JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 2, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i2.9361

Analisis Data RNA-Seq Homo sapiens Menggunakan Machine Learning untuk Klasifikasi Ekspresi Gen Berbasis Data NCBI SRA

Siti Haliza Zamili[1*], Adinda Soleha[2], Alya Namira[3], Khodotun Hadawiyah Margolang[4]

[1,2,3,4]Prodi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, Jalan Willem Iskandar, Pasar V, Medan Estate, Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara, 20371.

Keywords:
RNA-Seq, Machine Learning, Random Forest, support Vector Machine, Klasifikasi Ekspresi Gen.

Corespondent Email:
sitihalizazamili08@gmail.com

Abstrak. Penelitian ini bertujuan untuk menganalisis data RNA-Seq Homo sapiens serta mengklasifikasikan ekspresi gen menggunakan metode machine learning. Data diperoleh dari Sequence Read Archive dalam format FASTQ yang terdiri dari lima sampel paired-end. Tahap awal dilakukan quality control menggunakan FastQC dan dirangkum dengan MultiQC. Selanjutnya dilakukan preprocessing melalui proses trimming untuk meningkatkan kualitas data. Data yang telah diproses kemudian diubah menjadi dataset numerik dan diklasifikasikan menggunakan Random Forest dan Support Vector Machine (SVM). Hasil menunjukkan bahwa SVM memiliki akurasi lebih tinggi sebesar 0,60 dibandingkan Random Forest sebesar 0,50, namun Random Forest menunjukkan performa yang lebih seimbang dalam mengklasifikasikan kedua kelas. Temuan ini menunjukkan bahwa pemilihan model tidak hanya bergantung pada akurasi, tetapi juga pada kemampuan dalam menangani data yang tidak seimbang.

Abstract. *This study aims to analyze Homo sapiens RNA-Seq data and classify gene expression using machine learning methods. Data were obtained from the Sequence Read Archive in FASTQ format consisting of five paired-end samples. The initial stage was quality control using FastQC and summarized using MultiQC. Next, preprocessing was carried out through a trimming process to improve data quality. The processed data were then converted into a numerical dataset and classified using Random Forest and Support Vector Machine (SVM). The results showed that SVM had a higher accuracy of 0.60 compared to Random Forest at 0.50, but Random Forest showed a more balanced performance in classifying both classes. This finding suggests that model selection depends not only on accuracy but also on the ability to handle imbalanced data.*

1. PENDAHULUAN

Perkembangan teknologi dalam bidang bioinformatika telah menghasilkan peningkatan signifikan pada data genomik, khususnya melalui teknologi RNA-Sequencing (RNA-Seq). Teknologi ini memungkinkan analisis ekspresi gen secara menyeluruh dalam suatu organisme, sehingga menghasilkan data dalam jumlah besar dengan kompleksitas tinggi. Data RNA-Seq umumnya disimpan dalam basis data publik seperti National Center for Biotechnology Information dan Sequence Read Archive, yang menyediakan akses terbuka terhadap data sequencing dari berbagai penelitian di seluruh dunia.

Meskipun ketersediaan data yang melimpah memberikan peluang besar dalam penelitian genomik, hal ini juga menimbulkan tantangan dalam pengolahan dan analisis data. Data RNA-Seq memiliki ukuran besar serta kualitas yang bervariasi, sehingga memerlukan proses preprocessing seperti *quality control* (QC) dan trimming sebelum dapat dianalisis lebih lanjut. Tanpa proses ini, data yang digunakan dapat mengandung noise yang mempengaruhi hasil analisis [1].

Dalam konteks Big Data, metode konvensional seringkali tidak cukup untuk menangani kompleksitas data RNA-Seq. Oleh karena itu, diperlukan pendekatan yang mampu mengolah dan menganalisis data secara efisien, salah

satunya adalah Machine Learning. Metode ini mampu mengenali pola dalam data dan melakukan klasifikasi secara otomatis, termasuk dalam analisis ekspresi gen [2].

