KLASIFIKASI SPESIES BERDASARKAN *DNA BARCODE SEQUENCE* MENGGUNAKAN RANDOM FERNS

SKRIPSI

diajukan untuk memenuhi sebagian syarat untuk memperoleh gelar
Sarjana Komputer Program Studi Ilmu Komputer

oleh
M AMMAR FADHLUR RAHMAN
NIM 1507506

PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN PENDIDIKAN ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
2022

KLASIFIKASI SPESIES BERDASARKAN DNA BARCODE SEQUENCE MENGGUNAKAN RANDOM FERNS

oleh
M Ammar Fadhlur Rahman
NIM 1507506

Sebuah skripsi yang diajukan untuk memenuhi sebagian syarat untuk memperoleh gelar Sarjana Komputer pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

M AMMAR FADHLUR RAHMAN

1507506

KLASIFIKASI SPESIES BERDASARKAN DNA BARCODE SEQUENCE MENGGUNAKAN RANDOM FERNS

DISETUJUI DAN DISAHKAN OLEH PEMBIMBING:

Pembimbing I,

Lala Septem Riza, M.T., Ph.D.

NIP. 197809262008121001

Pembimbing II,

Herbert Siregar, M.T.

NIP. 197005022008121001

Mengetahui,

Ketua Program Studi Ilmu Komputer

Dr. Rani Megasari, M.T.

NIP. 198705242014042002

PERNYATAAN

Dengan ini saya menyatakan bahwa skripsi dengan judul “Klasifikasi Spesies berdasarkan *DNA Barcode Sequence* menggunakan Random Ferns” ini beserta seluruh isinya adalah benar-benar karya saya sendiri. Saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya saya ini.

Bandung, Agustus 2022

M Ammar Fadhlur Rahman
1507506

KLASIFIKASI SPESIES BERDASARKAN DNA BARCODE SEQUENCE MENGGUNAKAN RANDOM FERNS

oleh
M Ammar Fadhlur Rahman
NIM 1507506

ABSTRAK

Machine learning telah diterapkan dalam berbagai domain, termasuk bioinformatika. Salah satu persoalan bioinformatika yang dapat diselesaikan dengan pendekatan *machine learning* adalah klasifikasi spesies. Penelitian ini berupaya mengklasifikasikan spesies ke dalam famili berdasarkan sekuens *DNA barcode* menggunakan pendekatan *supervised learning* dengan algoritma Random Ferns. Digunakan model komputasi dengan 13 tahapan, termasuk pengunduhan data, rangkaian *praproses* data, *model training*, prediksi, dan evaluasi. Gen *ribulose-1,5-bisphosphate carboxylase-oxygenase large sub-unit* (rbcL) yang merupakan salah satu lokus *DNA barcode* untuk tanaman, digunakan untuk merepresentasikan spesies dalam famili Amarilis dan Lili. Berdasarkan hasil eksperimen dengan 1.245 sekuens DNA *training* dan 220 sekuens testing menunjukkan bahwa Random Ferns dapat digunakan untuk mengklasifikasikan spesies ke dalam famili yang sesuai secara cepat dan akurat. Tercapai tingkat akurasi yang konsisten hingga 99,09% dalam waktu *training* selama 180ms dengan hanya menggunakan memori sebanyak 14,5MB. Perbandingan dengan algoritma Random Forest yang menjadi *state-of-the-art* menunjukkan Random Ferns dapat mencapai tingkat akurasi yang sepadan secara lebih efisien.

