan hasil terbaik didapatkan dengan metode *Naive Bayes* sebesar 90,7%. Penerapan SMOTE pada masalah klasifikasi pada *dataset* menggunakan *Fuzzy C-Means* [8] dapat mengatasi ketidakseimbangan kelas pada data set tersebut. Penerapan teknik *ensemble* dengan metode *bagging* dapat meningkatkan sensitifitas algoitma pada prediksi cacat *software* [9], penerapan teknik *ensemble AdaBoost* untuk menyelesaikan ketidakseimbangan kelas pada penentuan kelulusan mahassiswa [10] terbukti dapat meningkatkan akurasi sekaligus meminimalisasi *error* prediksi.

***Synthetic Minority Over-sampling Technique* (SMOTE)**

Metode SMOTE (*Syntetic Minority Oversampling Method*) merupakan salah satu metode yang sangat populer dalam menangani data tidak seimbang yang bekerja pada level data. Pendekatan ini bekerja dengan membuat replika dari data minoritas. Replika tersebut dikenal dengan data sintetis (*syntetic data*). Metode SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan atau sintetis [11]. Data buatan atau sintesis tersebut dibuat berdasarkan k-tetangga terdekat (*k-nearest neighbor*).

Bagging

Bagging atau *boostrap agregating* merupakan meta-algoritma yang dutujukan untuk meningkatkan performa dari algoritma *machine learning* [12]. *Bagging* terdiri dari dua tahap yakni tahap *boostrap* yakni pembuatan sub-dataset dari dataset dengan melakukan *resampling* data latih, kemudian tahap *agregating* dilakukan dengan menggabungkan banyak nilai prediksi menjadi satu nilai prediksi akhir.

Algoritma *bagging* :

1. Susun *k* buah pohon keputusan:
   a. Tahapan *bootstrap*, yakni ambil sampel acak berukuran *n* dari dataset *training*
   b. Susun pohon terbaik berdasarkan data tersebut
   c. Ulangi langkah a dan b sebanyak *l* kali sehingga diperoleh *l* buah pohon keputusan
2. Lakukan prediksi gabungan berdasarkan *l* buah pohon tersebut. Prediksi dapat dilakukan dengan menggunakan konsep pengambilan suara terbanyak (*majority vote*) pada kasus pohon klasifikasi, rata-rata pada kasus pohon regresi, dan penjumlahan prediksi peluang masing-masing kelas pada kasus pohon klasifikasi dan pohon regresi

SMOTEBagging

SMOTEBagging merupakan pengabungan dari SMOTE dan Bagging (*boostrap agregating*). Pada SMOTEBagging, setiap sub-dataset

diseimbangkan oleh SMOTE sebelum proses pemodelan. Ada dua parameter yang harus ditentukan pada SMOTE : jumlah tetangga terdekat atau *k-nearest neighbor* dan total jumlah *oversampling* dari kelas minor. Jumlah total *oversampling* ditentukan sejauh jumlah kelas minoritas dan kelas mayoritas seimbang [13]

METODOLOGI PENELITIAN

Prosedur kerja model

Prosedur kerja penelitian ditampilkan pada gambar 1

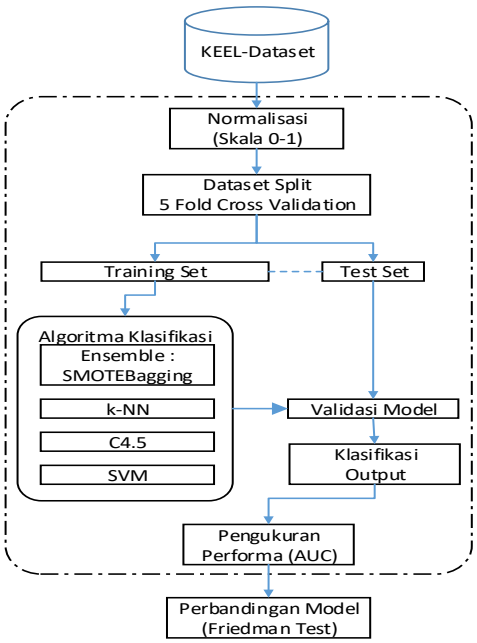
Dataset

Penelitian ini menggunakan 5 jenis dataset yang direkomendasikan oleh KEEL *Repository*.

