

ANALISIS DAN PERANCANGAN SISTEM UNTUK DIAGNOSA PENYAKIT PARKINSON MENGGUNAKAN METODE PCA DAN RANDOM FOREST

ANALYSIS AND SYSTEM DESIGN FOR DIAGNOSIS OF PARKINSON DISEASE USING PCA AND RANDOM FOREST METHODS

Reza Yudistira¹, Rita Purnamasari S.T.², M.T., Nur Ibrahim S.T., M.T.³,
Prodi S1 Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom.

¹rezayudistira721@student.telkomuniversity.ac.id ,

²ritapurnamasari@telkomuniversity.ac.id, ³nuribrahim@telkomuniversity.ac.id.

Abstrak

Penyakit parkinson adalah penyakit yang menyerang sistem motorik manusia dan termasuk salah satu penyakit yang sulit untuk didiagnosa. Untuk diagnosa penyakit parkinson dibutuhkan metode khusus yang dapat mendiagnosa penyakit parkinson. Hingga saat ini, belum ada satu tes spesifik untuk mendiagnosa adanya penyakit parkinson pada tubuh secara akurat. Penyebab penyakit ini adalah rusaknya sel syaraf yang bernama substantia nigra yang berfungsi menghasilkan senyawa bernama dopamin. Penderita parkinson biasanya mengalami kaku pada otot dan mengalami kesulitan dalam berjalan.

Penelitian ini dilakukan dengan cara mengklasifikasi rekaman sinyal yang dihasilkan oleh sensor yang bersumber dari database physiobank. Sensor VGRF yang berjumlah 16 sensor dipasang pada kaki saat berjalan, lalu hasil dari sensor inilah yang nantinya akan diproses dengan metode principal component analysis (PCA) untuk menentukan karakteristik data dan random forest untuk mengklasifikasi data. Banyaknya data dan atribut yang terdapat pada database akan diolah sehingga data lebih sederhana, namun variasi data tetap terjaga, karena variasi dan karakteristik data akan sangat berpengaruh dalam hasil akurasi sistem.

Penelitian ini bertujuan untuk membantu mendiagnosa sejak dini pada pasien yang diduga mengidap penyakit parkinson. Sistem yang dihasilkan pada penelitian ini diharapkan dapat membantu proses diagnosa pasien berdasarkan jejak rekam medis pasien. Reduksi dimensi data dapat dicapai menggunakan metode PCA, yaitu dengan mengubah dimensi data sebanyak 0,142% dari data utama. Metode random forest dapat mengklasifikasi data dengan baik, hal ini dicapai dengan kombinasi parameter $n_estimator = 31$ dan $max_depth = 9$. Akurasi terbaik yaitu sebesar 95,13 % namun proses komputasi memerlukan waktu 85 detik, sedangkan waktu komputasi terbaik didapatkan dengan mengubah parameter $max_depth = 2$ dengan waktu 68 detik namun dengan akurasi 71%.

Kata kunci: Penyakit Parkinson, Principal Component Analysis, Random Forest.

Abstract

Parkinson's disease is a disease that attacks the human motor system and is a disease that is difficult to diagnose. To diagnose Parkinson's disease, we need a special method that can diagnose Parkinson's disease. To date, there is no single specific test to accurately diagnose Parkinson's disease in the body. The cause of this disease is the destruction of nerve cells called substantia nigra which functions to produce a compound called dopamine. Parkinson's sufferers usually experience muscle stiffness and have difficulty walking.

This research was conducted by classifying the signal recordings generated by the sensor from the Physiobank database. The 16 VGRF sensors are mounted on the legs while walking, then the results of these sensors will be processed using the principal component analysis (PCA) method to determine the characteristics of the data and random forest to classify the data. The amount of data and attributes contained in the database will be processed so that the data is simpler, but data variations are maintained, because variations and data characteristics will greatly affect the results of system accuracy.

This study aims to help early diagnose patients who are suspected of having Parkinson's disease. The system produced in this research is expected to help the process of diagnosing patients based on the patient's medical record. Reduction of data dimensions can be achieved using the PCA

method, namely by changing the data dimensions by 0.142% of the main data. The random forest method can classify data well, this is achieved with a combination of the parameter $n_estimator = 31$ and $max_depth = 9$. The best accuracy is 95.13% but the computation process takes 85 seconds, while the best computation time is obtained by changing the max_depth parameter = 2 with a time of 68 seconds but with an accuracy of 71%.

