IJCIT (Indonesian Journal on Computer and Information Technology) 6 (1) (2021) 31-37
Diterima 8 Desembet 2020; Revisi 11 Februari 2021; Disetujui 24 Februari 2021

IJCIT
(Indonesian Journal on Computer and Information Technology)
Journal Homepage: http://ejournal.bsi.ac.id/ejurnal/index.php/ijcit

Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting

Nana Suryana[1], Pratiwi[2], Rizki Tri Prasetio[3]

[1,2]Sistem Informasi, Universitas Kebangsaan
Bandung, Indonesia
e-mail: nsuryana.dpt@gmail.com[1], pratiwi.wiie@gmail.com[2]

[3]Sistem Informasi, Universitas Adhirajasa Reswara Sanjaya
Bandung, Indonesia
e-mail: rizki@ars.ac.id

A B S T R A K

Industri telekomunikasi menghadapi persaingan yang ketat antara penyedia layanan (*service provider*). Persaingan ini mengakibatkan *customer churn* atau berpindahnya pelanggan dari satu layanan ke layanan lain. *Customer churn* menjadi masalah utama karena dapat mempengaruhi pendapatan perusahaan, profitabilitas, serta kelangsungan hidup perusahaan. Oleh karena itu, mengetahui pelanggan yang akan melakukan *churn* secara dini menjadi salah satu cara yang cukup efektif dilakukan, karena dapat membantu perusahaan dalam membuat rencana yang efektif untuk tetap mempertahankan pelanggannya. Jumlah pelanggan yang mengundurkan diri dari layanannya saat ini biasanya dimiliki perusahaan dalam jumlah yang sedikit. Kondisi kekurangan data ini menyebabkan kesulitan dalam memprediksi *customer churn*. Tujuan umum dari penelitian ini adalah memprediksi pelanggan yang akan berpindah ke layanan lain atau mengundurkan diri dari layanannya saat ini. Sementara tujuan khusus penelitian Penelitian ini berusaha menangani ketidakseimbangan data dalam prediksi *customer churn* menggunakan optimasi pada level data melalui metode sampling yaitu *Synthetic Minority Over Sampling*. Kemudian dikombinasikan dengan optimasi level algoritma melalui pendekatan teknik *Boosting*. Pada penelitian beberapa algoritma prediksi seperti *random forest, naïve bayes, decision tree, k-nearest neighbor* dan *deep learning* yang akan diimplementasikan untuk mengetahui algoritma yang paling baik setelah dilakukan optimasi menggunakan SMOTE dan *Boosting*. Metode penelitian yang digunakan pada penelitian ini adalah CRISP-DM, yang merupakan kerangka penelitian data mining untuk penelitian lintas industri. Hasil penelitian ini menunjukan bahwa algoritma *random forest* merupakan algoritma yang menghasilkan akurasi paling optimal setelah dioptimasi menggunakan SMOTE dan Boosting dengan hasil akurasi 89,19%.

Katakunci: *customer churn*, ketidakseimbangan data, SMOTE, b*oosting*

A B S T R A C T S

The telecommunications industry faces stiff competition between service providers. This competition results in customer churn. Customer churn is a major problem because it can affect company revenue, profitability, survival, and service quality of the company. Therefore, knowing which customers will churn in the future early is one of the most effective ways to do it, because it can help companies make an effective plan to keep their customers. The number of customers who withdrew from its current services is usually owned by a small number. This lack of data causes difficulties in predicting customer churn. This problem then becomes a challenging issue in machine learning. The general purpose of this research is to

predict customers who will churn. While the specific purpose of this research is to try to deal with data imbalances in predicting customer churn using optimization at the data level through the sampling method, namely Synthetic Minority Over Sampling (SMOTE). Then combined with algorithm level optimization through the Boosting technique approach. In this study, several prediction algorithms like the random forest, naïve Bayes, decision tree, k-nearest neighbor, and deep learning will be implemented to find out the best algorithm after optimization using SMOTE and Boosting. The method used in this study is CRISP-DM, which is a data mining research framework for cross-industry research. The results of this study indicate that the random forest algorithm is an algorithm that produces the most optimal accuracy after being optimized using SMOTE and Boosting with an accuracy of 89.19%.
***Keywords**: customer churn, imbalance data, SMOTE, boosting*