Kata kunci: *machine learning*, *random ferns*, klasifikasi *supervised*, klasifikasi spesies, *DNA barcode*, gen *rbcL*, analisa data, bioinformatika

SPECIES CLASSIFICATION BASED ON DNA BARCODE SEQUENCES USING RANDOM FERNS

arranged by
M Ammar Fadhlur Rahman
NIM 1507506

ABSTRACT

Machine learning has been applied in various domains, including bioinformatics. One of the bioinformatics problems that can be solved by using a machine learning approach is species classification. This study attempts to classify species into families based on their DNA barcode sequences using supervised learning approach, i.e., the Random Ferns algorithm. A computational model consisting of 13 steps was proposed, including data retrieval, a series of data preprocessing, model training, prediction, and evaluation. The ribulose-1,5-bisphosphate carboxylase-oxygenase large sub-unit (rbcL) gene that has been selected as one of the DNA barcode loci for plants is used to represent species in the Amaryllidaceae and Liliaceae families. By using 1,245 DNA sequences for training and 220 sequences as testing data, the experiment results show that Random Ferns can be used to classify species sequences quickly and accurately into appropriate families based on their DNA barcode sequences. The trained model could achieve persistent accuracy result as high as 99,09% within 180ms of training time and using only 14,5 MB of memory. A comparison against the state-of-the-art Random Forest algorithm showed Random Ferns was able to achieve the same level of accuracy more efficiently.

Keywords: machine learning, random ferns, supervised classification, species classification, DNA barcode, rbcL gene, data analysis, bioinformatics.

KATA PENGANTAR

Puji dan syukur ke hadirat Allah SWT yang sudah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi "Klasifikasi Spesies berdasarkan *DNA Barcode Sequence* menggunakan Random Ferns" ini dengan sebaik-baiknya. Skripsi ini disusun dalam rangka memenuhi sebagian syarat untuk memperoleh gelar Sarjana Komputer pada jenjang studi strata 1 (S1) di Program Studi Ilmu Komputer, Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia. Penulis menyadari bahwa dalam penulisan skripsi ini masih terdapat kesalahan dan kekurangan yang disebabkan oleh keterbatasan bahan yang diperoleh dan kemampuan yang penulis miliki. Oleh karena itu, penulis mengharapkan kritik dan saran sebagai bahan masukan bagi penulis di masa yang akan datang. Semoga Allah SWT dapat memberikan yang terbaik untuk semua pihak termasuk pembaca, semoga skripsi ini dapat memberikan manfaat bagi penulis khususnya, dan umumnya untuk semua pihak yang membacanya.

Bandung, Agustus 2022

Penulis

UCAPAN TERIMA KASIH

Dalam proses penulisan penelitian ini, penulis mendapat bimbingan, dorongan, dan bantuan yang diberikan baik secara langsung maupun tidak langsung dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan rasa terima kasih serta penghargaan yang setinggi-tingginya kepada:

1. Kedua orang tua, ayahanda (alm.) Pandiyo dan ibunda Didah Rosidah yang selalu memberikan doa dan dukungan moral dan materi, serta selalu menjadi penyemangat utama dalam menempuh pendidikan tinggi sehingga penulis dapat menyelesaikan skripsi ini.
2. Bapak Lala Septem Riza, Ph.D., selaku pembimbing I, pembimbing akademik, dan Kepala Departemen Pendidikan Ilmu Komputer atas segala waktu yang dicurahkan untuk membimbing penulis demi terselesaikannya penelitian skripsi ini.
3. Bapak Herbert Siregar, M.T., selakuo pembimbing II yang telah memberikan berbagai saran kepada penulis selama proses penulisan skripsi.
4. Bapak Topik Hidayat, Ph.D., selaku Dosen Departemen Pendidikan Biologi yang telah membantu penulis salah satunya dalam memahami konsep dasar dari penelitian ini.
5. Ibu Dr. Rani Megasari, M.T., selaku Ketua Program Studi Ilmu Komputer atas dukungannya kepada seluruh mahasiswa skripsi 2021/2022.
6. Ibu Rosa Ariani Sukamto, M.T., selaku Dosen Pemrograman Dasar dan Bapak Eddy Prasetyo Nugroho, M.T., selaku Dosen Rekayasa Perangkat Lunak yang telah memberikan motivasi kepada penulis untuk selalu belajar dan membagikannya saat kuliah.
7. Bapak dan Ibu Dosen Program Studi Pendidikan Ilmu Komputer dan Ilmu Komputer yang telah berbagi ilmu yang sangat bermanfaat kepada penulis.
8. Sahabat “Itok Meti3”: Adie, Adit, Arga, Dimas, Fakhri, Farhan, Fiko, Hafidz, Rahman, Trisna, dan Yogi yang senantiasa memberikan dukungan,

semangat, canda dan tawa kepada penulis baik selama proses perkuliahan maupun selama proses pengerjaan skripsi ini.