Keywords: Parkinson Disease, Principal Component Analysis, Random Forest.

1. Pendahuluan

Penyakit parkinson adalah degenerasi sel saraf secara bertahap pada otak bagian tengah yang berfungsi mengatur pergerakan tubuh. Gejala yang banyak diketahui orang dari penyakit ini adalah terjadinya tremor atau gemeteran. Sedangkan gejala awalnya biasanya sulit dikenali. Contoh gejala awal penyakit parkinson adalah merasa lemah atau terasa lebih kaku pada sebagian tubuh, gemeteran halus pada salah satu tangan saat beristirahat. Penyakit parkinson memengaruhi bagian kecil dari otak tengah yang bernama substantia nigra. Fungsi dari substantia nigra adalah mengirim pesan ke berbagai saraf di tulang belakang yang berfungsi mengendalikan otot-otot pada tubuh. Pesan akan dikirimkan dari sel otak ke saraf dan otot dengan memanfaatkan senyawa kimia yang disebut neurotransmitter. Salah satu neurotransmitter utama yang dihasilkan oleh sel otak di substantia nigra adalah dopamine. Saat jumlah dopamine menurun akan menyebabkan aktivitas otak akan terganggu. Inilah yang menyebabkan munculnya tanda-tanda dan gejala penyakit parkinson [1].

Penelitian ini bertujuan untuk menganalisa penggunaan metode Principal Component Analysis (PCA) sebagai ekstraksi ciri dan random forest sebagai klasifikasi data dalam mengolah data pasien parkinson untuk membantu diagnosa apakah pasien positif terjangkit parkinson. Kedua metode ini dipilih karena tingkat efisiensi algoritma yang tinggi, selain itu parameter yang digunakan pada metode ini tidak banyak, sehingga akan membuat waktu komputasi data semakin cepat namun tingkat akurasi klasifikasi data tetap tinggi.

2. Landasan teori

2.1 Parkinson

Orang berumur lebih dari 60 tahun rentan terhadap penyakit ini. Namun, orang yang lebih muda mungkin juga mengalaminya. Penyakit ini umumnya mempengaruhi lebih banyak pria dari pada wanita. Nama penyakit ini sebenarnya berasal dari nama seorang dokter Inggris James Parkinson yang mempublikasikan pertama kali penyakit kerusakan otak ini. Penyebab pasti dari penyakit ini sangat beragam, meskipun gejalanya hampir sama di setiap orang, tetapi mungkin ada perbedaan penting yang menyebabkan penyakit ini. Penyakit Parkinson menyebabkan degeneratif (kemunduran) saraf pusat terutama otak sehingga terjadi ketidakseimbangan neurotransmitter di otak dan seseorang akan terkena tremor atau kekakuan dalam pergerakan tubuh. Penyakit ini bersifat progresif, yang artinya semakin memburuk dari waktu ke waktu.

2.3 PCA

Metode ini mengubah sebagian besar variabel asli yang berkorelasi menjadi satu himpunan variabel baru yang lebih kecil dan saling bebas. Data inputan yang berupa matriks yang berdimensi besar memerlukan tenaga dan waktu untuk menganalisisnya, sehingga dibutuhkan penurunan ukuran dimensi yang tanpa mengurangi atau menghilangkan informasi yang penting didalamnya. Data tersebut akan diekstraksi fiturnya menggunakan vektor eigen dan nilai eigen yang telah diperoleh, sehingga dengan menggunakan PCA diharapkan data inputan tersebut dapat menunjang kinerja sistem untuk bisa bekerja optimal. Tahapan untuk melakukan reduksi dimensi dengan menggunakan PCA [8] [9], yaitu:

Data diatur sebagai satu set n data vektor $X_1 \dots X_n$, dimana matriks input untuk PCA adalah $X(i,j)$ untuk proses training, dengan i merupakan baris dan j merupakan kolom.