Beberapa penelitian sebelumnya telah menunjukkan bahwa algoritma Machine Learning seperti Random Forest mampu meningkatkan akurasi dalam berbagai kasus klasifikasi [3], [4]. Namun, masih terdapat kesenjangan dalam integrasi proses pengolahan data RNA-Seq secara menyeluruh, mulai dari pengambilan data, preprocessing, hingga klasifikasi menggunakan algoritma tertentu. Selain itu, perbandingan performa antar algoritma Machine Learning seperti Random Forest dan Support Vector Machine (SVM) dalam konteks klasifikasi ekspresi gen masih perlu dikaji lebih lanjut.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengolah data RNA-Seq berbasis Big Data, melakukan preprocessing untuk meningkatkan kualitas data, serta menerapkan metode Machine Learning dalam klasifikasi ekspresi gen. Selain itu, penelitian ini juga bertujuan untuk membandingkan performa algoritma Random Forest dan Support Vector Machine (SVM) dalam menghasilkan klasifikasi yang optimal.

2. TINJAUAN PUSTAKA

2.1 Machine Learning Dan Klasifikasi

Teknologi machine learning telah menciptakan kesempatan baru dalam pengembangan data RNA-seq. Model-model machine learning memberikan cara yang lebih lincah dan dapat disesuaikan dibandingkan dengan metode statistik konvensional [10].

Pembelajaran Mesin atau Machine Learning adalah metode yang berasal dari Kecerdasan Buatan (AI) yang digunakan untuk meniru dan mengambil alih fungsi manusia dalam menyelesaikan berbagai permasalahan. Dengan kata lain, Pembelajaran Mesin adalah perangkat yang dirancang untuk belajar dan menyelesaikan tugas tanpa perlu instruksi dari penggunanya. Menurut Arthur Samuel, seorang perintis Amerika di bidang permainan komputer dan kecerdasan buatan, AI menyatakan bahwa pembelajaran mesin adalah cabang ilmu yang mempelajari bagaimana memberi komputer kemampuan untuk belajar tanpa diprogram secara eksplisit [5]. Machine Learning memiliki peran penting dalam penelitian ini karena mampu mengolah data genom yang berukuran besar dan kompleks. Dengan memanfaatkan algoritma Machine Learning, data RNA yang telah diubah menjadi bentuk numerik dapat dianalisis untuk menemukan pola tertentu. Pola tersebut kemudian digunakan untuk melakukan klasifikasi variasi genetik secara otomatis, sehingga proses analisis menjadi lebih cepat, akurat, dan efisien dibandingkan metode manual.

Model regresi linier dan pohon keputusan memberi kesempatan kepada komputer untuk memahami dan menemukan pola dari informasi yang ada serta membuat ramalan yang tepat. Dalam hal memperkirakan cuaca dan iklim, pembelajaran mesin dapat berperan dalam mengenali pola yang sering muncul dan menciptakan ramalan yang lebih tepat dengan menggunakan data masa lalu.[11]

Salah satu penggunaan machine learning adalah dalam klasifikasi. Klasifikasi adalah metode dalam Machine Learning yang berfungsi untuk membagi data ke dalam kategori tertentu berdasarkan pola yang ada dalam data tersebut. Proses klasifikasi melibatkan pembelajaran dari data yang sudah berlabel, sehingga model bisa digunakan untuk memperkirakan kelas data yang baru. Tujuan utama dari klasifikasi adalah untuk mengotomatisasi pengelompokan data, meningkatkan efisiensi analisis, dan membantu dalam pengambilan keputusan. Dalam studi ini, klasifikasi dimanfaatkan untuk mengorganisir variasi genetik berdasarkan pola yang terdapat dalam data RNA.

2.2 Algoritma Random Forest Dan SVM

Random Forest adalah salah satu algoritma machine learning yang sangat terkenal. Secara umum, Random Forest terdiri dari sekumpulan decision tree yang digabungkan untuk menciptakan model yang lebih tepat. Inilah mengapa istilah 'hutan' digunakan, karena terdiri dari beberapa decision tree. Algoritma ini menciptakan beberapa pohon berdasarkan data sampel, di mana setiap pohon yang dibangun selama proses pelatihan tidak terkait dengan pohon sebelumnya, dan keputusan diambil berdasarkan suara terbanyak. Dua konsep utama dalam Random Forest adalah penggabungan pohon melalui teknik bagging (bootstrap aggregating) dengan penggantian, serta pemilihan fitur secara acak untuk masing-