9. Rekan kuliah 7 tahun: Agung, Burhanudin, Khamal, Firmansyah, dan Rendy yang senantiasa mendukung satu sama lain hingga dapat lulus di waktu yang tepat ini.
10. DEPADSOSPOL BEM KEMAKOM 2016/2017, yang telah membentuk keluarga baru di KEMAKOM pada tahun pertama penulis berorganisasi, serta memberikan kedekatan luar biasa hingga saat ini.
11. Kelas C 2015, yang sama-sama berjuang dari awal perkuliahan dari awal hingga ke titik akhir perkuliahan.
12. Bapak Muhammad Dahlan, yang selalu memberikan doa dan dukungan kepada penulis untuk menyelesaikan salah satu tahap kehidupan ini.
13. Pembayar pajak di Republik Indonesia, atas kebaikannya melalui program Bidikmisi dari Kementerian Riset, Teknologi, dan Pendidikan Tinggi (Ristekdikti) yang telah memberikan kesempatan bagi penulis untuk menempuh pendidikan strata 1 ini.
14. Komodowan dan Komodowati r/indonesia yang senantiasa mengingatkan penulis untuk menyelesaikan studinya.
15. Berbagai pihak yang telah membantu menyelesaikan penulisan penelitian ini namun tidak bisa penulis sebutkan satu per satu.

Semoga semua kebaikan yang telah diberikan kepada penulis mendapatkan balasan yang berlipat dari Allah SWT. Aamiin.

Bandung, Agustus 2022

Penulis

DAFTAR ISI

ABSTRAK i

ABSTRACT ii

KATA PENGANTAR iii

UCAPAN TERIMA KASIH iv

DAFTAR ISI vi

DAFTAR TABEL x

DAFTAR GAMBAR xi

BAB I PENDAHULUAN 1

1.1 Latar Belakang Penelitian 1

1.2 Rumusan Masalah Penelitian 7

1.3 Tujuan Penelitian 7

1.4 Manfaat/Signifikansi Penelitian 8

1.5 Batasan Masalah 8

1.6 Struktur Organisasi Skripsi 8

BAB II KAJIAN PUSTAKA 10

2.1 Peta Literatur 10

2.2 Bioinformatika 11

2.3 *Deoxyribonucleic Acid* (DNA) 12

2.4 *DNA Barcode Database* 14

2.5 GenBank 15

2.6 *DNA Sequence Alignment* 16

2.7 Multiple Sequence Comparison by Log-Expectation (MUSCLE) 17

2.8 Sekuens Konsensus 18
2.9 *Sequence Distance Matrix* 19
2.10 Klasifikasi 20
2.11 Random Ferns 20
2.12 *Confusion Matrix* 26
2.13 Pengukuran *Performance* Model Klasifikasi 27
2.13.1 *Accuracy* dan *Error Rate* 27
2.13.2 *Recall* 27
2.13.3 *Precision* 28
2.13.4 $\boldsymbol{F1}$ *Score* 28
2.13.5 *Cohen's Kappa Coefficient* 28
BAB III METODE PENELITIAN 30
3.1 Desain dan Instrumen Penelitian 30
3.1.1 Alat Penelitian 30
3.1.2 Bahan Penelitian 32
3.1.3 Metode Penelitian 32
3.2 Prosedur Penelitian 34
BAB IV TEMUAN DAN PEMBAHASAN 36
4.1 Pengumpulan Data 36
4.1.1 Mengunduh Data dari GenBank 36
4.1.2 Penjelasan Isi File 39
4.2 Model Komputasi *Klasifikasi Spesies berdasarkan DNA Barcode Sequence* menggunakan *Random Ferns* 41
4.2.1 *Retrieve Dataset* 43
4.2.2 *Pre-Processing* 43
4.2.3 *Model Training* 46

