Analisis Performa Klasifikasi Algoritma Pada Pendeteksian Penyakit Kanker Dengan Partition Membership

Mustikasari*[1], ST. Aminah Dinayati Ghani[2]
[1]UIN Alauddin Makassar; H.M. Yasin Limpo No. 36, Kab.Gowa , Telp. (0411) 841879
Program Studi Teknik Informatika, UIN Alauddin, Makassar
[2]STMIK Dipanegara Makassar; Jl. Perintis Kemerdekaan No. 9, Telp. (0411) 587194
Program Studi Teknik Informatika, Universitas Dipa Makassar

e-mail: *[1]mustikasari@uin-alauddin.ac.id, [2]dinayati.amy@dipanegara.ac.id

Abstrak

Deteksi dini terhadap penyakit utamanya penyakit dengan resiko tinggi seperti kanker paru dan kanker payudara adalah cara paling menjanjikan untuk meningkatkan peluang pasien untuk bertahan hidup. Makalah ini menyajikan metode klasifikasi yang dikembangkan dengan menggunakan algoritma pengklasifikasi untuk mengatasi masalah tersebut. Dua jenis pengujian dilakukan, salah satunya dengan tidak menggunakan filter *partition membership* pada algoritma klasifikasi dan lainnya menggunakan filter *partition membership* yang mempekerjakan dua algoritma untuk membangkitkan nilai keanggotaan partisi yaitu algorithma *random forest* dan *C4.5*. Kemudian, *10-fold cross validation* dilakukan pada dataset yang diubah. Enam algoritma pengklasifikasi diuji akurasinya atas pengaruh *preprocessing* data dengan filter *partition membership* tersebut. Berdasarkan hasil yang diperoleh, akurasi klasifikasi yang lebih baik dicapai ketika partition membership dipasangkan dengan keenam algoritma, dan diperoleh peningkatan akurasi klasifikasi pada sebagian besar hasil eksperimen dibandingkan dengan hasil nilai standar masing-masing algoritma klasifikasi pada pengukuran nilai akurasi, statistik *kappa*, dan *F-measure*.

Kata kunci — kinerja, klasifikasi, *partition membership*, penyakit

Abstract

Early detection of diseases primarily high risk diseases such as lung cancer and breast cancer is the most promising way to increase a patient's chances of survival. This paper presents a classification method developed using a classification algorithm to solve this problem. Two types of tests were carried out, one by not using the partition membership filter in the classification algorithm and the other using the partition membership filter which employs two algorithms to generate partition membership values, namely the random forest algorithm and the C4.5. Then, 10-fold cross validation is performed on the modified dataset. Six classification algorithms were tested for accuracy on the effect of data preprocessing with the partition membership filter. Based on the results obtained, better classification accuracy is achieved when partition membership is paired with the six algorithms, and an increase in classification accuracy is obtained compared to the standard value results of each classification algorithm on the measurement of accuracy values, kappa statistics, and F-measure.

Keywords*— performance, classification, partition membership, disease*

1. PENDAHULUAN

Deteksi penyakit merupakan salah satu masalah klasifikasi yang sangat penting. Saat ini implementasi pendeteksian penyakit banyak diterapkan pada beberapa domain, tidak

terkecuali bidang kesehatan yang butuh untuk mengkomunikasikan layanan ataupun tindakan kesehatan yang diperlukan kepada kelompok resiko tertentu. Misalnya, seseorang pakar kesehatan atau unit kerja khusus dari institusi kesehatan dapat menargetkan materi pendidikan dan berita yang disesuaikan ke kelompok kecil, dalam populasi umum, yang memiliki tingkat resiko tinggi terdampak penyakit tertentu. Prediksi risiko penyakit, dan pendeteksian dini yang sejalan dengan komunikasi kesehatan yang disesuaikan dapat mengarah pada saluran yang efektif untuk menyampaikan informasi khusus tentang gejala dan penangan penyakit bagi orang-orang yang membutuhkannya.