1. PENDAHULUAN

Pelanggan merupakan bagian paling esensial dari bisnis telekomunikasi karena pelanggan merupakan sumber pendapatan utama. Banyaknya penyedia layanan telekomunikasi menciptakan peluang *customer churn* atau pelanggan yang berpindah dari satu penyedia layanan ke penyedia layanan lain (Dalvi, Khandge, Deomore, Bankar, & Kanade, 2016) sehingga menjadi masalah utama yang harus diselesaikan karena akan mengurangi pendapatan perusahaan. Penelitian prediksi *customer churn* dilakukan dengan mengklasifikasikan pelanggan yang berpotensi untuk berpindah layanan, sehingga pada akhirnya akan menunjukan apakah pelanggan akan berpindah layanan atau tidak, sehingga perusahaan dapat mengambil tindakan untuk menghindarinya.

Beberapa algoritma klasifikasi telah diterapkan pada penelitian terdahulu seperti *logistic regression*, *naïve bayes*, *decision tree*, *neural network* serta *support vector machine*, akan tetapi algoritma individual tidak dapat menghasilkan prediksi yang akurat (Xiao, Jiang, He, & Teng, 2016). Algoritma tersebut termasuk kedalam algoritma yang populer, efektif serta menghasilkan prediksi yang relatif baik (Riana, Ramdhani, Prasetio, & Hidayanto, 2018), akan tetapi algortima individual tersebut tidak memperhitungkan adanya ketidakseimbangan data yang dapat berpengaruh pada berkurangnya performa prediksi serta meningkatkan bias pada kelas yang memiliki data lebih banyak (*majority class*) (Dittman, Khoshgoftaar, & Napolitano, 2015).

Dengan demikian, penanganan terhadap ketidakseimbangan data dalam prediksi *customer churn* merupakan hal yang penting. Dalam penelitian ini, churn data atau pelanggan yang berpindah layanan sangatlah sedikit dengan persentase sekitar 2% dari sampel data.

Secara umum ketidakseimbangan kelas dapat ditangani dengan dua pendekatan, yaitu level data dan level algoritma (Prasetio & Riana, 2015). Pendekatan level algoritma dilakukan dengan memperbaiki algoritma atau menggabungkan (*ensemble*) pengklasifikasi tunggal agar menjadi lebih baik (Saifudin & Wahono, 2015). Metode ensemble seperti *bagging* dan *boosting* merupakan metode lain yang digunakan secara luas untuk menangani masalah ketidakseimbang kelas (Zhongbin, et al., 2015).

Pada level data, metode utamanya adalah memperbaiki dataset agar memiliki jumlah sample yang merata dengan cara menambahkan atau mengurangi sample yang sudah ada. Optimasi level data dapat digunakan beberapa teknik yang dapat digunakan, diantaranya *Synthetic Minority Over Sampling* (SMOTE), *Random Over Sampling* (ROS) dan *Random Under Sampling* (RUS). SMOTE dan ROS meningkatan jumlah kelas minoritas sehingga jumlah sampel dapat berimbang. SMOTE memiliki waktu prediksi yang lebih lambat, akan tetapi lebih optimal untuk mengatasi ketidakseimbangan kelas (Park, Oh, & Pedryez, 2013). Teknik SMOTE juga lebih mudah dan cepat diaplikasikan dibandingkan metode lain (Yu, Hu, Tang, Shen, & Yang, 2013).

Beberapa penelitian sebelumnya telah dilakukan untuk memprediksi *customer churn* serta upaya mengatasi ketidakseimbangan data yang dialaminya (Lariviere & Poel, 2005), menginvestigasi bagaimana membuat data yang lebih seimbang dalam prediksi churn, menerapkan empat dari 10 metode yaitu random sampling, advance under sampling, *boosting* dan *cost-sensitive learner*. Hasil penelitiannya menunjukan bahwa under sampling menghasilkan performa prediksi yang lebih baik.

Mengkombinasikan sampling dengan *weighted random forest* (Effendy, Adiwijaya, & Baizal, 2014) juga diusulkan untuk mengatasi

ketidakseimbangan data. Sampling yang digunakan pada penelitian ini adalah SMOTE yang digunakan untuk menghasilkan data sintetis untuk menyeimbangkan dua kelas (*churn* dan *non churn*). Penelitian ini menghasilkan prediksi yang lebih optimal dari under sampling.