Tabel 1. Deskripsi dataset

No	Dataset	#Ex	#Atts	IR
1	glass0	214	9	2,06
2	new-thyroid2	215	5	4,92
3	glass6	214	9	6,38
4	glass2	214	9	10,39
5	yeast-1_vs_7	459	8	11,87
6	ecoli4	336	7	13,84
7	abalone9-18	731	8	16,68
8	yeast4	1484	8	28,41
9	yeast5	1484	8	32,78
10	yeast6	1484	8	39,15

Deskripsi lengkap dataset-dataset tersebut ditampilkan pada tabel 1. Pada tabel tersebut dideskripsikan banyak data (#Ex.), banyaknya atribut (#Atts.) dan rasio ketidakseimbangan atau *Imbalance Ratio* (*IR*).



Gambar 1. Kerangka Kerja Model yang Diusulkan

Teknik validasi dan evaluasi

Teknik evaluasi dan estimasi performa pada penelitian ini menggunakan skema 5-*fold cross-validation*. Hal ini berarti, dataset dibagi menjadi 5 bagian atau *fold* yang sama, setiap *fold* berisi *20%* dataset, kemudian dilakukan proses *learning* sebanyak 5 kali. Pada tabel 2 ditampilkan hasil partisi dataset *glass0*. Teknik evaluasi dan pengukuran performa menerapkan *Area Under ROC Curve* (AUC). Pertimbangan penggunaan AUC karena AUC secara statistik lebih konsisten. AUC juga lebih baik dari metode akurasi (*accuracy*) dalam mengevaluasi perbandingan kinerja berbagai algoritma *classifier*

Tabel 2. Partisi dataset *glass0*

Partisi Dataset	Jumlah data	Fungsi
glass0-5-1tra.dat	171	*Training*
glass0-5-1tst.dat	43	*Testing*
glass0-5-2tra.dat	171	*Training*
glass0-5-2tst.dat	43	*Testing*
glass0-5-3tra.dat	171	*Training*
glass0-5-3tst.dat	43	*Testing*
glass0-5-4tra.dat	171	*Training*
glass0-5-4tst.dat	43	*Testing*
glass0-5-5tra.dat	171	*Training*
glass0-5-5tst.dat	43	*Testing*

Kerangka kerja penelitian

Penelitian ini melakukan klasifikasi terhadap dataset tidak seimbang menggunakan *ensemble learning* yaitu SMOTEBagging. Parameter SMOTEBagging yang diterapkan ditampilkan pada tabel 3 berikut. Parameter level konfiden 0.25, jumlah item-set minimum per *leaf* 2, banyak *classifier* sebanyak 10 dan *pruning* diterapkan untuk mendapatkan pohon akhir.

Tabel 3. Parameter SMOTEBagging

No	Parameter	Nilai
1	Pruning	TRUE
2	Konfiden	0,25
3	Instance per Leaf	2
4	Jumlah *classifier*	10

Sebagai pembanding, hasil klasifikasi yang diperoleh dengan SMOTEBagging dibandingkan dengan tiga algoritma klasifikasi yang konvensional yaitu *Support Vector Machine* (SVM), *Decision Tree C4.5* dan *k-Nearest*

Neighbor (k-NN). Ke tiga algoritma klasik tersebut memiliki karakteristik yang berbeda.

a. *Decision Tree* C4.5 : menerapkan parameter yaitu level konfiden 0,25, jumlah item-set minimum per *leaf* 2 dan *pruning* diterapkan untuk mendapatkan pohon akhir.
b. *Support Vector Machine* : menerapkan fungsi radial atau *Radial Base Function* (RBF), epsilon 0,001.
c. *k-Nearest Neighbor* : menerapkan nilai ketetanggaan (*k*) =1 serta *euclidean distance*

HASIL DAN PEMBAHASAN

Algoritma SMOTEBagging mengadopsi teori pohon pada tahap *aggregating.* Pohon pada SMOTEBagging direalisasikan dalam bentuk aturan-aturan atau *rules***.** Pada tabel 4 ditampilkan rerata jumlah *rule* pada setiap *fold*. Pada fold 1 dihasilkan sebanyak 10 *classifier* dengan rerata *rule* per *classifier* sebanyak 10 *rules*.