4.2.4 *Prediction* 46
4.2.5 *Evaluation* 46
4.3 Implementasi Klasifikasi Spesies berdasarkan DNA *Barcode Sequence* menggunakan Random Ferns 46
4.3.1 Analisis & Penentuan Kebutuhan 47
4.3.2 Desain Sistem dan Perangkat Lunak 48
4.3.3 Implementasi 48
4.3.4 Pengujian 65
4.4 Studi Kasus 66
4.4.1 Data yang digunakan 67
4.4.2 Skenario Eksperimen 68
4.4.3 Instrumen Evaluasi 69
4.5 Hasil Eksperimen 70
4.5.1 Skenario 1 70
4.5.2 Skenario 2 71
4.5.3 Skenario 3 72
4.6 Pembahasan 73
4.6.1 *Dataset Training* 74
4.6.2 Model Klasifikasi 74
4.6.3 Perbandingan dengan Random Forest 78
4.6.4 Prediksi pada Spesies Terpilih 81
BAB V KESIMPULAN, IMPLIKASI, DAN REKOMENDASI 84
5.1 Kesimpulan 84
5.2 Implikasi 84
5.3 Rekomendasi 85
DAFTAR PUSTAKA 86

LAMPIRAN 94

Lampiran 1 Query Lengkap Dataset Amaryllis Train 94

Lampiran 2 Query Lengkap Dataset Lily Train 94

Lampiran 3 Query Lengkap Dataset Amaryllis Test 95

Lampiran 4 Query Lengkap Dataset Lily Test 95

Lampiran 5 Hasil *10x10-fold Cross Validation* Skenario 1 96

Lampiran 6 Hasil *10x10-fold Cross Validation* Skenario 2 98

Lampiran 7 Hasil *10x10-fold Cross Validation* Skenario 3 100

Lampiran 8 Grafik Tingkat Akurasi Rata-Rata Hasil *10x10-Fold Cross-Validation* (Gambar 4.41) 102

Lampiran 9 Grafik Perbandingan Sebaran Tingkat Akurasi dan Kappa Algoritma Random Ferns dan Random Forest (Gambar 4.43) 105

Lampiran 10 Hasil Prediksi Famili 108

Lampiran 11 Sekuens Konsensus Famili Amarilis 120

Lampiran 12 Sekuens Konsensus Famili Lili 121

Lampiran 13 *Distance Matrix* Sekuens terhadap Konsensus 122

DAFTAR TABEL

Tabel 1.1 Perbandingan *performance* beberapa algoritma klasifikasi 6

Tabel 2.1 Kode IUPAC 14

Tabel 2.2 Hasil kalkulasi *distance matrix* terhadap sekuens DNA 20

Tabel 2.3 Struktur *confusion matrix* 26

Tabel 4.1 Rangkuman *dataset* 37

Tabel 4.2 Daftar spesies yang dipilih untuk data testing 38

Tabel 4.3 Daftar spesies yang terkumpul dalam *dataset* testing 39

Tabel 4.4 Pengujian dengan metode *Black Box* 66

Tabel 4.5 Konfigurasi eksperimen 68

Tabel 4.6 *Confusion matrix* skenario 1 terhadap label famili NCBI..... 71

Tabel 4.7 *Confusion matrix* skenario 1 terhadap label famili hasil konsensus 71

Tabel 4.8 *Confusion matrix* skenario 2 terhadap label famili NCBI..... 71

Tabel 4.9 *Confusion matrix* skenario 2 terhadap label famili hasil konsensus 72

Tabel 4.10 *Confusion matrix* skenario 3 terhadap label famili NCBI..... 72

Tabel 4.11 *Confusion matrix* skenario 3 terhadap label famili hasil konsensus ... 73

Tabel 4.12 Statistik hasil eksperimen..... 73

Tabel 4.13 Perbandingan Random Ferns dengan Random Forest 78

Tabel 4.14 Perbedaan pada label famili dari NCBI dan hasil konsensus sekuens 81

Tabel 4.15 Perbedaan hasil prediksi dengan label famili dari NCBI dan hasil konsensus 82