Dua referensi tersebut mencoba menyelesaikan ketidakseimbangan data dengan mengubah distribusi data. Beberapa peneliti lain mencoba teknik *ensemble* untuk meningkatkan performa algoritma prediksi (Prasetio & Pratiwi, 2015). Teknik *bagging* dan *boosting* yang diterapkan oleh (Galar & Fernandez, 2011) menghasilkan hasil prediksi yang lebih baik jika dibandingkan teknik sampling.

Terdapat juga beberapa peneliti yang mengkombinasikan teknik sampling dan *ensemble* guna lebih mengoptimalkan hasil prediksi *customer churn*. Teknik *random under sampling* dan *boosting* diterapkan oleh (Dwiyanti, Adiwijaya, & Ardiyanti, 2016) dan (Awalludin, Adiwijaya, & Bijaksana, 2017) yang mengkombinasikan SMOTE dan *Boosting* pada algoritma C4.5. kedua referensi tersebut berhasil menghasilkan prediksi yang lebih baik dari penelitian menggunakan sampling dan *ensemble*.

Berdasarkan uraian tersebut maka tujuan khusus dari penelitian ini adalah mengatasi masalah ketidakseimbangan data pada *customer churn* menggunakan kombinasi teknik sampling SMOTE dan teknik *ensemble boosting* yang diterapkan pada beberapa algoritma prediksi. Manfaat dari penelitian ini memberikan gambaran umum bagi perusahaan tentang kemungkinan pelanggan yang akan beralih.

2. METODE PENELITIAN

Pendekatan resampling dibagi menjadi tiga kategori: metode *over-sampling*, *undersampling*, dan hibrida yang menggabungkan kedua pendekatan sampling (Jian, Gao, & Ao, 2016). Salah satu teknik resampling yang umum digunakan yaitu undersampling yang secara acak memilih sampel di kelas mayoritas dan menambahkannya ke kelas minoritas, membentuk sebuah dataset pelatihan baru. Oversampling bertujuan untuk meningkatkan sampel kelas minoritas sampai sama dengan kelas mayoritas lain dengan menduplikasi secara acak sampel kelas minoritas, namun duplikasi data tetap dilakukan dengan tepat dan relevan (He, Zhang, & Zhang, 2018).

Synthetic Minority Oversampling Technique(SMOTE) adalah salah satu metode oversampling yang bekerja dengan meningkatkan jumlah kelas positif melalui replikasi data secara acak, sehingga jumlah data positif sama dengan data negatif. Cara menggunakan data sintetis adalah dengan mereplika data pada kelas yang kecil. Algoritma SMOTE bekerja dengan mencari tetangga terdekat k untuk kelas positif, kemudian membangun duplikasi data sintetis sebanyak persentase yang diinginkan antara kelas k yang dipilih secara acak dan positif. Dengan metode ini, dinilai dapat mengatasi masalah ketidakseimbangan data.

Boosting merupakan meta-algoritma dalam *machine learning* untuk melakukan *supervised learning* (Prasetio & Susanti, 2019). Kebanyakan algoritma *boosting* mengikuti sebuah rancangan. Secara umum *boosting* terjadi dalam iterasi, secacra incremental menambahkan *weak learner* ke dalam strong learner. Pada setiap iterasi, satu *weak learner* belajar dari suatu data latihan. Kemudian, *weak learner* ditambahkan ke dalam strong learner. Setelah *weak learner* ditambahkan, data-data kemudian diubah masing-masing bobotnya. Data-data yang mengalami kesalahan klasifikasi akan mengalami penambahan bobot, dan data-data yang terklasifikasi dengan benar akan mengalami penguranan bobot.

AdaBoost merupakan yang teknik *boosting* paling terkenal dan dalam sejarah perkembangannya, merupakan algoritma pertama yang dapat beradaptasi dengan *weak learner*. AdaBoost merupakan salah satu metode *boosting* yang dapat meningkatkan ketelitian dalam proses klasifikasi dan prediksi dengan cara membangkitkan kombinasi dari suatu model, tetapi hasil klasifikasi dan prediksi yang dipilih adalah model yang memiliki nilai bobot paling besar (Zieba, Tomzcak, Lubicz, & Swiatek, 2014).

