**e-ISSN: 2964-2531**
**DOI : 10.36499/psnst.v12i1.6997** **p-ISSN: 2964-5131**

# SINTESIS *FITUR DENSITY BASED FEATURE SELECTION* (DBFS) DAN *ADABOOTS* DENGAN *XGBOOST* UNTUK MENINGKATKAN PERFORMA MODEL PREDIKSI

**Slamet Sudaryanto N[1] dan Sudaryanto[1]**
Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang
Jl. Imam Bonjol No. 2015-207, Kota Semarang 50131
*Email : slametalica301@dsn.dinus.ac.id, msdr8047@dsn.dinus.ac.id

**Abstrak**

*Ketidak seimbangan kelas akan menghasilkan akurasi prediksi yang baik pada kelas mayoritas tetapi menjadi tidak konduktif dalam memprediksi kelas minoritas, sehingga nilai hasil akurasi pengklasifikasian (classifier) menjadi tidak optimal. Masalah ketidakseimbangan kelas secara umum dapat ditangani dengan dua pendekatan, yaitu level data dan level algoritma. Pendekatan level data ditujukan untuk memperbaiki keseimbangan kelas, sedangkan pendekatan level algoritma ditujukan untuk memperbaiki algoritma atau menggabungkan (ensemble) pengklasifikasi agar lebih konduktif terhadap kelas minoritas. Beberapa metode telah diusulkan para peneliti untuk memecahkan masalah tersebut seperti metode smote, sampling, cost-sensitive learning, bagging dan boosting. Kebanyakan metode yang dikembangkan hanya pada salah satu level data atau pada level algoritma saja. Maka pada penelitian ini, akan dilakukan kombinasi ensemble baik pada level data maupun pada level algoritma. Pada level data akan menggabungkan metode seleksi fitur (yaitu algoritma Adaptive Boosting (Adaboost) dan metode Density Based Feature Selection (DBFS). Sedangkan pada level algoritma menggunakan salah satu model ensemble klasifikasi XGBoost. Model kombinasi ensemble baik dari level data maupun pada level algoritma tersebut digunakan untuk menagani ketidak seimbangan kelas agar didapatkan performa model prediksi. Penerapan algoritma adaboost dalam seleksi fitur dilakukan untuk memberi bobot pada setiap fitur yang direkomendasikan, sehingga ditemukan fitur yang merupakan classifier yang kuat. Algoritma DBFS berfokus dalam mengidentifikasi kelas minoritas dan mengevaluasi dampak dari sebuah fitur yang bermanfaat berdasarkan rangking fitur. Hasil dari penggabungan (ensemble) kedua algoritma tersebut adalah dataset yang seimbang untuk selanjutnya disintesiskan dengan algoritma XGBoost dalam melakukan perhitungan model prediksi. Hasil prediksi akan di evaluasi dengan confusion matrix dan AUC-ROC.*

***Kata kunci:*** *Ensemble, DBFS, AdaBoost, XGBoost, Confusion Matrix, AUC-ROC*

.

## 1. PENDAHULUAN

Data dengan kelas tidak seimbang (*class imbalance*) merupakan keadaan dataset dimana distribusi kelas yang ada dari dataset tidak terdistribusi secara seimbang. Dengan jumlah kelas data (*instance*) yang satu lebih sedikit atau lebih banyak dibandingkan dengan jumlah kelas data lainnya. Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas, sedangkan kelompok data yang lebih banyak disebut kelompok mayoritas. Biasanya, pada level data akan dijalankan sebagai langkah pra-pemrosesan untuk memastikan dataset yang tidak seimbang dapat disesuaikan dan kemiringan distribusi dapat dikurangi, dengan menggunakan semua jenis metode pengambilan sampel. Salah satu pendekatan paling umum untuk mengurangi dampak negatif ketidakseimbangan data adalah pra-pemrosesan dataset asli dengan strategi tingkat data seperti *undersampling* dan *oversampling*. Pendekatan *undersampling* dan *oversampling* tersebut adalah teknik standar yang digunakan dalam menangani kelas yang tidak seimbang. Pendekatan tersebut masih tentunya bukan pendekatan yang sempurna karena masih ditemukan keterbatasan dari masing-masing metode tersebut. Kelamahan pada pendekatan undersampling menyebabkan lebih banyak penghapusan sampel data yang pada akhirnya menyebabkan masalah kekurangan data, dengan peningkatan kemungkinan kehilangan data penting. Kemudian kelemahan pada *oversampling* adalah menyebabkan duplikasi data asli, sehingga menyebabkan *overfitting* kelas minoritas. *Overfitting* adalah masalah di mana perilaku pengklasifikasi terlalu sesuai dengan data percobaan. Ini berdampak buruk pada kinerja testing data, karena tidak semua informasi