DAFTAR GAMBAR

Gambar 1.1 Proses pengolahan spesimen makhluk hidup hingga menjadi *DNA barcode* 4

Gambar 2.1 Peta literatur penelitian 11

Gambar 2.2 Interaksi disiplin ilmu yang berkontribusi pada pembentukan bidang ilmu bioinformatika 12

Gambar 2.3 Struktur asam *deoksiribonukleat* (DNA) 13

Gambar 2.4 Struktur dalam Internasional Nucleotide Sequence Database Collection (INSDC) 15

Gambar 2.5 Tampilan antarmuka *website* NCBI 15

Gambar 2.6 Sekuens DNA sebelum dan sesudah dilakukan *sequence alignment* 17

Gambar 2.7 *Flowchart* algoritma MUSCLE 18

Gambar 2.8 Konsensus Sekuens 19

Gambar 2.9 Sekuens DNA berisi 4 sekuens 19

Gambar 2.10 Perbedaan struktur model Decision Tree dan Fern. 21

Gambar 2.11 Pakis (*fern*) 21

Gambar 2.12 Hasil pemilihan *subset* dari *feature* data input untuk setiap *fern* 22

Gambar 2.13 Struktur *fern* dengan kedalaman 3 22

Gambar 2.14 Contoh penempatan data *input* dalam *fern* 23

Gambar 2.15 Contoh sebaran data *input* dengan satu kelas label yang sama dalam *fern* dan representasinya dalam grafik probabilitas 24

Gambar 2.16 Memilih *class* posterior untuk mengklasifikasikan data testing 25

Gambar 2.17 Struktur Random Ferns dengan 3 buah *fern*..... 25

Gambar 3.1 Model pengembangan perangkat lunak *Waterfall* (atas) dan desain penelitian Klasifikasi Spesies berdasarkan DNA Barcode Sequence menggunakan Random Ferns (bawah) 33

Gambar 4.1 Pencarian sekuens DNA melalui website GenBank 37

Gambar 4.2 Contoh sekuens DNA dalam format FASTA..... 40

Gambar 4.3 Model komputasi Klasifikasi Spesies berdasarkan DNA Barcode Sequence menggunakan Random Ferns..... 42

Gambar 4.4 Konversi sekuens DNA dari format FASTA ke tipe data DNAStringSet 44

Gambar 4.5 Sekuens DNA sebelum dan sesudah dilakukan *trimming*..... 44

Gambar 4.6 Sekuens DNA dalam *dataframe*..... 45

Gambar 4.7 Instalasi *package* yang digunakan dalam penelitian 49

Gambar 4.8 Membuat vektor nama spesies untuk testing set 49

Gambar 4.9 Membuat *query exclusion* dari vektor nama spesies 50

Gambar 4.10 Membuat query pencarian untuk *training* data 50

Gambar 4.11 Mencari sekuens di GenBank berdasarkan query pencarian..... 50

Gambar 4.12 Mengunduh sekuens DNA dari hasil pencarian 51

Gambar 4.13 Menyimpan sekuens DNA ke dalam *file* FASTA 51

Gambar 4.14 Membuat *query inclusion* dari vektor nama spesies 51

Gambar 4.15 Membuat query pencarian untuk testing data 52

Gambar 4.16 Membaca dan mengonversi *file* FASTA ke dalam tipe data DNAStringSet 53

Gambar 4.17 Melakukan *sequence alignment* pada data sekuens DNA..... 53

Gambar 4.18 Konversi sekuens DNA dari tipe data hasil *alignment* DNAMultipleAlignment ke tipe data DNAbin 53

Gambar 4.19 Proses *sequence trimming* secara manual 54

Gambar 4.20 *Sequence trimming* menggunakan fungsi trimEnds() dari *package* ips 55

Gambar 4.21 Konversi sekuens DNA dari tipe data *DNAbin* ke tipe data *dataframe* 55

