Journal of Science and Social Research ISSN 2615 – 4307 (Print)
May 2025, VIII (2): 1435 – 1440 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PERBANDINGAN ALGORITMA LOGISTIC REGRESSION DAN NAÏVE BAYES CLASSIFIER DALAM IDENTIFIKASI PENYAKIT LIVER

Yuni Handayani[1], Taufik Hidayat[2], Dian Novitaningrum[3], Abdul Rahman Ismail[4]
[1,2,3]Universitas Selamat Sri, Jawah Tengah
[4]Universitas Negeri Gorontalo, Gorontalo
e-mail: [1]yuni0406handayani@gmail.com,
[2]taufikhidayat.jc@gmail.com,[3]dnovitaningrum.uniss@gmail.com,
[4]abdulrahmanismail123@gmail.com

Abstract: *Liver disease is a condition caused by various factors that can damage liver function, such as viral infections and alcohol consumption. Additionally, obesity is closely associated with liver damage. Over time, liver damage can lead to serious consequences. The presence of experts in this field is crucial to addressing liver disease by identifying the symptoms experienced by patients, determining the type of liver disease affecting them, and providing appropriate treatment guidance. The severity of this disease in Indonesia is evident from various studies, research, and related observations. In this study, researchers utilized and compared two data mining classification methods, namely Logistic Regression and Naïve Bayes, to diagnose liver disease. The findings revealed that the Logistic Regression method achieved an accuracy rate of 84.62% with an area under the curve (AUC) value of 0.841, while the Naïve Bayes method achieved an accuracy rate of 83.71% with an AUC value of 0.816. Based on the t-test, it was found that there was no significant difference between the two methods, with a p-value of 0.821 > 0.05. This indicates that the performance of Logistic Regression is comparable to Naïve Bayes in diagnosing liver disease.*

Keywords: *Liver Disease, Logistic Regression, Naïve Bayes, Confusion Matrix, ROC Curve*

Abstrak: Penyakit hati atau liver adalah kondisi yang disebabkan oleh berbagai faktor yang dapat merusak fungsi hati, seperti infeksi virus dan konsumsi alkohol. Selain itu, obesitas juga memiliki kaitan erat dengan kerusakan hati. Dalam jangka panjang, kerusakan hati dapat menimbulkan konsekuensi serius. Kehadiran ahli di bidang ini sangat diperlukan untuk membantu menangani masalah penyakit hati dengan mengidentifikasi gejala yang dialami pasien, menentukan jenis penyakit hati yang diderita, serta memberikan panduan penanganan yang sesuai. Skala permasalahan penyakit ini di Indonesia dapat diamati melalui berbagai studi, penelitian, dan pengamatan yang telah dilakukan. Dalam penelitian ini, peneliti menerapkan serta membandingkan dua metode klasifikasi data mining, yaitu Logistic Regression dan Naïve Bayes, untuk mendeteksi penyakit liver. Hasil penelitian menunjukkan bahwa Logistic Regression memiliki tingkat akurasi sebesar 84,62% dengan nilai area under the curve (AUC) sebesar 0,841, sementara Naïve Bayes mencapai akurasi 83,71% dengan AUC sebesar 0,816. Berdasarkan hasil uji-t, tidak ditemukan perbedaan signifikan antara kedua metode tersebut, dengan nilai p = 0,821 yang lebih besar dari 0,05. Ini menunjukkan bahwa performa Logistic Regression sebanding dengan Naïve Bayes dalam proses diagnosis penyakit liver.

Kata kunci: Penyakit Liver, Logistic Regression, Naïve Bayes, Confusion Matrix, ROC Curve

PENDAHULUAN

Penyakit liver jenis gangguan kesehatan serius yang menjadi penyebab kematian signifikan di berbagai belahan dunia. Berdasarkan data dari Organisasi Kesehatan Dunia (WHO), penyakit ini termasuk dalam kategori penyakit degeneratif yang sering kali sulit terdeteksi pada tahap awal karena gejala yang muncul umumnya tidak spesifik atau terlihat sangat ringan (Gobel, 2018). Keterlambatan dalam diagnosis dapat menyebabkan kondisi semakin memburuk, berujung pada komplikasi seperti sirosis atau bahkan kanker hati.

