KLASIFIKASI GENOM HUMAN PAPILLOMAVIRUS GENUS BETA DAN GAMMA MENGGUNAKAN FITUR 3-MER DAN ALGORITMA MACHINE LEARNING

Nahda Hayu Hemalina[1], Trimono[2]
23083010066@student.upnjatim.ac.id[1] , trimono.stat@upnjatim.ac.id[2]
UPN Veteran Jawa Timur

ABSTRAK

Human Papillomavirus (HPV) merupakan virus DNA yang memiliki keragaman genetik tinggi dan dikelompokkan ke dalam beberapa genus, termasuk Beta dan Gamma. Meningkatnya ketersediaan data genom membuka peluang penerapan metode machine learning untuk melakukan klasifikasi genom secara otomatis berdasarkan karakteristik sekuens DNA. Penelitian ini bertujuan mengklasifikasikan genom HPV ke dalam genus Beta dan Gamma menggunakan algoritma machine learning dan fitur berbasis 3-mer. Sebanyak 259 sekuens genom HPV lengkap diperoleh dari basis data National Center for Biotechnology Information (NCBI) dan divalidasi menggunakan Basic Local Alignment Search Tool (BLAST). Karakteristik genom direpresentasikan melalui frekuensi kemunculan 3-mer sehingga menghasilkan 64 fitur numerik pada setiap sekuens. Untuk mengatasi ketidakseimbangan kelas, diterapkan metode Synthetic Minority Oversampling Technique (SMOTE) pada data pelatih. Tiga algoritma machine learning, yaitu Random Forest, Extra Trees, dan CatBoost, dibangun dan dievaluasi menggunakan stratified 5-fold cross-validation. Hasil penelitian menunjukkan bahwa seluruh model mampu menghasilkan performa klasifikasi yang tinggi, dengan Extra Trees memperoleh nilai rata-rata akurasi dan F1-score terbaik dibandingkan model lainnya. Hasil tersebut menunjukkan bahwa fitur genomik berbasis 3-mer mampu merepresentasikan pola sekuens DNA yang membedakan genom HPV genus Beta dan Gamma secara efektif. Penelitian ini menunjukkan bahwa machine learning berpotensi menjadi pendekatan yang cepat dan terotomatisasi dalam klasifikasi genom HPV serta dapat mendukung pengembangan analisis data genom pada bidang bioinformatika
Kata Kunci: Bioinformatika, Human Papillomavirus, Klasifikasi Genom, Machine Learning, 3-Mer.

ABSTRACT

Human Papillomavirus (HPV) is a DNA virus with high genetic diversity and is classified into several genera, including Beta and Gamma. The increasing availability of genomic data has created opportunities to apply machine learning techniques for automated genome classification based on DNA sequence characteristics. This study aimed to classify HPV genomes into Beta and Gamma genera using machine learning algorithms and 3-mer sequence features. A total of 259 complete HPV genome sequences were obtained from the National Center for Biotechnology Information (NCBI) database and validated using the Basic Local Alignment Search Tool (BLAST). Genomic features were extracted using 3-mer frequency representation, resulting in 64 numerical features for each sequence. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. Three machine learning algorithms, namely Random Forest, Extra Trees, and CatBoost, were developed and evaluated using stratified 5-fold cross-validation. The results demonstrated that all models achieved high classification performance, with Extra Trees obtaining the highest average accuracy and F1-score. The findings indicate that 3-mer-based genomic features effectively capture sequence patterns that distinguish Beta and Gamma HPV genomes. This study highlights the potential of machine learning as a rapid and automated approach for HPV genome classification and demonstrates its applicability in bioinformatics-based genomic data analysis.
***Keywords**: Bioinformatics, Genome Classification, Human Papillomavirus, Machine Learning, 3-Mer.*

