Volume 6 Nomor 2 Mei 2021

INFORMASI INTERAKTIF

JURNAL INFORMATIKA DAN TEKNOLOGI INFORMASI

PROGRAM STUDI INFORMATIKA – FAKULTAS TEKNIK -UNIVERSITAS JANABADRA

INFORMASI INTERAKTIF	Vol. 6	No. 2	Hal. 56 - 95	Yogyakarta Mei 2021	ISSN 2527-5240

ISSN 2527-5240

DEWAN EDITORIAL

JURNAL INFORMASI INTERAKTIF merupakan media komunikasi hasil penelitian, studi kasus, dan ulasan ilmiah bagi ilmuwan dan praktisi dibidang Teknik Iinformatika. Diterbitkan oleh Program Studi Teknik Informatika Fakultas Teknik Universitas Janabadra di Yogyakarta, tiga kali setahun pada bulan Januari, Mei dan September.

INFORMASI INTERAKTIF ISSN 2527-5240

DAFTAR ISI

ISSN 2527-5240

PENGANTAR REDAKSI

Puji syukur kami panjatkan kehadirat Allah Tuhan Yang Maha Kuasa atas terbitnya JURNAL INFORMASI INTERAKTIF Volume 6, Nomor 2, Edisi Mei 2021. Pada edisi kali ini memuat 5 (lima) tulisan hasil penelitian dalam bidang informatika.

Harapan kami semoga naskah yang tersaji dalam JURNAL INFORMASI INTERAKTIF edisi Mei tahun 2021 dapat menambah pengetahuan dan wawasan di bidangnya masing-masing dan bagi penulis, jurnal ini diharapkan menjadi salah satu wadah untuk berbagi hasil-hasil penelitian yang telah dilakukan kepada seluruh akademisi maupun masyarakat pada umumnya.

Redaksi

PERBANDINGAN ANALISIS DATA FITUR NOMINAL MULTI-KATEGORI MENGGUNAKAN METODE *ADAPTIVE SYNTHETIC NOMINAL* (*ADASYN-N*) SERTA *ADAPTIVE SYNTHETIC-KNN* (*ADASYN-KNN*)

Jeffry Andhika Putra[1], Sri Rahayu[2]

[12]Program Studi Informatika, Fakultas Teknik, Universitas Janabadra, Yogyakarta
Jalan Tentara Rakyat Mataram No. 55 – 57 Yogyakarta 55231

Email: [1]jeffry@janabadra.ac.id, [2]ayu.dj@janabadra.ac.id

ABSTRACT

Growing need for efficient algorithms for data manipulation, analysis, and intelligent use has been a very active research area in machine learning field. However, some research areas still not fully developed, especially when unbalanced data classification is needed. Datasets with this class imbalance occur because of an unbalanced ratio between one case and another. This class imbalance will be detrimental to data mining because machine learning in data mining has difficulty in classifying minority classes (small instances) correctly. There are several approaches to handling imbalances, one of which is by using the original data sampling method. The first sampling method approach to overcome class imbalance is undersampling which is a method to balance classes by randomly reducing the majority class instances. Over-sampling is a method of balancing class distribution by randomly replicating instances in minority classes.

This study presents comparison of over-sampling techniques to overcome problem of class imbalances in datasets with nominal-multi categories features between Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-kNN (ADASYN-KNN) methods. There are seven datasets with nominal-multi categories features which have an unbalanced class distribution. Then the dataset that has been over-sampled with both methods is classified using the Random Forest method. Furthermore, a comparison of the accuracy of the original dataset and the dataset of the ADASYN-N and ADASYN-KNN over-sampling techniques was carried out.

Keywords: *ADASYN-KNN, ADASYN-N, class imbalance, nominal, multi-category, over-sampling.*

1. PENDAHULUAN

Peningkatan kebutuhan otomasi algoritma yang efisien untuk manipulasi data, analisis, serta penggunaan cerdas menjadi area penelitian aktif dalam bidang *machine learning*. Namun beberapa area penelitian belum sepenuhnya berkembang, terutama ketika diperlukan klasifikasi data tidak seimbang. Klasifikasi data tidak seimbang umumnya diperlukan untuk diagnosa penyakit, deteksi kecurangan, pengendalian kualitas, serta kasus lainnya [1].

Dataset dengan ketidakseimbangan kelas terjadi karena rasio tidak seimbang antara satu kasus dengan kasus lain merugikan penelitian bidang data mining karena *machine learning* pada data mining memiliki kesulitan mengklasifikasikan kelas minoritas. Algoritma mengasumsikan distribusi kelas diuji seimbang sehingga dalam mengklasifikasikan hasil setiap kelas. Pada algoritma seperti *Decision Tree*, *Nearest Neighbor*, dan *Support Vector Machine* (*SVM*) menggeneralisasi data uji sama serta menghasilkan hipotesis paling sederhana. Hal ini mengakibatkan kesalahan klasifikasi kelas minoritas karena ketidakseimbangan kelas yang cenderung berfokus pada mayoritas serta mengabaikan minoritas saat klasifikasi.

Terdapat dua pendekatan penanganan ketidakseimbangan kelas, yaitu data-level menggunakan metode *sampling* data asli baik mayoritas (*undersampling*), maupun minoritas (*over-sampling*). Kedua, pendekatan algoritma dengan mendesain algoritma baru atau meningkatkan algoritma. Pendekatan metode *sampling* untuk mengatasi ketidakseimbangan kelas adalah *undersampling*; metode menyeimbangkan kelas dengan mengurangi *instance* kelas mayoritas secara acak. Namun, metode *undersampling* memiliki kekurangan, yaitu hilangnya informasi serta data penting untuk proses pengambilan keputusan oleh *machine learning*.