Di Indonesia, hepatitis B dan C merupakan penyebab utama penyakit liver kronis. Berdasarkan data Riset Kesehatan Dasar (Riskesdas) serta pemeriksaan darah donor Palang Merah Indonesia (PMI), sekitar 10% populasi pernah mengalami infeksi hepatitis B atau C, dengan 28 juta orang hidup dengan infeksi kronis dan sekitar 1,4 juta orang berisiko terkena kanker hati (Kementerian Kesehatan RI, 2020). Kondisi ini tidak hanya berdampak pada kesehatan individu tetapi juga berpengaruh terhadap produktivitas masyarakat dan aspek sosial ekonomi secara luas. Oleh karena itu, deteksi dini dan diagnosis yang akurat sangat diperlukan untuk mengurangi angka kematian akibat penyakit liver.

Seiring dengan perkembangan teknologi, metode diagnostik berbasis data semakin berkembang dan mulai diterapkan dalam dunia medis. Salah satu pendekatan yang menjanjikan adalah penggunaan algoritma machine learning (Karo Karo & Hendriyana, 2022) dalam identifikasi penyakit liver. Algoritma machine learning memungkinkan analisis data klinis secara cepat dan akurat untuk membantu proses pengambilan keputusan medis. Dalam penelitian ini, dua algoritma klasifikasi yang akan dibandingkan adalah Logistic Regression dan Naïve Bayes Classifier.

Logistic Regression merupakan metode statistik yang banyak digunakan untuk analisis data biner, memodelkan hubungan antara variabel independen dan probabilitas suatu kejadian (A'yunan et al., 2023). Metode ini telah lama digunakan dalam bidang medis untuk diagnosis penyakit. Sementara itu, Naïve Bayes Classifier adalah algoritma berbasis probabilitas yang mengandalkan teorema Bayes dengan asumsi independensi antar fitur (Suryadi & Haris, 2015). Meskipun sederhana, Naïve Bayes sering menunjukkan performa yang baik dalam tugas klasifikasi medis.

Beberapa penelitian sebelumnya telah mengkaji berbagai algoritma klasifikasi dalam diagnosis penyakit liver. Penelitian oleh Rahmawati et al (Rahmawati et al., 2018) menunjukkan bahwa metode Decision Tree memiliki akurasi lebih tinggi dibandingkan Logistic Regression dalam mengklasifikasikan penyakit hepatitis. Penelitian lain oleh Nivaan et al (Nivaan & Emanuel, 2020) menunjukkan bahwa Logistic Regression cukup efektif dalam memprediksi penyakit hepatitis dengan akurasi 83,33%. Sementara itu, penelitian oleh Amrin membandingkan beberapa metode klasifikasi dan menemukan bahwa algoritma C4.5 memiliki performa terbaik dibandingkan k-Nearest Neighbor dan Naïve Bayes.

Para peneliti juga melihat penelitian lain yang telah meneliti cara untuk memprediksi penyakit hati di masa lalu sebagai referensi. Desi Rahmawati melakukan penelitian yang relevan. (Rahmawati et al., 2018) et al dengan judul Pemilihan Metode Klasifikasi Terbaik antara Regresi Logistik dan Decision Tree pada Dataset Hepatitis. Penelitian ini membahas tentang penerapan teknik klasifikasi seperti Regresi Logistik dan Decision Tree untuk mengklasifikasikan penyakit hepatitis. Hasil penelitian menunjukkan bahwa perbandingan menggunakan metode klasifikasi Regresi Logistik memperoleh nilai akurasi sebesar 80,207% sedangkan menggunakan algoritma Decision Tree sebesar 83,195%, dengan hasil tersebut

dapat disimpulkan bahwa hasil perbandingan metode terbaik adalah Decision Tree.

Penelitian lain oleh Nivaan et all (Nivaan & Emanuel, 2020) berjudul Analytic Predictive of Hepatitis Using The Regression Logic Algorithm menggunakan Regresi Logistik untuk memprediksi penyakit hepatitis dengan dataset dari UCI Machine Learning Repository. Hasil penelitian menunjukkan bahwa Regresi Logistik cukup baik sebagai metode prediksi, dengan akurasi sebesar 83,33%.