masing pohon yang dibentuk. Random Forest memiliki dua parameter utama, yaitu parameter m yang menunjukkan proporsi jumlah pohon yang akan digunakan, dan parameter k yang menggambarkan jumlah maksimum fitur yang dipertimbangkan saat proses percabangan dalam pohon [6]. Dalam studi ini, penulis menggunakan algoritma Random Forest untuk mendeteksi dan mengklasifikasikan data RNA-seq. Algoritma ini dipilih karena kemampuannya dalam menangani data dalam jumlah besar, waktu pelatihan yang cepat, hasil prediksi yang tepat, serta kemampuannya mengurangi risiko overfitting, sehingga dianggap sebagai algoritma yang paling sesuai untuk digunakan dalam penelitian ini. [12]

Random Forest memiliki sejumlah keuntungan, seperti meningkatkan ketepatan saat menghadapi data yang tidak lengkap, mengurangi kemungkinan terjadinya kesalahan, menyimpan data dengan cara yang efisien, serta mampu mengelola hasil keluaran dan penyimpanan data secara efektif..[13]

Support Vector Machine (SVM) adalah salah satu metode dalam pembelajaran yang dikenal memiliki performa yang lebih baik dibandingkan algoritma lainnya, sehingga sering diterapkan dalam berbagai penelitian. [7].

SVM, yaitu model Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang bersifat diawasi, yang membutuhkan data pelatihan yang telah dilengkapi dengan label atau kategori. SVM memanfaatkan kernel dalam ruang pencarian yang memiliki dimensi tinggi. Dua kategori kelas dipisahkan berdasarkan pola tertentu sehingga hyperplane yang dipengaruhi oleh margin di antara kelas tersebut dapat terbentuk secara optimal [8].

Salah satu keuntungan dari pendekatan ini adalah kemampuannya untuk melakukan klasifikasi dan menangani regresi baik dalam bentuk linear maupun non-linear. SVM menunjukkan tingkat akurasi dalam klasifikasi yang lebih tinggi jika dibandingkan dengan metode klasifikasi lainnya [14]. SVM juga dapat menangani isu overfitting, sebab ia berupaya untuk mengurangi kesalahan dalam klasifikasi sekaligus memperbesar jarak antara berbagai kelas data. Artinya, SVM tidak hanya fokus pada klasifikasi data yang ada dengan akurat, tetapi juga menjamin agar model yang dibentuk tidak terlalu disesuaikan dengan data pelatihan, sehingga meningkatkan kemampuan untuk beradaptasi terhadap data yang baru. [15]. Karena alasan tersebut, SVM sering dipakai dalam berbagai aplikasi pengelompokan, termasuk dalam melakukan analisis data RNA-Seq.

2.3 Variasi Genetik

Keanekaragaman spesies manusia, hewan, dan tumbuhan tidak terpisahkan dari faktor variasi genetik. Dalam konteks manusia, perbedaan genetis dapat dilihat melalui variasi fisik, kecenderungan terhadap penyakit, dan cara tubuh bereaksi terhadap obat-obatan tertentu. Salah satu contoh dari variasi genetik adalah alergi makanan. Reaksi alergi terhadap makanan dipengaruhi oleh varian genetik dan juga faktor lingkungan. Salah satu gen yang terlibat dalam alergi makanan adalah gen HLA (antigen leukosit manusia) yang berperan dalam sistem kekebalan tubuh. Variasi gen ini dapat memengaruhi cara tubuh mengatur dan merespons alergen. [9]. Peran variasi genetik dalam penelitian adalah untuk membantu mengidentifikasi perbedaan genetik antara individu yang berkaitan dengan karakteristik biologis dan kemampuan tubuh dalam menghadapi penyakit. Penelitian variasi genetik juga berguna untuk menjelaskan hubungan antara gen dan reaksi terhadap pengobatan, serta mendukung kemajuan di bidang genomik, bioinformatika, dan keberagaman hayati.

