Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

Analisa Kinerja Algoritma Random Forest dan XGBoost dalam Klasifikasi Penyakit Cacar Monyet (Monkeypox)

Mohammad Dito Dwi Krisna, Firman Noor Hasan*

Fakultas Teknologi Industri dan Informatika, Teknik Informatika, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta
Jl. Tanah Merdeka No.20, RT.11/RW.2, Rambutan, Kec. Ciracas, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia
Email: [1]ditozkrisna@gmail.com, [2,*]firman.noorhasan@uhamka.ac.id
Email Penulis Korespondensi: firman.noorhasan@uhamka.ac.id
Submitted: **01/04/2025;** Accepted: **30/04/2025;** Published: **30/04/2025**

Abstrak−Cacar monyet merupakan penyakit menular yang membutuhkan penanganan yang cepat dan akurat, khususnya dalam proses diagnosis. Namun, proses identifikasi gejala secara manual masih memakan waktu dan rentan terhadap kesalahan. Berdasarkan hal tersebut, penelitian ini dilakukan untuk membangun model klasifikasi berbasis machine learning yang diharapkan mampu membantu proses diagnosis menjadi lebih efisien. Dalam penelitian ini, digunakan dua algoritma machine learning, yaitu XGBoost regression dan Random Forest regression. Keduanya diterapkan untuk mengklasifikasikan pasien yang terinfeksi maupun tidak terinfeksi cacar monyet berdasarkan gejala klinis yang ada. Fokus utama penelitian ini adalah mengukur seberapa baik kedua algoritma tersebut dalam membedakan kelas positif dan negatif, terutama saat data yang digunakan tidak seimbang atau memiliki fitur yang serupa. Data yang digunakan bersumber dari Kaggle dengan jumlah 25.000 entri, masing-masing berisi informasi klinis terkait monkeypox. Sebelum pemodelan dilakukan, data melalui tahap eksplorasi (EDA) dan pra-pemrosesan, termasuk penanganan data hilang. Selanjutnya, digunakan teknik validasi silang dan penyetelan parameter untuk mengoptimalkan hasil klasifikasi. Dari hasil pengujian, XGBoost menunjukkan performa terbaik dengan akurasi 68%, presisi 69%, recall 89%, dan F1-score 78%. Sementara Random Forest mencatatkan hasil yang sedikit lebih rendah. Evaluasi menggunakan kurva ROC menunjukkan nilai AUC sebesar 0,60 pada kedua model, yang menandakan bahwa kemampuan keduanya dalam membedakan kelas positif dan negatif masih perlu ditingkatkan.

Kata Kunci: Klasifikasi; Random Forest; XGBoost; Machine Learning; Cacar Monyet

Abstract−Monkeypox is a contagious disease that requires prompt and accurate handling, particularly in the diagnostic process. However, identifying symptomps manually often takes time and is prone to error. In response to this challenge, this study aims to develop a machine learning based classification model to support a more efficient diagnosis process. This research applies two machine learning algorithms XGBoost regression and Random Forest regression to classify patients as infected or uninfected with monkeypox based on clinical symptoms. The study focuses on assessing how well each algorithm can distinguish between positive and negative cases, especially when dealing with imbalanced data or overlapping features. The dataset used consists of 25.000 entries sourced from Kaggle, each containing clinical indicators related to monkeypox. Before modeling, the data underwent exploratory data analysis (EDA) and preprocessing, including handling missing values. Furthermore, cross-validation and parameter tuning techniques were implemented to optimize model performance. The results indicate that XGBoost outperformed Random Forest, achieving 68% accuracy, 69% precision, 89% recall, and a 78% F1-score. In contrast, Random Forest yielded slightly lower scores. Both models were evaluated using the ROC curve, where each reached an AUC values of 0.60. This suggests that wile both models show potential, their ability to clearly distinguish between classes positive and negative remains limited and can be improves in future work.

Keywords: Classification; Random Forest; XGBoost; Machine Learning; Monkeypox

1. PENDAHULUAN

Monkeypox virus (MPXV) adalah virus zoonotic dari genus Orthopoxvirus yang memiliki gejala mirip dengan smallpox, seperti demam, nyeri otot, pembengkakan kelenjar getah bening,serta ruam di kulit [1], meskipun tingkat keparahannya lebih ringan dibandingkan smallpox, penyebaran monkeypox tetap perlu diwaspadai karena bisa menular melalui kontak langsung dengan luka terbuka, cairan tubuh, atau benda yang terkontaminasi [2]. Sejak mei 2022, penyakit ini mulai menyebar ke luar wilayah endemik di Afrika dan tercatat di berbagai negara termasuk Indonesia, Eropa, dan Amerika [3].

Pada awal 1970-an, Republik Demokratik, kongo, melaporkan kasus pertama cacar monyet pada manusia [4]. Penyakit monkeypox disebabkan oleh infeksi virus dari kelompok Orthopoxvirus, yaitu virus DNA yang dikenal dapat menimbulkan gejala serupa dengan penyakit cacar pada manusia [5][6]. Salah satu ciri khas dari virus dalam genus ini adalah kemampuannya menghindari sistem kekebalan tubuh inang, sehingga infeksi bisa terjadi tanpa langsung dikenali oleh sistem imun [7].

Secara umum, monkeypox terbagi menjadi dua jenis atau klade yang berbeda secara genetik dan geografis. Klade 1 umumnya ditemukan di wilayah Afrika Tengah dan cenderung menyebabkan infeksi yang lebih serius, dengan risiko kematian yang lebih tinggi. Sementara itu, Klade 2 yang lebih sering muncul di Afrika Barat, biasanya menimbulkan gejala yang lebih ringan. Perbedaan antara kedua klade ini turut memengaruhi pola penyebaran dan tingkat keparahan penyakit pada manusia [8].