1. Hitung mean (X_{mean}) dari data:

$$X_{mean} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.1)$$

dengan: n = jumlah sampel atau jumlah data observasi, i = jumlah dimensi, X = data observasi.

2. Menentukan matriks kovarians (C):

$$C(x) = \frac{1}{n-1} \sum_{m=1}^n (X_m - X_{mean})^T (X_m - X_{mean}) \quad (2.2)$$

dengan: n = jumlah sampel atau jumlah data observasi, m = jumlah dimensi, X = data observasi, X_{mean} = mean data.

3. Menghitung nilai eigen dan vektor eigen:

$$C_{xvm} = \lambda vm \quad (2.3)$$

dan

$$|Cx - \lambda|vm = 0 \quad (2.4)$$

dengan: vm = vektor eigen, λ = nilai eigen, Cx = matriks kovariansi, m = jumlah dimensi.

4. Ekstrak nilai diagonal dari nilai eigen dan urutkan nilai diagonal tersebut secara descending. Pada penelitian ini, adapun pemilihan jumlah vektor eigen yang akan digunakan yaitu dengan menggunakan proporsi kumulatif varians (nilai eigen). Proporsi varians menunjukkan besarnya persentase informasi variabel-variabel asal yang terkandung dalam setiap vektor ciri (vektor eigen) berdasarkan nilai eigen dan memberikan interpretasi mengenai seberapa besar data dapat terwakili dalam dimensi yang telah direduksi [10].
5. Adapun Perhitungan Proporsi Varians (PPV) untuk setiap komponen utama (vektor eigen) adalah sebagai berikut:

$$PPV = \frac{\sum_{i=0}^k \lambda_i}{\sum_{i=0}^m \lambda_i} \times 100\% \quad (2.5)$$

dengan λ = nilai eigen, m = jumlah atribut data awal, k = jumlah atribut yang akan dipilih Principal Component (PC) dipilih dari k kolom eigen vector yang memiliki nilai diagonal tertinggi.

6. Kemudian didapat data baru dengan mengalikan PC dengan data.

$$\text{Data baru} = \text{PC} \times \text{data} \quad (2.6)$$

2.4 Random Forest

Terdapat beberapa metode decision tree yang dapat digunakan salah satunya CART (Classification and Regression Tree) dan salah satu teknik impurity yang dapat digunakan yaitu gini index. Gini index merupakan ukuran statistik yang menyatakan persebaran dari suatu data. Pada decision tree, gini index mengukur tingkat ketidaksamaan antara setiap atribut terhadap kelas. Semakin besar gini index yang didapatkan suatu atribut terhadap kelas maka atribut tersebut dianggap penting. Gini index mengukur tingkat homogenitas dari data D dengan persamaan [12]:

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2 \quad (2.7)$$

dimana m adalah banyaknya atribut kelas pada D , p_i adalah probabilitas suatu tuple pada D memiliki atribut kelas C_i dan dihitung dengan persamaan $\frac{|C_i D|}{|D|}$, yaitu membagi banyaknya atribut kelas C_i pada D dengan banyaknya tuple pada D .

Untuk membelah data dengan homogenitas maksimal, diperlukan splitting attribute yang akan membagi data menjadi 2 partisi dan memiliki nilai gini index terendah. Oleh karena itu, setiap atribut harus dicari nilai gini index-nya. Perhitungan gini index dari atribut pembelah A dilakukan dengan perkalian antara gini index dari tiap hasil partisi dengan bobotnya masing-masing. Hal ini dilakukan karena CART bersifat binary. Maka, untuk data D yang terpartisi oleh atribut pembelah A akan menjadi D_1 dan D_2 dengan persamaan:

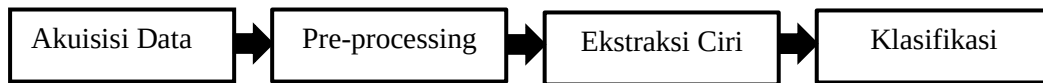
$$Gini_A = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (2.8)$$

Untuk setiap atribut, setiap kemungkinan partisi akan dihitung. Terdapat 2 macam tipe data atribut, diskrit dan kontinu. Perhitungan untuk 2 tipe data tersebut berbeda, tetapi inti dari perhitungannya adalah untuk mengambil subset yang memiliki gini index terendah sebagai kandidat atribut yang akan dipartisi.