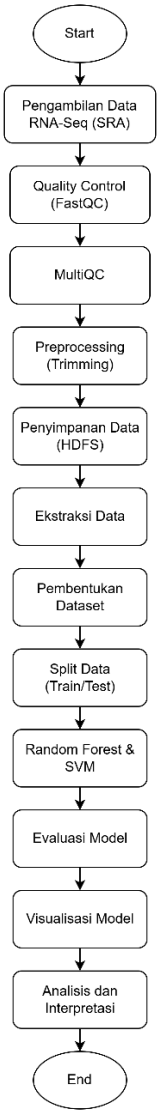
3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksperimen dengan menerapkan teknik *machine learning* untuk mengklasifikasikan data RNA-Seq. Proses penelitian dilakukan secara bertahap mulai dari pengambilan data, *preprocessing*, hingga evaluasi model guna memperoleh hasil klasifikasi yang optimal.

3.1. Alur Penelitian

Alur penelitian meliputi beberapa tahapan utama, yaitu pengambilan data dari Sequence Read Archive (SRA), *quality control*, *preprocessing*, penyimpanan data, pembentukan dataset, implementasi algoritma

machine learning, serta evaluasi dan visualisasi hasil.



Gambar 1. Alur Penelitian Klasifikasi Data RNA-Seq Menggunakan Metode *Machine Learning*

Gambar tersebut menunjukkan tahapan penelitian secara keseluruhan dari proses awal hingga diperoleh hasil klasifikasi.

3.2. Pengambilan dan Pengolahan Data

Data yang digunakan berupa RNA-Seq *Homo sapiens* yang diperoleh dari SRA dalam format FASTQ dengan lima sampel bertipe *paired-end*. Data kemudian melalui tahap *quality control* menggunakan FastQC dan dirangkum dengan MultiQC untuk memastikan kualitas awal.

Selanjutnya dilakukan preprocessing melalui proses *trimming* untuk menghilangkan bagian dengan kualitas rendah sehingga mengurangi *noise* pada data. Data hasil *preprocessing* disimpan menggunakan Hadoop Distributed File System (HDFS) pada direktori yang terstruktur untuk memudahkan pengelolaan.

3.3. Pembentukan Dataset dan Model

Data yang telah diproses kemudian dikonversi menjadi dataset numerik melalui ekstraksi fitur dan dibagi menjadi data latih dan data uji. Proses klasifikasi dilakukan menggunakan algoritma Random Forest dan Support Vector Machine (SVM).

Random Forest membangun beberapa *decision tree* dari subset data dan menentukan hasil klasifikasi melalui mekanisme voting. Sementara itu, SVM bekerja dengan membentuk *hyperplane* optimal untuk memisahkan kelas berdasarkan pola data yang telah dipetakan ke ruang berdimensi lebih tinggi.

3.4. Evaluasi dan Visualisasi

Kinerja model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai performa klasifikasi secara menyeluruh. Hasil evaluasi kemudian divisualisasikan dalam bentuk grafik, confusion matrix, dan feature importance untuk mempermudah interpretasi hasil.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Quality Control Preprocessing

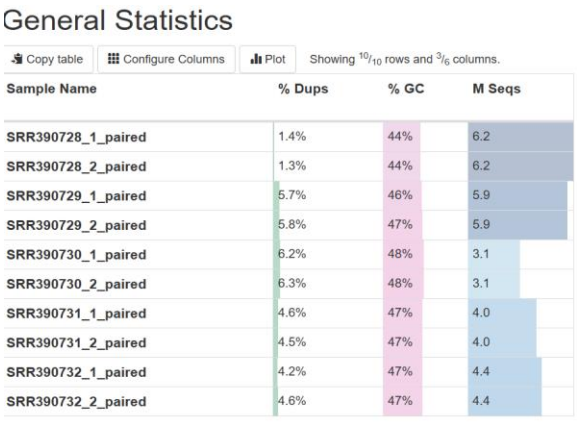
Berdasarkan tahapan metodologi yang telah dilakukan, data RNA-Seq *Homo sapiens* yang diperoleh dari Sequence Read Archive terlebih dahulu dianalisis menggunakan FastQC dan dirangkum menggunakan MultiQC untuk mengevaluasi kualitas data awal.