Berbeda dengan penyakit cacar atau cacar air yang hanya menular antar manusia melalui kontak langsung, monkeypox juga bisa menyebar dari hewan ke manusia. Penularan ini dapat terjadi melalui darah, luka terbuka, atau cairan tubuh lainnya. Sifat zoonotic inilah yang membuat monkeypox lebih sulit dikendalikan, karena berpotensi menyebar lintas spesies [9].

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

Pada 20 Agustus 2022, Kementerian Kesehatan RI mengumumkan kasus pertama MonkeyPox di Indonesia, pada seseorang pria 27 tahun yang baru kembali dari Eropa, pasien mulai menunjukkan gejala pada 11 Agustus dari hasil tes PCR pada 19 Agustus di konfirmasi infeksi tersebut, meskipun penularan MonkeyPox lebih rendah dibandingkan COVID-19, masyarakat diimbau tetap waspada dan menjaga kebersihan [10]. Oleh sebab itu, penggunaan teknologi berbasis data seperti machine learning dibutuhkan untuk membantu proses diagnosis secara lebih cepat dan tepat,

Banyak penelitian telah dilakukan untuk mendeteksi penyakit MonkeyPox, salah satunya menggunakan pendekatan Machine Learning [11]. Machine Learning merupakan bagian dari Artificial Intelligence (AI) yang memungkinkan mesin belajar dan menyelesaikan tugas secara fleksibel dalam berbagai bidang, dengan memanfaatkan data yang ada untuk diolah agar bisa mengenali pola dan membuat keputusan [12].

Machine Learning memberikan banyak manfaat di bidang kesehatan, seperti: membantu identifikasi, diagnosis, dan prediksi penyakit, serta meningkatkan analisis data rekam medis dan pengelolaan gambar hasil pemeriksaan kesehatan, Teknologi ini menawarkan solusi yang efisien dalam aspek pengelolaan kesehatan [13]. Maka dari itu, Machine Learning memiliki manfaat besar, banyak metode yang diberikan untuk memproses data yang terbagi menjadi empat jenis, yaitu: pembelajaran terbimbing, pembelajaran tidak terbimbing, pembelajaran semi-terbimbing, dan pembelajaran penguatan [14].

Pada penelitian ini, pembahasan akan difokuskan untuk menerapkan dua metode yang popular, yaitu XGBoost dan Random Forest untuk mengklasifikasikan penyakit cacar monyet (MonkeyPox), XGBoost atau Extreme Gradient Boosting adalah metode machine learning yang sangat efektif untuk tugas klasifikasi dan regresi, Algoritma ini menggunakan teknik ensemble learning, dimana model pohon keputusan dibuat secara bertahap, dengan setiap model baru bertujuan untuk memperbaiki kesalahan yang terjadi pada model sebelumnya, XGBoost juga memiliki mekanisme optimasi yang agresif untuk mengurangi kesalahan prediksi, sambil menggunakan teknik untuk membatasi kerumitan model agar tidak terlalu rumit atau sulit digunakan (overfitting), selain dari itu, algoritma ini memungkinkan evaluasi penting setiap fitur dalam proses prediksi, berkat kinerjanya yang tinggi dan fleksibel, XGBoost menjadi pilihan utama di berbagai kompetisi data dan pemanfaatan di berbagai bidang [15].

Random Forest adalah algoritma machine learning yang menggabungkan sejumlah pohon keputusan, di mana setiap pohon dibuat secara acak dari data pelatihan, hasil prediksi dari semua pohon ini kemudian digabungkan untuk menghasilkan prediksi akhir yang lebih akurat, keunggulan utama Random Forest adalah kemampuannya menangani data yang besar dan kompleks tanpa mudah mengalami overfitting, Algoritma ini juga memberikan hasil yang lebih stabil dan memungkinkan kita menilai sejauh mana setiap fitur berkontribusi terhadap prediksi cacar monyet [16].

Penelitian ini bertujuan untuk menganalisis kinerja metode XGBoost dan Random Forest dalam mengklasifikasikan penyakit cacar monyet (MonkeyPox), Dataset yang digunakan berisi berbagai fitur klinis, seperti: nyeri rektal, pembengkakan kelenjar getah bening, dan lesi kulit, yang relevan untuk diagnosis cacar monyet [17]. Data ini kemudian dilatih menggunakan kedua algoritma untuk membandingkan efektivitasnya dalam melakukan klasifikasi [18]. Berbagai penelitian tentang perbandingan algoritma dalam klasifikasi penyakit telah dilakukan dengan menggunakan berbagai metode untuk fokus menguji pada analisis kinerja algoritma untuk menentukan keunggulan masing-masing dalam menghasilkan prediksi yang akurat dan dapat diandalkan [19]. Beberapa penelitian sebelumnya, antara lain:

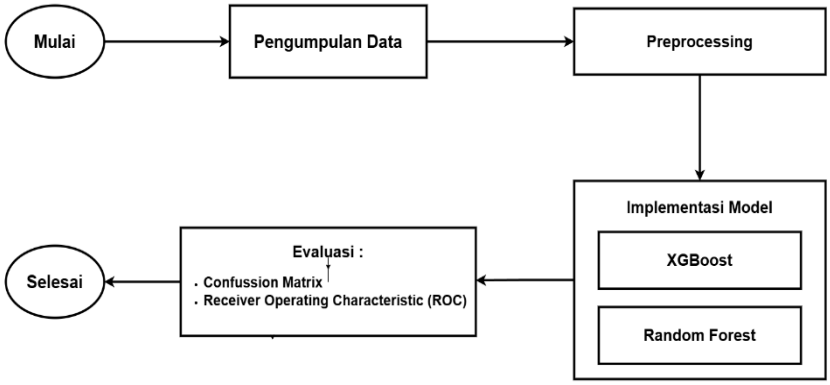
Sejumlah penelitian sebelumnya telah menggunakan algoritma machine learning untuk klasifikasi dalam bidang medis, khususnya dalam memprediksi tingkat keparahan suatu penyakit. Salah satu studi yang memanfaatkan data klinis pasien COVID-19 menemukan bahwa algoritma XGBoost mampu memberikan hasil paling akurat, dengan nilai AUC mencapai 0,93. Sementara itu, Random Forest menyusul dengan AUC sebesar 0,89. Temuan ini menunjukkan bahwa XGBoost memiliki keunggulan dalam membedakan tingkat keparahan pasien dibandingkan algoritma lainnya [20].