Penelitian lain oleh Amrin (Pahlevi & Amrin, 2020) antara lain pada tahun 2020 dengan judul Model Penambangan Data untuk Pengembangan Aplikasi Diagnostik Untuk mendiagnosis penyakit hati inflamasi, Penyakit Hati Inflamasi membandingkan berbagai metode klasifikasi penambangan data, seperti algoritma C4.5, Naïve Bayes, dan k-Nearest Neighbour. Metode C4.5 mengungguli yang lain dalam hal akurasi (70,99% dengan AUC 0,950), k-Nearest Neighbour (67,19 persen dengan AUC 0,873), dan Naïve Bayes (66,14 persen dengan AUC 0,742), menurut evaluasi kinerja yang dilakukan menggunakan metode Validasi Silang, Matriks, dan Kurva ROC.

Selanjutnya, penelitian oleh Wahyugi Fadri (Rafsanjani, 2018) Sistem ini layak diaplikasikan, dibuktikan dengan tingkat akurasi sebesar 60% yang diperoleh dari 45 data latih dan 5 data uji dalam pengkategorian penyakit liver dengan pendekatan Naïve Bayes.

Meskipun kedua algoritma, Logistic Regression dan Naïve Bayes, telah diterapkan dalam berbagai penelitian medis, perbandingan kinerjanya dalam diagnosis penyakit liver masih belum banyak dikaji secara mendalam. Dengan demikian, penelitian ini bertujuan untuk menganalisis serta membandingkan tingkat akurasi, presisi, dan recall dari kedua algoritma dalam mendeteksi penyakit liver berdasarkan dataset klinis yang tersedia. Hasil penelitian ini diharapkan dapat memberikan rekomendasi algoritma yang lebih efektif dan efisien dalam sistem diagnosis berbasis machine learning, sehingga dapat diterapkan dalam praktik medis guna meningkatkan kualitas layanan kesehatan. Penelitian ini akan memanfaatkan dataset yang berisi informasi klinis terkait penyakit liver, seperti hasil tes laboratorium dan pemeriksaan fisik, untuk menganalisis dan membandingkan akurasi, presisi, dan recall dari masing-masing algoritma. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam pengembangan sistem diagnosis berbasis machine learning yang lebih baik dan dapat diterapkan dalam praktik medis untuk meningkatkan kualitas layanan kesehatan. Sebagai solusi dari permasalahan tersebut, maka penulis melakukan penelitian dengan judul "Perbandingan Algoritma Logistic Regression dan Naïve Bayes Classifier dalam Identifikasi Penyakit Liver" untuk memberikan rekomendasi algoritma yang lebih efektif dan efisien dalam identifikasi penyakit liver untuk dapat diterapkan dalam sistem diagnosis berbasis machine learning.

METODE

Gambar 1 menunjukkan kerangka kerja penelitian ini, yang terdiri dari beberapa tahap. Keterbatasan utama penelitian ini adalah tidak adanya algoritma definitif untuk diagnosis penyakit hati. Jadi, untuk mengatasi masalah ini, peneliti merancang model menggunakan algoritma Regresi Logistik dan Naïve Bayes, lalu peneliti menguji seberapa baik kedua metode tersebut bekerja. Pengujian berulang dilakukan dengan menggunakan pendekatan Kurva ROC, Validasi Silang, dan Matriks. Model yang dikembangkan diterapkan

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
May 2025, VIII (2): 1435 – 1440 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

menggunakan Rapid Miner selama tahap pengembangan aplikasi.