Tabel 4. Rerata jumlah *rule* pada SMOTEBagging

Fold	**Jumlah *Classifier***	***Rules* per *Classifier***
Fold 1	10	10
Fold 2	10	11
Fold 3	10	11
Fold 4	10	9
Fold 5	10	12

Salah satu bentuk *rule* yang dihasilkan pada proses klasifikasi dengan SMOTEBagging ditampilkan pada gambar 2 berikut. Pada gambar tersebut terdapat 6 aturan, pada setiap aturan terdapat kondisi atribut-atribut dataset (Mg, RI, Fe, K, Si) serta kelas yang diprediksi (*negative/positive*)

```
Rule[1]: IF  Mg <= 2.681
  THEN Class = negative
Rule[2]: IF  Mg > 2.681 AND
             RI <= 1.517 AND
             Fe <= 0.112
  THEN Class = negative
Rule[3]: IF  Mg > 2.681 AND
             RI <= 1.517 AND
             Fe > 0.112 AND
             K <= 0.655
  THEN Class = positive
Rule[4]: IF  Mg > 2.681 AND
             RI <= 1.517 AND
             Fe > 0.112 AND
             K > 0.655
THEN Class = negative
Rule[5]: IF  Mg > 2.681 AND
             RI > 1.517 AND
             Si <= 71.501
THEN Class = negative
Rule[6]: IF  Mg > 2.681 AND
             RI > 1.517 AND
             Si > 71.501
THEN Class = positive
```

Rule atau aturan diterapkan untuk melakukan klasifikasi. Pada tabel 5 berikut adalah hasil klasifikasi pada data *Testing fold 1* (43 *record* data) dataset *glass0*. Kelas aktual adalah label yang tertera pada dataset, kelas prediksi adalah hasil klasifikasi dengan menggunakan algoritma *ensemble learning* yaitu SMOTEBagging.

Tabel 5. Hasil klasifikasi

Data	**Kelas Aktual**	**Kelas Prediksi**
1	positive	positive
2	positive	positive
3	positive	negative
4	positive	positive
5	positive	positive
6	positive	positive
7	positive	positive
8	positive	positive
9	positive	positive
10	positive	positive
11	positive	positive
12	positive	positive
13	positive	negative
14	positive	positive
15	negative	negative
16	negative	positive
17	negative	negative
18	negative	negative
19	negative	positive
20	negative	negative
43	negative	negative

AUC merupakan evaluasi yang diterapkan untuk mengukur performa SMOTEBagging pada

penelitian ini. Pada tabel 6 ditampilkan nilai AUC pada setiap fold dataset *glass0*, baik untuk data *testing* maupun data *training*. Rata-rata AUC untuk dataset glass0 menggunakan SMOTEBagging adalah 0,883 dan rerata standar deviasi 0,045

Tabel 6. Performa AUC pada dataset *glass0*

Partisi	Fold	AUC	StDev
Testing	Fold 1	0.842	0.076
	Fold 2	0.734	
	Fold 3	0.895	
	Fold 4	0.719	
	Fold 5	0.893	
Training	Fold 1	0.939	0.013
	Fold 2	0.970	
	Fold 3	0.961	
	Fold 4	0.939	
	Fold 5	0.939	
Rata-rata		**0.883**	**0.045**

Penelitian ini diuji pada 10 jenis dataset tidak seimbang. Hasil pencapaian AUC untuk keseluruhan dataset ditampilkan pada tabel 7 berikut. Nilai AUC tertinggi diperoleh sebesar 0.919 pada dataset *yeast6* dan terkecil sebesar 0.753 pada dataset *abalone19*.