Penyeimbangan kelas dapat juga dengan melakukan *sampling* pada kelas minoritas (*over-sampling*). *Over-sampling* merupakan metode penyeimbangan distribusi kelas dengan

mereplikasi *instance* kelas minoritas secara acak. Namun, *over-sampling* memungkinkan muncul *overfitting* karena duplikasi *instance*. Tahun 2002, Chawla [2] mengajukan solusi menangani *overfitting* pada metode *over-sampling*, yaitu *SMOTE* (*Synthetic Minority Over-sampling Technique*). *SMOTE* memanfaatkan *Nearest Neighbors* serta jumlah *over-sampling* diinginkan. *SMOTE* digunakan untuk pendekatan data numerik.

Selain *SMOTE*, terdapat metode pendekatan *sampling* pada pembelajaran dengan *dataset* tidak seimbang dengan fitur numerik yang diajukan, yaitu *ADASYN* [3]. Ide utama *ADASYN* menggunakan bobot distribusi data pada kelas minoritas berdasarkan pada tingkat kesulitan belajar, dimana data sintesis dihasilkan dari kelas minoritas kesulitan untuk belajar dibandingkan dengan data minoritas lebih mudah belajar. Untuk penanganan data fitur nominal, Chawla mengajukan metode *SMOTE*-N merupakan pengembangan *SMOTE* [2]. Pada *SMOTE*-N, *Nearest Neighbor* dihitung menggunakan versi modifikasi *Value Difference Metric* (*VDM*) yang diajukan Cost dan Salzberg.

2. TINJAUAN PUSTAKA

Beberapa studi penelitian terdahulu diantaranya *Statistical Learning with Imbalanced Data* [1] mengimplementasikan beberapa metode *sampling* untuk pembelajaran statistik dengan data tidak seimbang serta menguji performa metode pengukuran akurasi ketidakseimbangan. Beberapa metode *sampling* yang diimplementasikan antara lain *Random Undersampling*, *Tomek Links*, *Condensed Nearest Neighbor Rule*, *One-Sided Undersampling*, *NearMiss Undersampling*, *Cluster Based Undersampling*, *Random Over-sampling*, *SMOTE* [2], *Borderline-SMOTE*, *ADASYN* [3], serta *Cluster Based Over-sampling*. Penelitian menggunakan beberapa algoritma klasifikasi pengujian; *Naive Bayes*, *Classification Tree*s, Neural Network, *Random Forest*, *SVM*, *K-Nearest Neighbors*, serta *Extreme Gradient Boosting*.

Penelitian yang dilakukan Shipitsyn [1] mengemukakan kesimpulan performa algoritma klasifikasi berbeda tidak selalu menunjukkan perubahan kepada algoritma *sampling* berbeda. Terdapat beberapa *classifier* memiliki sensitivitas tinggi terhadap algoritma *sampling* seperti *SVM*, *Random Forest*, *K-Nearest Neighbors*, serta *Classification Tree*. Shipitsyn menyebutkan algoritma *over-sampling* menunjukkan performa lebih baik daripada *undersampling*. Hal serupa dikemukakan Garcia [5] yang melakukan penelitian untuk mengetahui pengaruh *imbalance ratio* serta *classifier* kinerja beberapa stategi *re-sampling* untuk *dataset* tidak seimbang [6].

Huang [7] dalam penelitian *Classification of Imbalanced Data Using Synthetic Over-Sampling Techniques* membandingkan performa kombinasi metode *over-sampling* dengan pengujian *classifier* pada tujuh *UCI-Datasets*. Perbandingan performa bukan menggunakan pengukuran umum seperti *Overall Accuration*, *Precision*, maupun *Recall*, melainkan dengan *F-Measure*, *G-Mean*, serta AUC. Huang menggunakan empat *classifier*, yaitu *Naive Bayes*, *Decision Tree*, *Random Forest*, serta SVM yang dikombinasikan dengan *Over-sampling with Replacement*, *SMOTE*, *ADASYN*, serta *Dataset* tanpa *over-sampling*. Hasil penelitian menunjukkan dengan menggunakan metode *over-sampling* menghasilkan performa lebih baik, namun tidak ada metode *over-sampling* yang mengungguli metode lainnya. Kinerja metode *sampling* ini bervariasi antar *dataset* serta *classifier*.

Penelitian berjudul *Multi-Class Imbalance Learning* dengan metode *Adaptive Synthetic-Nominal* (*ADASYN-N*) serta *Adaptive Synthetic-kNN* (*ADASYN*- KNN) untuk *re-sampling* data hasil tes Pap-Smear [4] mengembangkan metode *ADASYN-N* serta *ADASYN-KNN* merupakan pengembangan metode *ADASYN*. Metode *ADASYN-N* serta *ADASYN-KNN* mampu ketidakseimbangan data dengan fitur nominal. Penelitian tersebut menguji metode *ADASYN-N* serta *ADASYN-KNN* pada satu *dataset* hasil patologi anatomi tes PapSmear, dilakukan klasifikasi dengan *Naïve Bayes Classifier* (NBC). Hasil klasifikasi kedua metode tersebut dibandingkan dengan *SMOTE*-N [2] menunjukkan *ADASYN-N* meningkatkan akurasi lebih baik dari *SMOTE*-N, metode *ADASYN-KNN* menunjukkan performa akurasi kedua metode tersebut.