Penelitian ini menggunakan dataset yang diperoleh dari platform Kaggle, yang terdiri dari 25.000 data dengan berbagai indikator klinis yang berkaitan dengan penyakit monkeypox. Untuk melakukan prediksi, diterapkan dua algorima machine learning, yaitu Random Forest dan XGBoost. Proses penelitian diawali dengan tahap preprocessing data, kemudian dilanjutkan dengan pengujian model menggunakan teknik K-Fold Cross Validation. Hasil dari model dievaluasi dengan beberapa metrik umum, seperti akurasi, presisi, recall, dan F1-score. Kedua algoritma dipilih karena dinilai mampu mengelola data dalam jumlah besar secara efisien, serta memiliki kemampuan yang baik dalam mengurangi risiko overfitting, sehingga sangat sesuai untuk diterapkan pada kasus klasifikasi seperti ini.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan algoritma Random Forest dan XGBoost sebagai metode utama. Proses dalam metode ini terdiri dari beberapa tahap, seperti yang ditunjukkan pada Gambar 1.

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167



Gambar 1. Tahapan Penelitian

Dalam penelitian ini, tahapan awal yang dilakukan adalah proses pengumpulan data, di mana penelitian menggunakan dataset dari Kaggle yang memuat sekitar 25.000 entri data terkait kondisi klinis pasien monkeypox, mencakup atribut seperti nyeri rektal, pembengkakan amandel, infeksi HIV, dan sejumlah gejala lainnya [21]. Setelah data diperoleh, langkah berikutnya adalah melakukan Exploratory Data Analysis (EDA). Tahapan ini sangat penting karena memberikan gambaran awal mengenai struktur data, seperti nilai rata-rata, median, hingga penyebaran data melalui standar deviasi dan varians. EDA juga mebantu mendeteksi anomaly atau outlier yang dapat memengaruhi akurasi model secara signifikan [22].

Data yang sudah dianalisis kemudian dibersihkan melalui tahap preprocessing. Pada tahap ini, peneliti menghilangkan duplikasi, nilai yang hilang, serta mengonversi data kategorikal ke bentuk numerik. Beberapa fitur dikodekan menggunakan teknik mapping sederhana maupun one-hot encoding, tergantung pada jumlah kategori yang dimiliki [23][24]. Proses ini bertujuan agar data menjadi lebih terstruktur dan mudah diproses oleh algoritma klasifikasi. Setelah data lengkap, dilakukan pembagian dataset ke dalam dua bagian menggunakan teknik stratified sampling, yakni 80% untuk pelatihan dan 20% untuk pengujian, agar distribusi antara label positif dan negatif tetap seimbang [25].

Selanjutnya, dua algoritma machine learning diterapkan, yaitu XGBoost dan Random Forest. XGBoost yang dikenal sebagai Extreme Gradient Boosting, bekerja dengan cara membangun model secara bertahap untuk memperbaiki kesalahan dari model sebelumnya, dengan memanfaatkan optimasi berbasis gradien dan hessian, serta dikenal sangat efisien dalam menangani data kompleks [26]. Di sisi lain, Random Forest adalah algoritma ensemble yang menggabungkan banyak pohon Keputusan, lalu mengambil hasil voting sebagai output akhir. Algoritma ini dikenal stabil dan mampu menghindari overfitting, meskipun membutuhkan sumber daya komputasi yang relatif besar [27]. Seperti pada rumus perhitungan nomor 1, 2, 3, dan 4.

$$\text{Accuracy} = \frac{\text{TP+FN}}{\text{TP+FP+FN+TN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

Untuk menguji keandalan kedua model, digunakan teknik 10 fold cross-validation, yang membagi data pelatihan menjadi 10 bagian dan secara bergiliran digunakan untuk validasi [28]. Hasil model kemudian dievaluasi menggunakan confusion matrix, yang mencakup empat kategori utama, yaitu True Positive, True Negative, False Positive, dan False Negative. Dari sini dihitung metrik evaluasi seperti akurasi, presisi, recall, dan F1-score [29][30]. Selain itu, analisis dilengkapi dengan kurva ROC (Receiver Operating Characteristic) yang menggambarkan keseimbangan antara true positive rate dan false positive rate, serta menghasilkan nilai AUC sebagai ukuran seberapa baik model dapat membedakan antara dua kelas. Kurva ROC juga memberikan visualisasi trade-off yang penting untuk dipertimbangkan dalam pemilihan model [31]. Untuk mengevaluasi kinerja model klasifikasi, digunakan confusion matrix sebagai alat pengujian hasil prediksi. Melalui confusion matrix ini, performa model dapat dianalasis berdasarkan empat komponen utama, yang ditampilkan secara rinci pada Tabel 1 berikut.