Meskipun beberapa catatan klinis berfitur lengkap sulit diakses karena masalah privasi tentang beberapa penyakit, telah tersedia kumpulan dataset dunia nyata yang berhasil merangkum kombinasi gejala dari sejarah diagnosis medis yang timbul dan dibuat dalam data publik di repository data penelitian. Data yang dapat dengan mudah diakses para peneliti data mining. Beberapa teknik data mining juga telah diterapkan pada kumpulan data medis kesehatan tersebut untuk mendeteksi penyakit tertentu. Nur Salman, dkk [1], menerapkan Bayesian Network dalam penelitiannya dan mengeksplorasi ruang pencarian dari algoritma Tree Augmented Network (TAN) untuk memperlihatkan hubungan antar fitur dalam sebuah struktur. Hasil eksperimen penelitian tersebut menunjukkan bahwa algoritma BN-TAN mengungguli algoritma menggunakan struktur pembelajaran lainnya untuk akurasi dan waktu konstruksi pada dataset hipotiroid. Nilashi, Mehrbakhshi, dkk.[2], telah mengerjakan penelitian yang lebih spesifik pada klasifikasi kanker payudara menggunakan pengelompokan Expectation-Maximization (EM), noise removal, dan Regression Trees (CART).

Bagaimanapun juga, disadari bahwa kumpulan data publik penyakit yang tersedia, tidak memiliki semua variabel dari rekam medis asli, namun, data ini tetap mempertahankan beberapa karakteristik utamanya yaitu *mixed data type* atau terdiri dari data campuran seperti numerik dan kategorikal, rentan terdapat ketidakseimbangan data dan penggunaan terminologi medis terkontrol. Oleh karena itu, mempertimbangkan keragaman data dan kompleksitas data yang terkait tipe data dan attribut, serta banyaknya jumlah data serta kelengkapan data penyakit itu sendiri, maka ketepatan pemilihan algoritma dan penggunaan filter spesifik juga menjadi pertimbangan yang tidak kalah pentingnya dalam mendukung kinerja dengan hasil yang optimal dari pekerjaan ini.

Fokus utama pada makalah ini yaitu deteksi penyakit kanker paru, dan kanker payudara menggunakan algoritma pengklasifikasi dengan kinerja yang dioptimalkan. Untuk menghadapi tantangan tersebut, dalam makalah ini mengusulkan klasifikasi penyakit kanker menggunakan metode *preprocessing* data berbasis proposisionalisasi. Teknik proposisionalisasi menggunakan pohon keputusan pada metode preprocessing data ini telah dipilih untuk meningkatkan akurasi klasifikasi. Untuk tujuan tersebut, filter partition membership dipekerjakan pada awal pemrosesan (*preprocessing*) proses klasifikasi penyakit kanker paru dan payudara, lalu dilanjutkan dengan menjalankan eksperimen menggunakan enam algoritma pengklasifikasi umum pada dataset kanker paru (*lung cancer and breast cancer)* dari repositori UCI Machine learning. Perkerjaan penelitian ini adalah untuk membuktikan hipotesis kami bahwa filter partiton membership menggunakan pohon keputusan menghasilkan akurasi klasifikasi yang lebih baik pada pengklasifikasian beberapa jenis penyakit kanker.

2. METODE

Penelitian ini merupakan metode eksperimental. Metode ini dapat digambarkan sebagai metode kuantitatif dengan melakukan percobaan untuk melihat hasil dan menyelidiki serta mengkonfirmasi hasilnya serta menganalisa hubungan kausal yang terjadi. Kontrol yang ketat dilakukan untuk menganalisa hasil yang diperoleh.

2.1 Data

Dataset yang kami gunakan dikumpulkan dari University of California di Irvine (UCI) Machine Learning Repository [3], secara khusus kami menggunakan dua kumpulan data patokan, yang berhubungan dengan dua penyakit.

Dua dataset penyakit kanker digunakan yaitu *lung cancer* dataset dan *breast cancer* dataset. Dataset pertama terdiri atas terdiri atas 286 data (instance), dan tersusun atas 9 attribut ditambah satu attribut class, Kumpulan data ini mencakup 201 data dari satu kelas dan 85 data dari kelas lain. Data dijelaskan oleh 9 atribut, beberapa di antaranya linier dan beberapa nominal. Dataset kedua menggambarkan 3 jenis kanker paru-paru patologis. Jumlah Instance adalah 32, dan jumlah atribut 57 (1 atribut kelas, 56 prediktif). Semua atribut prediktif adalah nominal, mengambil nilai integer 0-3.

2. 2 Evaluasi Model Klasifikasi

Masalah klasifikasi [4] adalah mempelajari struktur kumpulan data contoh, yang sudah dipartisi menjadi beberapa kelompok, yang disebut sebagai kategori atau kelas. Pembelajaran kategori ini biasanya dicapai dengan model. Model ini digunakan untuk memperkirakan pengidentifikasi grup (atau label kelas) dari satu atau lebih contoh data yang sebelumnya tidak terlihat dengan label yang tidak diketahui. Oleh karena itu, salah satu masukan untuk masalah klasifikasi adalah contoh kumpulan data yang telah dipartisi ke dalam kelas yang berbeda. Ini disebut sebagai data pelatihan, dan pengidentifikasi grup kelas ini disebut sebagai label kelas. Himpunan kelas yang memungkinkan telah diketahui sebelumnya.