PENDAHULUAN

Human Papillomavirus (HPV) merupakan kelompok virus DNA yang menginfeksi jaringan epitel manusia dan terdiri atas berbagai genus dengan karakteristik biologis yang berbeda. Di antara genus yang telah diidentifikasi, kelompok Betapapillomavirus dan Gammapapillomavirus merupakan genus yang banyak ditemukan pada jaringan kulit manusia. Beberapa penelitian menunjukkan bahwa HPV genus Beta memiliki keterkaitan dengan perkembangan lesi kulit dan kanker kulit non-melanoma pada kondisi tertentu, sedangkan HPV genus Gamma umumnya ditemukan sebagai flora virus kulit dengan tingkat patogenisitas yang lebih rendah. Perbedaan karakteristik biologis tersebut menyebabkan identifikasi dan klasifikasi genom HPV menjadi penting untuk mendukung penelitian virologi, epidemiologi molekuler, serta pengembangan sistem deteksi berbasis komputasi. Perkembangan teknologi Next Generation Sequencing (NGS) menghasilkan peningkatan jumlah data genom virus yang tersimpan pada berbagai basis data biologis, termasuk NCBI Virus. Kondisi tersebut mendorong kebutuhan terhadap metode analisis yang mampu mengolah data genom dalam jumlah besar secara cepat dan akurat. Salah satu pendekatan yang banyak digunakan adalah machine learning, yaitu metode kecerdasan buatan yang mampu mengenali pola biologis dari data genom melalui proses pembelajaran otomatis (Libbrecht & Noble, 2015). Dalam bidang bioinformatika, machine learning telah dimanfaatkan untuk klasifikasi genom, identifikasi gen, prediksi fungsi protein, hingga analisis filogenetik (Greener et al., 2022).

Representasi genom menggunakan fitur k-mer merupakan salah satu teknik yang umum digunakan dalam klasifikasi sekuens DNA. Metode ini mengubah urutan nukleotida menjadi representasi numerik berdasarkan frekuensi kemunculan kombinasi nukleotida tertentu sehingga dapat diproses oleh algoritma machine learning. Beberapa penelitian menunjukkan bahwa fitur k-mer mampu menghasilkan performa klasifikasi yang baik pada berbagai kasus analisis genomik karena dapat merepresentasikan karakteristik biologis suatu organisme tanpa memerlukan proses penyelarasan (alignment) yang kompleks (Zielezinski et al., 2017). Berbagai penelitian terdahulu telah menerapkan algoritma machine learning untuk analisis genom virus. Namun, sebagian besar penelitian berfokus pada identifikasi tipe HPV tertentu atau klasifikasi berbasis gen spesifik, sedangkan penelitian yang membandingkan performa beberapa algoritma ensemble learning pada klasifikasi genom HPV genus Beta dan Gamma masih relatif terbatas. Selain itu, sebagian penelitian menggunakan dataset dengan ukuran besar dan fitur yang kompleks sehingga sulit direplikasi pada lingkungan komputasi sederhana.

Berdasarkan kondisi tersebut, penelitian ini bertujuan membangun model klasifikasi genom HPV genus Beta dan Gamma menggunakan fitur 3-mer yang diekstraksi dari sekuens DNA genom lengkap. Tiga algoritma machine learning, yaitu Random Forest, Extra Trees, dan CatBoost, dibandingkan untuk menentukan model yang memiliki performa terbaik. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan metode klasifikasi genom virus berbasis machine learning yang sederhana, efisien, dan mudah direplikasi pada penelitian bioinformatika selanjutnya.

KAJIAN TEORITIS

Human Papillomavirus (HPV) merupakan virus DNA untai ganda yang menginfeksi jaringan epitel manusia dan termasuk dalam famili Papillomaviridae. Berdasarkan karakteristik genomiknya, HPV diklasifikasikan ke dalam beberapa genus, di antaranya Alpha, Beta, dan Gamma. Genus Beta diketahui berasosiasi dengan berbagai lesi kulit dan berpotensi berperan dalam perkembangan kanker kulit non-melanoma pada kondisi tertentu, sedangkan genus Gamma umumnya ditemukan sebagai bagian dari mikrobioma

kulit manusia dengan tingkat patogenisitas yang relatif lebih rendah (Bzhalava et al., 2015). Perbedaan karakteristik biologis antar genus tersebut menjadikan klasifikasi genom HPV penting untuk mendukung penelitian epidemiologi, diagnosis molekuler, serta pemahaman hubungan evolusi virus. Seiring berkembangnya teknologi sekuensing modern, jumlah data genom HPV yang tersedia pada basis data biologis terus meningkat sehingga diperlukan metode komputasi yang mampu menganalisis data genom secara cepat dan akurat.