Tabel 1. Confusion Matrix

	Prediksi Positive	**Prediksi Negative**
Aktual Positive	True Positive (TP)	False Positive (FP)
Aktual Negative	False Negative (FN)	True Negative (TN)

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pengumpulan data ialah suatui proses ataupun kegiatan yang dilakukan oleh peneliti dengan tujuan mengetahui atau menjaring berbagai fenomena, informasi atau kondisi lokasi penelitian yang sesuai dengan lingkup penelitian

Tabel 2. Daftar fitur dalam dataset klasifikasi monkeypox

No	Fitur Dataset	Penjelasan
1	Systemic Ilness	Indikasi apakah pasien memiliki penyakit sistemik
2	Rectal Pain	Indikasi apakah pasien mengalami nyeri rektal
3	Sore Throat	Indikasi apakah pasien mengalami sakit tenggorokan
4	Penile Oedema	Indikasi apakah pasien mengalami pembengkakan pada penis
5	Oral Lesions	Indikasi apakah pasien memiliki lesi atau luka di mulut
6	Solitary Lesion	Indikasi apakah pasien memiliki lesi tunggal pada tubuh
7	Swollen Tonsils	Indikasi apakah pasien mengalami pembengkakan amandel
8	HIV infection	Indikasi apakah pasien terinfeksi HIV
9	Sexually Transmitted Infection	Indikasi apakah pasien memiliki infeksi menular seksual lainnya

Penelitian ini memakai dataset dari Kaggle, yang berisi 25000 data dengan 9 atribut klinis, seperti Nyeri Rektal, Pembengkakan Kelenjar Getah Bening, Lesi Kulit, dan gejala lainnya, data ini terbagi dalam dua kategori, yaitu positif dan negative pada penyakit Monkeypox. Melalui penggunaan dataset ini, peneliti dapat membangun model klasifikasi yang akurat untuk memprediksi risiko infeksi Monkeypox pada pasien. Hasil analisis ditampilkan pada Tabel 2.

3.2 Preprocessing

Sebagai Langkah awal dalam proses pemodelan, dilakukan Exploratory Data Analysis (EDA) untuk memperoleh pemahaman menyeluruh terhadap struktur dan karakteristik data yang digunakan. Dataset terdiri dari 25.000 entri yang merepresentasikan kondisi klinis pasien terkait infeksi monkeypox, dengan total 11 fitur yang mencakup gejala seperti Sore Throat, Penile Oedema, HIV Infection, hingga label target MonkeyPox.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Patient_ID                    25000 non-null  object
 1   Systemic Illness              25000 non-null  object
 2   Rectal Pain                   25000 non-null  bool
 3   Sore Throat                   25000 non-null  bool
 4   Penile Oedema                 25000 non-null  bool
 5   Oral Lesions                  25000 non-null  bool
 6   Solitary Lesion               25000 non-null  bool
 7   Swollen Tonsils               25000 non-null  bool
 8   HIV Infection                 25000 non-null  bool
 9   Sexually Transmitted Infection  25000 non-null  bool
 10  MonkeyPox                     25000 non-null  object
dtypes: bool(8), object(3)
memory usage: 781.4+ KB
```

Gambar 2. Tampilan struktur dataframe

Pada Gambar 2 ditampilkan ringkasan struktur data yang menunjukkan bahwa seluruh kolom memiliki nilai lengkap (non-null), sehingga tidak diperlukan proses penggantian nilai terhadap data yang hilang. Sebagian besar fitur bertipe boolean, sementara tiga fitur lainnya bertipe objek yaitu Patient_ID, Systemic Illness, dan MonkeyPox. Struktur ini menunjukkan bahwa data telah cukup siap untuk di proses lebih lanjut oleh algoritma klasifikasi.

MonkeyPox	count
Positive	15909
Negative	9091

dtype: int64

Gambar 3. Distribusi label positif dan negatif

Pada Gambar 3 diketahui bahwa sebanyak 15909 data masuk dalam kategori psoitif (terinfeksi) dan 9091 termasuk negatif (tidak terinfeksi). Ketidakseimbangan ini mengindikasikan dominasi kelas positif hingga sekitar

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

64% dari total data, yang dapat memengaruhi performa model, khususnya dalam mengenali kelas minoritas. Penyesuaian bobot kelas menjadi pertimbangan penting dalam proses latih dan uji.

3.2.1 Mapping

Setelah mendapatkan gambaran umum tentang data, dilakukan proses mapping terhadap fitur-fitur kategorikal yang memiliki nilai teks, seperti pada kolom Systemic Illnes.

	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	MonkeyPox
0	None	0	1	1	1	0	1	0	0	0
1	Fever	1	0	1	1	0	0	1	0	1
2	Fever	0	1	1	0	0	0	1	0	1
3	None	1	0	0	0	1	1	1	0	1
4	Swollen Lymph Nodes	1	1	1	0	0	1	1	0	1

Gambar 4. Hasil mapping nilai kategorikal

Pada Gambar 4 ditunjukkan bagaimana nilai-nilai seperti “Fever”, “None”, atau “Swollen Lymph Nodes” dikonversi menjadi representasi numerik sederhana, seperti 0 dan 1. Langkah ini bertujuan agar nilai teks tersebut dapat dibaca oleh algoritma machine learning, sekaligus menjaga efisiensi dan konsistensi dalam pengolahan data.

3.2.2 One-hot encoding

Selanjutnya, untuk fitur kategorikal yang memiliki lebih dari dua nilai unik, diterapkan teknik one-hot encoding.

	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	MonkeyPox	Systemic Illness_Muscle Aches and Pain	Systemic Illness_None	Systemic Illness_Swollen Lymph Nodes
0	0	1	1	1	0	1	0	0	0	0	1	0
1	1	0	1	1	0	0	1	0	1	0	0	0
2	0	1	1	0	0	0	1	0	1	0	0	0
3	1	0	0	0	1	1	1	0	1	0	1	0
4	1	1	1	0	0	1	1	0	1	0	0	1

Gambar 5. Hasil one-hot encoding fitur systemic illness

Pada Gambar 5 metode ini memecah satu kolom menjadi beberapa kolom baru, di mana masing-masing kolom mewakili satu kategori dalam format biner. Sebagai contoh, kolom Systemic Illness dipecah menjadi Systemic Ilness_None, Systemic Ilness_Muscle Aches and Pain, dan Systemic Ilness_Swollen Lymph Nodes. Pendekatan ini digunakan untuk mencegah algoritma salah memahami hubungan antar kategori sebagai urutan atau tingkat tertentu dan sekaligus membantu memperjelas struktur data bagi model klasifikasi.

3.3 Implementasi Model

Data yang sudah melewati tahap processing, dibagi menjadi dua bagian, yaitu data latih 80% dan data uji 20%. Pembagian dilakukan dengan teknik stratified sampling, untuk memastikan proporsi label positif dan negatif tetap seimbang di kedua subset dengan masing-masing memiliki 11 fitur.

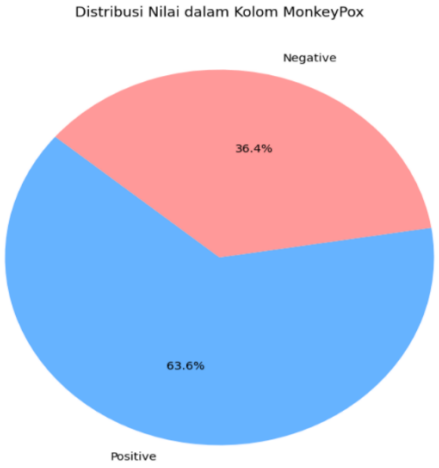
```
y train: MonkeyPox
1    0.63635
0    0.36365
Name: proportion, dtype: float64