Hasil *quality control* menunjukkan bahwa sebagian besar sampel memiliki kualitas sekuens yang baik, namun ditemukan penurunan kualitas pada bagian akhir read. Selain itu, parameter seperti GC content, distribusi panjang read, serta tingkat duplikasi masih berada dalam rentang yang dapat diterima, sehingga data dinilai layak untuk diproses lebih lanjut.

Selanjutnya dilakukan tahap preprocessing melalui proses *trimming* untuk menghilangkan bagian read dengan kualitas rendah. Hasil setelah *preprocessing* menunjukkan adanya peningkatan kualitas data yang signifikan, ditandai dengan berkurangnya noise serta meningkatnya kestabilan kualitas basa. Proses ini juga menyebabkan sedikit penurunan panjang read karena bagian dengan kualitas rendah telah dihapus.

Hasil ini menunjukkan bahwa *preprocessing* memiliki peran penting dalam meningkatkan kualitas data sebelum digunakan dalam analisis lanjutan. Tanpa *preprocessing*, keberadaan noise dalam data dapat mengganggu proses pembelajaran model dan menurunkan akurasi klasifikasi.

Gambar 2. Sebelum *Preprocessing*



Gambar 3. Sesudah *Preprocessing*

4.2. Hasil Penyimpanan Data

Data RNA-Seq disimpan menggunakan Apache Hadoop melalui Hadoop Distributed File System (HDFS). Data yang digunakan terdiri dari lima sampel *paired-end* yang menghasilkan sepuluh file FASTQ yang tersimpan pada direktori `/genomic/raw_data`.

Penggunaan HDFS memungkinkan penyimpanan data dalam jumlah besar secara terstruktur dan terdistribusi, sehingga memudahkan proses pengelolaan serta akses data selama penelitian berlangsung. Meskipun tidak secara langsung mempengaruhi hasil klasifikasi, penyimpanan berbasis Hadoop memberikan keunggulan dalam hal skalabilitas dan efisiensi pengolahan data.

Gambar 4. Penyimpanan Data

4.3. Hasil Klasifikasi Machine Learning

Tahap klasifikasi dilakukan menggunakan algoritma Random Forest dan Support Vector Machine (SVM) terhadap dataset yang telah melalui proses preprocessing dan ekstraksi fitur.

Hasil dari pengujian menunjukkan bahwa model SVM memiliki tingkat akurasi mencapai 0.60, lebih unggul dibandingkan Random Forest dengan akurasi sebesar 0.50. secara keseluruhan hasil ini mengindikasikan bahwa SVM lebih efektif dalam melakukan prediksi kelas secara umum.

Namun, analisis tambahan mengungkapkan bahwa SVM cenderung menghasilkan prediksi yang tidak seimbang, di mana model ini lebih sering mengklasifikasikan satu kelas ketimbang kelas lainnya. Hal ini tampak dari nilai *precision* dan *recall* yang tidak konsisten dan juga didukung oleh hasil confusion matrix.

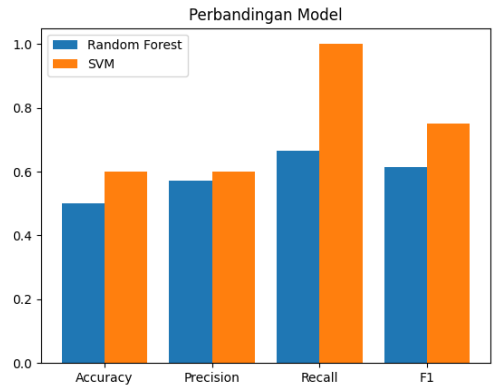
Di sisi lain, Random Forest menunjukkan kinerja yang lebih konsisten dengan distribusi prediksi yang lebih seimbang antara kedua

kelas. Meskipun nilai akurasinya lebih rendah, model ini mampu mengenali kedua kelas dengan lebih baik dibandingkan SVM.