3. LANDASAN TEORI

3.1 K-Nearest Neighbors

K-Nearest Neighbors adalah algoritma *supervised learning* dimana hasil *instance* baru diklasifikasi berdasarkan mayoritas kategori *K-Neighbors* terdekat. Menurut Gorunescu, algoritma KNN merupakan metode klasifikasi objek dimana objek baru diberi label berdasarkan objek tetangga terdekat serta merupakan algoritma paling sederhana diantara algoritma *machine learning*, karena klasifikasi objek berdasarkan *majority vote* tetangga (*neighbors*) [12].

$$D(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (1)$$

Dengan *D* adalah jarak antara titik x serta y yang diklasifikasikan, dimana $x = x_{/}, x_{.}, x_1, \dots, x_*$; $y = y_{/}, y_{.}, y_1, \dots, x_*$; k merepresentasikan nilai atribut, serta n merupakan dimensi atribut.

3.2 *Adaptive Synthetic Nominal* (ADASYN-N) dan *Adaptive Synthetic-kNN* (ADASYN-kNN)

Pada ADASYN-N, *Nearest Neighbor* dihitung menggunakan versi modifikasi *Value Difference Metric* (VDM) seperti SMOTE-N [2]. Chawla menggunakan versi modifikasi VDM yang diajukan oleh Cost serta Salzberg, VDM melihat pada nilai fitur yang overlap terhadap semua vektor fitur. Matriks mendefinisikan jarak antara nilai fitur yang sesuai untuk vektor fitur yang dibuat. Jarak δ antara dua nilai fitur sesuai didefinisikan seperti pada persamaan (1).

ADASYN-kNN merupakan pengembangan metode ADASYN-N dengan pengembangan prosedur untuk menghasilkan *instance* data sintesis sebanyak g_6. Berdasarkan gambar 2.2 pada ADASYN-KNN, data sintesis dihasilkan dari *Nearest Neighbor instance* yang dievaluasi. Atribut sintesis dihasilkan dengan melakukan *Voting* berdasarkan kemunculan atribut *Nearest Neighbor*. Kemudian, *instance* sintesis yang dihasilkan diduplikasi sebanyak g_6.

3.3 Penyusunan *Dataset*

Dataset yang diolah pada penelitian ini berupa tujuh *Dataset* dari sumber *UCI-Datasets* dengan rincian sebagai berikut:

Tabel 1. Detail *Dataset*

Dataset	*Instances*	Kelas	Distribusi Kelas
Audiology	26	6	mixed_cochlear_age_fixation 1 cochlear_age 11 normal_ear 2 cochlear_poss_noise 4 cochlear_age_and_noise 4 mixed_cochlear_unk_fixation 4
Balance Scale	626	3	L 288 B 49 R 288
Breast Cancer	286	2	no-recurrence-events 201 recurrence-events 85
Car	1728	4	unacc 1210 acc 384 good 69 vgood 65
Lenses	24	3	hard-contact-*Lenses* 4 soft-contact-*Lenses* 5 no-contact-*Lenses* 15
Lymphography	148	4	normal 2 metastases 81 malign_lymph 61 fibrosis 4

3.4 Penghitungan kNN

Untuk menghitung kNN setiap data, dilakukan penghitungan menggunakan persamaan *Euclidean Distance* sebagai berikut:

$$D(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (2)$$

Dengan fitur nominal *multi categories*, maka rumusan *Euclidean Distance* menjadi:

$$D(D_1, D_2) = \sqrt{\sum_{k=1}^{n}\left(D_{1,k} - D_{2,k}\right)^2} \quad (3)$$

Dengan D_1 serta D_2 adalah data yang diukur jarak *Euclidean*, k adalah fitur pada data. Pada kasus nominal *multi categories* penghitungan $D_{/,*} - D_{.,*}$ menggunakan persamaan $\delta\_F_{6,bc}$, $F_{6,bd}$` dimana $F_{6,bc}$ fitur ke-i dengan kategori a. Selanjutnya, menghitung *distance* setiap fitur menggunakan persamaan:

$$\delta(F_k Ca, F_k Cb) = \sum_{i=1}^{n}\left|\frac{Ca_i}{Ca} - \frac{Cb_i}{Cb}\right| \quad (4)$$

Dengan C_a dan C_b adalah kategori fitur k yang terdapat pada data dan dihitung jaraknya, i adalah distribusi kelas pada data. Dari model perhitungan, pemberian parameter k berbeda memungkinkan untuk memberikan performansi berbeda pada klasifikasi *dataset* dengan distribusi tidak seimbang.

3.5 Penghitungan *ADASYN*

Untuk contoh penghitungan *ADASYN dataset*, dilakukan evaluasi *Nearest Neighbor* data pertama (D_1), jika nilai K = 5, maka ditunjukkan pada tabel 2.

Tabel 2. *Nearest Neighbor* Data 1 (D_1)

Data	Kelas	Jarak
D_2	A	0,167
D_4	B	1,667
D_3	A	1,675
D_6	B	2,0605
D_{10}	B	2,236

Tabel 2 menunjukkan tiga data dengan kelas selain A, sehingga $\Delta_{D1} = 3$. Maka, rasio untuk data D_1 adalah:

$$r_{D1} = \frac{\Delta_{D1}}{K}$$

$$r_{D1} = \frac{3}{5}$$

$$r_{D1} = 0{,}6$$

Dengan cara sama, diperoleh nilai Δ untuk setiap data seperti pada tabel 3.

Tabel 3. NN Data Kelas A

Data Evaluasi	Data terdekat	Δ_i	r_i
D_1	D_2, D_4, D_3, D_6, D_{10}	3	0,6
D_2	D_1, D_3, D_4, D_6, D_{10}	3	0,6
D_3	D_4, D_{10}, D_6, D_9, D_5/D_8	5	1

Normalisasi r_i untuk mendapatkan *Density Distribution* ($\hat{r}$), sehingga didapatkan hasil pada tabel 4.