Gambar 4.22 Mengambil label nama dari *file* sekuens DNA *dataset* testing 56

Gambar 4.23 Membuat *dataframe* baru berisi sekuens DNA untuk *training* 56

Gambar 4.24 Membuat *dataframe* baru berisi sekuens DNA untuk testing 56

Gambar 4.25 *Deduplikasi* baris dalam *dataframe* 57

Gambar 4.26 Mengonversi tipe data dalam kolom *dataframe* menjadi *factor* 58

Gambar 4.27 Mengisi sel *dataframe* yang kosong dengan simbol gap 58

Gambar 4.28 Membuat label famili pada nama sekuens Amarilis 59

Gambar 4.29 Membuat label famili pada nama sekuens Amarilis 59

Gambar 4.30 Menggabungkan label ke *dataframe training* 60

Gambar 4.31 Menghapus nama baris dari *dataframe* 60

Gambar 4.32 Training model Random Ferns 60

Gambar 4.33 Menyesuaikan model Random Ferns 61

Gambar 4.34 Konfigurasi cross-validation dan parameter untuk algoritma Random Ferns 62

Gambar 4.35 Menjalankan training model Random Ferns dengan konfigurasi cross-validation dan parameter tuning 62

Gambar 4.36 Konfigurasi *parallel processing* di komputer server 63

Gambar 4.37 Prediksi label famili pada data *test* menggunakan model hasil *training* 63

Gambar 4.38 Membuat *confusion matrix* hasil prediksi data testing..... 64

Gambar 4.39 Kalkulasi nilai *error* hasil prediksi data testing 64

Gambar 4.40 Mengekspor data tes disertai hasil prediksi 65

Gambar 4.41 Tingkat akurasi rata-rata skenario eksperimen 75

Gambar 4.42 Nilai kappa rata-rata skenario eksperimen..... 77

Gambar 4.43 Perbandingan Algoritma Random Ferns (RFe) dan Random Forest (RF) berdasarkan nilai akurasi dan kappa dari hasil *10x10-fold cross validation* 79

Gambar 4.44 Perbandingan Algoritma Random Ferns dan Random Forest berdasarkan aspek durasi *training* (kiri) dan *memory* yang digunakan (kanan) pada masing-masing skenario..... 80

DAFTAR PUSTAKA

Bao, L., & Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, *21*(10), 2185–2190. https://doi.org/10.1093/bioinformatics/bti365

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ*, *324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of disease in childhood - Education & practice edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Belson, W. A. (1959). Matching and Prediction on the Principle of Biological Classification. *Applied Statistics*, *8*(2), 65. https://doi.org/10.2307/2985543

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127. https://doi.org/10.1561/2200000006

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Centers for Disease Control and Prevention. (2016). *Whole Genome Sequencing (WGS)*. Centers for Disease Control and Prevention. https://www.cdc.gov/pulsenet/pathogens/wgs.html

Chang, W., Luraschi, J., & Mastny, T. (2020). *profvis: Interactive Visualizations for Profiling R Code*. https://cran.r-project.org/package=profvis

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation

Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access*, *9*, 78368–78381. https://doi.org/10.1109/ACCESS.2021.3084050

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research*, *13*(9), 3021–3030. https://doi.org/10.1093/nar/13.9.3021

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242. http://www.jstor.org/stable/2983890

Dahl, C. A., & Strausberg, R. L. (1996). *Human genome project: revolutionizing biology through leveraging technology* (G. E. Cohn, S. A. Soper, & C. H. W. Chen (ed.); hlm. 190–201). https://doi.org/10.1117/12.237605

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Fix, E., & Hodges, J. L. (1989). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 238. https://doi.org/10.2307/1403797

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, *269*(5223), 496–512. https://doi.org/10.1126/science.7542800

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro, A., Sciaini, Marco, Scherer, & Cédric. (2021). *{viridis} - Colorblind-Friendly Color Maps for R*. https://doi.org/10.5281/zenodo.4679424

Global Biodiversity Information Facility. (2021). *Accelerating biodiversity research through DNA barcodes, collection and observation data*. https://docs.gbif.org/course-dna-barcoding

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Ratnasingham, S., & DeWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological sciences*, *270 Suppl*, S96-9. https://doi.org/10.1098/rsbl.2003.0025

Heibl, C. (2008). *PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages*. http://www.christophheibl.de/Rpackages.html

Hesper, B., & Hogeweg, P. (1970). Bioinformatica: een werkconcept. *Kameleon*, *1*(6), 28–29.