Algoritma klasifikasi yang berbeda (SGD, SMO, Random Tree, NBTree, RepTree, Filtered Classifier) digunakan untuk mengklasifikasikan penyakit kanker dalam penelitian ini. Hasil Klasifikasi dibandingkan berdasarkan kategori berikut:

a. *Akurasi (accuracy)* - Akurasi pengklasifikasi didefinisikan sebagai persentase kumpulan data yang diklasifikasikan dengan benar oleh metode. Akurasi semua pengklasifikasi yang digunakan untuk mengklasifikasikan penyakit kanker. Akurasi sendiri adalah rasio dari jumlah nilai elemen diagonal confusion matriks terhadap jumlah seluruh elemen matriks.
b. Statistik Kappa *(kappa statistic)* - Untuk mengukur keakuratan algoritma pembelajaran pada data yang gunakan, diterapkan 10-fold cross-validation untuk memperkirakan nilai statistik kappa, yang dapat dilihat sebagai versi akurasi klasifikasi yang dinormalisasi yang sangat berguna ketika kelas tidak seimbang. Rumus Kappa dihitung sebagai berikut:

$$Kappa = ((\alpha - \alpha_r)/(1 - \alpha_r)) \qquad (1)$$

dengan α adalah perkiraan akurasi klasifikasi dari algoritma pembelajaran yang ingin kita evaluasi, dan α_r adalah akurasi yang diharapkan dari pengklasifikasi acak yang memberikan contoh secara acak ke kelas sedemikian rupa sehingga menetapkan jumlah contoh yang sama untuk setiap kelas sebagai pembelajaran algoritma yang kami evaluasi. Jika kappa lebih besar dari nol, algoritma pembelajaran menunjukkan akurasi yang lebih besar dari yang diharapkan dengan menetapkan klasifikasi secara acak ke kantong yang terjadi di lipatan uji validasi silang. Nilai satu adalah nilai maksimal yang bisa dicapa

c. F-measure - F-measure atau F1-Score merupakan perbandingan rata-rata presisi dan recall yang dibobotkan.

$$F\_measure = (2 * \text{Recall} * \text{Precision})/((\text{Recall} + \text{Precision})) \qquad (2)$$

2. 3 *Partition Membership*

Filter digunakan untuk memproses data untuk menghapus data yang berisik di data mining. Ada dua jenis filter *supervised* dan *unsupervised* filter. Filter dapat diterapkan ke set data pelatihan dan pengujian. Pada penerapan *supervised Partition Membership*, filter mengubah nilai numerik menjadi nominal dan mendistribusikannya ke dalam jumlah bin yang dipilih secara merata. Motivasi dibalik pemilihan filter ini adalah bahwa pendekatan ini membagi ruang instance menjadi beberapa wilayah dan mengukur kepemilikan, teknik ini memungkinkan untuk mendeskripsikan distribusi kantong di ruang kejadian. Sekaligus memberikan alternatif untuk pendekatan proposisionalisasi [5] yang lebih sederhana yang menghitung statistik ringkasan seperti mean dan standar deviasi dari nilai atribut dalam kantong, di mana kantong data diubah menjadi vektor pasangan nilai-atribut sehingga algoritma pembelajaran proposisional standar (yaitu nilai atribut) dapat diterapkan.

Kantong (*Bag*)		
x0	*x1*	*x2*
tidak	ya	tidak
tidak	ya	ya
tidak	ya	ya
ya	tidak	tidak
ya	tidak	ya
ya	tidak	ya
ya	tidak	tidak
ya	ya	ya
tidak	tidak	tidak

Data (Instance)												
y1	y2	y9	y3	y6	y10	y13	y4	y5	y7	y8	y11	y12
9	4	5	1	3	4	1	1	0	1	2	2	2

Gambar 1. Kantong data (bag of instance) dan bentuk proposisional

2. 4 *Random Forest* (RF)