Tabel 7. Performa SMOTEBagging

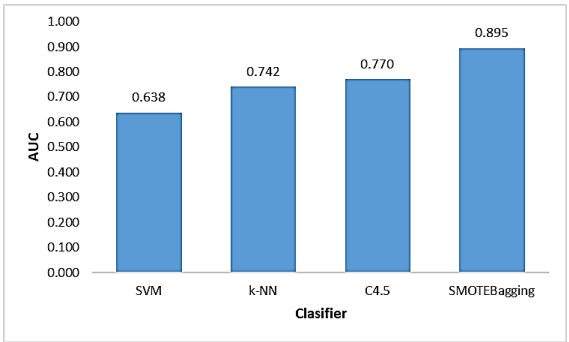
Dataset	Global AUC	Standar Deviasi
glass0	**0.883**	**0.045**
new-thyroid2	0.972	0.024
glass6	0.952	0.033
glass2	0.873	0.031
yeast-1_vs_7	0.833	0.022
ecoli4	0.948	0.030
abalone9-18	0.866	0.059
yeast4	0.868	0.044
yeast5	0.975	0.015
yeast6	0.919	0.040
abalone19	0.753	0.018

Hasil AUC yang diperoleh dengan SMOTEBagging untuk keseluruhan dataset dibandingkan dengan hasil yang diperoleh oleh *classifier* lain yaitu SVM, C4.5 dan k-NN sebagaimana dijabarkan pada tabel 8 berikut. Pada tabel tersebut, nilai AUC terbaik adalah angka yang diberi tanda penebalan atau **bold.** Pada tabel tersebut, SMOTEBagging memperoleh nilai AUC yang tertinggi pada 10 dataset sementara SVM memperoleh nilai tertinggi pada satu dataset.

Tabel 8. Hasil AUC pada semua *classifier*

Dataset	SVM	C4.5	k-NN	SMOTE Bagging
glass0	0.767	0.874	0.815	**0.883**
new-thyroid2	**0.983**	0.967	0.972	0.972
glass6	0.948	0.880	0.888	**0.952**
glass2	0.500	0.767	0.602	**0.873**
yeast-1_vs_7	0.500	0.665	0.660	**0.833**
ecoli4	0.806	0.866	0.871	**0.948**
abalone9-18	0.517	0.628	0.597	**0.866**
yeast4	0.500	0.664	0.667	**0.868**
yeast5	0.500	0.914	0.840	**0.975**
yeast6	0.500	0.744	0.751	**0.919**
abalone19	0.500	0.500	0.496	**0.753**

Selain membandingkan nilai AUC per dataset, nilai *mean* AUC ditampilkan pada gambar 3 berikut. Berdasarkan gambar tersebut, *mean* AUC dari terendah sampai *mean* AUC tertinggi diperoleh pada *classifier* SVM, k-NN, C4.5 dan SMOTEBagging, sehingga SMOTEBagging menjadi *classifier* terbaik pada penelitian ini.



Gambar 2. Perbandingan *Mean* AUC

Perbandingan Model

Untuk menguji apakah model SMOTEBagging memiliki keunggulan yang signifikan dibandingkan dengan ketiga model lain maka dilakukan uji perbandingan dengan menerapkan uji *Friedman* dengan

mengikutsertakan metode *post hoc* yaitu *shaffer*. Hasil uji perbandingan dijabarkan pada tabel 9 berikut.

Tabel 9. Nilai *p-value* untuk α=0,05

i	**algorithms**	***z* = (*R0* - *Ri*)/*SE***	***p***	**Shaffer**
6	SVM vs. SMOTEBagging	4.129	**0.000**	0.008
5	k-NN vs. SMOTEBagging	2.973	**0.003**	0.0167
4	C4.5 vs. SMOTEBagging	2.807	**0.005**	0.0167
3	SVM vs. C4.5	1.321	0.186	0.0167
2	SVM vs. SMOTEBagging	1.156	0.248	0.025
1	C4.5 vs. SMOTEBagging	0.165	0.869	0.050