3. Analisis dan perancangan sistem

Secara garis besar perancangan sistem pada penelitian ini dibagi menjadi 4 bagian utama, yaitu

akuisisi data, preprocessing, reduksi data atau ekstraksi ciri menggunakan metode PCA dan klasifikasi data menggunakan metode random forest.



Gambar 3-1 Diagram blok sistem

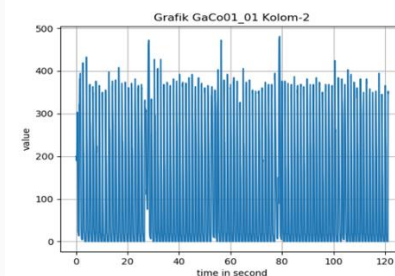
3.1 Akuisisi data

Data akan disiapkan terlebih dahulu untuk menjadi database utama pada sistem ini, Data diambil dari physiobank, database ini berisikan banyak data dan atribut yang nantinya akan diproses dalam sistem. Data sinyal VGRF terdiri atas 19 kolom dan berupa rekaman sinyal VGRF. Data sinyal masukan yang didapat dari physiobank akan digunakan dalam format text document (*.txt) [5].

Tabel 2.1 Format rekaman data sinyal VGRF

Kolom	Keterangan
1	Waktu dalam satuan detik
2-9	Nilai 8 sensor VGRF yang diletakan pada kaki sebelah kiri
10-17	Nilai 8 sensor VGRF yang diletakan pada kaki sebelah kanan
18	Total nilai sensor VGRF yang ada pada kaki kiri
19	Total nilai sensor VGRF yang ada pada kaki kanan

Dalam mendiagnosa data yang berasal dari sinyal VGRF terdeteksi parkinson atau tidak, dibutuhkan teknik pensinyalan yang berfokus pada total nilai sensor VGRF yang dipasang pada kaki sebelah kanan dan kiri [1]. Terdapat 2 jenis tipe file yaitu (.txt berisikan rekaman sinyal VGRF) dan (.xls berisikan keterangan data pasien). Total data yang diambil berjumlah 306 data. Data tersebut merupakan hasil percobaan dari 165 orang (93 parkinson dan 72 normal), namun setiap orang ada yang melakukan 2 kali percobaan sehingga total data menjadi 306. Data setiap pasien diambil dalam waktu 121,1715 detik dan menghasilkan data berupa 12119 baris dan 18 kolom atau matriks 12119x18.



Gambar 3-2 Grafik pasien pada salah satu sensor di 70000 data pertama (120 detik)

3.2 Preprocessing

Sebelum dilakukan proses reduksi data menggunakan PCA, data setiap pasien yang berupa matriks 12119x18 dirubah menjadi matriks 1x218142 namun didalam program dibatasi menjadi 70000 kolom dengan menghilangkan kolom data yang variasi datanya sedikit, hal ini bertujuan agar data lebih sederhana dan bervariasi. Total data berjumlah 304 dengan 92 data adalah data normal dan 214 data terjangkit parkinson. Hal selanjutnya adalah menyeimbangkan data, hal ini bertujuan untuk menghilangkan imbalance data dan meningkatkan hasil akurasi, maka perlu dilakukan sampling pada data pasien normal agar jumlah data pasien normal dan pasien parkinson sama. Data pasien berjumlah 214 pasien normal dan 214 pasien parkinson dengan total 428 data akan dibagi menjadi 70% sebagai data training dan 30% data testing yang dimana setiap data diacak secara urutan data agar setiap run program urutan data berbeda dan menghasilkan hasil yang lebih valid pada klasifikasinya. Lalu dilakukan proses Feature scaling through standardization atau normalisasi skor Z untuk standarisasi data.