Randomforests adalah gagasan teknik umum hutan keputusan acak yang merupakan teknik pembelajaran ensembel untuk klasifikasi, regresi, dan tugas-tugas lainnya, yang mengontrol dengan membangun banyak pohon keputusan pada waktu pelatihan dan mengeluarkan kelas yang merupakan mode kelas (klasifikasi) atau prediksi rata-rata (regresi) dari masing-masing pohon. Hutan keputusan acak adalah solusi akurat untuk masalah pohon keputusan yang overfitting ke set pelatihan. RF adalah pembelajar*an* ensemble, metode yang menghasilkan banyak pengklasifikasi dan menggabungkan hasil-hasilnya. Setiap pohon di RF akan memberikan "vote" untuk beberapa masukan *x*, kemudian keluaran dari pengklasifikasi ditentukan oleh pemungutan suara terbanyak dari pohon. RF dapat menangani data berdimensi tinggi dan menggunakan sejumlah besar pohon dalam ansambel. Beberapa fitur penting RF adalah:

a. RF memiliki metode yang efektif untuk memperkirakan data yang hilang.
b. Memiliki metode, weighted random forest (WRF), untuk menyeimbangkan kesalahan dalam data yang tidak seimbang [6].
c. RF memperkirakan pentingnya variabel yang digunakan dalam klasifikasi.

2.4.1 *Variabel Importance*

Salah satu fitur RF yang paling penting adalah output dari *Variabel Importance*. *Variabel Importance* mengukur tingkat hubungan antara variabel tertentu dan hasil klasifikasi. Untuk memperkirakan *Variabel Importance* untuk beberapa variabel j, sampel out-of-bag (OOB) diturunkan ke pohon dan akurasi prediksi dicatat. Kemudian nilai variabel j diubah dalam sampel OOB dan akurasi diukur kembali. Perhitungan ini dilakukan pohon demi pohon saat RF dibangun. Penurunan rata-rata dalam akurasi permutasi ini kemudian dirata-ratakan pada semua pohon dan digunakan untuk mengukur pentingnya variabel j.

2.4.2 *Splitting Criterion*

RF menggunakan ukuran Gini Impurity untuk memilih pemisahan dengan impurity terendah di setiap node. Gini impurity adalah ukuran distribusi label kelas di node. Gini impurity mengambil nilai dalam [0, 1], di mana 0 diperoleh ketika semua elemen dalam sebuah node memiliki kelas yang sama. Secara formal, ukuran Gini impurity untuk variabel X = $\{x_1, x_2, ..., x_j\}$ pada node *t*, dimana *j* adalah jumlah anak pada node *t*, N adalah jumlah sampel, *n* c_i adalah jumlah sampel dengan nilai x_i termasuk kelas *c*, a_i adalah jumlah sampel dengan nilai x_i pada node *t*.

Indeks Gini dari sebuah split adalah rata-rata tertimbang dari ukuran Gini di atas nilai-nilai yang berbeda dari variabel X. Keputusan kriteria pemisahan akan didasarkan pada nilai Gini impurity terendah yang dihitung di antara variabel M. Dalam RF, setiap pohon menggunakan sekumpulan variabel *m* yang berbeda untuk membuat aturan pemisahan.

2. 5 *C4.5*

Pohon keputusan (*decision tree*) adalah salah satu metode klasifikasi yang paling populer karena algoritma ini mengubah data menjadi pohon keputusan dan aturan keputusan dinyatakan dalam bentuk tabel dengan atribut dan catatan. Selain itu, pohon keputusan menggabungkan eksplorasi data dan pemodelan sehingga sangat baik sebagai langkah awal pemodelan bahkan ketika digunakan sebagai model akhir dari beberapa teknik lainnya. Pohon keputusan merupakan salah satu teknik klasifikasi yang dilakukan berdasarkan kriteria pemisahan. Pohon keputusan adalah diagram alir seperti struktur pohon yang mengklasifikasikan data dengan mengurutkannya berdasarkan nilai fitur (atribut). Setiap node dalam pohon keputusan mewakili fitur dalam sebuah instance yang akan diklasifikasikan. Semua cabang menunjukkan hasil tes, setiap simpul daun memegang label kelas. Data diklasifikasikan dari awal berdasarkan nilai fiturnya. Pohon keputusan menghasilkan aturan untuk klasifikasi kumpulan data. Tiga algoritma dasar yang banyak digunakan yaitu ID3, C4.5, dan CART.

C4.5 adalah algoritma pohon keputusan yang dihasilkan oleh Quinlan. Ini adalah perluasan dari algoritma ID3. Algoritma C4.5 banyak digunakan karena klasifikasinya yang cepat dan presisi yang tinggi. C4.5 didasarkan pada rasio perolehan informasi yang dievaluasi oleh entropi. Ukuran rasio perolehan informasi digunakan untuk memilih fitur pengujian di setiap node dalam pohon. Ukuran seperti itu disebut sebagai ukuran pemilihan fitur (atribut). Atribut dengan rasio perolehan informasi tertinggi dipilih sebagai fitur pengujian untuk node saat ini.