y test: MonkeyPox
1    0.6364
0    0.3636
Name: proportion, dtype: float64
```

Gambar 6. Distribusi label pada data train dan test

Pada Gambar 6, distribusi kelas pada data latih menunjukkan bahwa 63,6% merupakan label positif (terinfeksi monkeypox) dan 36,4% adalah negative, begitu juga dengan data uji yang memiliki proporsi serupa. Untuk menjaga keseimbangan model selama proses pelatihan.

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167



Gambar 7. Diagram lingkaran pada kolom distribusi label monkeypox

Visualisai pada Gambar 7 menujukkan sebaran data pda kolom Monkeypox dalam bentuk diagram lingkaran. Terlihat bahwa Sebagian besar data, yaitu sekitar 63,6% merupakan kasus positif, sementara sisanya sebesar 36,4% termasuk dalam kategori negative.

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
1	XGBClassifier	67.52	1.11	67.90	0.79
0	Random Forest	65.94	1.05	66.96	0.87

Gambar 8. Hasil evaluasi model AUC dan akurasi

Pada Gambar 8 penggunaan model Random Forest dan XGBoost, kedua model dilatih menggunakan data hasil preprocessing, kemudian diuji menggunakan teknik 10-fold cross validation untuk mendapatkan rata-rata performa dan standar deviasi dari setiap metrik evaluasi. Hasil validasi silang pada data latih menunjukkan bahwa XGBoost memiliki nilai ROC AUC rata-rata sebesar 67,52% dan akurasi rata-rata sebesar 67,90%, dengan deviasi yang cukup rendah yaitu 1,11 untuk ROC dan 0,79 untuk akurasi. Sedangkan Random Forest memiliki ROC AUC sebesar 65,94% dan akurasi rata-rata sebesar 66,96% dengan standar deviasi yang juga rendah. Nilai menunjukkan kedua model relative stabil dan tidak overfitting.

Tabel 3. Hasil dari proses pemodelan algoritma.

Metode	Accuracy	Precision	Recall	F1-Score
XGBoost	68%	69%	89%	78%
Random Forest	67%	69%	87%	77%

Pada Tabel 3 hasil setelah proses pelatihan, kedua model diuji menggunakan 5.000 data uji yang telah dipisahkan sebelumnya. Hasil ini menunjukkan bahwa XGBoost unggul dalam metrik recall dan F1-score, yang berarti lebih mampu dalam mendeteksi kasus positif secara benar. Nilai recall yang tinggi sangat penting dalam konteks penyakit menular seperti monkeypox, karena lebih baik mengidentifikasi sebanyak mungkin kasus positif meskipun ada kemungkinan beberapa false positive. Namun, Random Forest sedikit lebih baik dari segi presisi dan ROC-AUC, yang berarti cenderung memberikan prediksi positf yang lebih akurat. Kedua model memiliki akurasi yang hampir setara dan perbedaan performa sangat tipis.

3.4 Evaluasi

Sebagian dari hasil proses analisis, data juga telah diuji menggunakan dua algoritma klasifikasi, yakni XGBoost dan Random Forest. Hasil penerapan keduanya dapat dilihat pada gambar 9 dan 10 diperoleh hasil evaluasi berupa confusion matrix. Dari matrix tersebut dihitung nilai akurasi, presisi, recall, dan F1-score untuk menilai performa masing-masing metode, serta menentukan algoritma mana yang paling efektif digunakan dalam penelitian ini.

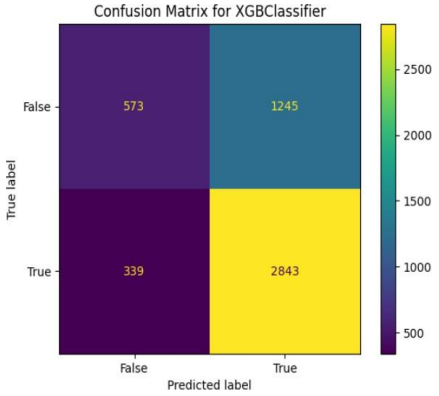
a. XGBoost
   Data yang telah diproses menggunakan metode XGBoost, selanjutnya di evaluasi melalui confusion matrix. Dari hasil tersebut, diperoleh nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) yang ditampilkan pada Tabel 4

Tabel 4. Tabel confusion matrix untuk model XGBoost

	True Positive	True Negative
False Positive	2.843	339
False Negative	1.245	573

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

Berdasarkan implementasi dari metode XGBoost yang terdapat pada gambar 9, menunjukkan bahwa accuracy yang dihasilkan sebesar 68.32%, precision sebesar 69.57%, recall sebesar 89.35% dan F1-score 78%. Nilai recall yang tinggi mengindikasikan bahwa XGBoost memiliki kemampuan sangat baik dalam mendeteksi kasus positif, meskipun terdapat trade-off dengan precision.



Gambar 9. Confusion matrix hasil prediksi model XGBoost [ada data uji.

$$Accuracy = \frac{2843+573}{2843+573+1245+339} = 68\% \quad (1)$$

$$Precision = \frac{2843}{2843+1245} = 69.5\% \quad (2)$$

$$Recall = \frac{2843}{2843+339} = 89.3\% \quad (3)$$

$$F1-score = 2 \times \frac{0.695 \times 0.893}{0.695+0.893} = 0.78 \quad (4)$$