3.3 Ekstraksi Ciri

Setelah data terorientasi secara rapi, data akan diolah menggunakan Principle Component Analysis (PCA). Didalam PCA terdapat parameter untuk menentukan berapa hasil output atau jumlah data, dimana data tersebut merupakan hasil dari proses reduksi dimensi dan setiap data tersebut merupakan representasi dari data matriks 1×70000 . Data yang diperoleh dari hasil PCA adalah berupa data 2 dimensi. Jadi misalkan parameter yang ditentukan adalah 200, maka data matriks 1×70000 akan dikalkulasi dengan perhitungan PCA dalam sistem menjadi matriks 1×200 , dimana hasil matriks tersebut merupakan representasi dari matriks sebelumnya. Matriks 1×200 juga merepresentasikan data satu pasien, maka dari itu jumlah data yang akan diproses dan diklasifikasi adalah matriks 428×200 karena terdapat 428 pasien.

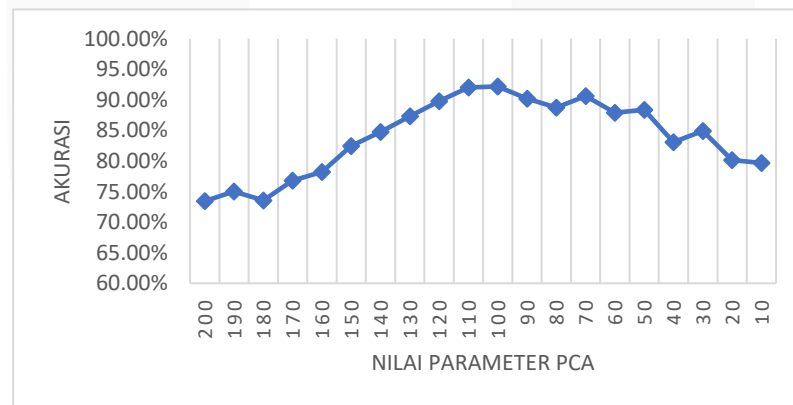
3.4 Klasifikasi data

Pada tahap sebelumnya telah dilakukan proses reduksi data menggunakan PCA untuk mengkarakteristik data. Terdapat dua tahap untuk proses klasifikasinya yaitu proses pembentukan model klasifikasinya dengan memanfaatkan data training serta pengujian terhadap model yang telah dibuat dengan memanfaatkan data testing. Untuk pembentukan setiap tree akan menerapkan metode bootstrap sehingga dapat menghasilkan tree yang berbeda ataupun sama karena metode ini mengambil beberapa sampel data secara acak. Dalam pembangkitan tree, gini index akan digunakan sebagai teknik impurity-nya. Hasil dari pembangkitan sejumlah tree akan digunakan sebagai model klasifikasi.

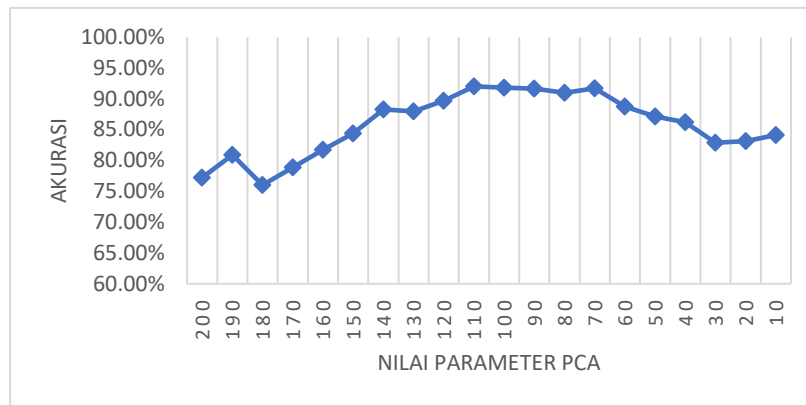
4. Sistem pengujian dan analisis

4.1 PCA

Grafik dibawah menunjukkan hasil pencarian parameter PCA terbaik. Ada 2 skenario grafik yang akan ditampilkan, perbedaan kedua grafik ini adalah pada parameter random forest dimana parameter random forest akan dipilih secara acak untuk dijadikan sampel. Pada skenario 1 parameter yang dijadikan sampel adalah $n_estimator = 21$ dan $max\ depth = 7$, sedangkan pada skenario 2 parameternya adalah $n_estimator = 31$ dan $max\ depth = 9$. Karakteristik yang akan diambil dari parameter PCA adalah dengan menilai hasil akurasi dan waktu komputasi, dimana akurasi yang akan ditampilkan adalah rata-rata dari hasil 3 kali percobaan.