Salah satu pendekatan yang banyak digunakan dalam bioinformatika adalah representasi sekuens DNA menggunakan metode k-mer. Metode ini mengubah urutan nukleotida menjadi fitur numerik berdasarkan frekuensi kemunculan kombinasi nukleotida sepanjang k karakter pada suatu genom. Pada penelitian ini digunakan fitur 3-mer sehingga setiap genom direpresentasikan ke dalam 64 kombinasi nukleotida yang mungkin terbentuk dari basa Adenin (A), Timin (T), Guanin (G), dan Sitosin (C). Pendekatan tersebut dipilih karena mampu mempertahankan informasi komposisi genom sekaligus menghasilkan jumlah fitur yang masih efisien untuk dianalisis. Selain itu, metode k-mer termasuk pendekatan alignment-free yang tidak memerlukan proses penyelarasan sekuens secara langsung sehingga lebih cepat diterapkan pada data genom dalam jumlah besar.

Perkembangan machine learning turut mendorong kemajuan analisis genomik karena mampu mengenali pola kompleks pada data biologis secara otomatis. Dalam bidang bioinformatika, machine learning telah banyak dimanfaatkan untuk klasifikasi genom, identifikasi gen, prediksi fungsi protein, hingga analisis penyakit berbasis data molekuler. Pada penelitian ini digunakan tiga algoritma klasifikasi, yaitu Random Forest, Extra Trees, dan CatBoost, untuk mengelompokkan genom HPV ke dalam genus Beta dan Gamma berdasarkan karakteristik sekuens DNA yang direpresentasikan melalui fitur 3-mer. Kinerja model dievaluasi menggunakan metrik accuracy, precision, recall, dan F1-score, serta divalidasi menggunakan metode Stratified 5-Fold Cross Validation guna memperoleh hasil yang lebih stabil dan mengurangi potensi overfitting. Dengan pendekatan tersebut, klasifikasi genom HPV diharapkan dapat dilakukan secara akurat serta memberikan gambaran mengenai kemampuan algoritma machine learning dalam mendukung analisis bioinformatika berbasis data genom.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental untuk mengklasifikasikan genom Human Papillomavirus (HPV) berdasarkan karakteristik sekuens DNA. Data penelitian diperoleh dari basis data NCBI Virus berupa 259 genom HPV dalam format FASTA dan metadata pendukung dalam format TSV. Sebelum dilakukan analisis, identitas sekuens diverifikasi menggunakan Basic Local Alignment Search Tool (BLAST) untuk memastikan kesesuaian genom dengan data referensi. Selanjutnya, sekuens DNA direpresentasikan menjadi fitur numerik menggunakan metode 3-mer frequency sehingga pola komposisi nukleotida pada setiap genom dapat dianalisis menggunakan pendekatan machine learning.

Proses klasifikasi dilakukan menggunakan algoritma Random Forest, Extra Trees Classifier, dan CatBoost Classifier. Data dibagi menjadi data latih dan data uji dengan rasio 80:20, kemudian ketidakseimbangan kelas ditangani menggunakan Synthetic Minority Oversampling Technique (SMOTE). Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score, serta divalidasi menggunakan metode Stratified 5-Fold Cross Validation untuk memperoleh hasil yang lebih stabil dan objektif. Seluruh proses analisis dilakukan menggunakan bahasa pemrograman Python pada lingkungan Kaggle Notebook.