dalam data percobaan berguna. Problem umum suatu algoritma klasifikasi pada machine learning adalah kelas pada dataset yang memiliki distribusi kelas yang tidak seimbang (*class imbalance*) menyebabkan terjadinya akurasi klasifikasi lebih condong ke kelas mayoritas daripada kelas minoritas. Ketidakseimbangan kelas dalam sekumpulan dataset merupakan sebuah permasalahan utama pada machine learning, dimana jumlah kelas mayoritas (negatif) lebih besar daripada jumlah kelas minoritas (positif) (Gonzalez-Abril et al., 2014). Dampak negatif dari ketidakseimbangan data pada kinerja pengklasifikasi lebih diperburuk oleh faktor aliran data yang tidak seimbang melekat seperti tumpang tindih kelas, terputus-putus kecil, adanya kebisingan, dan jumlah pengamatan pelatihan yang tidak mencukupi (Zyblewski et al., 2021). *Class imbalance* muncul terkait dengan evolusi pengembangan ilmu pengetahuan menjadi teknologi terapan dalam *machine learning*. Masalah *class imbalance* adalah permasalahan yang umum ditemukan pada kumpulan dataset dari berbagai bidang, termasuk aliran data yang tidak seimbang dari transaksi onlines (Czarnowski, 2022).

Dalam banyak aplikasi industri terdapat beberapa kasus dimana dataset diakitkan dengan label kelas yang tidak seimbang dalam kumpulan data training. Kumpulan data yang tidak seimbang merupakan akibat dari distribusi dari dataset yang didominasi oleh satu kelas tertentu. Ketika kelas mayoritas positif melebihi kelas lain (kelas mayoritas atau negatif) maka dapat diasumsilan sebagi input yang tidak seimbang sehingga dapat mengurangi keakuratan prediksi kelas. Sebelum menganalisa data memerlukan teknik pemrosesan data yang tepat agar menghasilakn data input untuk model machine learning yang seimbang. Teknik *downsampling* mrupakan strategi yang digunakan untuk meningkatkan jumlah sampel dalam minoritas atau mengurangi jumlah sampel secara mayoritas. Dengan cara memilih sampel yang paling informatif dalam kumpulan data tidak seimbang yang diberikan melalui strategi pembelajaran aktif untuk mengurangi efek label kelas yang tidak seimbang. Sebagai catatan penting bahwa sampel informatif dapat berupa minoritas atau mayoritas, bukan hanya memilih sampel mayoritas. Hasil dari pengembangan algorima menunjukkan kinerja yang lebih baik dibandingkan dengan metode *resampling* lainnya dengan ukuran sampel yang lebih kecil (Lee & Seo, 2022).