Penelitian ini mengkombinasikan optimasi level data menggunakan SMOTE dan optimasi level algoritma menggunakan AdaBoost untuk mengatasi ketidakseimbangan data pada dataset *customer churn*. Metode yang diusulkan ini akan diterapkan pada beberapa algoritma klasifikasi diantaranya *random forest, naive bayes, decision tree, k-nearest neighbor* dan *deep learning*. Algoritma tersebut dipilih berdasarkan algoritma yang telah diteliti pada penelitian sebelumnya.

3. HASIL DAN PEMBAHASAN

Hasil dalam penelitian dilakukan dalam empat eksperimen yaitu eksperimen terhadap

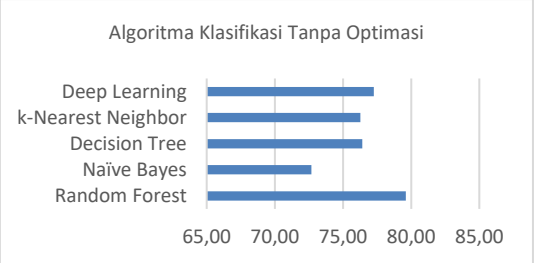
algoritma klasifikasi(*random forest, naïve bayes, decision tree, deep learning* dan *k-nearest neighbour*) sederhana tanpa optimasi, algortima klasifikasi dengan *upsampling* menggunakan SMOTE, dengan *boosting* menggunakan ADABoost dan kombinasi *upsampling* menggunakan SMOTE dikombinasikan dengan *boosting* menggunakan ADABoost pada algoritma klasifikasi.

Eksperimen pertama menggunakan algoritma klasifikasi sederhana tanpa dilakukan optimasi baik pada dataset maupun pada algoritma. Eksperimen pertama dilakukan sebanyak lima kali sejumlah algoritma klasifikasi yang digunakan. Validasi yang digunakan pada eksperimen ini menggunakaan *cross validation* dengan jumlah *fold* sebanyak 10. Evaluasi menggunakan *confussion matrix*.

Hasil eksperimen pada Tabel 1 menunjukan bahwa akurasi yang dihasilkan dari algoritma *random forest* merupakan akurasi terbaik yang didapatkan pada eksperimen tanpa optimasi dengan akurasi 79,62%. Sementara akurasi terendah diperoleh algoritma naïve bayes dengan akurasi 72,67%. Rata-rata akurasi yang dapat dihasilkan oleh algoritma klasifikasi tanpa optimasi sebesar 76,45%. Perbandingan akurasi dapat dilihat pada Gambar 1.

Tabel 1. Hasil Eksperimen Algoritma Individual

Algoritma	Akurasi
Random Forest	**79,62%**
Naïve Bayes	72,67%
Decision Tree	76,42%
k-Nearest Neighbor	76,27%
Deep Learning	77,25%



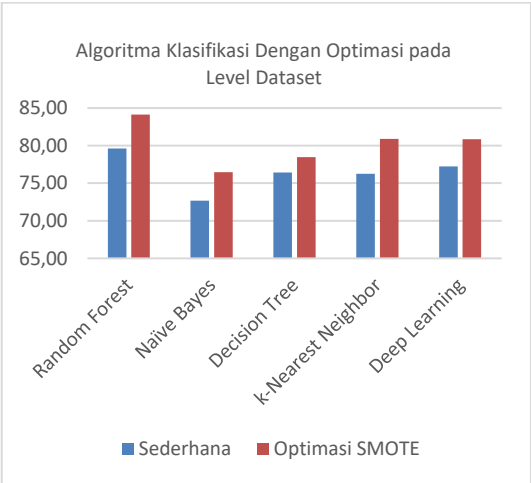
Gambar 1. Grafik Perbandingan Performa Algoritma Individual

Eksperimen kedua menggunakan algoritma klasifikasi dikombinasikan dengan optimasi pada level dataset menggunakan metode *upsampling* menggunakan SMOTE. Eksperimen kedua dilakukan sebanyak lima kali sejumlah algoritma klasifikasi yang digunakan yakni, *random forest, naïve bayes, decision tree, k-nearest neighbor* dan *deep learning* yang telah dikombinasikan dengan optimasi level dataset. Validasi yang digunakan pada eksperimen ini menggunakaan *cross validation* dengan jumlah *fold* sebanyak 10. Indikator untuk mengetahui hasil terbaik ditunjukan oleh besarnya nilai akurasi yang dihitung menggunakan *confussion matrix* untuk masing-masing algoritma klasifikasi.