HASIL DAN PEMBAHASAN

Data genom HPV diperoleh dari basis data NCBI Virus dalam format FASTA dan metadata TSV. Hasil proses prapengolahan menghasilkan 208 genom yang memiliki informasi genus lengkap, terdiri atas 194 henom Gamma dan 14 genus Beta. Distribusi tersebut menunjukkan adanya ketidakseimbangan kelas karena jumlah genom Gamma jauh lebih banyak dibandingkan genom Beta. Oleh karena itu, pada tahap pelatihan model diterapkan metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengurangi potensi bias terhadap kelas mayoritas. Kemudian, setiap genom direpresentasikan menggunakan metode 3-mer *frequency*.

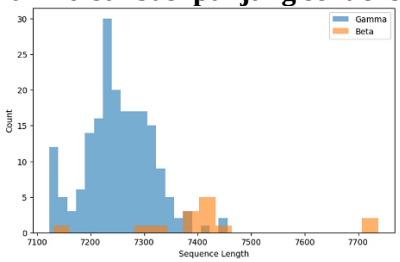
$$f_i = \frac{n_i}{L - k + 1}$$

f_i berarti frekuensi kemunculan 3-mer ke-i, n_i menunjukkan jumlah kemunculan 3-mer ke-i dalam genom, L berarti panjang sekuens DNA, dan k menunjukkan ukuran k-mer $(k = 3)$. Metode ini menghitung frekuensi kemunculan seluruh kombinasi dari tiga nukleotida yang mungkin terbentuk dari basa Adenine (A), Thymine (T), Guanine (G), dan Cytosine (C) sehingga menghasilkan 64 fitur numerik. Representasi tersebut digunakan sebagai masukan bagi algoritma *machine learning* untuk mengidentifikasi pola karakteristik yang membedakan genus Beta dan Gamma.

Evaluasi Awal Model

Proses klasifikasi dilakukan menggunakan tiga algoritma, yaitu *Random Forest*, *Extra Trees Classifier*, dan *CatBoost Classifier*. Pengujian awal menggunakan data uji menunjukkan bahwa ketiga model menghasilkan performa yang hampir sama dengan tingkat akurasi sekitar 97.6%. Hasil tersebut menunjukkan bahwa fitur berbasis 3-mer *frequency* mampu merepresentasikan karakteristik genom HPV dengan baik sehingga ketiga model dapat melakukan klasifikasi secara efektif.

Gambar 1 Grafik distribusi panjang sekuens genom HPV



Sumber: Hasil Analisis Panjang Sekuens

Visualisasi histogram di atas menunjukkan distribusi panjang genom HPV berdasarkan genus Beta dan Gamma. Sebagian besar genom Gamma memiliki panjang sekitar 7.200-7.350 pasangan basa (*bp/base pair*). Sementara itu, genom Beta memiliki rentang panjang genom sekitar 7.380-7.450 bp, serta beberapa genom memiliki panjang hingga sekitar 7.700 bp. Distribusi data genus Gamma terlihat jauh lebih banyak dibandingkan genus Beta, yang mengindikasikan adanya ketidakseimbangan kelas (*class imbalance*). Selain itu, rentang panjang genom kedua genus relatif mirip sehingga panjang genom saja belum cukup untuk membedakan genus HPV secara akurat. Oleh karena itu, diperlukan ekstraksi fitur yang lebih representatif dari pola penyusun sekuens DNA.

1. Ekstraksi fitur genomik menggunakan metode 3-mer

Algoritma *machine learning* diketahui tidak dapat memproses data DNA secara langsung, akibatnya setiap sekuens perlu diubah menjadi numerik. Metode yang digunakan adalah k-mer *frequency* dengan nilai k=3. Pada metode ini, setiap kombinasi tiga nukleotida dihitung frkuensi kemunculannya dalam genom. Dengan adanya empat jenis nukleotida (A, T, G, dan C), jumlah kombinasi 3-mer yang mungkin terbentuk adalah $4^3 =$