Tabel 4. *Density Distribution* Data Kelas A

Data Evaluasi	$\hat{r}_1$
D_1	0,2727
D_2	0,2727
D_3	0,4545

Menghitung jumlah duplikasi *instance* sintesis setiap data ke-i (x_i). Sebagai contoh data ke-1 maka:

$$g_1 = \hat{r}_1 x G_c$$

$$g_1 = 0{,}2727x4$$

$$g_1 = 1{,}0908$$

Untuk data diperoleh g_i dan dilakukan pembulatan untuk mendapatkan jumlah duplikasi data sintesis seperti ditunjukkan tabel 5.

Tabel 5. Jumlah Duplikasi Data Sintesis

Data	$\hat{r}_1$	g_i	Sintesis
D_1	0,2727	1,0908	1
D_2	0,2727	1,0908	1
D_3	0,4545	1,818	2

Untuk *ADASYN-N*, replikasi dilakukan secara langsung sejumlah nilai sintesis yang didapat dan ditambahkan *dataset* awal sehingga menghasilkan *dataset* baru seperti pada tabel 6.

Tabel 6. *Dataset* Setelah Penambahan Data Sintesis Metode *ADASYN-N*

No.	F_1	F_2	F_3	F_4	*Class*
1.	C_1	C_1	C_1	C_1	A
2.	C_1	C_2	C_1	C_2	A
3.	C_2	C_2	C_1	C_3	A
4.	C_2	C_2	C_1	C_1	B
5.	C_2	C_3	C_3	C_2	B
6.	C_2	C_1	C_2	C_3	B
7.	C_2	C_3	C_2	C_1	B
8.	C_2	C_3	C_3	C_2	B
9.	C_3	C_2	C_3	C_3	B
10.	C_3	C_3	C_1	C_1	B
11.	C_1	C_1	C_1	C_1	A
12.	C_1	C_2	C_1	C_2	A
13.	C_2	C_2	C_1	C_3	A
14.	C_2	C_2	C_1	C_3	A

Sedangkan metode *ADASYN-KNN* memanfaatkan *voting neighbor* setiap data, sebagai berikut:

Tabel 7. Evaluasi *Voting* Data D_1

No.	F_1	F_2	F_3	F_4	*Class*
Evaluasi D_1	C_1	C_1	C_1	C_1	A
NN-1	C_1	C_2	C_1	C_2	A
NN-2	C_2	C_2	C_1	C_1	B
NN-3	C_2	C_2	C_1	C_3	A
NN-4	C_2	C_1	C_2	C_3	B
NN-5	C_3	C_3	C_1	C_1	B
Voting	C_2	C_2	C_1	C_1	A

Tabel 8. Evaluasi *Voting* Data D_2

No.	F_1	F_2	F_3	F_4	Class
Evaluasi D_2	C_1	C_2	C_1	C_2	A
NN-1	C_1	C_1	C_1	C_1	A
NN-2	C_2	C_2	C_1	C_3	A
NN-3	C_2	C_2	C_1	C_1	B
NN-4	C_2	C_1	C_2	C_3	B
NN-5	C_3	C_3	C_1	C_1	B
Voting	C_2	C_1	C_1	C_1	A

Tabel 9. Evaluasi *Voting* Data D_3

No.	F_1	F_2	F_3	F_4	CLASS
Evaluasi D_3	C_2	C_2	C_1	C_3	A
NN-1	C_2	C_2	C_1	C_1	B
NN-2	C_3	C_3	C_1	C_1	B
NN-3	C_2	C_1	C_2	C_3	B
NN-4	C_3	C_2	C_3	C_3	B
NN-5	C_2	C_3	C_3	C_2	B
Voting	C_2	C_2	C_1	C_1	A

Data sintetis yang dihasilkan proses tersebut ditambahkan pada *dataset* asli sehingga membentuk *dataset* baru.

Tabel 10. Hasil Penambahan *Dataset* Asli Dengan Data Sintesis Dengan Metode *ADASYN-KNN*

No.	F_1	F_2	F_3	F_4	Class
1	C_1	C_1	C_1	C_1	A
2	C_1	C_2	C_1	C_2	A
3	C_2	C_2	C_1	C_3	A
4	C_2	C_2	C_1	C_1	B
5	C_2	C_3	C_3	C_2	B
6	C_2	C_1	C_2	C_3	B
7	C_2	C_3	C_2	C_1	B
8	C_2	C_3	C_3	C_2	B
9	C_3	C_2	C_3	C_3	B
10	C_3	C_3	C_1	C_1	B
11	C_2	C_2	C_1	C_1	A
12	C_2	C_1	C_1	C_1	A
13	C_2	C_2	C_1	C_1	A
14	C_2	C_2	C_1	C_1	A

3.6 Pembuatan *Instance* Sintesis dengan *ADASYN-N*

ADASYN-N diimplementasikan dengan bahasa *Python*. Proses awal pembuatan *instance* sintesis *ADASYN-N* adalah menghitung jumlah kemunculan suatu nilai kategori fitur. Jumlah ini digunakan dalam perhitungan *Euclidean Distance* untuk mencari *Nearest Neighbor*.

Tabel 11. Jumlah Kemunculan Kategori Fitur *Dataset Balance Scale*

Fitur	Kemunculan				
	1	2	3	4	5
Left-Weight	125	125	125	125	125
Left-Distance	125	125	125	125	125
Right-Weight	125	125	125	125	125
Right-Distance	125	125	125	125	125

Selanjutnya menghitung jarak setiap fitur. Jarak setiap fitur didapatkan dengan persamaan 3. Berdasarkan persamaan tersebut, diperoleh jarak setiap fitur yang tersaji pada tabel 12.