2. 6 Validasi

Eksperimen menggunakan *tool* data minig weka. Pada tahapan ini dilakukan perbandingkan kinerja klasifikasi dari enam algoritma, dimana kami juga menganalisis pengaruh penerapan *partition membership* yang mempekerjakan algoritma decision tree yaitu

model *Random Forest* dan *C4.5* untuk menghasilkan nilai keanggotaan partisi pada tahap preprocessing.

Pada tahap pengujian, kami memilih menggunakan validasi silang 10 kali lipat (*10-fold cross validation*) sebagai mode pengujian untuk mencatat akurasi klasifikasi. Pendekatan ini cocok untuk menghindari hasil yang bias dan memberikan ketegasan pada klasifikasi. Selain daripada itu, parameter algoritma klasifikasi ditetapkan ke nilai standar atau defaultnya. Cross-validasi berbeda dari metrik penilaian lokal dalam hal kualitas. Kami menyajikan hasil pada data dunia nyata menggunakan dataset lung-cancer dari UCI machine Learning repository. Langkah-langkah berikut diterapkan untuk menghasilkan data eksperimental untuk menarik kesimpulan:

1. Lakukan percobaan untuk mengukur kinerja klasifikasi file pengklasifikasi dalam dataset analisa dan konfirmasi hasilnya.
2. Lakukan percobaan untuk mengukur kinerja klasifikasi menggunakan filter partition membership, analisa dan konfirmasi hasilnya serta gambarkan hubungan kausal yang terjadi.

3. HASIL DAN PEMBAHASAN

Untuk mengevaluasi kinerja dari pendekatan filter *partisi membership* dengan algoritma C4.5 dan RF, kami melakukan percobaan pada dataset kanker paru dan kanker payudara *(lung cancer and breast cancer dataset)* yang sebelumnya telah disebutkan dalam makalah ini.

Tabel 1: Kinerja klasifikasi (akurasi) dari enam algoritma tanpa filter.

Algoritma Klasifikasi	***No Partition Membership***		
	Acc (%)	***Kappa***	***F_measure***
SGD (Lung_Cancer)	62.5	0.005	0.60
SGD (Breast_Cancer)	69.93	0.21	0.68
SMO(Lung_Cancer)	65.62	0.12	0.65
SMO (Breast_Cancer)	69.58	0.20	0.68
RandomTree(Lung_Cancer)	75	0.38	0.75
Random_Tree (Breast_Cancer)	66.78	0.18	0.66
NB_Tree(Lung_Cancer)	78.12	0.44	0.78
NB_Tree (Breast_Cancer)	70.98	0.25	0.69
Rep_Tree(Lung_Cancer)	78.12	0.29	0.73
Rep_Tree (Breast_Cancer)	66.78	0.18	0.66
FilteredClassifier(Lung_Cancer)	78.12	0.40	0.77
FilteredClassifier (Breast_Cancer)	75.52	0.28	0.72

Pada skenario pertama sebagaimana yang disajikan pada Tabel 1, dengan memilih opsi uji menggunakan 10-fold cross validation, ke enam algoritma pengklasifikasi bekerja untuk mengklasifikasi data kanker paru *(lung cancer dataset)*. Dengan perolehan hasil yang beragam namun cenderung masih dibawah 80% atau tertinggi pada perolehan 78.12 % untuk algoritma NBTree, RepTree dan Filtered Classifier sedangkan nilai akurasi terendah menggunakan algoritma SGD yaitu 62.5%. Sementara pada pengklasifikasian data kanker payudara *(breast cancer dataset*) nilai akurasi tertinggi dicapai algoritma Filtered Classifier yaitu 75.52% dan terendah oleh algoritmaRandom Tree dan Rep Tree dengan nilai yang sama yaitu 66.78% . Ini

dapat dilatarbelakangi oleh kehadiran missing value pada kedua data set yang digunakan, dimana algoritma berbasis tree dapat lebih kuat menangani kondisi tersebut.

Pada skenario kedua, ke enam dataset pengklasifikasi di uji dengan memasukkan tahapan prepocessing data. Pada tahap tersebut, filter partition membership digunakan untuk menerapkan konsep proposisionalization pada attribut. Hasil yang diperoleh dari eksperimen tersebut kami sajikan sebagiman pada Tabel 2.