64 kombinasi sehingga diperoleh 64 fitur yang merepresentasikan karakteristik genom HPV. Setiap genom dihitung frekuensi kemunculan seluruh kemungkinan kombinasi tiga nukleotida yang berjumlah 64 pola. Setelah seluruh kombinasi 3-mer dibentuk, setiap genom diubah menjadi vektor numerik berdasarkan frekuensi kemunculan masing-masing pola 3-mer. Nilai frekuensi diperoleh dengan membagi jumlah kemunculan suatu 3-mer dengan total seluruh 3-mer dalam genom, sehingga perbedaan panjang genom tidak memengaruhi hasil perhitungan. Hasil transformasi menghasilkan matriks fitur berukuran 259 × 64, yang berarti setiap genom HPV direpresentasikan oleh 64 fitur numerik. Lalu, seluruh data tanpa label genotype telah berhasil dihapus sehingga dataset akhir terdiri atas 208 genom HPV yang terbagi ke dalam 194 genom Gamma dan 14 genom Beta. Matriks inilah yang selanjutnya digunakan sebagai input pada tahap pembangunan model *machine learning.* Label kelas yang masih berbentuk teks (*Beta Gamma*) perlu diubah menjadi format numerik menggunakan *Label Encoding.* Proses berikut ini mengonversi genus *Beta* menjadi 0 dan genus *Gamma* menjadi 1 sehingga dapat diproses oleh algoritma *machine learning.* Hasil yang didapatkan adalah 14 genom *Beta* dan 194 genom *Gamma,* yang berarti adanya ketidakseimbangan jumlah sampel antar kelas.

2. Pembagian data latih dan data uji

Dataset dibagi menjadi data latih (*training set*) dan data uji (*testing set*) menggunakan rasio 80:20. Data latih digunakan untuk membangun model *machine learning*, sedangkan data uji digunakan untuk mengevaluasi kemampuan model pada data yang belum pernah dilihat sebelumnya. Pembagian dilakukan menggunakan parameter 'stratify=y*'* agar proporsi kelas *Beta* dan *Gamma* tetap terjaga pada kedua kelompok data. Hasil pembagian menghasilkan 166 data latih dan 42 data uji dengan distribusi kelas yang tetap merepresentasikan kondisi dataset asli, sehingga proses pelatihan dan evaluasi model dapat dilakukan secara lebih konsisten dan tidak bias akibat perubahan proporsi kelas.

3. Penanganan ketidakseimbangan data menggunakan SMOTE

Berdasarkan hasil pembagian data, jumlah genom Gamma pada data latih jauh lebih banyak dibandingkan genom *Beta,* yaitu 155 berbanding 11. Kondisi ini dapat menyebabkan model *machine learning* lebih cenderung memprediksi kelas mayoritas (Gamma) dan mengabaikan karakteristik kelas minoritas (Beta). Untuk mengatasi permasalahan tersebut digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE bekerja dengan membuat data sintetis baru pada kelas minoritas berdasarkan kemiripan karakteristik sampel yang sudah ada. Pada penelitian ini digunakan parameter *k_neighbors =3,* yang berarti setiap sampel *Beta* akan dibandingkan dengan tiga tetangga terdekatnya untuk menghasilkan data sintetis baru. Setelah prooses SMOTE dilakukan, jumlah data pada kedua kelas menjadi seimbang, yaitu masing-masing 155 sampel.

Tabel 1 Frekuensi SMOTE genus beta dan gamma

Encode	**Sebelum**	**Sesudah**
0 (Beta)	11	155
1 (Gamma)	155	155

Sumber: Hasil SMOTE.

4. Model *machine learning* dan evaluasi

Model yang digunakan terdiri atas *random forest, extra trees classifier,* dan *catboost classifier*. Ketiga model dipilih karena mampu menangani data berdimensi tinggi, memiliki performa yang baik pada tugas klasifikasi biologis, serta mampu mengurangi risiko *overfitting* melalui pendekatan *ensemble.* Pada penelitian ini digunakan 200 pohon keputusan (*n_estimators = 200*) untuk meningkatkan stabilitas prediksi. Setiap model yang

dibangun akan dievaluasi berdasarkan metrik pengukuran terkait performanya dalam hasil klasifikasi menggunakan data uji sehingga mengasilkan nilai performa sebagai berikut.