Dari tabel di atas, *p-value* untuk baris 1, 2 dan 3 (ditebalkan) lebih kecil dari 0,05. Hal tersebut mengindikasikan bahwa SMOTEBagging berbeda secara signifikan dengan SVM, C4.5 dan k-NN. Sementara *p-value* untuk baris 4, 5 dan 6 lebih besar dari 0,05. Hal tersebut mengindikasikan bahwa SVM, C4.5 maupun k-NN tidak berbeda secara signifikan. Nilai shaffer terkecil adalah 0,008 sementara *p-value* untuk baris ,2 dan 3 masih lebih kecil dari 0,008 sehingga hal tersebut menunjukkan bahwa dengan metode shaffer, SMOTEBagging tetap berbeda secara signifikan dengan ketiga metode lain.

KESIMPULAN

Berdasarkan hasil eksperimen di atas diperoleh kesimpulan bahwa penanganan ketidakseimbangan kelas memberikan dampak yang baik bagi perbaikan performa *machine learning*. Algortima *classifier* SVM, k-NN dan C4.5 masih belum optimal jika bekerja pada data tidak seimbang. Metode SMOTEBagging merupakan metode *ensemble* yang memiliki keunggulan dan kemampuan melakukan klasifikasi dan generalisasi terhadap data tidak seimbang. Dari sebanyak 11 dataset yang diuji, SMOTEBagging memiliki nilai AUC yang lebih baik pada 10 dataset. Dengan menerapkan *Friedman test* diperoleh fakta bahwa performa AUC SMOTEBagging berbeda secara signifikan dengan ketiga metode konvensional yang lain yaitu SVM, k-NN dan C4.5

DAFTAR PUSTAKA

[1] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: a review," *Int J Adv. Soft Compu Appl*, vol. 7, no. 3, 2015.

[2] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *ArXiv Prepr. ArXiv171005381*, 2017.

[3] C. Zhang, Y. Chen, X. Liu, and X. Zhao, "Abstention-SMOTE: An over-sampling approach for imbalanced data classification," in *Proceedings of the 2017 International Conference on Information Technology*, 2017, pp. 17–21.

[4] G. Y. Wong, F. H. Leung, and S.-H. Ling, "A Hybrid Evolutionary Preprocessing Method for Imbalanced Datasets," *Inf. Sci.*, 2018.

[5] Q. Gu, X.-M. Wang, Z. Wu, B. Ning, and C.-S. Xin, "An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification," *J Dig Inf Manag*, vol. 14, no. 2, pp. 92–103, 2016.

[6] A. Mishra and U. S. Reddy, "A comparative study of customer churn prediction in telecom industry using ensemble based classifiers," in *Inventive Computing and Informatics (ICICI), International Conference on*, 2017, pp. 721–725.

[7] B. Karlik, A. Yibre, and K. Barış, *Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility*, vol. 3. 2016.

[8] R. Pruengkarn, K. W. Wong, and C. C. Fung, "Multiclass Imbalanced Classification Using Fuzzy C-Mean and SMOTE with Fuzzy Support Vector Machine," in *International Conference on Neural Information Processing*, 2017, pp. 67–75.

[9] A. Saifudin, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, p. 11, 2015.

[10] A. Bisri and R. S. Wahono, "Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada

Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree," *J. Intell. Syst.*, vol. 1, no. 1, p. 6, 2015.

[11] M. Beckmann, N. F. F. Ebecken, and B. S. L. Pires de Lima, "A KNN Undersampling Approach for Data Balancing," *J. Intell. Learn. Syst. Appl.*, vol. 07, no. 04, pp. 104–116, 2015.

[12] M. Moukhafi, K. E. Yassini, and S. Bri, "Mining network traffics for intrusion detection based on Bagging ensemble Multilayer perceptron with Genetic algorithm optimization," p. 8, 2018.

[13] L. Hakim, B. Sartono, and A. Saefuddin, "Bagging Based Ensemble Classification Method on Imbalance Datasets," vol. 6, no. 6, p. 7, 2017.