Perbedaan ini menunjukkan bawa akurasi tidak selalu mencerminkan performa model secara keseluruhan, terutama pada dataset dengan distribusi yang tidak seimbang.

```
=== Random Forest ===
              precision    recall  f1-score   support

           0       0.33      0.25      0.29         8
           1       0.57      0.67      0.62        12

    accuracy                           0.50        20
   macro avg       0.45      0.46      0.45        20
weighted avg       0.48      0.50      0.48        20

=== SVM ===
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         8
           1       0.60      1.00      0.75        12

    accuracy                           0.60        20
   macro avg       0.30      0.50      0.38        20
weighted avg       0.36      0.60      0.45        20
```

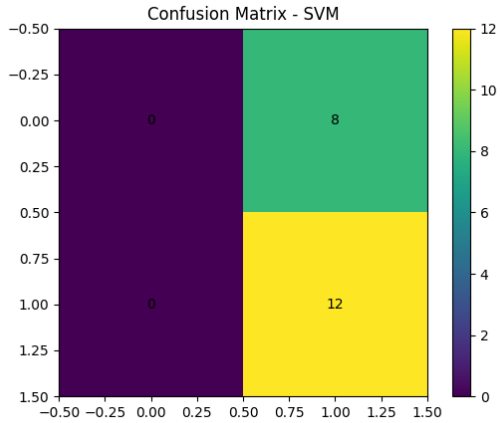
Gambar 5. Hasil *Machine Learning*



Gambar 6. Perbandingan Model Random Forest dan SVM

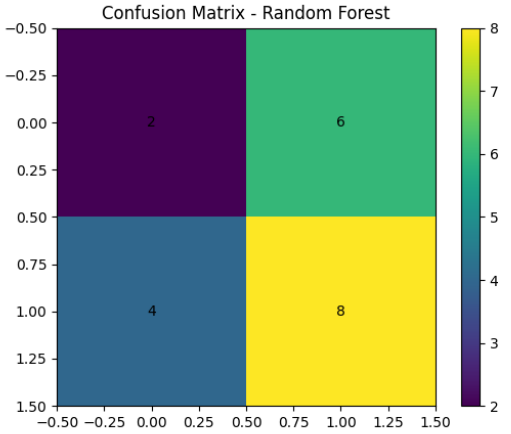
4.4. Visualisasi dan Interpretasi Hasil

Visualisasi digunakan untuk memberikan pemahaman yang lebih mendalam terhadap performa model klasifikasi. Confusion matrix menunjukkan bahwa model SVM memiliki kecenderungan untuk memprediksi data ke dalam satu kelas dominan, sehingga menyebabkan kesalahan klasifikasi pada kelas lainnya.



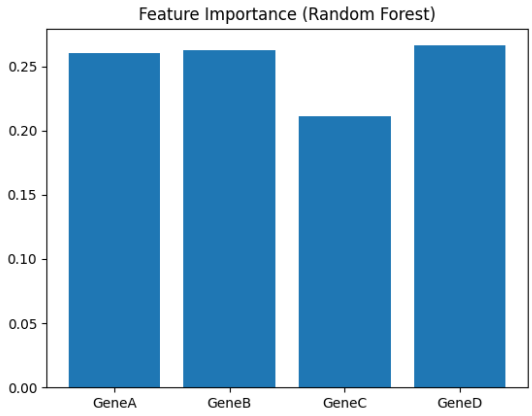
Gambar 7. Confusion Matrix SVM

Sebaliknya, Random Forest menunjukkan distribusi prediksi yang lebih merata, yang mengindikasikan kemampuan model dalam mengenali pola data secara lebih menyeluruh. Ini membuktikan bahwa Random Forest memiliki kemampuan generalisasi yang lebih baik dibandingkan SVM pada dataset yang digunakan.



Gambar 8. Confusion Matrix Random Forest

Selain itu, visualisasi feature importance menunjukkan bahwa hanya beberapa fitur yang memiliki kontribusi signifikan terhadap proses klasifikasi. Fitur-fitur tersebut menjadi faktor utama dalam menentukan hasil prediksi model, sementara fitur lainnya memiliki pengaruh yang relatif kecil.



Gambar 9. Feature Impoetance Random Forest

Hasil ini menunjukkan pentingnya proses seleksi fitur dalam meningkatkan performa model, karena penggunaan fitur yang relevan dapat membantu model dalam mengenali pola data dengan lebih baik.