Gambar 1. Alur penelitian

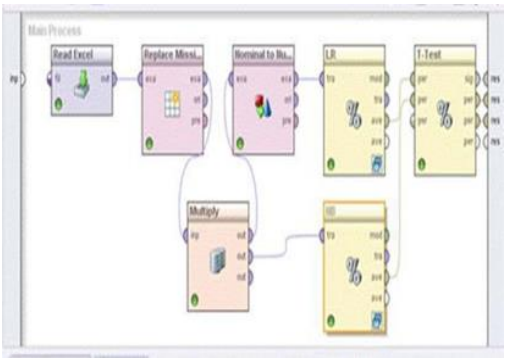
HASIL DAN PEMBAHASAN

Data pasien penyakit hati dari situs web UCI Machine Learning Repository digunakan sebagai kumpulan data dalam studi ini. Ada 310 catatan dalam kumpulan data, yang mewakili 246 kasus penyakit liver yang aktif dan 64 kematian. Sebanyak 19 atribut masukan yang menggambarkan gejala dan 1 atribut keluaran yang membuat keputusan atau kelas membentuk kumpulan data tersebut. Nilai 2 menunjukkan "hidup" dan nilai 1 menunjukkan "mati" dalam properti keluaran.

Tabel 1 di bawah ini menjelaskan atribut kumpulan data:

Atribut	Nilai Atribut	Keterangan
Age	Angka Numerik	Umur Pasien
Sex	Male, Female	Jenis Kelamin Pasien
Steroid	No, Yes	Apakah mendapatkan terapi steroid?
Antivirals	No, Yes	Apakah mendapatkan terapi antiviral?
Fatigue	No, Yes	Apakah mengalami symptoms atau gejala kelelahan akut?
Malaise	No, Yes	Apakah mengalami symptoms atau gejala malaise (rasa tidak nyaman)?
Anorexia	No, Yes	Apakah mengalami symptoms atau gejala anorexia (muntah setiap makan)?
Liver_Big	No, Yes	Apakah kondisi hati/liver membesar?
Liver_Firm	No, Yes	Apakah kondisi hati/liver mengeras?
Spleen_Palpable	No, Yes	Apakah ada gejala spleen palpable / limfa lebih jelas / besar dari normal?
Spiders	No, Yes	Apakah ada gejala spider/ pembuluh darah upnormal pada kulit (pembuluh darah mengumpul dan menonjol pada permukaan kulit)?
Ascites	No, Yes	Terjadi penumpukan cairan pada rongga perut?
Varices	No, Yes	Terjadi pembekakan vena esophagus (varises)?
Bilirubin	No, Yes	Nilai kadar bilirubin dalam darah
Alk_Phospate	No, Yes	Kadar Alkalin Phospate dalam liver
Sgot	No, Yes	Nilai sgot
Albumin	No, Yes	Kadar Albumin
Protime	No, Yes	Uji Masa protrombhine
Histology	No, Yes	Apakah dilakukan pemeriksaan dengan histology (biopsy hati)?
Class	No, Yes	Class apakah pasien positif liver atau tidak?

Model yang dibuat diuji dalam penelitian ini melalui eksperimen. Selain itu, peneliti memvalidasi dan menilai model untuk menemukan akurasi dan nilai AUC-nya. Untuk menemukan akurasi dan nilai AUC dari setiap algoritma yang diuji, Rapid Miner digunakan bersama dengan operator validasi silang 10 kali lipat. Pada Gambar 2 di bawah ini dijelaskan model sebegai berikut :



Gambar 2. Pemodelan yang diusulkan Area di Bawah Kurva (AUC) dan Kurva Karakteristik Operasi Penerima (ROC) (Matriks Kebingungan) digunakan untuk evaluasi.

Confusion Matrix

Algoritma Logistic Regression

Tabel 2 menunjukkan matriks konfusi untuk metode regresi logistik. Sebanyak 226 data diklasifikasikan sebagai "hidup (2)" dan sesuai dengan data aktual. Sebanyak 20 data diprediksi sebagai "mati (1)", tetapi sebenarnya diklasifikasikan sebagai "hidup". Selanjutnya, 36 data diklasifikasikan sebagai "mati (1)" sesuai dengan data aktual, sedangkan 28 data diprediksi sebagai "hidup (2)", tetapi sebenarnya diklasifikasikan sebagai "mati(1)".