Hidayat, T., & Azzahra, A. (2021). *Genom Kloroplas Mendukung Penggabungan Amaryllidaceae Kedalam Liliaceae*. [Unpublished Manusript], Universitas Pendidikan Indonesia.

Hu, Y., Da, Q., Zeng, A., Yu, Y., & Xu, Y. (2018). Reinforcement Learning to Rank in E-Commerce Search Engine. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 368–377. https://doi.org/10.1145/3219819.3219846

Ilman, M. N. (2019). *Algoritma Optimasi Spiral Dynamics menggunakan Bahasa Pemrograman R untuk DNA Barcoding* [Universitas Pendidikan Indonesia]. http://repository.upi.edu/38393/

Jacobson, K., Murali, V., Newett, E., Whitman, B., & Yon, R. (2016). Music Personalization at Spotify. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, 373–373. https://doi.org/10.1145/2959100.2959120

Johnson, A. D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics (Oxford, England)*, *26*(10), 1386–1389. https://doi.org/10.1093/bioinformatics/btq098

Kuhn, M. (2022). *caret: Classification and Regression Training*. https://cran.r-project.org/package=caret

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier

Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, *51*(2), 181–207. https://doi.org/10.1023/A:1022859003006

Kursa, M. B. (2014a). rFerns : An Implementation of the Random Ferns Method for General-Purpose Machine Learning. *Journal of Statistical Software*, *61*(10), 1–13. https://doi.org/10.18637/jss.v061.i10

Kursa, M. B. (2014b). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, *15*(1), 8. https://doi.org/10.1186/1471-2105-15-8

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Li, Y., Wang, Z., You, Z.-H. H., Li, L.-P. P., & Hu, X. (2022). Predicting Protein-Protein Interactions via Random Ferns with Evolutionary Matrix Representation. *Computational and Mathematical Methods in Medicine*, *2022*, 1–11. https://doi.org/10.1155/2022/7191684

Liljas, L. (2013). Consensus Sequences. Dalam S. Maloy & K. Hughes (Ed.), *Brenner's Encyclopedia of Genetics (Second Edition)* (Second Edition, hlm. 163–164). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374984-0.00325-9

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Manwell, C., & Baker, C. M. A. (1963). A sibling species of sea cucumber discovered by starch gel electrophoresis. *Comparative Biochemistry and Physiology*, *10*(1), 39–53. https://doi.org/10.1016/0010-406X(63)90101-4

Meher, P. K., Sahu, T. K., Gahoi, S., Tomar, R., & Rao, A. R. (2019). funbarRF: DNA barcode-based fungal species prediction using multiclass Random Forest supervised learning model. *BMC Genetics*, *20*(1), 2. https://doi.org/10.1186/s12863-018-0710-z

Meher, P. K., Sahu, T. K., & Rao, A. R. (2016). Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. *Gene*, *592*(2), 316–324. https://doi.org/10.1016/j.gene.2016.07.010

Microsoft Corporation, & Weston, S. (2022). *doParallel: Foreach Parallel Adaptor for the "parallel" Package*. https://cran.r-project.org/package=doParallel

Microsoft, & Weston, S. (2022). *foreach: Provides Foreach Looping Construct*. https://cran.r-project.org/package=foreach

Mitchell, T. (1997). *Machine Learning* (1 ed.). McGraw-Hill.

Mount, D. (2001). *Bioinformatics: Sequence and Genome Analysis* (1st ed.). Cold Spring Harbor Laboratory Press.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists* (1 ed.). O'Reilly Media.