Tabel 2 Evaluasi model

Model	Accuracy	Precision	Recall	F-1 Score
Random Forest	0.97619	0.982143	0.97619	0.977737
Extra Trees	0.97619	0.982143	0.97619	0.977737
CatBoost	0.97619	0.982143	0.97619	0.977737

Sumber: Hasil evaluasi model.

Hasil pengujian menunjukkan bahwa semua model menghasilkan performa yang sama di masing-masing metrik evaluasi. Selain itu, model mampu mengidentifikasi seluruh sampel Beta pada data uji dengan nilai *recall* sebesar 100%, sementara sebagian besar sampel Gamma juga berhasil diklasifikasikan dengan benar. Berdasarkan hasil tersebut, ketiga model memiliki klasifikasi yang sangat baik. Namun, evaluasi ini hanya menggunakan satu kali pembagian data dengan jumlah data uji yang relatif kecil, yaitu 42 genom dan tidak dapat menjadi penentu model terbaik. Maka dari itu, akan dilakukan evaluasi menggunakan *stratified 5-fold cross validation* sehingga performa model dapat diuji pada beberapa kombinasi pembagian data yang berbeda.

Hasil *Stratified 5-Fold Cross Validation*

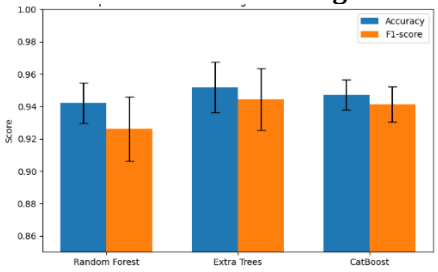
Untuk memperoleh evaluasi yang lebih stabil dan representatif, digunakan metode *stratified 5-fold ross validation. Cross-validation* merupakan teknik validasi yang banyak digunakan untuk mengukur kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya dan menghasilkan estimasi performa yang lebih reliabel dibandingkan pengujian dengan satu kali pembagian data (Berrar, 2019). Pada metode ini, dataset dibagi menjadi lima bagian dengan proporsi kelas *Beta* dan *Gamma* yang tetap terjaga pada setiap *fold.* Secara bergantian, empat bagian digunakan sebagai data latih dan satu bagian digunakan sebagai data validasi hingga seluruh data memperoleh kesempatan menjadi data validasi. Pendekatan ini memungkinkan model dievaluasi pada beberapa kombinasi data yang berbeda sehingga hasil yang diperoleh lebih jelas dalam membedakan performa antar model.

Tabel 3 Hasil Evaluasi Tiga Model Klasifikasi

Model	Accuracy Mean	Accuracy Std	F1 Mean	F1 Std
Random Forest	0.942160	0.012518	0.925989	0.019907
Extra Trees	0.951800	0.015617	0.944432	0.019130
CatBoost	0.947154	0.009306	0.941240	0.010964

Sumber: Hasil evaluasi model.

Gambar 2 Grafik Perbandingan Model



Sumber: Hasil evaluasi model.

Hasil validasi silang menunjukkan adanya perbedaan performa antar model. *Extra Trees Classifier* menghasilkan rata-rata akurasi tertinggi sebesar 95.18% dan rata-rata *F1-Score* sebesar 94.44%. sementara itu, *CatBoost Classifier* memperoleh rata-rata akurasi sebesar 94.71% dan *F1-Score* sebesar 94.12%. Adapun *Random Forest* menghasilkan performa yang sedikit lebih rendah dibandingkan kedua model tersebut. *Stratified 5-fold cross validation* melakukan evaluasi pada beberapa kombinasi pembagian data yang

berbeda sehingga mampu memberikan gambaran performa model yang lebih stabil dan objektif. Oleh karena itu, perbedaan kemampuan masing-masing model baru terlihat setelah dilakukan validasi silang, di mana *extra trees* menunjukkan performa rata-rata yang lebih tinggi dibandingkan *random forest* dan *catboost.*

Gambar 3 Grafik Perbandingan Model

Sumber: Hasil evaluasi model.