Tabel 2: Kinerja klasifikasi enam algoritma dengan filter.

Algoritma Klasifikasi	***Partition Generator (Random Forest)***			***Partition Generator (C4.5)***		
	Acc (%)	***Kappa***	***F_measure***	***Acc (%)***	***Kappa***	***F_measure***
SGD (Lung_Cancer)	81.25	0.50	0.80	93.75	0.84	0.94
SGD (Breast_Cancer)	71.68	0.28	0.70	75.52	0.29	0.71
SMO(Lung_Cancer)	78.12	0.44	0.78	90.625	0.54	0.90
SMO (Breast_Cancer)	72.73	0.32	0.72	75.52	0.29	0.71
RandomTree (Lung_Cancer)	78.12	0.44	0.78	90.625	0.76	0.90
Random_Tree (Breast_Cancer)	67.83	0.19	0.67	75.52	0.29	0.71
NB_Tree (Lung_Cancer)	84.37	0.56	0.83	87.5	0.64	0.86
NB_Tree (Breast_Cancer)	90.62	0.76	0.93	75.87	0.29	0.72
Rep_Tree (Lung_Cancer)	75 (-3.125)	0.22 (-0.07)	0.70 (-0.03)	81.25	0.50	0.80
Rep_Tree (Breast_Cancer)	68.88	0.15 (-0.03)	0.66	74.13	0.22	0.68
FilteredClassifier (Lung_Cancer)	65.62 (-12.5)	0.05 (-0.35)	0.62 (-0.15)	87.5	0.64	0.86
FilteredClassifier (Breast_Cancer)	73.08 (-2.44)	0.19 (-0.09)	0.68 (-0.04)	75.87	0.29	0.72

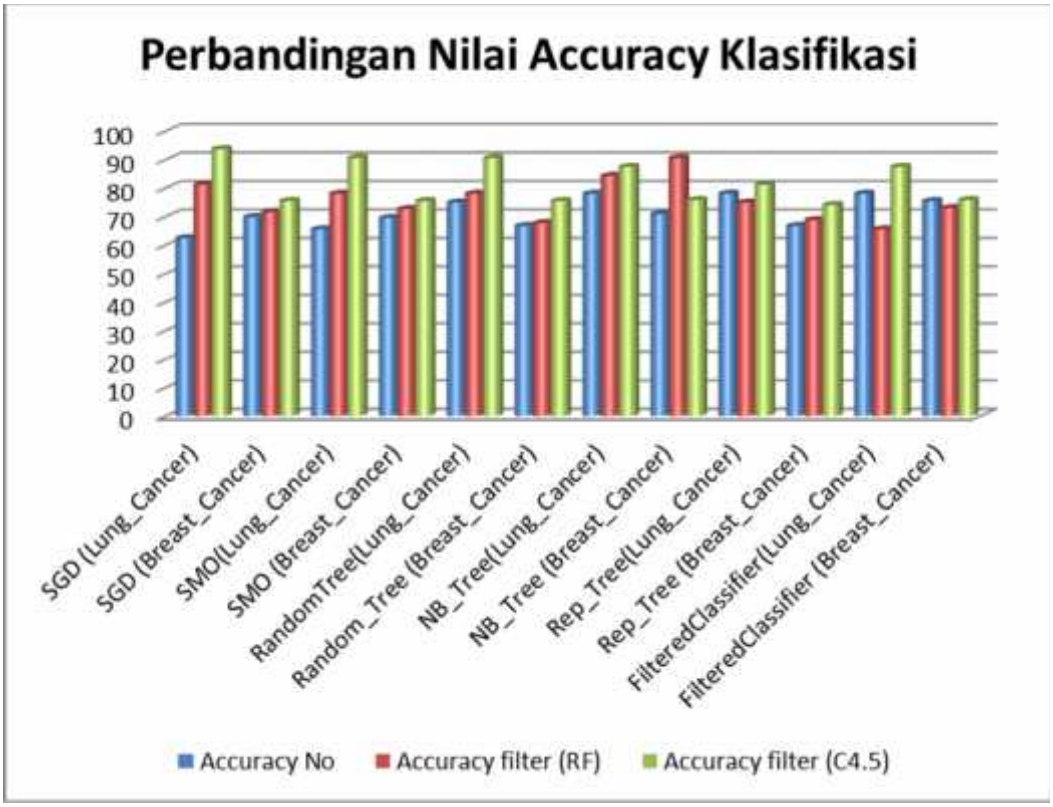
Hasil eksperimen pada skenario dua menunjukkan bahwa penerapan filter partition membership memang mampu meningkatkan nilai akurasi klasifikasi pada sebagian besar algoritma. Terutama dengan pada penggunaan C4.5 demikian juga dengan RF meskipun harus tersubtitusi oleh runtime proses partisi dan klasifikasi.