Tabel 12. Jarak Antar Kategori Setiap Fitur Pada *Dataset Balance Scale*

Distance	Fitur			
	Left-Weight	*Left-Distance*	*Right-Weight*	*Right-Distance*
D[1-2]	0.432000000000000005	0.432000000000000005	0.432000000000000005	0.432000000000000005
D[1-3]	0.736	0.736	0.736	0.736
D[1-4]	0.96	0.96	0.96	0.96
D[1-5]	1.136000000000000001	1.136000000000000001	1.136000000000000001	1.136000000000000001
D[2-3]	0.32	0.32	0.32	0.32
D[2-4]	0.544	0.544	0.544	0.544
D[2-5]	0.72	0.72	0.72	0.72
D[3-4]	0.24	0.24	0.24	0.24
D[3-5]	0.399999999999999999	0.399999999999999999	0.399999999999999999	0.399999999999999999
D[4-5]	0.175999999999999996	0.175999999999999996	0.175999999999999996	0.175999999999999996

Setelah dibangkitkan jumlah sampel *instance* sintesis, proses *ADASYN-KNN* untuk mendapatkan *instance* sintesis, *instance* yang dipertimbangkan dicari lima dan sembilan *nearest neighbor* kemudian fitur setiap data *nearest neighbor* dilakukan *voting* terhadap fitur untuk mendapatkan fitur sintesis, *instance* sintesis digandakan sejumlah g_i (jumlah duplikasi *instance* sintesis setiap data kelas minoritas).

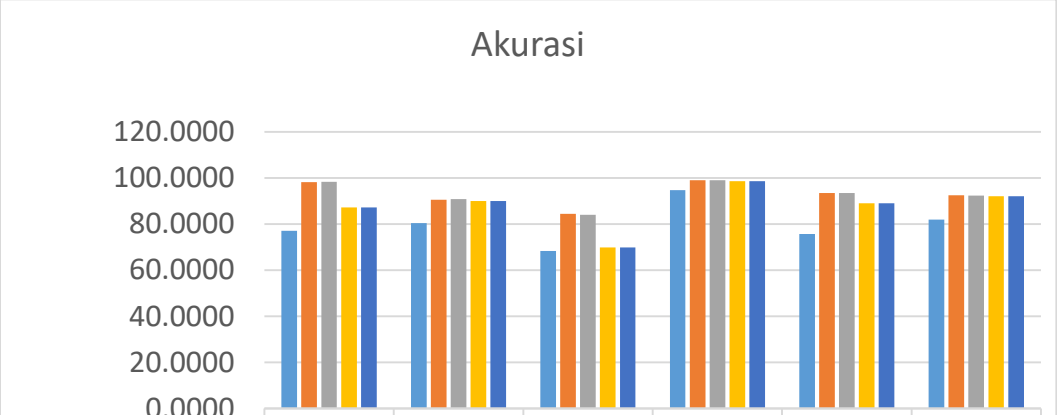
4. HASIL PENELITIAN

4.1 Penggabungan *Dataset* Baru

Hasil proses pembuatan *instance* sintesis dengan metode *ADASYN-N* dan *ADASYN-KNN* dengan k=5 dan k=9 digabungkan dengan *dataset* asli menghasilkan *dataset* baru. Untuk mengetahui kinerja metode hasil penelitian ini, digunakan *classifier Random Forest*. Hasil *dataset* dibandingkan dengan berbagai *performance metrics*, yaitu akurasi, *G-Mean*, *F-Score*, dan *ROC Area*. Keenam *dataset* yang digunakan penelitian ini memiliki fitur nominal multi kategori dengan empat perlakuan metode *re-sampling* yang berbeda.

4.2 Akurasi

Nilai akurasi menunjukkan rasio *dataset* terklasifikasi benar oleh sistem dari keseluruhan dokumen. Perbandingan nilai rerata (*mean*) akurasi *30-Stages 10-Cross Fold Validation* metode *ADASYN-N* serta *ADASYN-KNN* setiap *dataset* dengan metode klasifikasi *Random Forest* dapat dilihat pada gambar 2.1.



	Audiology	Balance Scale	Breast Cancer	Car	Lenses	Lymphography
NONE	77.0556	80.4194	68.3526	94.7379	75.7778	82.0032
ADASYN-N 5NN	98.2619	90.5419	84.4518	99.0152	93.4833	92.5495
ADASYN-N 9NN	98.3968	90.8675	84.0372	99.0917	93.4833	92.4176
ADASYN-5NN	87.2460	90.0414	69.9384	98.6480	89.0833	92.1337
ADASYN-9NN	87.2460	90.0414	69.9384	98.6480	89.0833	92.1337

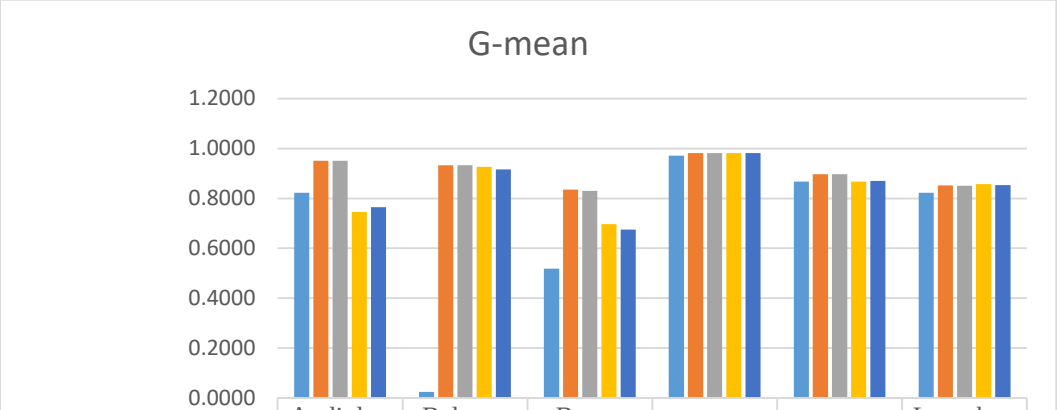
Gambar 1. Perbandingan Nilai Rerata Akurasi

Dari gambar 1 terlihat metode *ADASYN-N* dan *ADASYN-KNN* memberikan peningkatan akurasi dominan pada *dataset* kecil seperti *Audiology* serta *Lenses* akurasinya meningkat 10,19% dan 13,31% dari *Dataset* asli dengan metode *ADASYN-KNN* serta meningkat sekitar 21,34% dan 17,71% dengan metode *ADASYN-N*. Meskipun parameter k=5 memiliki nilai akurasi sedikit lebih baik, namun pemberian parameter berbeda masing-masing metode tidak terlalu mempengaruhi nilai akurasi dari metode itu sendiri.

