

DIAGNOSIS PENYAKIT PARKINSON BERDASARKAN KOMBINASI ALGORITME DATA MINING DAN SELEKSI FITUR

Tri Astuti¹, Tri Ferinanto²

¹Jurusan Teknik Informatika, ²Jurusan Sistem Informasi
STMIK Amikom Purwokerto, Indonesia
¹tri_astuti@amikompurwokerto.ac.id, ²feriixell@gmail.com

Abstract

Parkinson's is a degenerative disorder of the central nervous system that causes disturbances in the motor system due to a lack of dopamine in the brain. The brain is an organ that is extremely vital in humans that contains millions of nerve cells that coordinates all the activities performed by the body. While dopamine is a liquid that is used as the sender of signals in nerves. Parkinson's disease can be treated with a variety of efforts, including the administration of drugs and brain transplants. However, handling the administration of drug continuously for long periods of time adversely affect other organs. Disease risk can be minimized with early accurate diagnosis methods. Besides diagnosis using medical devices that have a lot of risk, diagnosis can also be done with a computer-based method of diagnosis. Computer-based diagnostic methods can be implemented in the classification of machine learning. However, to get an accurate classification results, it is necessary to do a comparison between classifiers to get the best results. This study aimed to compare the performance between classifiers. Classification algorithms used is Naive Bayes and J48 using the same feature selection is CFS. From the comparison between the classification algorithms, algorithms J48 has the best accuracy value to the value of 96.923% with a mean absolute error 0.0554, ROC 0.982 and the running time of 0.1 seconds.

Keywords: Parkinson, diagnosis, method, classification.

I. PENDAHULUAN

Penyakit Parkinson (Parkinson's disease) adalah kelainan degeneratif dari sistem saraf pusat yang menyebabkan gangguan pada sistem motorik dan biasanya penderita mengalami tremor, kaku otot, sulit berjalan, gangguan keseimbangan dan gerak menjadi lambat (bradykinesia). Penyakit ini biasanya dialami pada usia 60 tahun keatas, walaupun ditemukan juga pada beberapa penderita Parkinson yang berusia dibawah 50 tahun.[1]

Menurut Dorsey, salah satu faktor yang mengakibatkan pertambahan penderita parkinson adalah semakin banyaknya penduduk usia lanjut di negara-negara besar dunia, terutama di Tiongkok. Di negara tersebut, penderita parkinson pada 2030 diperkirakan mencapai 5 juta jiwa karena dinegara tersebut penyakit parkinson tidak dianggap sebagai masalah yang serius. Penyakit parkinson paling banyak terjadi pada masyarakat yang memiliki infrastruktur terbatas hingga tidak bisa mendiagnosis pasien dengan tepat atau tidak dapat melayani penyakit tersebut secara serius.[2]

Selain diagnosis dengan menggunakan teknologi kedokteran, diagnosis penyakit parkinson juga dapat dilakukan dengan mesin pembelajaran (machine learning) dan data mining.

Metode diagnosis penyakit dengan teknologi pemindai *CT (Computerized tomography) Scan*,

PET (Positron Emission Tomography) Scan dan *MRI (Magnetic Resonance Imaging)* merupakan alat yang penting dalam dunia medis. Namun *CT Scan* hanya dianjurkan oleh dokter ketika kondisi pasien benar-benar memerlukannya. Hal itu disebabkan karena penggunaan *CT Scan* juga memiliki risiko meningkatkan risiko kanker dan kedalaman pemindainya tidak cukup untuk memindai penyakit parkinson. Sedangkan teknologi *MRI* dan *PET Scan* sendiri masih jarang ditemui di rumah sakit rumah sakit karena harga alat dan biaya untuk memindainya yang mahal. Selain diagnosis dengan menggunakan teknologi kedokteran, diagnosis penyakit parkinson juga dapat dilakukan dengan mesin pembelajaran (*machine learning*) dan *data mining*. Teknik yang dapat dilakukan menggunakan klasifikasi penyakit parkinson. Teknik klasifikasi ini guna meningkatkan produktifitas dan konsistensi diagnosis. Banyak penelitian telah menggunakan teknik klasifikasi. Hal tersebut telah meningkatkan konsistensi hasil dan mendorong terciptanya metode-metode baru dalam melakukan klasifikasi sehingga membuat teknik ini terus berkembang.

Sektor medis menyediakan banyak informasi yang terstruktur untuk para peneliti melakukan pengembangan dalam penelitian yang mereka lakukan. Penggunaan *machine learning* menjadi salah satu indikasi yang digunakan dalam pengembangan akurasi medis dalam teknologi. Saat

ini tersedia lebih dari 28% data set “*life sciences*” dari total data set yang ada di UCI.