Temuan dalam eksperimen ini adalah bahwa dua algoritma dari 6 algoritma (yaitu RepTree dan Filtered Classfier) gagal atau bahkan mengalami penurunan nilai akurasi, *kappa statistic* dan f-measure pada eksperimen tersebut. Bagaimanapun, kantong proposisional telah menghasilkan vektor attribut yang sangat jarang sehingga beberapa wilayah yang didefenisikan oleh pohon keputusan tidak memiliki / tidak akan berisi contoh kantong tertentu.

Angka minus pada Tabel 2. Merupakan informasi penurunan nilai akurasi dibandingkan hasil akurasi tanpa penerapan filter. Walau demikian, kita dapat lihat pada Gambar 2, 3 dan 4

bahwa kecenderungan peningkatan nilai akurasi klasifikasi secara dominan diperoleh setelah filter *partition member*ship dterapkan.

Dapat kita amati pula pada Gambar 2 bahwa C4.5 sebagai partition generator memiliki dominansi dalam meningkatkan nilai kinerja klasifikasi dan mengurangi waktu runtime menjadi lebih rendah dibandingkan penggunaan RF sebagai pembanding.

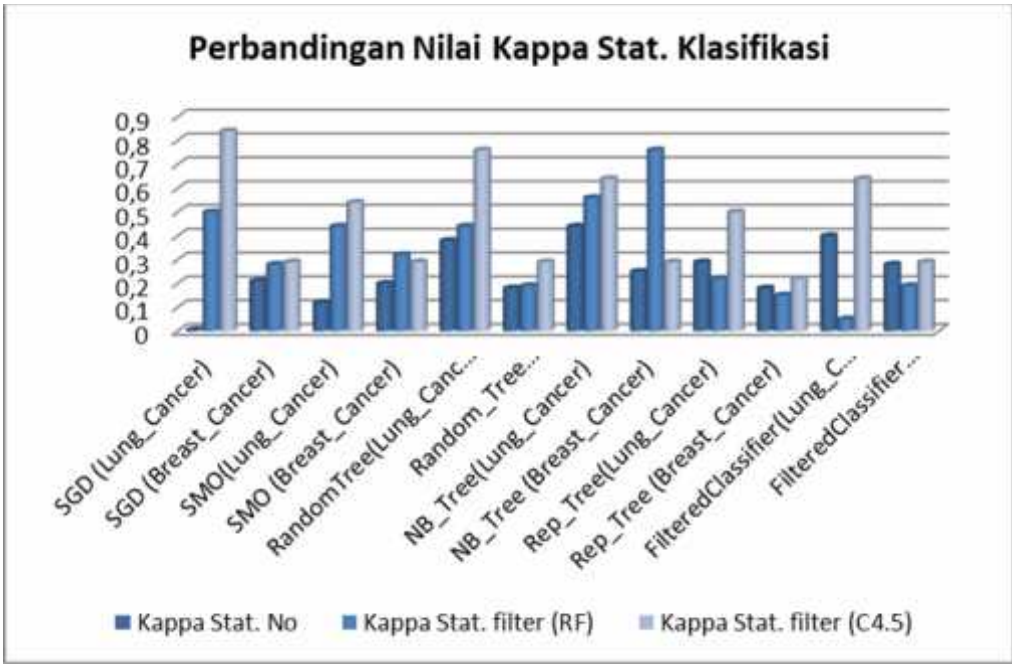


Gambar 2. Perbandingan Akurasi

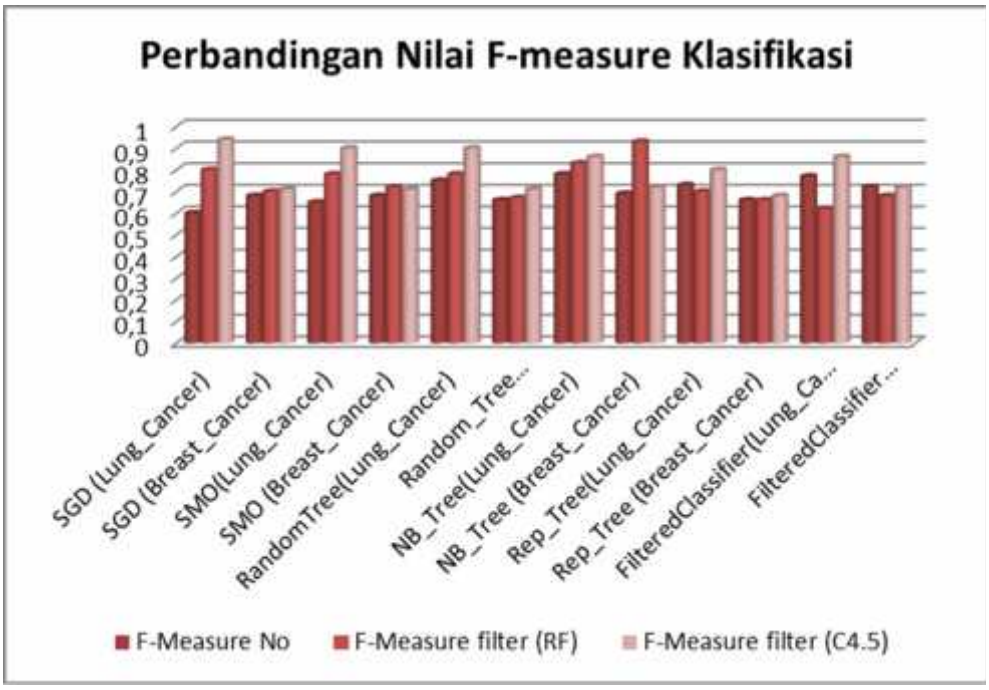
Dapat dilihat bahwa penggunaan random forest untuk membuat proposisional data secara signifikan meningkatkan waktu pelatihan di semua kasus. Salah satu alasannya adalah bahwa ansambel pohon perlu ditanam, bukan hanya berupa satu pohon. Alasan lain adalah bahwa contoh dalam data yang diproposionalisasi memiliki lebih banyak atribut saat menggunakan hutan acak daripada saat menggunakan pohon deterministik tunggal karena ansambel pohon digunakan sebagai pengganti satu pohon dan satu pohon di hutan acak umumnya lebih besar dari satu pohon.

Pohon tunggal yang tumbuh secara deterministik, dimana pemilihan atribut menggunakan perolehan informasi bertujuan untuk memperkecil ukuran pohon. Bagaimanapun juga kinerja akurasi klasifikasi masih lebih diprioritaskan dalam pendeteksian penyakit daripada pertimbangan atas waktu pelatihan.