Hasil eksperimen pada Tabel 2 menunjukan bahwa akurasi yang dihasilkan dari algoritma *random forest* merupakan akurasi terbaik yang didapatkan pada eksperimen dengan optimasi pada level dataset menggunakan SMOTE dengan akurasi 84,13%. Sementara akurasi terendah diperoleh algoritma naïve bayes dengan akurasi 76,48%. Rata-rata akurasi yang dapat dihasilkan oleh algoritma klasifikasi tanpa optimasi sebesar 80,17%. Perbandingan akurasi dapat dilihat pada Gambar 2.

Tabel 2. Hasil Eksperimen Algoritma dengan Optimasi pada Level Dataset

Algoritma	Akurasi	
	Individual	Optimasi SMOTE
Random Forest	**79,62%**	**84,13%**
Naïve Bayes	72,67%	76,48%
Decision Tree	76,42%	78,48%
k-Nearest Neighbor	76,27%	80,89%
Deep Learning	77,25%	80,87%



Gambar 2. Grafik Perbandingan Performa Algoritma dengan Optimasi pada Level Dataset

Eksperimen ketiga menggunakan algoritma klasifikasi dikombinasikan dengan optimasi pada level algoritma menggunakan metode *boosting*

menggunakan algoritma AdaBoost. Eksperimen ketiga dilakukan sebanyak lima kali sejumlah algoritma klasifikasi yang digunakan dengan dikombinasikan dengan optimasi level algoritma. Validasi 10 *fold cross validation* digunakan pada eksperimen ini.

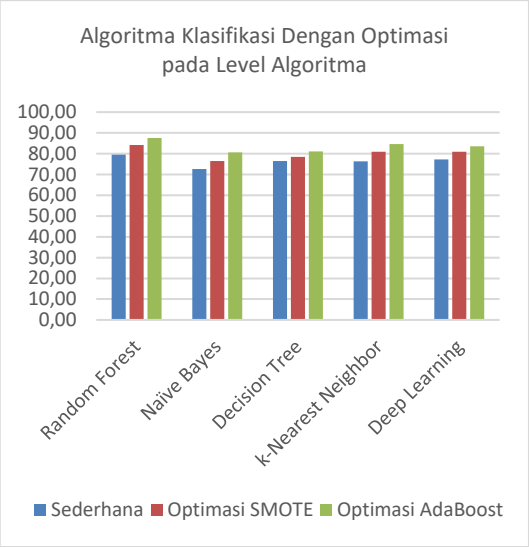
Hasil eksperimen pada Tabel 3 menunjukan bahwa akurasi yang dihasilkan dari algoritma *random forest* merupakan akurasi terbaik yang didapatkan pada eksperimen dengan optimasi pada level algoritma menggunakan AdaBoost dengan akurasi 87,52%. Sementara akurasi terendah diperoleh algoritma naïve bayes dengan akurasi 80,58%. Rata-rata akurasi yang dapat dihasilkan oleh algoritma klasifikasi tanpa optimasi sebesar 83,48%. Perbandingan akurasi dapat dilihat pada Gambar 3.

Eksperimen terakhir menggunakan algoritma klasifikasi dikombinasikan dengan optimasi pada level algoritma menggunakan metode *boosting* menggunakan algoritma AdaBoost dan optimasi pada level data menggunakan metode *upsampling* menggunakan SMOTE. Eksperimen ketiga dilakukan sebanyak lima kali sejumlah algoritma klasifikasi yang digunakan yang telah dikombinasikan dengan optimasi level algoritma dan level data. Validasi yang digunakan pada eksperimen ini menggunakaan *cross validation* dengan jumlah *fold* sebanyak 10. Indikator untuk mengetahui hasil terbaik ditunjukan oleh besarnya nilai akurasi yang dihitung menggunakan *confussion matrix* untuk masing-masing algoritma klasifikasi.

Hasil eksperimen pada Tabel 4 menunjukan bahwa akurasi yang dihasilkan dari algoritma *random forest* merupakan akurasi terbaik yang didapatkan pada eksperimen dengan metode yang diusulkan dengan akurasi hingga 89,19%. Sementara akurasi terendah diperoleh algoritma naïve bayes dengan akurasi 82,10%. Rata-rata akurasi yang dapat dihasilkan oleh algoritma klasifikasi tanpa optimasi sebesar 85,97%. Perbandingan akurasi dapat dilihat pada Gambar 4.

