ROUTERS: Jurnal Sistem dan Teknologi Informasi
Vol. 1, No. 1, Februari 2023, pp. 1-8
ISSN: 2962-1224, DOI: https://doi.org/10.20222/rt.v1i1.2665

Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN)

Zuriati Z[1], Qomariyah N[2]
[1,2] Jurusan Ekonomi dan Bisnis, Program Studi Manajemen Informatika, Politeknik Negeri Lampung, Kota Bandar Lampung, Lampung, 35141

INFORMASI ARTIKEL

Diterima : 20022/09/12
Direvisi : 2022/10/28
Diterbitkan : 2022/11/02

Kata kunci:

Algoritma;
KNN;
Penyakit;
Stroke

ABSTRAK

Penerapan algoritma klasifikasi merupakan salah satu solusi yang mampu mengklasifikasikan gejala penyakit stroke. Pengklasifikasian gejala ini dalam bentuk model prediksi dapat digunakan sebagai upaya deteksi dini penyakit stroke. Algoritma yang diterapkan untuk membangun model prediksi adalah K-Nearest Neighbor (KNN). Algoritma KNN terbukti mampu memprediksi sampel uji yang baru berdasarkan jarak *euclidean*. Dataset terdiri dari 5110 record, attribut yang digunakan adalah: *gender, age, hypertension, heart_disease, ever_married, bmi, work_type, residence_type, Avg_glucosa_level, smoking_status*, kelompok stroke. Tahap penelitian adalah: Pengumpulan Data, *Preprocessing* Data, *Split* Data, Penerapan Algoritma KNN dan Evaluasi kinerja KNN dengan *confusion matrix* dan penghitungan akurasi. Performa algoritma KNN terbaik didapatkan dengan nilai k=5 dan akurasi 93.54%.

Classification of Stroke Using the K-Nearest Neighbor (KNN) Algorithm

ARTICLE INFO

Received
Revised
Published

Keyword:

Algoritma;
KNN;
Desease;
Stroke

ABSTRACT

The application of the classification algorithm is one solution that is able to classify the symptoms of stroke. This symptom classification in the form of a predictive model can be used as an effort to detect stroke early. The algorithm applied to build the prediction model is K-Nearest Neighbor (KNN). The KNN algorithm is proven to be able to predict the new test sample based on the Euclidean distance. The dataset consists of 5110 records, the attributes used are: gender, age, hypertension, heart_disease, ever_married, bmi, work_type, residence_type, Avg_glucosa_level, smoking_status, stroke group. The research stages are: Data Collection, Data Preprocessing, Data Split, Application of the KNN Algorithm and Evaluation of KNN performance with confusion matrix and calculation of accuracy. The best KNN algorithm performance is obtained with a value of k = 5 and an accuracy of 93.54%.

Corresponding Author:
Zuriati, Politeknik Negeri Lampung
Email: zuriati_mi@polinela.ac.id

1. PENDAHULUAN

World Health Organization [1] mendefinisikan stroke sebagai tanda klinis yang terjadi akibat adanya gangguan fungsi otak lokal atau global, dengan gejala yang berlangsung selama 24 jam atau lebih, dapat menyebabkan kematian, tanpa adanya penyebab lain selain vaskuler. Penyakit stroke adalah penyebab kecacatan nomor satu dan penyebab kematian nomor tiga di dunia setelah penyakit jantung dan kanker. Beberapa faktor penyebab penyakit stroke adalah: tekanan darah, riwayat fibrilasi atrium, kolesterol, diabetes, dan lain sebagainya. Penyakit stroke terjadi karena terputusnya suplai darah menuju otak, dapat terjadi karena adanya sumbatan berupa darah yang menggumpal atau terdapat semburan pada pembuluh darah.