Penelitian [4] melakukan analisis penyakit Parkinson yang merupakan penyakit yang disebabkan oleh degenerasi progresif *dopamine* yang mengandung neuron dalam struktur sel otak yang disebut ganglia. Penelitian menggambarkan analisis yang dilakukan berdasarkan pada dua algoritme pelatihan *Levenberg-Marquardt* yaitu (LM) dan *Scaled Conjugate Gradient (SCG) Multilayer Perceptrons (MLPs) Neural Network* dalam mendiagnosis penyakit Parkinson dengan Analisis Varians (ANOVA) sebagai pilihan fitur. Dataset informasi dari proyek ini telah diambil dari Dataset Penyakit Parkinson. Hasil yang dicapai menegaskan bahwa LM memiliki kinerja yang baik dengan tingkat akurasi 90% sebelum dan sesudah seleksi fitur sementara SSG mencapai di atas 85% setelah implementasi ANOVA sebagai seleksi fitur.

[2] membandingkan dua buah algoritme yaitu algoritme *Bagging J48* dan *SMO* dengan menggunakan metode seleksi fitur *CFS*, *Gain Ratio*, *RELIEF* dan *Wrapper*. Dari keempat hasil seleksi fitur yang diperoleh, performa akurasi terbaik didapat dari subset seleksi fitur *CFS* dengan dengan algoritme *Bagging J48* yaitu sebesar 91% dengan nilai *sensitivity* 0,996, *specificity* 0,755, *ROC* 0,941 dan *running time* 0,01 detik.

Penelitian Rustempasic dan Mehmet (2013) dengan menggunakan dua metode, metode pertama bertujuan untuk membedakan kinerja antara dua kelas, yaitu penderita penyakit parkinson dengan yang sehat. Metode kedua menggunakan pengenalan pola. Penelitian ini menggunakan metode seleksi fitur *CFS* dan menghasilkan hasil yang memuaskan dengan nilai prediktif 80,88%.

Perbandingan dua buah algoritme yaitu algoritme *Bagging J48* dan *SMO* dengan menggunakan metode seleksi fitur *CFS*, *Gain Ratio*, *RELIEF* dan *Wrapper*. Dari keempat hasil seleksi fitur yang diperoleh, performa akurasi terbaik didapat dari subset seleksi fitur *CFS* dengan dengan algoritme *Bagging J48* yaitu sebesar 91% dengan nilai *sensitivity* 0,996, *specificity* 0,755, *ROC* 0,941 dan *running time* 0,01 detik [3].

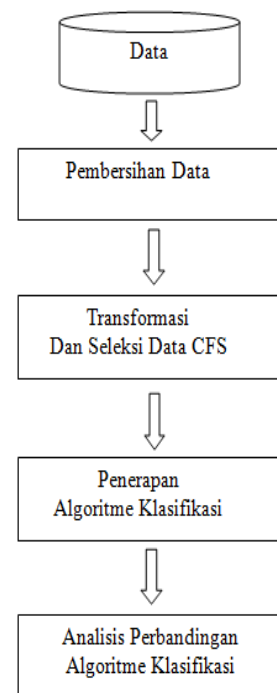
Teknik yang dapat dilakukan menggunakan klasifikasi penyakit parkinson. Teknik klasifikasi ini guna meningkatkan produktifitas dan konsistensi diagnosis. Banyak penelitian telah menggunakan teknik klasifikasi. Hal tersebut telah meningkatkan konsistensi hasil dan mendorong terciptanya metode-metode baru dalam melakukan klasifikasi sehingga membuat teknik ini terus berkembang.

Pemilihan algoritme Naive Bayes karena algoritme ini memiliki pola kerja dalam penentuan kasus berdasarkan propabilitas dan memiliki keunggulan dalam klasifikasi karena implementasi yang sederhana dan lebih cepat dalam proses penghitungan. Sedangkan algoritme *RIPPER* merupakan optimasi dari algoritme *IREP* yang merupakan salah satu algoritme klasifikasi berbasis *rules* dan dapat menangani dataset noise dalam jumlah besar secara efisien.

II. METODOLOGI

Pada penelitian ini, data yang digunakan adalah dataset parkinson yang diambil dari *UCI machine learning repository*. Atribut dataset parkinson meliputi 24 atribut kondisi dan 1 atribut kelas dengan jumlah *instance* sebanyak 195.