Dalam tahapan *preprocessing* seleksi fitur adalah salah satu teknik terpenting dan sering digunakan dalam membangun model dalam *machine learning*. Fokus seleksi fitur adalah untuk memilih subset variabel dari masukan yang bisa menggambarkan efisiensi input data dalam mengurangi dampak dari noise atau variabel yang tidak relevan dan tetap memberikan hasil prediksi yang baik (Chandrashekar & Sahin, 2014). Untuk dataset tidakseimbang (*imbalance*), metode seleksi fitur juga harus fokus pada atribut yang membantu dalam identifikasi kelas minoritas (Haro-garcía et al., 2020). Selain itu, kinerja metode seleksi fitur berkembang ketika rasio ketidakseimbangan meningkat. Hasil penelitian menunjukkan bahwa di berbagai rasio ketidakseimbangan kelas, metode DBFS (*Density Based Feature Selection*) melebihi metode saingan seleksi fitur lainnya terutama ketika lebih dari 0,5% dari fitur yang dipilih untuk tugas klasifikasi. Peningkatan ini lebih nyata sesuai dengan evaluasi statistik AUC (*area under curve*) terutama dengan rasio ketidakseimbangan tinggi (Alibeigi et al., 2012). Pendekatan untuk menggabungkan seleksi fitur dengan proses *boosting* fokus pada dua skenario yang berbeda yaitu seleksi fitur dilakukan sebelum proses *boosting* dan seleksi fitur yang dilakukan dalam proses *boosting*. Hasil percobaan menunjukkan bahwa melakukan seleksi fitur dalam boosting umumnya lebih baik daripada menggunakan seleksi fitur sebelum proses *boosting* (Yan et al., 2014). Adaboost adalah salah satu jenis Algoritma *boosting* yang bekerja iteratif dengan memberikan bobot yang berbeda pada distribusi training data di setiap iterasinya. Dalam setiap iterasi boosting menambahkan bobot pada setiap klasifikasi yang salah dan menurunkan bobot pada setiap kasus klasifikasi yang benar, sehingga secara efektif dapat merubah distribusi pada data training. Algoritma *adaptive boosting* (adaboost) juga telah telah memenuhi sebagai *framework* untuk mengatasi masalah ketidakseimbangan kelas (*class imbalance*) (Purnajiwa Arimbawa & Sanjaya ER, 2020)(Lee et al., 2017) . AdaBoost merupakan algoritma *machine learning* yang dirumuskan oleh Yoav Freund and Robert Schapire. AdaBoost secara teoritis dapat secara signifikan digunakan untuk mengurangi kesalahan dari beberapa algoritma pembelajaran yang secara konsisten menghasilkan kinerja pengklasifikasi yang lebih baik. Kinerja adaBoost lebih baik dari *random forest* untuk prediksi performansi dan dapat memperbaiki kinerja *classifier* (Bandara et al., 2020). Metode adaBoost terbukti efektif untuk penyelesaian masalah ketidakseimbangan kelas (*class*

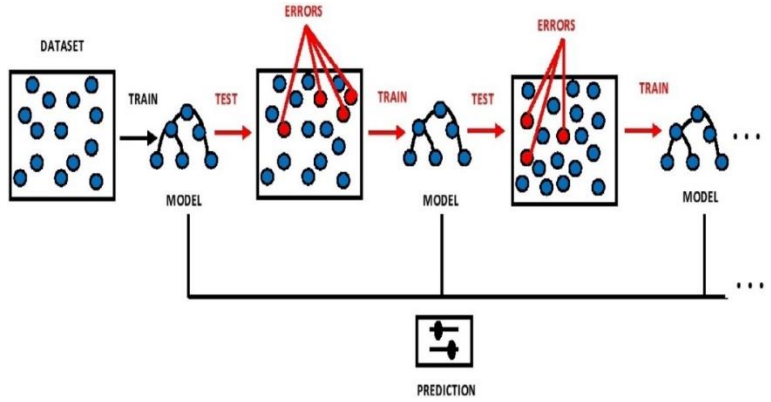
e-ISSN: 2964-2531
DOI : 10.36499/psnst.v12i1.6997
p-ISSN: 2964-5131

*imbalance*) pada penentuan *multi class* dengan metode *decision tree*, dan menghasilkan sebuah model arsitektur yang optimal dan hasil estimasi yang akurat. Adaboost berhasil meningkatkan akurasi pembelajaran yang lemah dan mengoptimalkan keragaman klasifikasi dasar (Cao et al., n.d.).