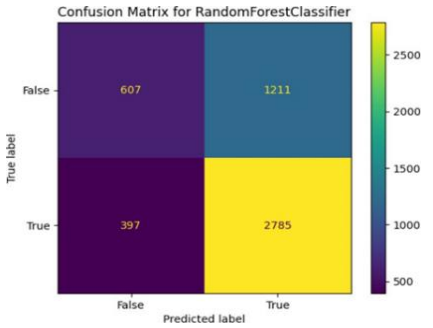
b. Random Forest

Data yang telah diproses menggunakan metode Random Forest, selanjutnya di evaluasi melalui confusion matrix. Dari hasil tersebut, diperoleh nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) yang ditampilkan pada Tabel 5

Tabel 5. Tabel confusion matrix untuk model Random Forest.

	True Positive	True Negative
False Positive	2.785	397
False Negative	1.211	607

Berdasarkan hasil evaluasi dari model Random Forest yang ditunjukkan pada gambar 10, diperoleh nilai accuracy sebesar 67.84%, precision sebesar 69%, recall sebesar 87.54%, dan F1-score sebesar 77.5%. meskipun sedikit lebih dibandingkan XGBoost dalam aspek recall dan F1-score, Random Forest menunjukkan presisi yang serupa, serta memiliki kemampuan lebih baik dalam mengurangi kesalahan klasifikasi negatif (False Positive)



Gambar 10. Confusion matrix hasil prediksi model Random Forest pada data uji

$$Accuracy = \frac{2785+607}{2785+607+1211+397} = 67\% \quad (1)$$

$$Precision = \frac{2785}{2785+1211} = 69.7\% \quad (2)$$

$$Recall = \frac{2875}{2785+397} = 87.5\% \quad (3)$$

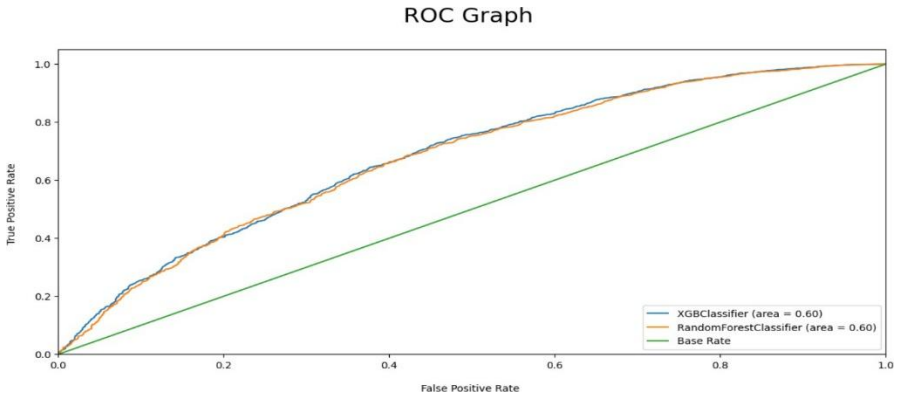
$$F1-score = 2 \times \frac{0.697 \times 0.875}{0.697+0.875} = 0.77 \quad (4)$$

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

Di sisi lain, algoritma Random Forest menunjukkan hasil yang sedikit berbeda, dengan recall sebesar 87%, sedikit lebih rendah dibandingkan XGBoost. Berdasarkan confusion matrix untuk Random Forest (Gambar 10), algoritma ini mencatat 607 True Negative (TN), 1211 False Positive (FP), 2785 True Positive (TP), dan 397 False Negative (FN). Secara keseluruhan, algoritma Random Forest walaupun deteksi data positif nya sedikit lebih rendah dibandingkan XGBoost, algoritma Random Forest lebih baik dalam meminimalkan kesalahan prediksi pada data negative (False Positive). Kedua algoritma ini memberikan hasil yang cukup baik dan dapat dipilih sesuai dengan kebutuhan spesifik dalam penerapannya.

3.5 ROC (Receiver Operating Characteristic)

Analisis ROC (Receiver Operating Characteristic) adalah metode untuk menggambarkan, mengatur, dan mengklasifikasikan beberapa kategori yang ditentukan pada sebuah model statistik berdasarkan kinerjanya.



Gambar 11. Kurva ROC untuk model XGBoost dan Random Forest pada klasifikasi penyakit monkeypox

Gambar 11 memperlihatkan kurva ROC dari dua model klasifikasi yang digunakan dalam penelitian ini, yaitu XGBoost dan Random Forest. Pada grafik tersebut, garis biru menggambarkan performa XGBoost, sementara garis oranye menunjukkan hasil dari Random Forest. Sebagai pembanding, garis hijau merepresentasikan baseline atau tingkat akurasi dari model acak (AUC = 0,50). Berdasarkan grafik, terlihat bahwa kedua model memiliki nilai AUC sebesar 0,60. Niali ini menunjukkan bahwa kemampuan kedua model dalam membedakan antara positif dan negatif masih tergolong rendah. Meski demikian, kurva ROC keduanya tetap berada di atas baseline, yang menandakan bahwa model memiliki performa yang lebih baik daripada tebakan acak. Rendahnya nilai AUC ini memungkinkan dipengaruhi oleh sejumlah faktor, seperti distribusi kelas yang tidak seimbang, kemiripan karakteristik antar fitur, atau belum optimalnya proses feature engineering yang dilakukan. Oleh karena itu, diperlukan pengembangan lebih lanjut agar performa klasifikasi dapat ditingkatkan secara signifikan.

4. KESIMPULAN

Pada penelitian machine learning ini yang bertujuan untuk menguji dan mengevaluasi efektivitasnya algoritma dalam menganalisis penyakit cacar monyet. Pengujian dilakukan dengan menggunakan dua algoritma sebagai alat klasifikasi. Melalui data latih dan data uji, penelitian ini mengevaluasi dengan menggunakan kinerja algoritma Random Forest (RF) dan XGBoost (GB) melalui pengukuran performa yang mencakup akurasi, presisi, recall, F1-score, serta penggunaan confusion matrix. Berdasarkan hasil penelitian, akurasi tertinggi sebesar 0.68% diperoleh oleh algoritma XGBoost, yang mengungguli Random Forest sebesar 0.67% dalam analisis penyakit