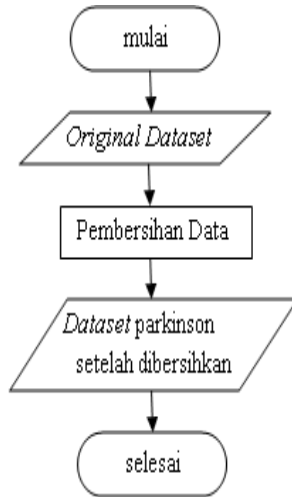
Tahapan pada penelitian ini meliputi proses pembersihan dataset, transformasi data, seleksi fitur dilanjutkan proses pengaplikasian algoritme *data mining* yang dipilih dilanjutkan dengan analisis hasil dan dibandingkan performa masing-masing algoritme. Evaluasi performa algoritme dilakukan berdasarkan *10-fold cross validation*. Pada Gambar 1 ditampilkan alur metode yang diusulkan.



Gambar 3.1. Alur metode yang diusulkan

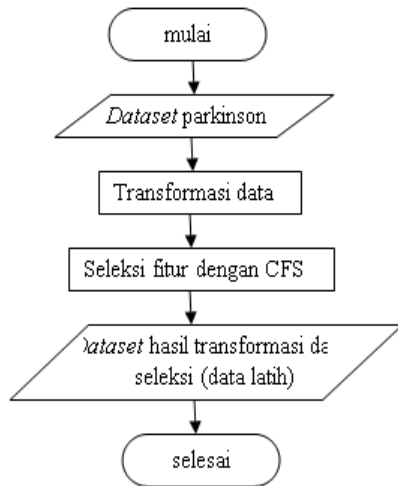
Pada tahap pembersihan data dilakukan untuk membuang data yang tidak konsisten dan *noise*. Pada tahap ini data yang tidak lengkap akan

dibuang. Alur pembersihan dataset tampak pada Gambar 3.2 berikut ini.



Gambar 3.2. Alur proses pembersihan dataset

Tahap transformasi dilakukan untuk mengubah data menjadi bentuk-bentuk yang sesuai untuk dilakukan proses data mining. Beberapa teknik data mining membutuhkan format data yang khusus sebelum diaplikasikan. Tahap seleksi dilakukan untuk menentukan data yang sesuai dengan format data yang dipersyaratkan.



Gambar 3.3. Alur proses transformasi dan seleksi fitur

Setelah data sudah siap selanjutnya penerapan algoritme klasifikasi *data mining*. Algoritme yang digunakan untuk klasifikasi yaitu algoritme Naive Bayes, RIPPER dan J48. Setelah dilakukan penerapan beberapa penerapan algoritme klasifikasi data mining, maka selanjutnya adalah melakukan perbandingan hasil penerapan. Pada tahap ini, dilakukan perbandingan performa algoritme klasifikasi. Analisis dilakukan untuk mengetahui algoritme klasifikasi data mining yang paling akurat

diterapkan untuk diagnosis penyakit parkinson. Selain itu juga akan diperoleh algoritme klasifikasi yang paling dominan. Perbandingan performa algoritme menggunakan tingkat akurasi dan waktu komputasi dan nilai ROC.

III. PEMBAHASAN

Pembersihan data dilakukan untuk membersihkan data yang tidak konsisten dan *noise*. Pada tahap ini data yang dihapus yakni data *name* atau nomor panggil. Data nomor panggil dibersihkan karena data tersebut tidak konsisten dengan data yang lain. Berikut ini tabel dataset parkinson

No	name	...	Status
1	phon_R01_S01_1	1
2	phon_R01_S01_2	1
3	phon_R01_S01_3	1
...
...
194	phon_R01_S50_5	0
195	phon_R01_S50_6	0

Tahap transformasi dilakukan untuk mengubah data yang kompleks dengan tidak menghilangkan isi, sehingga lebih mudah diolah dan diproses *data mining*. Beberapa teknik data mining membutuhkan format data yang khusus sebelum diaplikasikan. Tahap seleksi dilakukan untuk menentukan data yang sesuai dengan format data yang dipersyaratkan. Berikut penerapan langkah-langkahnya:

a. Transformasi

Berikut beberapa teknik yang dilakukan dalam transformasi data :

- 1) Mengubah format data, format data yang awalnya *.csv* diubah menjadi format data yang dapat dibaca oleh WEKA yaitu *.arff*.
- 2) Perubahan format atribut, perubahan struktur atribut dilakukan untuk mengubah struktur / urutan dari atribut *class(status)* atau target menjadi atribut paling belakang atau terakhir. Format data *real* merupakan bilangan kontinyu (*continuos*) yang berarti data diambil berdasarkan skala pengukuran. Untuk format label disesuaikan dari format biner(1,0), 1 untuk penderita parkinson, 0 untuk bukan penderita atau sehat (*healthy*).
- 3) Mengubah tipe data status atau target dari yang awalnya biner menjadi label. Untuk tipe data 1 menjadi *parkinson* (penderita) dan untuk 0 menjadi *healthy* (bukan penderita / sehat).