Tabel 2. Model Matriks Confusion untuk Regresi Logistik

Akurasi : 84.62% +/- 7.57% (mikro: 84.52%)		
	True 2.0	True 1.0
pred 2.0	226	28
Pred 1.0	20	36
recall	91.87%	56.25%

Algoritma Naïve Bayes

Tabel 3 menyajikan matriks konfusi untuk metode Naïve Bayes. Sebanyak 218 data diklasifikasikan sebagai “hidup (2)” dan sesuai dengan data aktual. Sebanyak 28 data diprediksi sebagai “mati (1)”, tetapi sebenarnya diklasifikasikan sebagai “hidup (2)”. Selain itu, 42 data diklasifikasikan sebagai “mati (1)” sesuai dengan data aktual, sedangkan 22 data

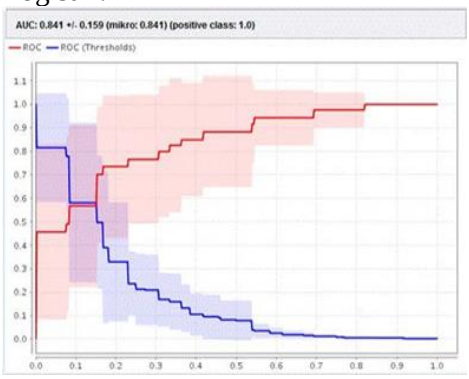
diprediksi sebagai "hidup (2)", tetapi sebenarnya diklasifikasikan sebagai "mati (1)".

Tabel 3. Model Matriks Konfusi untuk Algoritma Naïve Bayes

Akurasi : 83.71% +/- 9.95% (mikro: 83.87%)		
	True 2.0	True 1.0
pred 2.0	218	22
Pred 1.0	28	42
recall	88.62%	65.62%

Kurva Receiver Operating Characteristic Metode Logistic Regression

Di bawah pada Gambar 3 Anda dapat melihat kurva Karakteristik Operasi Penerima (ROC) metode Regresi Logistik.



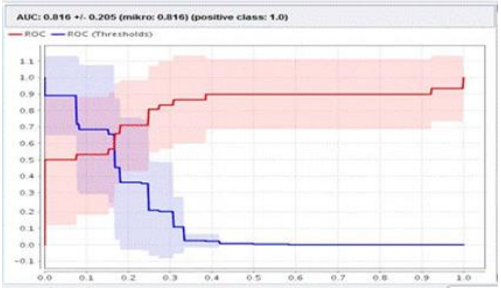
Gambar 3. Kurva Receiver Operating Characteristic

Metode Logistic Regression

Temuan matriks ditunjukkan pada Gambar 3, yang merupakan kurva Karakteristik Operasional Penerima (ROC). Positif palsu ditunjukkan pada sumbu horizontal dan positif nyata pada sumbu vertikal.

Metode Naïve Bayes

Gambar 4, yang ada di bawah, menampilkan kurva Karakteristik Operasi Penerima (ROC) untuk algoritma Naïve Bayes.



Gambar 4. Kurva Receiver Operating Characteristic

Metode Naïve Bayes

Analisis Hasil Komparasi

Di bawah ini, Anda dapat menemukan perbandingan akurasi algoritma Naïve Bayes dan Regresi Logistik serta nilai AUC pada Tabel 4.

Metode	Accuracy	AUC
Logistic Regression	84.62%	0.841
Naïve Bayes	83.71%	0.816

Tabel 4 membandingkan nilai akurasi dan AUC masing-masing metode. Secara keseluruhan dapat dilihat bahwa metode Logistic Regression memiliki nilai akurasi yang sedikit lebih tinggi dibandingkan Naïve Bayes, begitu pula dengan nilai AUC-nya. Untuk memastikan kebenaran perbedaan tersebut, maka dilakukan evaluasi dengan menggunakan metode statistika tradisional yang umum digunakan, yaitu t-Test. Hasil kinerja dari t-Test ditunjukkan pada Gambar 5. Berdasarkan hasil t-Test diketahui bahwa tidak terdapat perbedaan yang signifikan antara Logistic Regression dengan Naïve Bayes, karena nilai $\alpha = 0{,}821 > 0{,}05$. Hal ini menunjukkan bahwa metode Logistic Regression memiliki kinerja yang setara dengan metode Naïve Bayes.