*Extreme Gradient Boosting* (XGBoost) merupakan salah satu algoritma yang paling populer dan paling banyak digunakan karena algoritma ini termasuk algoritma yang *powerful*. Pada dasarnya, algoritma ini adalah pengembangan, diman algoritma *gradient boost* dikembangkan beberapa proses tambahan sehingga lebih *powerful*. Proses tambahan tersebut adalah pemangkasan, *newton boosting*, dan parameter pengacakan ekstra. Proses pemangkasan atau penyusutan proporsional simpul daun digunakan untuk meningkatkan generalisasi model. Pproses *newton boosting* adalah proses untuk menyediakan rute langsung sehingga tidak memerlukan penurunan *gradient*. Proses pengacakan parameter bertujuan untuk mengurangi korelasi antar tree sehingga dapat meningkatkan kekuatan algoritma *ensemble*. *Extreme Gradient Boosting* (XGBoost) sebagai algoritma klasifikasi yang dihasilakan oleh para ilmuwan atau peneliti *machine learning* yang mengikuti ajang kompetisi Kaggle pada tahun 2015. Algoritma XGBoost merupakan metode klasifikasi yang dikembangkan dari *gradient tree boosting* yang sangat efektif dan banyak digunakan pada *machine learning*. Algoritma XGboost saat ini merupakan metode klasifikasi yang efektif dengan pemakaian sumberdaya yang minimal. XGBoost semakin popular karena ada 17 dari 29 peserta ajang kompetis Kaggle meraih kemenangan dengan menggunakan metode XGBoost. Sebagai bagian dari kombinasi algoritmanya XGBoost menerapkan metode *ensemble* (*Gradient Tree Boosting*) yang diharapkan dapat memberikan prediksi yang lebih baik dengan data kelas yang tidak seimbang. Algoritma XGBoost juga merupakan peskalaan *tree boosting* yang mampu menyelesaikan masalah nyata dengan penggunaan sumberdaya yang minimal (Chen & Guestrin, 2016).

## 2. METODE PENELITIAN

*Ensemble learning* adalah proses yang menggabungkan beberapa prediktor dasar seperti algoritma pembelajaran individu untuk menghasilkan hasil yang lebih baik dalam hal akurasi atau stabilitas. Di bidang mesin pembelajaran (ML), berbagai peneliti telah memberikan teknik pembelajaran ansambel manfaat secara signifikan meningkatkan kinerja umum model ML dan memproduksi beberapa prototipe terbaik dari sistem pembelajaran. Untuk meningkatkan kinerja model ML, teknik pembelajaran *ensemble* mengekspos algoritma pembelajaran individu (prediktor dasar) untuk belajar dari perspektif yang berbeda dari dataset baik dengan pembelajaran ansambel heterogen (menggunakan algoritma pembelajaran yang berbeda) atau dengan pembelajaran ansambel homogen (menggunakan satu algoritma yang sama). Algoritma pembelajaran tunggal yang belajar pada himpunan bagian acak yang berasal dari dataset asli. Perlu disebutkan bahwa keragaman prediktor dasar berkontribusi pada akurasi hasil model pembelajaran ensemble menyarankan empat level katageori metode pembelajaran ansambel yang berbeda, yaitu: level pengklasifikasi atau regresi, level data, level fitur, dan level kombinasi (*ensemble*). Dalam penelitian ini, kami fokus pada dua level pertama (pengklasifikasi dan level data) yang terdiri dari *ensemble* algoritma DBFS dan algoritma AdaBoost. Sedangkan pada level algoritma mengggunakan XGBoost dan *Random Fprest* (*Entreme Gradien Random Forest Boost*) untuk merancang kerangka penelitian ini.




**Gambar 1. Prinsip Algoritma *Boosting* (Ramzai, 2019)**

2.1. ***Ensemble Adaboost***

Teknik ensemble model AdaBoost menggunakan persamaan sebagai berikut :

$$F(x) = \sum_{t=1}^{T} (\propto_t \ \ h_t\,(X) \qquad (1)$$

Dimana :

$h_t\,(X)$ = Pengklasifikasi dasar atau leamah (*weak*)
$\propto_t$ = Tingkat pembelajaran (*learning rate*)
$F(x)$ = Hasil pengkalsifikaasi akhir (kuat)

Sedangaka algoritama AdaBoost dapat dijelaskan sebagai berikut :

*Input :*
Dataset D = {(x1,y1), (x2,y2), ....., (xm,ym)};
Algoritma pembelajaran lemah (Weak Learner) L;
Sebuah integer T menyatakan banyaknya iterasi. Proses:

*Inisialisasi bobot distribusi:*
1. Berikan data training dari instance space
   S = { (x1, y1), …(xm,ym)} dimana xi **ϵ** X dan yi **ϵ** Y = {-1, +1}.
2. Inisial distribusi $D1\,(i) = 1\ m$ untuk semua $i = 1, \dots, m$