National Human Genome Research Institute. (2022). *Talking Glossary of Genomic and Genetic Terms*. https://www.genome.gov/genetics-glossary

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nieuwenhuysen, P. (2018). Information Discovery and Images A Case Study of Google Photos. *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, 16–21. https://doi.org/10.1109/ETTLIS.2018.8485238

Nurfathiya, M. I. (2020). *DNA Barcoding dengan Algoritma Particle Swarm Optimization menggunakan Apache Spark SQL* [Universitas Pendidikan Indonesia]. http://repository.upi.edu/49774/

Özuysal, M., Calonder, M., Lepetit, V., & Fua, P. (2010). Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 448–461. https://doi.org/10.1109/TPAMI.2009.23

Ozuysal, M., Fua, P., & Lepetit, V. (2007). Fast Keypoint Recognition in Ten Lines of Code. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383123

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2022). *Biostrings: Efficient manipulation of biological strings*. https://bioconductor.org/packages/Biostrings

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rudis, B. (2020). *hrbrthemes: Additional Themes, Theme Components and Utilities for "ggplot2."* https://cran.r-project.org/package=hrbrthemes

Russel, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach, Global Edition* (4 ed.). Pearson.

Salemi, M., Lemey, P., & Vandamme, A. M. (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press. https://books.google.co.id/books?id=DeD%5C_lQ-kBPQC

Salzberg, S. (1995). Locating protein coding regions in human DNA using a decision tree algorithm. *Journal of computational biology : a journal of computational molecular cell biology*, *2*(3), 473–485. https://doi.org/10.1089/cmb.1995.2.473

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer. http://lmdvr.r-forge.r-project.org

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Bolchacova, E., Voigt, K., Crous, P. W., Miller, A. N., Wingfield, M. J., Aime, M. C., An, K.-D., Bai, F.-Y., Barreto, R. W., Begerow, D., Bergeron, M.-J., Blackwell, M., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–

6246. https://doi.org/10.1073/pnas.1117018109

Selzer, P. M., Marhöfer, R. J., & Koch, O. (2018). *Applied Bioinformatics. An Introduction* (2nd ed.). Springer.

Sheng, Q., Moreau, Y., Smet, F. De, Marchal, K., & Moor, B. De. (2005). Advances in Cluster Analysis of Microarray Data. Dalam *Data Analysis and Visualization in Genomics and Proteomics* (hlm. 153–173). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470094419.ch10

Sohsah, G. N., Ibrahimzada, A. R., Ayaz, H., & Cakmak, A. (2020). Scalable classification of organisms into a taxonomy using hierarchical supervised learners. *Journal of Bioinformatics and Computational Biology*, *18*(05), 2050026. https://doi.org/10.1142/S0219720020500262

Sommerville, I. (2015). *Software Engineering* (10th ed.). Pearson Education.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

van Bemmelen van der Plaat, A., van Treuren, R., & van Hintum, T. J. L. (2021). Reliable genomic strategies for species classification of plant genetic resources. *BMC Bioinformatics*, *22*(1), 173. https://doi.org/10.1186/s12859-021-04018-6

Wang, C., Zhang, Y., & Han, S. (2020). Its2vec: Fungal Species Identification Using Sequence Embedding and Random Forest Classification. *BioMed Research International*, *2020*, 1–11. https://doi.org/10.1155/2020/2468789

Wankhede, K., Wukkadada, B., & Nadar, V. (2018). Just Walk-Out Technology and its Challenges: A Case of Amazon Go. *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 254–257. https://doi.org/10.1109/ICIRCA.2018.8597403

Weitschek, E., Fiscon, G., & Felici, G. (2014). Supervised DNA Barcodes species classification: analysis, comparisons and results. *BioData Mining*, *7*(1), 4. https://doi.org/10.1186/1756-0381-7-4

Weyenberg, G., & Yoshida, R. (2015). Chapter 12 - Reconstructing the Phylogeny: Computational Methods. Dalam R. S. Robeva (Ed.), *Algebraic and Discrete Mathematical Methods for Modern Biology* (hlm. 293–319). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-801213-0.00012-5

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/package=dplyr

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*(2), 520. https://doi.org/10.32614/RJ-2017-058

Wright, E. S. (2016). Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*, *8*(1), 352–359.

Xu, R., & Wunsch, D. C. (2008). *Clustering* (illustrate). Wiley-IEEE Press.

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.01032

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, *82*, 128–150. https://doi.org/10.1016/j.eswa.2017.04.003

INDEKS