Tabel 3. Hasil Eksperimen Algoritma dengan Optimasi pada Level Algoritma

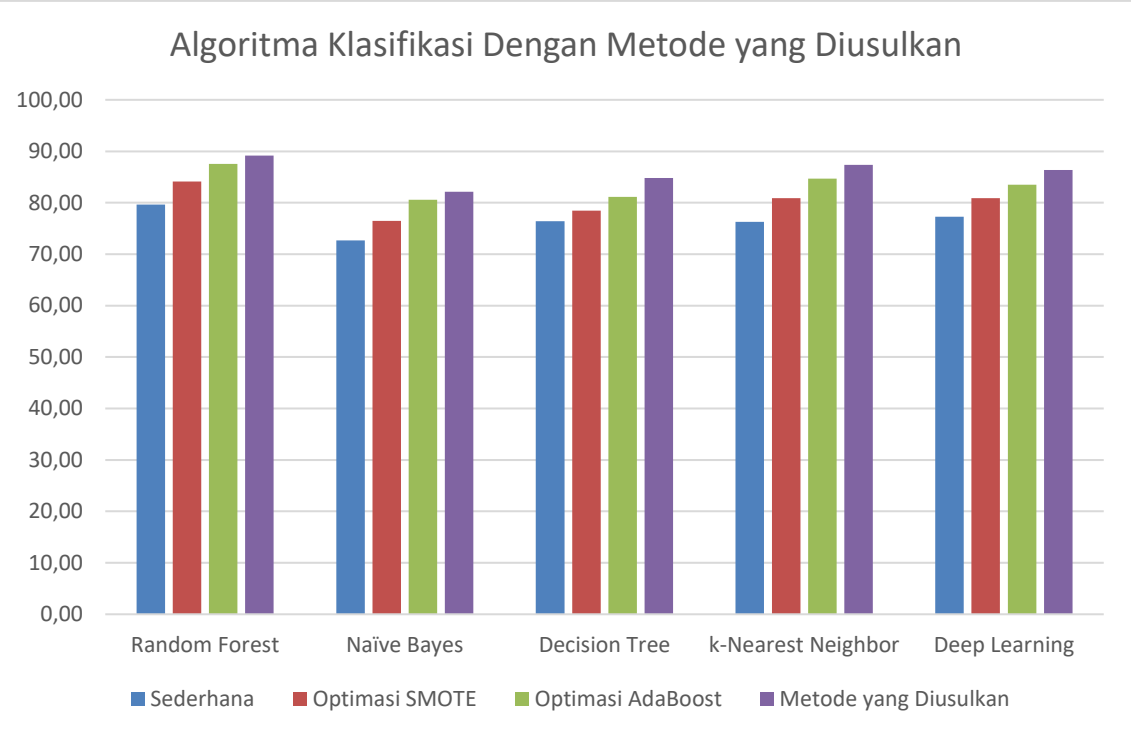
Algoritma	Akurasi		
	Individual	SMOTE	AdaBoost
Random Forest	79,62%	84,13%	87,52
Naïve Bayes	72,67%	76,48%	80,58
Decision Tree	76,42%	78,48%	81,12
k-Nearest Neighbor	76,27%	80,89%	84,68
Deep Learning	77,25%	80,87%	83,52



Gambar 3. Grafik Perbandingan Performa Algoritma dengan Optimasi pada Level Algoritma

Tabel 4. Hasil Eksperimen dengan Metode yang Diusulkan

Algoritma	Akurasi			
	Individual	Optimasi SMOTE	Optimasi AdaBoost	Metode yang Diusulkan
Random Forest	79,62%	84,13%	87,52	89,19
Naïve Bayes	72,67%	76,48%	80,58	82,10
Decision Tree	76,42%	78,48%	81,12	84,81
k-Nearest Neighbor	76,27%	80,89%	84,68	87,39
Deep Learning	77,25%	80,87%	83,52	86,37



Gambar 4. Grafik Perbandingan Performa Menggunakan Metode yang Diusulkan

4. KESIMPULAN

Berdasarkan hasil eksperimen yang telah dilakukan sebanyak empat kali, dapat disimpulkan bahwa algoritma individual hanya mampu memberikan akurasi paling baik sebesar 79,62% yang diperoleh oleh algoritma *random forest.* Kemudian dilakukan optimasi individu baik optimasi level data maupun optimasi level algoritma terbukti dapat meningkatkan hampir seluruh algoritma klasifikasi yang digunakan pada eksperimen.