Gambar 4-1 Grafik akurasi dengan skenario 1



Gambar 4-2 Grafik akurasi dengan skenario 2

Grafik diatas adalah hasil akurasi dari percobaan pada skenario 1 dan skenario 2 dengan menganalisa nilai parameter PCA dari range 400 sampai 10.

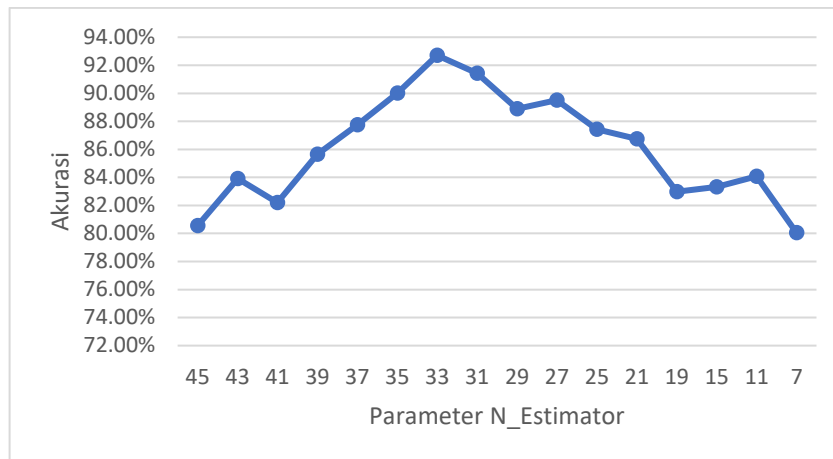
Hasil analisa dari dua skenario ini adalah ketika nilai parameter PCA terbaik pada jarak antara nilai 150 sampai 50, untuk memperkuat analisa ini penulis melakukan percobaan lebih lanjut dengan melakukan running program lebih dari 5 kali pada parameter-parameter tersebut dan hasilnya adalah ketika nilai parameter disekitar nilai 100, tingkat akurasi yang didapat lebih besar dan stabil yaitu diatas 90% dan nilai tingkat akurasi tertinggi didapat pada nilai parameter 100 dimana tingkat akurasinya adalah 92,82% dengan waktu komputasi terbaik yaitu 102 detik. Dengan mempertimbangkan hal tersebut maka penulis menentukan nilai pada parameter PCA adalah 100, dimana hal ini juga mengindikasikan bahwa data yang reduksi oleh PCA dari data sebelumnya adalah menjadi 100 data pada setiap pasien.

4.2 Random Forest

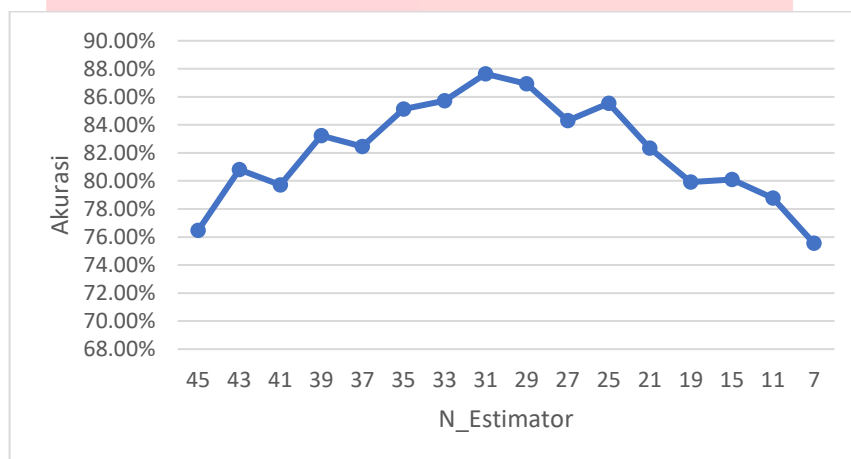
Dalam klasifikasi data menggunakan random forest, terdapat 2 parameter yang dapat menentukan hasil tingkat akurasi dan waktu komputasi. Tiga parameter tersebut yaitu:

1. **N_estimator**
N_estimator adalah parameter untuk mengatur jumlah tree pada voting tree, dimana semakin besar indeks yang dimasukkan pada parameter n_estimator maka akan semakin banyak jumlah tree yang ada pada voting tree random forest.
2. **Max Depth**
Max depth adalah parameter yang mengatur kedalam/tingkat level tree pada voting tree random forest. Semakin besar indeks yang dimasukkan akan mengakibatkan semakin banyak tingkatan level voting tree.