Penanganan penyakit stroke dilakukan melalui pemeriksaan oleh dokter spesialis penyakit syaraf. Dokter melakukan diagnosis pada pasien dengan cara mengajukan pertanyaan untuk menggali keluhan yang dirasakan oleh pasien serta faktor-faktor yang dapat memicu terjadinya stroke, sehingga akan didapatkan suatu kesimpulan tingkat risiko penyakit stroke pada pasien. Dengan deteksi dini dan pengendalian faktor risiko penyakit stroke dapat dicegah [2]. Kegiatan ini membutuhkan biaya dan waktu karena harus berkunjung ke rumah sakit. Untuk membantu pasien dan masyarakat umum diperlukan suatu model yang dapat memprediksi penyakit stroke agar dapat terhindar dari stroke. Model ini juga dapat digunakan untuk membantu dokter atau tenaga kesehatan dalam melakukan diagnosa penyakit stroke. Pemilihan algoritma yang tepat untuk pengembangan model prediksi penyakit stroke penting karena berpengaruh pada hasil yang didapatkan.

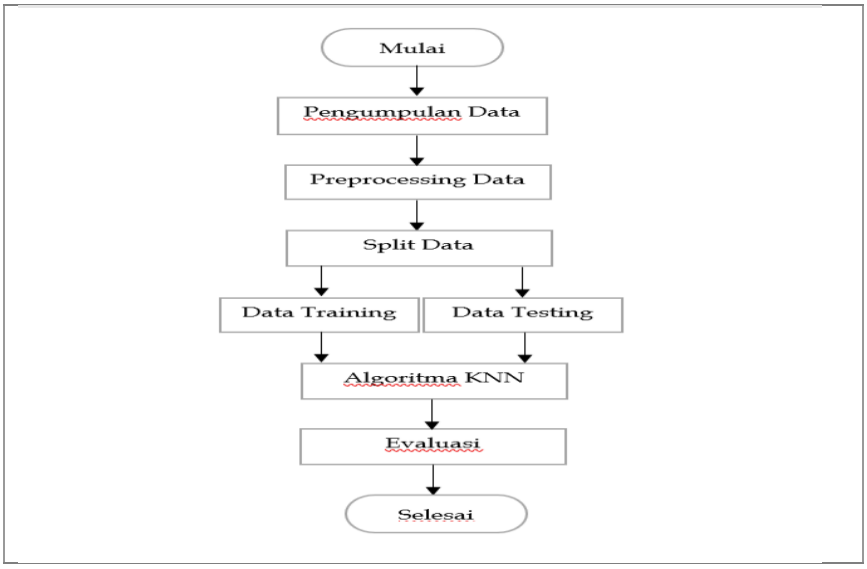
K-Nearest Neighbor (KNN) adalah algoritma yang sederhana tapi kuat dan efektif untuk mengklasifikasikan data [3], [4], [5]. Algoritma KNN melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [6], perhitungan jarak menggunakan persamaan *euclidean distance*. Selanjutnya nilai mayoritas dari KNN dijadikan dasar untuk menentukan kategori dari data berikutnya.

Algoritma KNN telah diterapkan pada berbagai kasus dengan akurasi yang baik. Penelitian terkait menggunakan algoritma KNN untuk data penyakit stroke telah dilakukan oleh [7] dengan akurasi 100% dengan jumlah data latih 60, data uji 30 dan nilai k=20-30, sedangkan [8] mendapatkan nilai akurasi 68.30%. Penerapan algoritma KNN untuk klasifikasi kematangan buah tomat dengan akurasi 92.5% dan nilai k=3 [9], untuk prediksi hasil studi siswa nilai akurasi yang didapatkan 72.28 % untuk data yang tidak di-*preprocessing* terlebih dahulu, dan 98.42 % untuk data yang telah di-*preprocessing* [10]. Penerapan KNN untuk klasifikasi beasiswa dengan akurasi 85.56% [11], penentuan kualitas restoran [12], pengelolaan data layanan medis dan perawatan kesehatan [4], klasifikasi kebisingan suara di perkotaan [13], klasifikasi kredit perbankan [14] dan klasifikasi bantuan untuk pedesaan [15].