4.5. Pembahasan

Berdasarkan hasil penelitian yang telah diperoleh, dapat diketahui bahwa setiap tahapan dalam metodologi memiliki peran yang saling berkaitan dalam menentukan hasil akhir analisis data RNA-Seq *Homo sapiens*.

Tahap *quality control* dan *preprocessing* terbukti menjadi faktor penting dalam meningkatkan kualitas data. Penurunan kualitas pada bagian akhir read yang terdeteksi pada tahap awal dapat diatasi melalui proses *trimming*, sehingga menghasilkan data yang lebih bersih dan siap digunakan dalam analisis. Peningkatan kualitas ini berkontribusi langsung terhadap performa model machine learning.

Pada tahap klasifikasi, perbedaan performa antara Random Forest dan SVM menunjukkan bahwa karakteristik algoritma sangat mempengaruhi hasil yang diperoleh. SVM menghasilkan akurasi yang lebih tinggi karena kemampuannya dalam menemukan batas pemisah optimal antar kelas. Namun, model ini cenderung bias terhadap satu kelas pada dataset yang tidak seimbang, sehingga mengurangi kemampuannya dalam melakukan generalisasi.

Sebaliknya, Random Forest yang menggunakan pendekatan ensemble mampu menghasilkan performa yang lebih stabil. Dengan menggabungkan banyak *decision tree*, model ini mampu mengurangi variansi serta meningkatkan kemampuan dalam mengenali pola data secara menyeluruh. Hal ini menjadikan Random Forest lebih robust terhadap variasi data dibandingkan SVM.

Selain itu, hasil visualisasi menunjukkan bahwa tidak semua fitur memiliki pengaruh yang sama dalam proses klasifikasi. Beberapa fitur memiliki kontribusi dominan, yang menunjukkan bahwa pemilihan fitur menjadi faktor penting dalam meningkatkan performa model. Dengan melakukan optimasi fitur, performa model dapat ditingkatkan lebih lanjut.

Penelitian ini juga menunjukkan bahwa nilai akurasi tidak dapat dijadikan satu-satunya indikator dalam mengevaluasi model. Pada dataset yang tidak seimbang, ukuran seperti precision, recall, dan F1-score memberikan pandangan yang lebih lengkap mengenai kinerja model terhadap. Oleh karena itu, evaluasi model harus dilaksanakan secara menyeluruh untuk mendapatkan hasil yang lebih representatif.

Implikasi dari penelitian ini adalah bahwa dalam analisis data RNA-Seq, diperlukan pendekatan yang komprehensif yang mencakup *preprocessing* data, pemilihan model, serta evaluasi yang tepat. Untuk pengembangan selanjutnya, disarankan untuk menggunakan jumlah kumpulan data yang lebih luas, melakukan tuning parameter model, serta mengeksplorasi metode klasifikasi lain guna meningkatkan akurasi dan stabilitas model.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat ditarik kesimpulan bahwa:

a. Data RNA-Seq *Homo sapiens* yang diperoleh dari Sequence Read Archive berhasil diproses melalui tahap *quality control* dan preprocessing sehingga menghasilkan data dengan kualitas yang lebih baik. Proses trimming terbukti efektif dalam mengurangi noise dan meningkatkan kualitas data.
b. Model Support Vector Machine (SVM) mencapai akurasi yang lebih tinggi yaitu sebesar 0,60 dibandingkan Random Forest

sebesar 0,50. Namun, SVM cenderung bias terhadap satu kelas sehingga kurang mampu melakukan klasifikasi secara seimbang.

c. Random Forest menunjukkan performa yang lebih stabil dalam mengklasifikasikan kedua kelas, meskipun memiliki akurasi yang lebih rendah. Hal ini menunjukkan bahwa pendekatan ensemble lebih robust terhadap variasi data. Terutama di dalam kumpulan data yang tidak seimbang. Metrik lain seperti precision, recall, dan F1-score perlu dipertimbangkan.

d. Hasil penelitian menunjukkan bahwa akurasi tidak dapat dijadikan satu-satunya indikator dalam mengevaluasi model,

e. Penelitian ini memiliki keterbatasan pada jumlah dataset yang relatif kecil, sehingga mempengaruhi performa model. Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar, melakukan optimasi parameter, dan juga mengeksplorasi algoritma lain sebagai upaya untuk meningkatkan hasil klasifikasi

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] R. Permana dan F. A. Herdiana, “Analisis Klasifikasi dan Prediksi Pola Publikasi Berita Menggunakan Machine Learning,” Jurnal Infortech, vol. 7, no. 1, pp. 50–55, 2025.