Dengan mempertimbangkan analisis yang dilakukan sebelumnya, maka nilai parameter PCA yang akan digunakan pada proses klasifikasi random forest adalah 100. Grafik dibawah ini adalah hasil dari percobaan yang dilakukan dengan melakukan running program dengan kombinasi ketiga parameter yang berbeda-beda dengan tujuan menemukan kombinasi terbaik, dimana kombinasi ini diharapkan dapat menghasilkan hasil tingkat akurasi dan waktu komputasi terbaik. Ada 2 skenario grafik yang akan ditampilkan untuk mencari parameter n_estimator terbaik. Perbedaan kedua grafik ini adalah pada parameter random forest max depth, dimana parameter tersebut akan dipilih secara acak untuk dijadikan sampel. Pada skenario 3 parameter yang dijadikan sampel adalah PCA = 100 dan max depth = 7, sedangkan pada skenario 4 parameternya adalah PCA = 100 dan max depth = 11. Parameter n_estimator diukur dari jarak nilai 7 sampai 45.



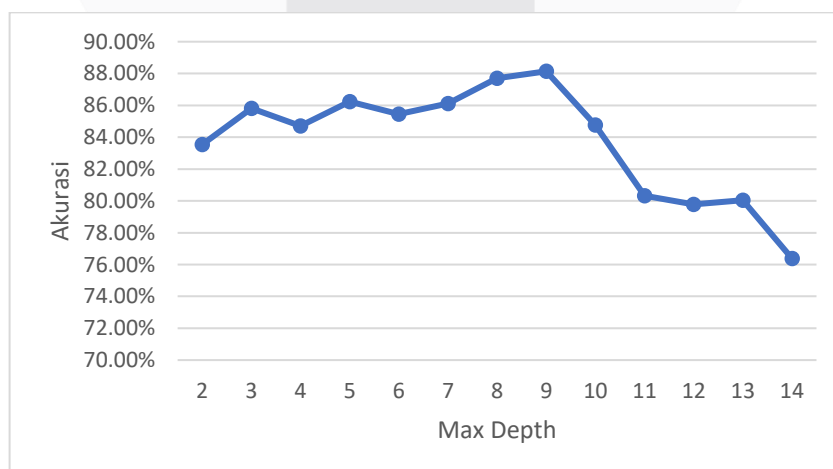
Gambar 4-3 Hasil akurasi random forest skenario 3



Gambar 4-4 Hasil akurasi random forest skenario 4

Dengan menganalisa hasil dari kedua grafik dapat disimpulkan bahwa parameter terbaik untuk PCA pada sistem ini adalah parameter diantara 31, 33 dan 35. Berdasarkan kesimpulan tersebut maka penulis memilih nilai indeks parameter untuk $n_{estimator}$ adalah 31 karena pada nilai tersebut hasil akurasi lebih stabil dibandingkan dengan 2 parameter lainnya.

Selanjutnya, untuk mencari parameter max depth akan dilakukan percobaan dengan menggunakan sampel nilai PCA = 100 dan $n_{estimator}$ = 31. Pada nilai parameter max depth = 9, didapatkan hasil akurasi rata-rata 88,14%.



Gambar 4-5 Hasil akurasi random forest skenario 5

Dari hasil analisis yang dilakukan berdasarkan running sistem dengan berbagai skema kombinasi didapatkan akurasi terbaik untuk klasifikasi parkinson berada pada kombinasi menggunakan parameter PCA = 100, n_estimator = 31 dan max depth = 9 dengan akurasi lebih dari 89% dan rata-rata waktu estimasi adalah 120 detik. Dimana waktu komputasi dan akurasi terbaik dicapai dengan tingkat akurasinya adalah 94,13% dengan waktu estimasi terbaik yaitu 102 detik.