Gambar 4. Kurva Receiver Operating Characteristic Metode Naïve Bayes

T-Test Significance

	0.846+/-0.076	0.837+/- 0.099
0.846+/- 0.076		0.821
0.837 +/- 0.099		

Probabilitas untuk nilai acak dengan hasil yang sama menunjukkan adanya perbedaan signifikan antara rata-rata aktual jika nilainya lebih kecil dari alpha = 0,050. Nilai tersebut menandakan kemungkinan adanya perbedaan yang signifikan secara statistik.

Kelompok nilai AUC berikut dapat diidentifikasi dalam klasifikasi penambangan data, menurut Gorunescu (2011) yang dikutip dalam Amrin et al. (2021):

a. 0,90-1,00 = klasifikasi sangat baik
b. 0,80-0,90 = klasifikasi baik
c. 0,70-0,80 = klasifikasi cukup

d. 0,60-0,70 = klasifikasi kurang
e. 0,50-0,60 = klasifikasi keliru

Metode Regresi Logistik dan Naïve Bayes termasuk dalam klasifikasi baik, menurut kriteria ini.

SIMPULAN

Kesimpulan yang dapat diambil dari penelitian ini adalah metode Regresi Logistik untuk memprediksi penyakit liver memiliki tingkat akurasi sebesar 84,62% dengan nilai area under the curve (AUC) sebesar 0,841. Sementara itu, metode Naïve Bayes menunjukkan tingkat akurasi sebesar 83,71% dengan nilai AUC sebesar 0,816.

Hasil uji t menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara metode Regresi Logistik dan Naïve Bayes, dengan nilai $\alpha = 0,821 > 0,05$. Hal ini menunjukkan bahwa kedua metode tersebut memiliki kinerja yang setara dalam memprediksi penyakit liver.

Untuk pengembangan penelitian selanjutnya, model dapat disempurnakan dengan menambahkan algoritma optimasi pada metode Regresi Logistik dan Naïve Bayes. Selain itu, penelitian selanjutnya juga dapat dilengkapi dengan informasi tentang penyakit liver yang lebih rinci dan mendetail.

DAFTAR PUSTAKA

C. Y. Gobel, "Sistem Pakar Penyakit Liver Menggunakan K- Nearest Neighbors Algoritm Berbasis Website," Ilk. J. Ilm., vol. 10, no. 2, pp. 152–159, 2018, doi: 10.33096/ilkom.v10i2.296.152-159.

Kementerian Kesehatan RI, Rencana Aksi Nasional Pencegahan dan Pengendalian Hepatitis. 2020. [Online]. Available: https://www.globalhep.org/sites/default/files/content/action_plan_article/files/2022-05/RAN HEP 2020-2024 KDT_0.pdf

I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," J. Teknol. Terpadu, vol. 8, no. 2, pp. 94–99, 2022, doi: 10.54914/jtt.v8i2.564.

Y. A. D. K. A'yunan, U. Indahyanti, and S. Busono, "Implementasi Data Mining dalam Klasifikasi Diagnosa Kanker Payudara menggunakan Algoritma Logistic Regression," J. TEKINKOM, vol. 6, no. 2, pp. 400–407, 2023, doi: 10.37600/tekinkom.v6i2.948.

U. T. Suryadi and R. Haris, "Implementasi Metode Naive Bayes Untuk Mendiagnosa Penyakit Pasien," J. Teknol. Inf. dan Komun. STMIK Subang, no. April, pp. 100–110, 2015.

D. Rahmawati, A. E. Maruruk, C. Bintang, and G. Allo, "PEMILIHAN METODE KLASIFIKASI TERBAIK ANTARA LOGISTIC REGRESSION DAN DECISION TREE PADA DATASET HEPATITIS," pp. 56–61, 2018.

G. V. Nivaan and A. W. R. Emanuel, "Analytic Predictive of Hepatitis using the Regression Logic Algorithm," 2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020, pp. 106–110, 2020, doi: 10.1109/ISRITI51436.2020.9315365.

O. Pahlevi and A. Amrin, "Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease," SinkrOn, vol. 5, no. 1, p. 51, 2020, doi: 10.33395/sinkron.v5i1.10589.

R. G. et al Rafsanjani, "Diagnosis Penyakit Hati Menggunakan Metode Naive Bayes Dan Certainty Factor," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 11, pp. 4478–4482, 2018.