Pada *dataset* lebih besar, peningkatan akurasinya tidak terlalu besar. Seperti pada *dataset Car*, akurasi metode *ADASYN-KNN* serta *ADASYN-N* hanya meningkat 3,91% dan sekitar 4,35% dari data asli. Hal ini disebabkan pada *dataset* besar, pola fragmentasi data masih terbentuk dengan baik meskipun dari kelas minoritas.

4.3 *G-Mean*

G-Mean menunjukkan evaluasi tingkat bias induktif dalam rasio akurasi positif dan akurasi negatif. Perbandingan nilai rerata (*mean*) *G-Mean 30-Stages 10-Cross Fold Validation* metode *ADASYN-N* serta *ADASYN-KNN* setiap *dataset* dengan metode klasifikasi *Random Forest* dilihat pada gambar 2.2.



	Audiology	Balance Scale	Breast Cancer	Car	Lenses	Lymphography
NONE	0.8231	0.0246	0.5187	0.9708	0.8675	0.8222
ADASYN-N 5NN	0.9509	0.9329	0.8358	0.9819	0.8970	0.8521
ADASYN-N 9NN	0.9503	0.9333	0.8302	0.9822	0.8970	0.8505
ADASYN-5NN	0.7461	0.9262	0.6966	0.9821	0.8675	0.8570
ADASYN-9NN	0.7648	0.9166	0.6750	0.9819	0.8703	0.8530

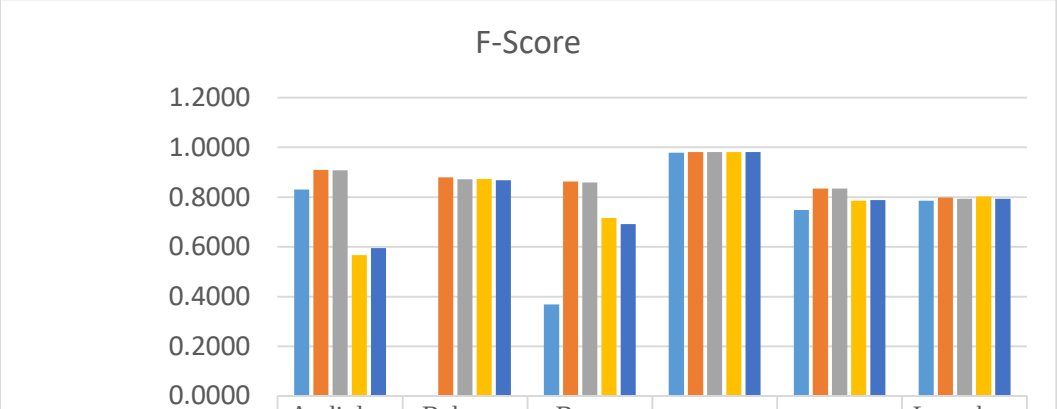
Gambar 2. Perbandingan Nilai Rerata *G-Mean*

Gambar 2. menunjukkan perbandingan rerata *G-Mean* semua *dataset* yang diuji, peningkatan nilai *G-Mean* paling dominan terjadi pada *dataset Balance Scale*, yaitu 90,8% dengan *ADASYN-N* serta 89,2% dengan *ADASYN-KNN*. Peningkatan dominan terjadi pada *dataset Breast Cancer* yang memiliki kelas biner, yaitu 31% dengan metode *ADASYN-N* serta 15,6% dengan *ADASYN-KNN* (k = 9) serta 17,8% dengan *ADASYN-KNN* (k = 5). Hal ini disebabkan data asli *dataset Balance Scale* dan *Breast Cancer* memiliki satu kelas minoritas, sehingga performa proses identifikasi kelas sangat kecil. Dengan adanya penyeimbangan kelas, proses identifikasi kelas pun menjadi lebih baik.

Rerata *G-Mean* dengan metode *ADASYN-KNN* justru menurun dari data asli pada *dataset Audiology* yang memiliki data kecil dengan distribusi kelas banyak. Hal ini dapat disebabkan proses *voting* untuk menghasilkan *instance* sintesis metode tersebut, sehingga justru menurunkan performa identifikasi kelas *dataset*. Secara garis besar nilai *G-Mean* seluruh data, metode *ADASYN-N* memberikan peningkatan proses identifikasi kelas lebih baik daripada metode *ADASYN-KNN*. Meskipun parameter k = 5 memberikan peningkatan nilai *G-Mean* sedikit lebih baik, namun pemberian parameter berbeda pada metode tidak memberikan pengaruh dominan terhadap nilai *G-Mean* dari metode itu sendiri.

4.4 *F-Score*

F-Score menunjukkan rerata tertimbang presisi dan *Recall*. Perbandingan nilai rerata (*mean*) *F-Score 30-Stages 10-Cross Fold Validation* setiap metode *ADASYN-N* serta metode *ADASYN-KNN* setiap *dataset* dengan metode klasifikasi *Random Forest* dapat dilihat pada gambar 3.