b. Seleksi Fitur CFS

Untuk memaksimalkan performa klasifikasi, maka dilakukan seleksi data dimana tahap ini bertujuan untuk menghilangkan atribut dan data yang kurang mempengaruhi atau mengurangi akurasi dari proses klasifikasi dan mengurangi beban komputasi yang dikarenakan beban fitur yang banyak. Dari proses seleksi oleh CFS dengan target status menghasilkan delapan subset atribut kondisi. Hasil proses seleksi fitur dengan sebelum diseleksi fitur dapat dilihat pada Tabel 4.1 beserta 1 atribut kelas status.

Tabel 4.1 Daftar atribut hasil seleksi fitur CFS

No.	Atribut hasil seleksi fitur CFS
1	MDVP:Fo(Hz)
2	MDVP:Fhi(Hz)
3	MDVP:Flo(Hz)
4	MDVP:RAP
5	MDVP:APQ
6	spread1
7	spread2
8	PPE
9	status

Tahapan analisis digunakan delapan atribut hasil seleksi fitur dengan CFS. Delapan atribut yang dipakai diantaranya sebagai *predictor* dan satu atribut sebagai target. Ada satu kelas yang memiliki dua tujuan/respon/target dari hasil klasifikasi yang terdapat pada atribut status yaitu *parkinson* (penderita parkinson) dan *healthy* (bukan penderita/sehat). Algoritme yang digunakan untuk melakukan klasifikasi yaitu *Naive Bayes*, *RIPPER(JRIP)* dan *J48*. Hasil keseluruhan evaluasi dari kinerja algoritme *Naive Bayes*, *RIPPER(JRIP)* dan *J48* dapat dilihat pada tabel 4.2. Mode tes yang digunakan yaitu *use training set*. Jumlah total data yang diuji yaitu 195 data (*instances*), spesifikasi hardware yang digunakan yaitu AMD C-60 APU 1.0 GHz dengan RAM 4 GB dan perangkat lunak yang digunakan adalah WEKA 3.7.

Tabel 4.2 Hasil evaluasi dari kinerja algoritme

Algoritme	Running Time	Akurasi (%)	ROC
Naive Bayes	0.04 detik	76.923	0.896
RIPPER	0.16 detik	94.871	0.917
J48	0.10 detik	96.923	0.982

Berdasarkan Tabel 4.2 diketahui algoritme yang memiliki tingkat akurasi paling baik adalah algoritme *J48*. Berdasarkan nilai ROC diketahui ketiga algoritme masuk dalam kategori baik. Terlihat bahwa *J48* memiliki nilai ROC yang tertinggi. Sementara berdasarkan waktu komputasi diketahui *Naive Bayes* memiliki kecepatan yang lebih baik sebesar 0.04 detik.

IV. KESIMPULAN

Penelitian mengkaji tentang perbandingan performa algoritme klasifikasi data mining dengan seleksi fitur CFS untuk analisis diagnosis penyakit parkinson dapat disimpulkan bahwa algoritme *J48* memiliki nilai akurasi terbaik yaitu sebesar 96.923% dengan, ROC 0.982 dan running time 0.10 detik. Sedangkan untuk algoritme klasifikasi dengan performa terendah dimiliki oleh algoritme *Naive Bayes* dengan nilai 76.923%.

DAFTAR PUSTAKA

- [1] Sunaryati, Titiek. 2015. Penyakit Parkinson
- [2] Dorsey, ER dkk. "Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030" 68. Diakses 28 November 2015. <http://www.neurology.org/content/68/5/384>
- [3] Dewi, Sarini Vita. Analisis Performa Klasifikasi Untuk Diagnosis Penyakit Parkinson. Universitas Gadjah Mada. 2014.
- [4] Bakar, Zahari Bin Abu. (2010). "CLASSIFICATION OF PARKINSON'S DISEASE (PD) BASED ON MULTILAYER PERCEPTRONS (MLPs) NEURAL NETWORK." *universiti teknologi MARA*. Diakses 27 November 2015. <http://ir.uitm.edu.my/5281/>.
- [5] Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram, Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection, *International Journal of Information Technology and Knowledge Management*, (2) 2, 2010, pp. 271-277
- [6] Zhao Jing-Jing, HUANG Xiao-Hong, SUN Qion, MA Yan, Real time feature selection in traffic classification, *The Journal of China Universities of Posts and Telecommunication*, 2008.