Proses Agorithma
for t=1, ... , T: do
Melatih pembelajar dasar/lemah ht : X→ R menggunakan distribusi $D_t$
Menentukan bobot $\propto t$ dari $ht$
Update kelebihan distribusi data (kelas) dari training set :

$$D_{t+1}\,(i) = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$$

$$D_{t+1}\,(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

Dimana $Z_t$ adalah factor normalisasi terpilih jadi $D_{t+1}$ akan menjadi sebuah distribusi.
End for
Nilai Akhir :

$$f(x) = \sum_{t=0}^{T} \alpha_t h_t(x) \text{ dan } H(x) = sign(f(x))$$

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v12i1.6997
p-ISSN: 2964-5131

Kesalahan diukur dengan memperhatikan distribusi $D_t$ di mana algoritma pembelajar lemah dilatih. Dalam prakteknya, algoritma pembelajar lemah mungkin merupakan suatu algoritma yang dapat menggunakan bobot Dt pada sampel pelatihan. Atau, bila hal ini tidak memungkinan, bagian dari sampel pelatihan dapat di-resampling menurut $D_t$, dan hasil dari *resampling* yang tidak berbobot (*unweighted*) dapat digunakan untuk melatih algoritma pembelajar yang lemah

### 2.2. Klasifikasi Menggunakan *Extreme Garadient Boosting* (*XGBOOST*)

Setelah proses normalisasi, hasilnya berupa dataset yang sudah terstandar dan seimbang. Proses selanjutnya adalah membagi data menjadi dua bagian, yakni data latih dan data test dengan menggunakan dengan prosentase 70 % data digunakan sebagai data latih dan 30 % data tes. Data latih digunakan untuk membangun model klasifikasi sedangkan data tes digunakan untuk menguji model. Proses pemodelan dilakukan dengan metode *Extreme Gradient Boosting* (XGBoost).

XGBoost adalah algoritma yang ditingkatkan berdasarkan *gradient boosting decision tree* dan dapat membangun *boosted trees* secara efisien dan beroperasi secara paralel. XGBoost merupakan salah satu teknik pembelajaran mesin untuk mengatasi permasalahan regresi dan klasifikasi berdasarkan *Gradient Boosting Decision Tree* (GBDT). XGBoost pada dasarnya adalah metode *ensemble* yang didasarkan pada *gradient boosting tree*. Didalam pohon regresi, *nodes* bagian dalam mewakili nilai nilai untuk tes atribut dan leaf nodes dengan skor mewakili keputusan. Hasil prediksi adalah jumlah skor yang diprediksi oleh pohon K, seperti ditunjukkan pada persamaan berikut:

$$\hat{y} = \sum_{k}^{K} f_k(x_i), f_k \in F \quad (2)$$

Metode penelitian berisikan tentang bagaimana penelitian dikerjakan yang dijelaskan secara detail. Pada setip parragraf bisa tediri dari beberapa *subparagraph* yang ditunjukkan pada persamaan:

$$obj\,(\theta) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_l) + \sum_{k}^{K} \Omega\,(f_k) \quad (3)$$

Dimana $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ adalah differentiable loss function untuk mengukur apakah model tersebut cocok untuk set data pelatihan dan $\sum_{k}^{K} \Omega\,(f_k)$ adalah item yang menentukan kompleksitas model. Ketika kompleksitas model meningkat skor yang sesuai dikurangi nilainya. Sebelum membangun model prediki, dilakukan optimasi parameter *turning* XGBoost dengan beberapa parameter

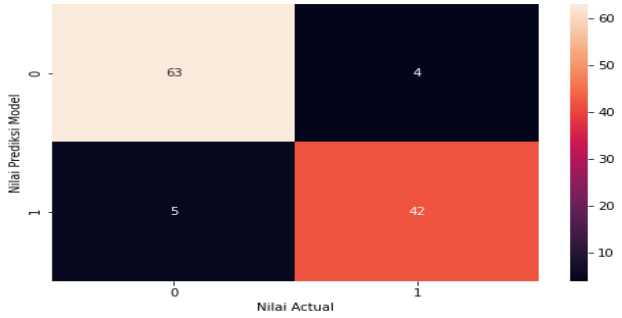
## 3. HASIL DAN PEMBAHASAN

Model prediksi *multilevel ensemble* didesain dengan tugas atau fungsi memprediksi apakah seorang mengarah dalam kelompok kanker ganas (*malignant*) atau jinak (*benign*). Oleh karena itu, masalah prediksi ini adalah masalah klasifikasi biner, dan harus diselesaikan dengan metode pembelajaran terawasi. Di sini, model prediksi kami dibangun berdasarkan teknik *multi-level ensemble*. Teknik tersebut terdiri dari ensemble level 1 yang berkaitan penaganan data yang tidak seimbang. Metode yang digunakan adala gabungan metode DBFS dan AdaBoost. Kemudian teknik *ensemble* level 2 berkaitan dengan algoritma klasifikasi dengan menggunakan *ensemble* XGBoost.