	Audiology	Balance Scale	Breast Cancer	Car	Lenses	Lymphography
NONE	0.8302	0.0031	0.3688	0.9793	0.7484	0.7855
ADASYN-N 5NN	0.9089	0.8799	0.8622	0.9812	0.8344	0.7979
ADASYN-N 9NN	0.9078	0.8715	0.8595	0.9816	0.8344	0.7930
ADASYN-5NN	0.5667	0.8727	0.7163	0.9819	0.7854	0.8025
ADASYN-9NN	0.5956	0.8683	0.6914	0.9809	0.7888	0.7935

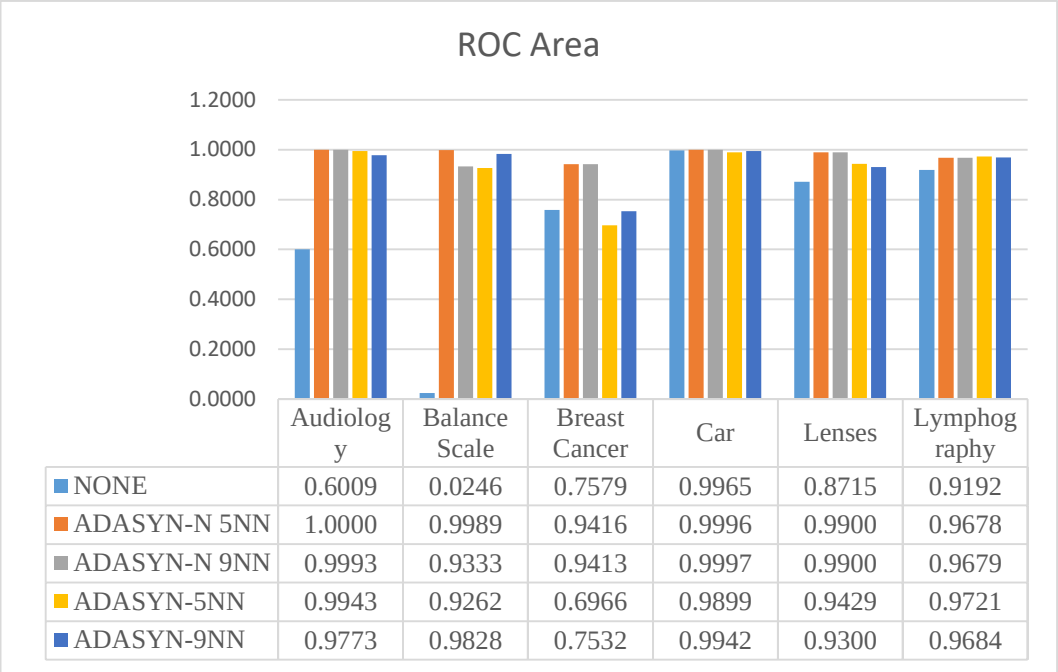
Gambar 3. Perbandingan Nilai Rerata *F-Score*

Dari perbandingan nilai rerata *F-Score* yang ditampilkan pada gambar 3, peningkatan nilai *F-Score* paling dominan terjadi pada *dataset Balance Scale*, yaitu 87% dengan metode *ADASYN-N* serta 86% dengan metode *ADASYN-KNN*. Peningkatan dominan terjadi pada *dataset Breast Cancer* yang memiliki kelas biner, yaitu 49% dengan metode *ADASYN-N*, 34% dengan *ADASYN-KNN* (k = 9), serta 32% dengan *ADASYN-KNN* (k = 5). Hal ini dapat disebabkan, data asli *dataset Balance Scale* dan *Breast Cancer* memiliki satu kelas minoritas, sehingga nilai *Recall* dan presisi tidak seimbang. Dengan adanya penyeimbangan kelas ini, akan meningkatkan keseimbangan nilai *Recall* serta presisi.

Rerata *F-Score* dengan metode *ADASYN-KNN* justru menurun dari data asli pada *dataset Audiology* yang memiliki data kecil dengan distribusi kelas banyak. Hal ini dapat disebabkan proses *voting* untuk menghasilkan *instance* sintesis pada metode tersebut, sehingga menurunkan keseimbangan *Recall* serta presisi pada *dataset*. Secara garis besar, nilai *F-Score* seluruh data, metode *ADASYN-N* memberikan peningkatan keseimbangan nilai *Recall* dan presisi lebih baik daripada metode *ADASYN-KNN*. Meskipun parameter k=5 memberikan peningkatan nilai *F-Score* yang lebih baik, namun pemberian parameter berbeda setiap metode tidak memberikan pengaruh dominan terhadap nilai *F-Score* dari metode itu sendiri.

4.5 *ROC Area*

ROC Area mengkuantifikasi keseluruhan performa *classifier* dengan distribusi kelas berbeda. Perbandingan nilai rerata (*mean*) *ROC Area 30-Stages 10-Cross Fold Validation* setiap metode *ADASYN-N* serta *ADASYN-KNN* setiap *dataset* dengan metode klasifikasi *Random Forest* dapat dilihat pada gambar 4.



	Audiology	Balance Scale	Breast Cancer	Car	Lenses	Lymphography
NONE	0.6009	0.0246	0.7579	0.9965	0.8715	0.9192
ADASYN-N 5NN	1.0000	0.9989	0.9416	0.9996	0.9900	0.9678
ADASYN-N 9NN	0.9993	0.9333	0.9413	0.9997	0.9900	0.9679
ADASYN-5NN	0.9943	0.9262	0.6966	0.9899	0.9429	0.9721
ADASYN-9NN	0.9773	0.9828	0.7532	0.9942	0.9300	0.9684

Gambar 4. Perbandingan Nilai Rerata *ROC Area*

Nilai *ROC Area* paling dominan terjadi pada *dataset Balance Scale* dengan peningkatan 97,4% dengan metode *ADASYN-N* (k = 5) serta 95,8% dengan metode *ADASYN-KNN* (k = 9). Hal ini disebabkan data asli *dataset Balance Scale* memiliki satu kelas minoritas dengan dua kelas mayoritas, sehingga performa klasifikasi *Random Forest* untuk *dataset* ini sangat rendah. Dengan adanya penyeimbangan data, performa klasifikasi *Random Forest* untuk *dataset* meningkat.