5. Kesimpulan

Sistem ini dapat diaplikasikan dengan baik dalam membantu mendeteksi penyakit parkinson dengan input data berupa sinyal VGRF. Reduksi dimensi data dapat dicapai menggunakan metode PCA, yaitu dengan mengubah dimensi data menjadi matriks 1x100 (parameter PCA = 100) dari data utama 1x70000 atau 0,142% dari data utama. Parameter tersebut dipilih karena dapat mewakili karakteristik dari data utama, hal ini dibuktikan dengan tingkat akurasi dan waktu komputasi yang bagus dan stabil, terutama jika dianalisa dengan jumlah literasi yang banyak. Metode random forest dapat mengklasifikasi data dengan baik, hal ini dicapai dalam skenario percobaan 7 yaitu dengan kombinasi parameter n_estimator = 31, max dan depth = 9. Dengan kombinasi tersebut semua data pasien dapat diproses secara efisien, terutama pada jumlah cabang tree yang dikomputasi sehingga menghasilkan klasifikasi data yang bagus. Di dalam penelitian ini didapat akurasi terbaik yaitu sebesar 95,13 % namun proses komputasi memerlukan waktu 85 detik yang didapatkan pada skenario 7, sedangkan waktu komputasi terbaik didapatkan dengan mengubah parameter max depth = 2 dengan waktu 68 detik namun dengan akurasi 71%. Dari analisa diatas, terjadi trade off antara hasil akurasi dan waktu komputasi.



Daftar Pustaka

- [1] NINDS, Hope Through Research: Parkinson's Disease, USA: National Institutes of Health, 1994.
- [2] P. J. Jankovic, Parkinson's Disease Center and Movement Disorders Clinic, Texas, USA.: Baylor College of Medicine, Department of Neurology Texas, 2008, p. Department of Neurology.
- [3] G. A. Arrazaq, "Diagnosis Of Parkinson's Disease Through Gait Analysis Based On Vgrf Using Wavelet And Support Vector Machine," vol. Vol 4, p. 1855, 2017.
- [4] L. Medeiros, "A Gait Analysis Approach to Track Parkinson's Disease Evolution Using Principal Component Analysis," *IEEE 29th Int. Symp. Comput. Med. Syst.*, vol. 29, p. 6, 2016.
- [5] M. Margaret, "Parkinsonism: Onset, Progression, And Mortality," *Neurology* 17.5, vol. 5, p. 427, 1967.
- [6] A. Goldberger and L. Amaral, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation [Online]*, vol. 101 (23), pp. 215-220, 2000.
- [7] P. e. al, "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825-2830, 2011.
- [8] I. Smith, A tutorial on Principal Components Analysis, New York: John Wiley & Sons Inc, 2002.
- [9] I. Bajwa, "PCA Based Classification Of Single-Layered Cloud Types," *Market Forces*, vol. 1, 2005.
- [10] I. Jolliffe, Principal Component Analysis, Berlin: Springer, 2011.
- [11] H. Varian, "Bootstrap Tutorial," *Mathematics Journal*, vol. 9.4, pp. 768-775, 2005.
- [12] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, pp. 5-22, 2001.
- [13] J. Friedman, "The elements of statistical learning," *Machine Learning*, vol. 1, pp. 5-32, 2001.
- [14] A. Ramadhan and B. Susetyo, "PENERAPAN METODE KLASIFIKASI RANDOM FOREST DALAM MENGIDENTIFIKASI FAKTOR PENTING PENILAIAN MUTU PENDIDIKAN," *Jurnal Pendidikan Dan Kebudayaan*, vol. 4, pp. 169-182, 2019.
- [15] M. F. Setiadi, "mfikri.com," December 2017. [Online]. Available: <http://mfikri.com/artikel/aturan-dan-teknik-dalam-melakukan-normalisasi-data.html>. [Accessed April 2019].
- [16] R. Purnamasari and I. Zakia, "FAST FOURIER TRANSFORM SPARSITY FOR HIGH QUALITY WEATHER RADAR RECONSTRUCTION," *2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7748-7751, 2019.