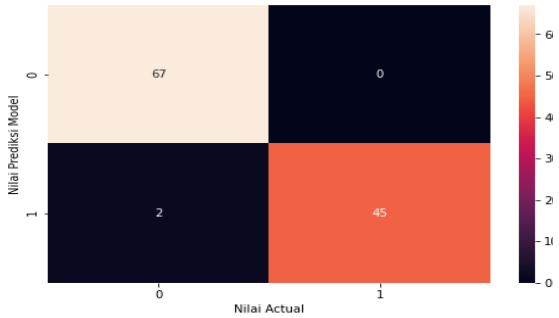
Dalam melakukan eksperimen dalam penelitian ini menggunakan data set terbuka dari UCI *Repository* yaitu *Breast CancerWisconsin Diagnostic* (BCWD). Dataset tersebut berisi 569 baris data, 31 (30 fitur variable bebas 1 fitur sebagai label kelas) fitur yang berkaitan dengan kanker payudara, dan dua jenis golongan, yaitu ganas (*malignant*) dan jinak (*benign*). Pada eksperimen akan dilakukan perlakuan data dan algoritma dengan 2 teknik *ensemble*. Tujuan dari model prediksi multilevel *ensemble* adalah untuk memprediksi apakah seorang pasien mengarah dalam dalam kelompok kanker ganas (*malignant*) atau jinak (*benign*). Oleh karena itu, masalah prediksi ini adalah masalah klasifikasi biner, dan harus diselesaikan dengan metode pembelajaran terawasi. Model prediksi kami dibangun

berdasarkan hasil seleksi fitur dari teknik *ensemble* level 1 sebagai input model pada *ensemble* level 2. Hasil evaluasi secara umum sebelum penggabungan dan setelah penggabungan 2 level *ensemble* tersebut memiliki perbedaan *accuracy* sebesar 0,03.

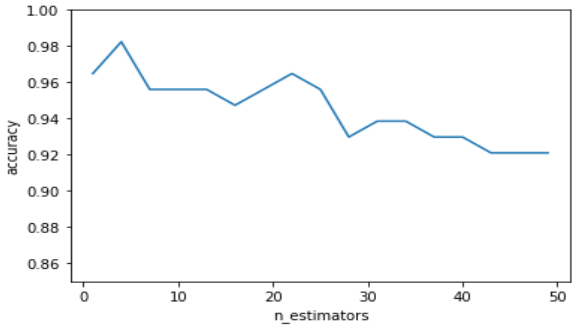
Perbandingan evaluasi antara penerapan model ensemble level 1 dan model *ensemble* level 2 dengan 80 % adalah data *latigh* dan 20% data *train*.sebagai berikut.