Visualisasi *boxplot* di atas digunakan untuk melihat sebaran nilai *F1-Score* yang diperoleh setiap model pada lima *fold* validasi silang. Setiap kotak menunjukkan distribusi nilai *F1-Score* yang dihasilkan oleh model selama proses *cross validation*, sedangkan garis di tengah kotak menunjukkan nilai median. Berdasarkan grafik, ketiga model menghasilkan performa yang relatif tinggi dan stabil dengan nilai *F1-Score* berada pada kisaran 0,94–0,98. Median ketiga model terlihat hampir sama, menunjukkan bahwa seluruh model mampu melakukan klasifikasi genom HPV dengan baik. Namun, hasil evaluasi rata-rata sebelumnya menunjukkan bahwa *extra trees classifier* memperoleh nilai *F1-Score* rata-rata tertinggi sebesar 94,44%, diikuti *catboost* sebesar 94,12% dan *random forest* sebesar 92,60%. Dengan demikian, meskipun perbedaan performa antar model tidak terlalu besar, *extra trees* tetap dipilih sebagai model terbaik karena memberikan performa klasifikasi paling tinggi pada evaluasi *stratified 5-fold cross validation.*

KESIMPULAN

Penelitian ini berhasil mengimplementasikan pendekatan machine learning untuk mengklasifikasikan genom Human Papillomavirus (HPV) berdasarkan karakteristik sekuens DNA menggunakan fitur 3-mer frequency. Hasil penelitian menunjukkan bahwa representasi genom menggnakan 64 fitur 3-mer mampu membedakan genus Beta dan Gamma dengan baik. Berdasarkan evaluasi menggunakan Stratified 5-Fold Cross Validation, model Extra Trees Classifier menghasilkan performa terbaik dengan nilai accuracy sebesar 95,18% dan F1-Score sebesar 94,44%. Selain itu, hasil confusion matrix menunjukkan bahwa model mampu mengklasifikasikan 41 dari 42 data uji dengan benar. Temuan ini menunjukkan bahwa pola frekuensi 3-mer dapat dimanfaatkan sebagai representasi genom yang efektif untuk mendukung proses klasifikasi HPV berbasis data genom.

Penelitian ini masih memiliki keterbatasan pada jumlah sampel yang relatif sedikit dan distribusi kelas yang tidak seimbang antara genus Beta dan Gamma. Oleh karena itu, penelitian selanjutnya disarankan menggunakan jumlah genom yang lebih besar, melibatkan lebih banyak genus HPV, serta membandingkan berbagai metode ekstraksi fitur dan algoritma klasifikasi lainnya. Pengembangan model berbasis deep learning dan pemanfaatan data genom yang lebih beragam juga berpotensi meningkatkan performa klasifikasi serta memperluas penerapan bioinformatika dalam identifikasi dan karakterisasi virus.

DAFTAR PUSTAKA

Asensio-Puig, L., Alemany, L., & Pavón, M. A. (2022). A straightforward HPV16 lineage classification based on machine learning. Frontiers in Artificial Intelligence, 5, 851841. https://doi.org/10.3389/frai.2022.851841

Berrar, D. (2019). Cross-validation. In Encyclopedia of Bioinformatics and Computational Biology (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bzhalava, D., Eklund, C., & Dillner, J. (2015). International standardization and classification of human papillomavirus types. Virology, 476, 341–344. https://doi.org/10.1016/j.virol.2014.12.028

Chen, Z., Utro, F., Platt, D., DeSalle, R., Parida, L., Chan, P. K. S., & Burk, R. D. (2021). K-mer analyses reveal different evolutionary histories of Alpha, Beta, and Gamma papillomaviruses. International Journal of Molecular Sciences, 22(17), 9657. https://doi.org/10.3390/ijms22179657

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology, 23(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920

Peng, L., Yuan, R., & Shen, L. (2021). LPI-EnEDT: An ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification. Genomics, Proteomics & Bioinformatics, 14, 50. https://doi.org/10.1186/s13040-021-00277-4

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biology, 18(1), 186. https://doi.org/10.1186/s13059-017-1319-7