cacar monyet. Selain itu, hasil AUC pada algoritma XGBoost dan Random Forest memiliki nilai 0,60. Kurva ROC dari kedua model terlihat berada sedikit di atas garis baseline namun hasil hampir saling tumpang tindih, menandakan bahwa kedua model algoritma memiliki performa hampir sama. Metodologi analisis penyakit cacar monyet ini diharapkan Khususnya di bidang Kesehatan, untuk menilai dampak penyakit cacar monyet dan langkah-langkah pencegahan terhadap kesehatan pasien serta masyarakat. Dengan demikian, tindakan cepat dapat diambil untuk menangani penyakit ini sebelum lebih luas dan memberi dampak negatif pada masyarakat. Peneliti menyarankan untuk mengembangkan model yang telah diuji dan memperluas ruang lingkup penelitian, termasuk membandingkan dengan kinerja algoritma machine learning (ML) yang lain, dalam upaya dapat mendukung proses diagnose penyakit secara lebih cepat dan akurat, khusunya dalam konteks penanganan penyakit menular seperti monkeypox yang memerlukan respon dini dan cepat

REFERENCES

[1] J. Kwong, K. C. McNabb, J. G. Voss, A. Bergman, K. McGee, and J. Farley, “Monkeypox Virus Outbreak 2022: Key Epidemiologic, Clinical, Diagnostic, and Prevention Considerations,” J. Assoc. Nurses AIDS Care, vol. 33, no. 6, pp. 657–667, 2022, doi: 10.1097/JNC.0000000000000365.

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

[2] E. Sherwood et al., "Invasive group A streptococcal disease in pregnant women and young children: a systematic review and meta-analysis," Lancet Infect. Dis., vol. 22, no. 7, pp. 1076–1088, Jul. 2022, doi: 10.1016/S1473-3099(21)00672-1.
[3] A. R. A. Saied, M. Dhawan, A. A. Metwally, M. L. Fahrni, P. Choudhary, and O. P. Choudhary, "Disease History, Pathogenesis, Diagnostics, and Therapeutics for Human Monkeypox Disease: A Comprehensive Review," MDPI, Vol 10, No 12, doi: 10.3390/vaccines10122091.
[4] F. Wei et al., "Study and prediction of the 2022 global monkeypox epidemic," J. Biosaf. Biosecurity, vol. 4, no. 2, pp. 158–162, Dec. 2022, doi: 10.1016/j.jobb.2022.12.001.
[5] J. Lu et al., "Mpox (formerly monkeypox): pathogenesis, prevention, and treatment," Dec. 27, 2023, Springer Nature. doi: 10.1038/s41392-023-01675-2.
[6] H. Harapan et al., "Monkeypox: A Comprehensive Review," Sep. 29, 2022, MDPI. doi: 10.3390/v14102155.
[7] D. L. Fink et al., "Clinical features and management of individuals admitted to hospital with monkeypox and associated complications across the UK: a retrospective cohort study," Lancet Infect. Dis., vol. 23, no. 5, pp. 589–597, May 2023, doi: 10.1016/S1473-3099(22)00806-4.
[8] F. Aldi, I. Nozomi, R. B. Sentosa, and A. Junaidi, "Machine Learning to Identify Monkey Pox Disease," Sinkron, vol. 8, no. 3, pp. 1335–1347, Jul. 2023, doi: 10.33395/sinkron.v8i3.12524.
[9] sehatnegeriku.kemkes.go.id, "Kasus monkeypox pertama di Indonesia terkonfirmasi. Sehat Negeriku," https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220820/3140968/kasus-monkeypox-pertama-di-indonesia-terkonfirmasi-2/ .
[10] M. M. Ahsan, M. R. Uddin, and S. A. Luna, "Monkeypox Image Data collection," arxiv, Jun. 2022, doi: https://doi.org/10.48550/arxiv.2206.01774.
[11] A. Wijoyo, A. Y. Saputra, S. Ristanti, R. Sya'ban, M. Amalia, and R. Febriansyah, "Pembelajaran Machine Learning," OKTAL, vol. 3, pp. 375–380, Feb. 2024, Accessed: Feb. 05, 2024. [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal/article/view/2305
[12] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," in Machine Learning and the Internet of Medical Things in Healthcare, Elsevier, 2021, pp. 89–111. doi: 10.1016/B978-0-12-821229-5.00003-3.
[13] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms," MDPI, Vol 12, No 8, 2023, MDPI. doi: 10.3390/electronics12081789.
[14] N. Nyoman, P. Pinata, M. Sukarsa, N. Kadek, and D. Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," J. Ilm. MERPATI, vol. 8, pp. 188–196, Dec. 2020, doi: 10.24843/jim.2020.v08.i03.p04,2020.
[15] F. ANISHA, Dodi Vionanda, Nonong amalita, and Zilrahmi, "Application of Random Forest for The Classification Diabetes Mellitus Disease in RSUP Dr. M. Jamil Padang," UNP J. Stat. Data Sci., vol. 1, no. 2, pp. 45–52, Mar. 2023, doi: 10.24036/ujsds/vol1-iss2/30.