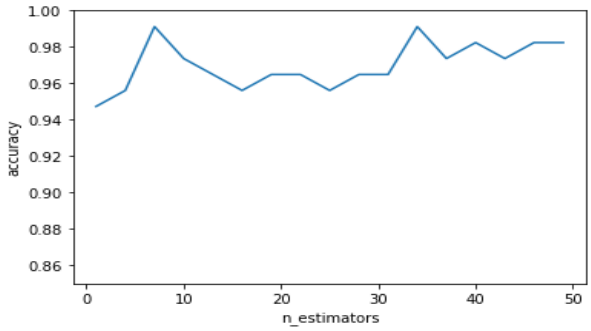

**Gambar 2. *Confusion Matrix Model Ensemble* Level 1**



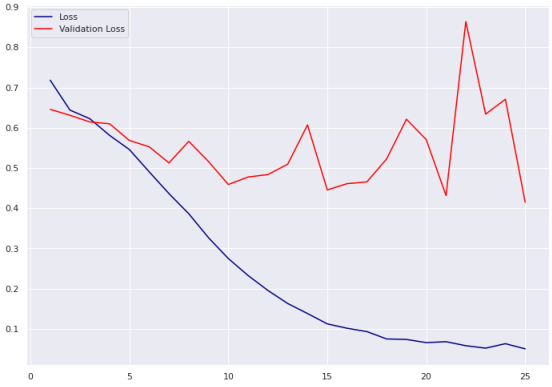

***Gambar 3. Confusion Matrix Model Model Ensemble Level 2***



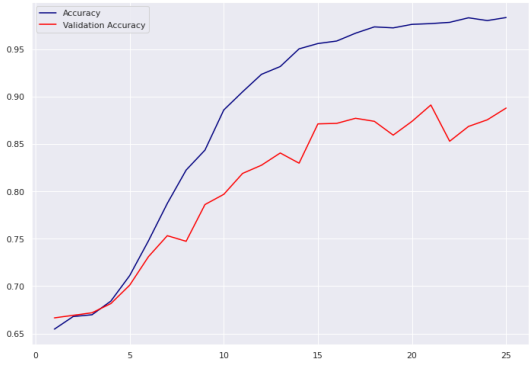

**Gambar 4. Grafik *Acurracy Model Ensemble* Level 1 (Cenderung Penurunan)**




**Gambar 5. Grafik *Acurracy Model Ensemble* Level 1 & 2**




**Gambar 6. Grafik Loos Model Ensemble Level 1 & 2**




**Gambar 7. Grafik Accuracy Model Ensemble Level 2 (Cenderung Naik)**

**e-ISSN: 2964-2531**
**DOI : 10.36499/psnst.v12i1.6997** **p-ISSN: 2964-5131**

Hasil perbandingan metode esnsemble level 1 dan level 2 berkaitan dengan performansi model, dimana model ensevel level 2 memiliki perbaikan nilai accuracy, precission, recall dan KPPA-Stats seperti table 1 dibawah ini.

**Tabel 1. Nilai Akurasi, Prssisi, *Recall* & F-*Measure***

| Metode Ensembel | Accuracy | Precission | Recall | KPPA-Stats |
|---|---|---|---|---|
| Model Ensemble Level 1 | 0.95 | 0.95 | 0.93 | 0.90 |
| Model Ensemble Level 2 | 0.98 | 1.0 | 0.95 | 0.96 |

Hasil perbandingan metode esnsemble level 1 dan level 2 berkaitan dengan performansi model, dimana model ensevel level 2 memiliki perbaikan nilai mean absolute error, mean squared error dan F-measure table 2 dibawah ini.

**Tabel 2. Nilai *Errors* & F-*Measure***

| Metode Ensembel | Mean Absolute Error | Mean Squared Error | F-Measure |
|---|---|---|---|
| Model Ensemble Level 1 | 0.04 | 0.04 | 0.93 |
| Model Ensemble Level 2 | 0.01 | 0.01 | 0.95 |

Hasil perbandingan metode esnsemble level 1 dan level 2 berkaitan dengan performansi model, dimana model ensevel level 2 memiliki perbaikan nilai accuracy berdasarkan penggunaan jumlah data test (50%, 30%, 20 %) seperti table 3 dibawah ini.

**Tabel 3. *Accuracy* Berdasar % Data Tes**

| Metode Ensembel | Data Tes 50% | Data Tes 30% | Data Tes 20% |
|---|---|---|---|
| Model Ensemble Level 1 | 0.94 | 0.94 | 0.95 |
| Model Ensemble Level 2 | 0.98 | 0.98 | 0.98 |

Hasil perbandingan metode esnsemble level 1 dan level 2 berkaitan dengan performansi model, dimana model ensevel level 2 memiliki perbaikan nilai accuracy berdasarkan jumlah n_estimator (50%,40, 30%, 20 %, 10%) seperti table 4 dibawah ini.