Secara umum, melalui optimasi level data menggunakan SMOTE mampu meningkatkan kemampuan algoritma klasifikasi untuk mengatasi ketidakseimbangan kelas dengan peningkatan akurasi rata-rata hingga 3%. Melalui optimasi level algoritma menggunakan AdaBoost, peningkatan kemampuan algoritma klasifikasi untuk mengatasi ketidakseimbangan kelas mampu meningkat drastis dengan peningkatan akurasi rata-rata hingga 8%.

Kombinasi optimasi level data dan level algoritma yang dilakukan pada eksperimen menggunakan metode yang diusulkan mampu menghasilkan akurasi yang sangat memuaskan dengan peningkatan akurasi rata-rata hingga 11% dengan akurasi tertinggi diperoleh algoritma *random forest* dengan akurasi sebesar 89,19%.

5. REFERENSI

Awalludin, Adiwijaya, & Bijaksana, M. (2017). Churn Prediction on Fix Broadband Internet Using Combined Feed Forward Neural Network and SMOTEBoost Algorithm.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, P. V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *IEEE- Symposium on Colossal Data Analysis and Networking (CDAN)*.

Dittman, D. J., Khoshgoftaar, T. M., & Napolitano, A. (2015). The effect of data sampling when using random forest on imbalanced bioinformatics data. *IEEE 16th International Conference on Information Reuse and Integration*.

Dwiyanti, E., Adiwijaya, & Ardiyanti, A. (2016). Handling Imbalanced Data in Churn Prediction Using RUSBoost and Feature Selection. *International Conference Soft Computeing and Data Mining*.

Effendy, V., Adiwijaya, & Baizal, Z. A. (2014). Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. *Information and Communication Technology*.

Galar, M., & Fernandez, A. (2011). A review on ensembles for the class imbalance problem : bagging-, boosting-, and hybrid-based approaches. *IEEE Transc. On System, MAN and Cybernetics-Part C: Application and Review*.

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*.

Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*.

Lariviere, B., & Poel, D. V. (2005). Predicting Customer Retention and Profitability by Using Random Forest and Regression Forest Techniques. *Expert System and Applications*.

Park, B. J., Oh, S. K., & Pedryez, W. (2013). The Design of Polynomial Function-Based Neural Network Predictors for Detection of Software Defects. *Information Sciences*.

Prasetio, R. T., & Pratiwi, P. (2015). PENERAPAN TEKNIK BAGGING PADA ALGORITMA KLASIFIKASI UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS DATASET MEDIS. *Jurnal Informatika*.

Prasetio, R. T., & Riana, D. (2015). A comparison of classification methods in vertebral column disorder with the application of genetic algorithm and bagging. *2015 4th international conference on instrumentation, communications, information technology, and biomedical engineering (ICICI-BME)* (hal. 163-168). Bandung: IEEE.

Prasetio, R. T., & Susanti, S. (2019). Implementasi Algoritma Genetika pada k-nearest neighbours untuk Klasifikasi Kerusakan Tulang Belakang. *Jurnal Responsif*, 64-69.

Riana, D., Ramdhani, Y., Prasetio, R. T., & Hidayanto, A. N. (2018). Improving Hierarchical Decision Approach for Single Image Classification of Pap Smear. *International Journal of Electrical and Computer Engineering*.

Saifudin, A., & Wahono, R. S. (2015). Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. *Journal of Software Engineering*.

Xiao, J., Jiang, X., He, C., & Teng, G. (2016). Churn prediction in customer relationship management via gmdh-based multiple classifers ensemble. *IEEE Computer Society*.

Yu, D., Hu, J., Tang, Z., Shen, H., & Yang, J. (2013). Neurocomputing Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing*.

Zhongbin, S., Qinbao, S., Xiaoyan, Z., Heli, S., Baowen, X., & Yuming, Z. (2015). A novel ensemble method for classifying imbalanced data . *Elsevier Pattern Recognition*.

Zieba, M., Tomzcak, J. M., Lubicz, M., & Swiatek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*.