Peningkatan nilai *ROC Area* cukup dominan terlihat pada *dataset Audiology* yang memiliki *dataset* kecil dengan distribusi kelas banyak. peningkatan pada *dataset* ini sebesar 39,9% dengan metode *ADASYN-N*, 39,3% dengan metode *ADASYN-KNN* (k = 5), serta 37,6% dengan metode *ADASYN-KNN* (k = 9). Pada *dataset Breast Cancer* yang memiliki kelas biner, peningkatan nilai *ROC Area* cukup dominan dengan metode *ADASYN-N*, yaitu 18,3%. Sedangkan nilai *ROC Area* justru lebih rendah dari data asli tanpa *over-sampling* dengan metode *ADASYN-KNN*. Hal ini dapat disebabkan prosedur pemilihan data sintetis dari kedua metode menyebabkan pengaruh signifikan terhadap performa klasifikasi *Random Forest*. Secara garis besar dari nilai *ROC Area* pada seluruh data, metode *ADASYN-N* memberikan peningkatan performa klasifikasi lebih baik daripada metode *ADASYN-KNN*. Meskipun parameter k = 5 memberikan peningkatan nilai *ROC Area* sedikit lebih baik, namun pemberian parameter berbeda pada setiap metode tidak memberikan pengaruh dominan terhadap nilai *ROC Area* dari metode itu sendiri.

5. KESIMPULAN

Berdasarkan pengamatan serangkaian pengujian yang telah dilakukan, kesimpulan yang diperoleh adalah sebagai berikut:

1. Dengan metode klasifikasi *Random Forest*, metode ADASYN-N memberikan nilai akurasi lebih baik dari metode ADASYN-KNN pada semua *dataset* yang diuji dalam penelitian ini meskipun peningkatan nilai yang diberikan bervariasi antar *dataset* dengan jumlah *instances* berbeda. *Dataset* kecil seperti *Audiology* dan *Lenses* yang akurasinya meningkat 10,19% dan 13,31% dari *dataset* asli dengan metode ADASYN-KNN serta meningkat sekitar 21,34% dan 17,71% dengan metode ADASYN-

N. Sedangkan *dataset* besar seperti *Car*, peningkatan akurasi hanya 3,91% dengan metode ADASYN-KNN dan 4,35% dengan metode ADASYN-N, kedua metode tidak memberikan pengaruh besar terhadap nilai *G-Mean*, *F-Score*, dan *ROC Area*.

2. Nilai *G-Mean* dengan metode ADASYN-KNN tidak meningkat dari data asli tanpa *over-sampling* pada *dataset* kecil seperti *Lenses*. Bahkan nilai *G-Mean* dengan metode ADASYN-KNN menurun 7,7% dari data asli pada *dataset Audiology* yang memiliki distribusi kelas yang lebih banyak. Nilai *F-Score*, dan *ROC Area* metode ADASYN-N dan metode ADASYN-KNN juga memberikan nilai bervariasi dikarenakan *dataset* yang diuji dalam penelitian ini memiliki jumlah *instance*s dan distribusi kelas berbeda.
3. Metode ADASYN-N memberikan kinerja lebih baik daripada metode ADASYN-KNN pada *dataset* berfitur nominal dengan kategori dua dan/atau lebih, namun besar kecilnya peningkatan kinerja dipengaruhi jumlah *instance*s dan banyak distribusi kelas pada *dataset*. Meskipun parameter k=5 memiliki nilai akurasi yang lebih baik, namun pemberian parameter berbeda masing-masing metode tidak terlalu mempengaruhi nilai akurasi dari metode itu sendiri.

Berdasarkan kesimpulan, disarankan penelitian selanjutnya melakukan eksplorasi algoritma ADASYN-N ataupun ADASYN-KNN untuk *dataset* yang memiliki fitur dengan tipe data kombinasi, numeris dan nominal serta *dataset* dengan jumlah *instance*s besar. Selain itu, kedua metode ini dapat diberikan parameter k yang lebih bervariasi, dikombinasikan dengan metode *sampling* lain dan dapat diklasifikasi dengan metode selain *Random Forest*.

DAFTAR PUSTAKA

[1] A. Shipitsyn, "*Statistical Learning with Imbalanced Data*," Linköping University.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: *Synthetic Minority Over-sampling Technique*," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[3] H. He and Y. Ma, *Imbalanced* Learning: Foundations, Algorithms, and Applications, 1st ed. Wiley-IEEE Press, 2013.

[4] Y. E. Kurniawati, "*Multi-Class Imbalance Learning* dengan Adaptive Synthetic – Nominal (ADASYN-N) dan Adaptive Synthetic – KNN (ADASYN-KNN) untuk Re*sampling* Data pada Data Hasil Tes Pap Smear," Universitas Gadjah Mada, 2017.

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adaptive Synthetic *Sampling* Approach for *Imbalanced* Learning," Int. Jt. Conf. Neural Networks, no. 3, pp. 1322–1328, 2008.

[6] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class *imbalance*," Knowledge-Based Syst., vol. 25, no. 1, pp. 13–21, 2012.

[7] P. J. Huang, "*Classification of Imbalanced Data Using Synthetic Over-Sampling Techniques*," University of California, 2015.