**Tabel 4. Accuracy Berdasar Jumlah *n_estimator***

| Metode Ensembel | 10 estimator | 20 estimator | 30 estimator | 40 estimator | 50 estimator |
|---|---|---|---|---|---|
| Model Ensemble Level 1 | 0.96 | 0.95 | 0.94 | 0.93 | 0.92 |
| Model Ensemble Level 2 | 0.98 | 0.97 | 0.97 | 0.98 | 0.97 |

## 4. KESIMPULAN

*Feature importance* yang dihasilkan dengan teknik *ensemble* (DBFS dan AdaBoost) membantu meningkatkan kualitas klasifikasi klasifikasi *ensemble* XGBoost. Dari 31 *feature*, terdapat 27 *feature* yang memiliki pengaruh terbesar dalam menghasilkan model klasifikasi *multilevel ensemble*. Dengan membandikan hasil Model Ensembel Level 1 (murni menggunakan XGBoost) dan Model Ensembel Level 2 (gabungan *ensemble* DBFS dan AdaBosst dengan *Ensemble* XGBoost) didapatkan hasil performa klasifikasi yang lebih baik. Untuk akurasi Model Ensembel Level 2 memiliki akurasi 0,98 dibanding Model Ensembel Level 1 yang memiliki akurasi 0,95. Selain itu akurasi Model Ensembel Level 2 lebih stabil baik dari jumlah *n_estimator* maupun dari proporsi pembagian data *training*. Model Ensembel Level 1 memiliki kecenderungan penurunan akurusi baik dari jumlah *n_estimator* maupun dari jumlah proporsi pembagian data training. Dari performa *recall* dan *precession* dan KPPA-*Stats* Model Ensembel Level 2 juga menunjukan performa klasifikasi yang lebih baik dan stabil dari pada Model Ensembel Level 1. Model Ensembel Level 2 juga memiliki tingkat kesalahan yang lebih kecil daripada Model Ensembel Level 1, nilai MAE dan MSE memiliki sesilih 0.03, F-1 *Score* atau F-*Measure* memiliki selisih nilai 0,02. Secar keseluruhan Model Ensembel Level 2 lebih akurat, lebih setabil dan lebih general untuk klasifikasi data tidak seimbang *binary* klasifikasi.

## DAFTAR PUSTAKA

Alibeigi, M., Hashemi, S., & Hamzeh, A. (2012). Data & Knowledge Engineering DBFS : An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets. *DATAK*, *81–82*, 67–103. https://doi.org/10.1016/j.datak.2012.08.001

Bandara, D., Grant, T., Hirshfield, L., & Velipasalar, S. (2020). Identification of Potential Task Shedding Events Using Brain Activity Data. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00034-y

Cao, Y., Miao, Q., Liu, J., & Gao, L. (n.d.). Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica*, *39*(6), 745–758. https://doi.org/10.1016/S1874-1029(13)60052-X

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods q. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. 785–794.

Czarnowski, I. (2022). Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams. *Journal of Computational Science*, *61*(October 2021), 101614. https://doi.org/10.1016/j.jocs.2022.101614

**e-ISSN: 2964-2531**
**DOI : 10.36499/psnst.v12i1.6997** **p-ISSN: 2964-5131**

Gonzalez-Abril, L., Nuñez, H., Angulo, C., & Velasco, F. (2014). GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems. *Applied Soft Computing Journal*, *17*, 23–31. https://doi.org/10.1016/j.asoc.2013.12.013

Haro-garcía, A. De, Cerruela-garcía, G., & García-pedrajas, N. (2020). A proposal and comparative study Ensembles of Feature Selectors for dealing with Class-Imbalanced Datasets : A proposal and comparative study ⋆. *Information Sciences*. https://doi.org/10.1016/j.ins.2020.05.077

Lee, W., Jun, C., & Lee, J. (2017). Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification. *Information Sciences*, *381*, 92–103. https://doi.org/10.1016/j.ins.2016.11.014

Lee, W., & Seo, K. (2022). Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning. *Big Data Research*, *28*, 100314. https://doi.org/10.1016/j.bdr.2022.100314

Purnajiwa Arimbawa, I. G. A., & Sanjaya ER, N. A. (2020). Penerapan Metode Adaboost Untuk Multi-Label Classification Pada Dokumen Teks. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, *9*(1), 127. https://doi.org/10.24843/jlk.2020.v09.i01.p13

Yan, Y., Shen, H., Liu, G., Ma, Z., Gao, C., & Sebe, N. (2014). GLocal tells you more : Coupling GLocal structural for feature selection with sparsity for image and video classification. *Computer Vision and Image Understanding*, *124*, 99–109. https://doi.org/10.1016/j.cviu.2014.02.006

Zyblewski, P., Sabourin, R., & Woźniak, M. (2021). Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. *Information Fusion*, *66*(August 2017), 138–154. https://doi.org/10.1016/j.inffus.2020.09.004