[2] G. Airlangga, “Application of Traditional Machine Learning Techniques for the Classification of Human DNA Sequences,” Jurnal Informatika Universitas Pamulang, vol. 9, no. 1, pp. 23–30, 2024.

[3] A. Priandika dan A. R. Isnain, “Penerapan Teknik Ensemble Learning untuk Klasifikasi Jenis-jenis Anemia,” MALCOM: Indonesian Journal of Machine Learning, vol. 5, no. 3, pp. 972–980, 2025.

[4] A. Wantoro, et al., “Evaluasi Kinerja Algoritma Machine Learning Menggunakan Seleksi Fitur pada Klasifikasi Diabetes,” Jurnal Informatika Polinema, vol. 11, no. 3, pp. 311–316, 2025.

[5] Wijoyo, A., Saputra, A. Y., Ristanti, S., Sya’Ban, S. R., Amalia, M., Febriansyah, R. (2024). Pembelajaran Machine Learning. *Jurnal Ilmu Komputer dan Science*. 3(2):375-380)

[6] Alvanof, M. M., Bustami, & Dinata, R. K. (2024). Penerapan algoritma Random Forest dalam deteksi dan klasifikasi ransomware. *Jurnal Elektronika dan Teknologi Informasi*, 5(2):23-31

[7] Nurrokhman, M. Z. (2023). Perbandingan Algoritma Support Vector Machine dan Neural Network untuk Klasifikasi Penyakit Hati Ma’mur Zaky Nurrokhman. Indonesian Journal of Computer Science Attribution, 12(4), 2096–2106.

[8] Risnantoyo, R., Nugroho, A., & Mandara, K. (2021). Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm. Journal Of Informatics And Telecommunication Engineering, 4(1), 86–96.

[9] Johansson, E., & Mersha, T. B. (2021). Genetics of food allergy. PMC, 41(2), 301–319. Nusantari, E. (2014). *GENETIKA: Belajar Genetika dengan Mudah & Komprehensif*. Deepublish Publisher

[10] sutaryani, A., Sunarno, S., Djuniadi. (2024). PERBANDINGAN PERFORMA MODEL MACHINE LEARNING DALAM PREDIKSI SUHU DI SEMARANG. JITET (Jurnal Informatika dan Teknik Elektro Terapan). 13(3):2770-2775

[11] F. Indriaharti Harida and N. Khazizah, “Analisis CuacaDi Kota Jakarta Bulan Januari Tahun 2018 Menggunakan Algoritma Decision Tree,” Jurnal Poros Teknik, vol. 14, no. 1, pp. 33–37, 2022

[12] Alvanof, M. M., Bustami, & Dinata, R. K. (2024). Penerapan algoritma Random Forest dalam deteksi dan klasifikasi ransomware. *Jurnal Elektronika dan Teknologi Informasi*, 5(2).

[13] Novianto, E., Suhirman, & Prasetyo, D. (2024). Perbandingan metode klasifikasi Random Forest dan Support Vector Machine dalam memprediksi capaian studi mahasiswa. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI)*, 9(4), 1821–1833.

[14] Dasmasela, R., Tomasouw, B. P., & Leleury, Z. A. (2022). Penerapan metode Support Vector Machine (SVM) untuk mendeteksi penyalahgunaan narkoba. *Parameter: Jurnal Matematika, Statistika dan Terapannya*, 1(2), 111–122.

[15] Eldo, H., Ayuliana, A., Suryadi, D., Chrisnawati, G., & Judijanto, L. (2024). Penggunaan algoritma Support Vector Machine (SVM) untuk deteksi penipuan pada transaksi online. *Jurnal Minfo Polgan*, 13(2), 1627–1632