Dalam Gambar 2. Dapat disimpulkan bahwa RF paling sukses bekerja meningkatkan performa klasifikasi algoritma NB_Tree pada dataset kanker payudara (breast cancer). Dan C4.5 sebagai *partition generator* berhasil meningkatkan performa klasifikasi lima algoritma lainnya untuk kedua dataset yang digunakan.



Gambar 3. Perbandingan Nilai Kappa



Gambar 4. Perbandingan Nilai F-measure

4. KESIMPULAN

Hasil penelitian menunjukkan algoritma pengklasifikasi terbaik adalah pengklasifikasi C4.5 untuk dataset dari repository UCI Machine Learning, dan kinerja akurasi algoritma klasifikasi dari keenam algoritma (SGD, SMO, Random Tree, NBTree, RepTree dan Filtered Classifier) dapat ditingkatkan dengan data yang diproposisionalisasi lebih dulu dengan menerapkan filter *partition membership*.

DAFTAR PUSTAKA

[1] N. Salman dan M. Mustikasari, "STRUKTUR PEMBELAJARAN JARINGAN BAYESIAN MENGGUNAKAN ALGORITMA TREE AUGMENTED NETWORK (TAN) UNTUK DIAGNOSIS HIPOTIROID," in *SENSITIf: Seminar Nasional Sistem Informasi dan Teknologi Informasi*, 2019, hal. 1011–1018.

[2] M. Nilashi, O. Ibrahim, H. Ahmadi, dan L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telemat. Informatics*, vol. 34, no. 4, hal. 133–144, 2017.

[3] C. Blake, "UCI repository of machine learning databases," *http//www. ics. uci. edu/~ mlearn/MLRepository. html*, 1998.

[4] C. C. Aggarwal, "Data classification," in *Data Mining*, 2015, hal. 285–344.

[5] E. Frank dan B. Pfahringer, "Propositionalisation of multi-instance data using random forests," in *Australasian Joint Conference on Artificial Intelligence*, 2013, hal. 362–373.

[6] M. Khalilia, S. Chakraborty, dan M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, hal. 1–13, 2011.