[16] L. Hoang Huong, N. Hoang Khang, L. Nhat Quynh, L. Huu Thang, D. Minh Canh, and H. Phuoc Sang, "A Proposed Approach for Monkeypox Classification," International Journal of Advanced Computer Science and Applications(IJACSA), Vol 14, No 8, 2023. doi: http://dx.doi.org/10.14569/IJACSA.2023.0140871.
[17] M. E. Haque, M. R. Ahmed, R. S. Nila, and S. Islam, "Classification of Human Monkeypox Disease Using Deep Learning Models and Attention Mechanisms," arxiv, Nov. 2022, doi: https://doi.org/10.48550/arXiv.2211.15459.
[18] A. Khairunnisa, "Perbandingan Model Random Forest Dan Xgboost Untuk Prediksi Kejahatan Kesusilaan Di Provinsi Jawa Barat," JIKO (Jurnal Inform. dan Komputer), vol. 7, no. 2, p. 202, Sep. 2023, doi: 10.26798/jiko.v7i2.799.
[19] W. Hong et al., "A Comparison of XGBoost, Random Forest, and Nomograph for the Prediction of Disease Severity in Patients With COVID-19 Pneumonia: Implications of Cytokine and Immune Cell Profile," Front. Cell. Infect. Microbiol., vol. 12, Apr. 2022, doi: 10.3389/fcimb.2022.819267.
[20] M. Ahmed, "Monkey-Pox PATIENTS Dataset.," https://doi.org/10.34740/KAGGLE/DSV/4271503.
[21] H. Faisal, A. Febriandirza, and F. N. Hasan, "Analisis Sentimen Terkait Ulasan Pada Aplikasi PLN Mobile Menggunakan Metode Support Vector Machine," KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen), vol. 5, no. 1, pp. 303–312, Jan. 2024, doi: https://doi.org/10.30645/kesatria.v5i1.339.
[22] S. Desmalia, A. Mutoi Siregar, K. A. Baihaqi, and T. Rohana, "Comparison Model Optimal Machine Learning Model With Feature Extraction for Heart Attack Disease Classification," Sci. J. Informatics, vol. 11, no. 2, pp. 485–492, Mar. 2024, doi: 10.15294/sji.v11i2.4561.
[23] K. Nugroho and F. N. Hasan, "Analisis Sentimen Masyarakat Mengenai RUU Perampasan Aset Di Twitter Menggunakan Metode Naïve Bayes," SMATIKA J., vol. 13, no. 02, pp. 273–283, Dec. 2023, doi: 10.32664/smatika.v13i02.899.
[24] D. Baharudin, Pembelajaran Machine Learning. 2024. doi: https://books.google.co.id/books?id=bdouEQAAQBAJ&lpg=PP1&lr&pg=PP1#v=onepage&q&f=false.
[25] R. Chairunisa, Adiwijaya, and W. Astuti, "Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray," Jurnal Resti, , vol. 4, no. 5, pp. 805–812, Oct. 2020, doi: 10.29207/resti.v4i5.2083.
[26] F. Parsakh Nursyamsyi and F. Noor Hasan, "KLIK: Kajian Ilmiah Informatika dan Komputer Klasifikasi Sentimen Terhadap Aplikasi Identitas Kependudukan Digital Menggunakan Algoritma Naïve Bayes dan SVM," Media Online, vol. 4, no. 3, pp. 1788–1798, Dec. 2023, doi: 10.30865/klik.v4i3.1517.
[27] Muslih, A. Hilda Meutia, and M. Elly Jafar, "PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika Metode Klasifikasi Support Vector Machine (SVM) Untuk Analisis Sentimen Aplikasi Bing: Chat with AI & GPT-4 Di Google Play Store," PETIR J. Pengkaj. dan Penerapan Tek. Inform., vol. 17, pp. 68–76, Jun. 2024, doi: 10.33322/petir.v17i1.2283.
[28] A. P. Kirana, F. Dimas, and N. H. Firman, "Implementation of Data Mining to Predict Student Study Period with Decision Tree Algorithm (C4.5)," J. Sisfokom (Sistem Inf. dan Komputer), vol. 13, no. 1, pp. 31–39, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1943.
[29] S. Ramadani and N. H. Firman, "Analisis Sentimen Terhadap Program Makan Siang & Susu Gratis Menggunakan Algoritma Naive Bayes," J. Teknol. Dan Sist. Inf. Bisnis, vol. 6, no. 3, pp. 411–419, Jul. 2024, doi:

Journal of Information System Research (JOSH)
Volume 6, No. 3, April 2025, pp 1757−1766
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i3.7167

10.47233/jteksis.v6i3.1378.

[30] S. Rampogu, "A review on the use of machine learning techniques in monkeypox disease prediction," Elsevier B.V, Sep. 23, 2023, doi: 10.1016/j.soh.2023.100040.

[31] Hozairi, Anwari, and S. Alim, "Implementasi orange data mining untuk klasifikasi kelulusan mahasiswa dengan model K-Nearest Neighbor, Decision Tree serta Naive Bayes,"Jurnal Ilmiah Nero, Vol 6, No 2, 2021, doi: 10.21107/nero.v6i2.237.