Tujuan penelitian adalah untuk menerapkan algoritma KNN untuk klasifikasi data penyakit stroke. Percobaan dilakukan dengan melakukan 3 skenario split data training dan testing dengan perbandingan: 90%:10%, 80%:20% , 70%:30%. Selain itu juga melakukan percobaan untuk nilai k=3 dan k=5 untuk menemukan akurasi terbaik. Nilai k menyatakan berapa banyak jumlah neighbor atau data yang terdekat dengan suatu objek. Bahwa dengan uji coba nilai k yang berbeda-beda, menghasilkan tingkat akurasi yang berbeda pula [16]. Performa algoritma KNN terbaik didapatkan pada percobaan dengan pembagian data training dan data testing 90% : 10% dengan nilai k= 5 dan akurasi 93.54%.

2. METODE

1. Tahapan penelitian ini ditampilkan pada Gambar 1 berikut ini:



Gambar 1. Tahapan penelitian

2.1. Pengumpulan Data

Penelitian ini menggunakan dataset publik yang tersedia pada *kaggel dataset repository*, yaitu dataset *stroke prediction*. Dataset tersebut memiliki 5110 record data dan terdiri dari 11 attribut. 10 atribut digunakan sebagai Fitur dan 1 attribut untuk Label.

2.2. *Preprocessing* Data

Tujuan *preprocessing* data adalah untuk membersihkan dan melengkapi data agar dapat digunakan untuk membangun model prediksi yang diinginkan.

2.3. Split Data

Tujuan dari proses split data adalah untuk mengelompokkan data, sehingga didapatkan kelompok data untuk data training dan data testing.

2.4. Algoritma KNN

Algoritma KNN digunakan untuk klasifikasi data stroke. Algoritma KNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Jarak antara data testing dengan data training dihitung dengan cara mengukur jarak antara titik yang merepresentasikan data testing dengan semua titik yang merepresentasikan data training dengan rumus *Euclidean Distance*. Persamaan atau rumus *Euclidean Distance* adalah (1):

$$dist = \sum_{i=1}^{p} \sqrt{(x2 - x1)^2} \quad (1)$$

Keterangan :
dist = Jarak
x1 = Data *Training*
x2 = Data *testing*
i = Variable Data

p = Jumlah Atribut

Semakin besar nilai *dist* maka semakin jauh tingkat keserupaan antara kedua individu dan sebaliknya jika nilai *dist* semakin kecil maka akan semakin dekat tingkat keserupaan antar individu tersebut. Nilai k yang terbaik untuk algoritma ini tergantung pada data. Secara umum, nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi semakin kabur. Penelitian ini mencoba nilai k = 3 dan k= 5.

2.5. Evaluasi Algoritma KNN

Evaluasi algoritma KNN menggunakan *confusion matrix*, dengan tujuan memetakan kinerja algoritma dalam bentuk tabulasi. *Confusion matrix* menunjukkan hubungan antara benar tidaknya sebuah data dikategorikan. *Confusion matrix* terdiri dari True positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). True positive merepresentasikan data yang berada pada kelas positif yang diprediksi secara benar oleh algoritma. False Positive merepresentasikan data yang seharusnya berada pada kelas positif diprediksi menjadi kelas negatif oleh algoritma. False Negative merupakan data yang seharusnya berada di kelas negatif diprediksi menjadi kelas positif oleh algoritma. True Negative merupakan data yang berada pada kelas negatif dan diprediksi secara benar oleh algoritma. *Confusion matrix* dapat dilihat pada Tabel 1.

Tabel 1. *Confusion matrix*

	Label	
	Positif	Negatif
Positif	True Positif	False Positif
Negatif	False Negatif	True Negatif

Berdasarkan *confusion matrix*, dapat diketahui berbagai parameter pengukuran kinerja algoritma, misalnya akurasi. Akurasi merupakan perhitungan yang umum digunakan untuk mengevaluasi kinerja dari sebuah algoritma. Akurasi dihitung berdasarkan rasio jumlah data yang diprediksi secara benar oleh algoritma dengan jumlah semua data yang ada pada dataset menggunakan persamaan (2) berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada Tabel 2 disajikan attribut dan tipe data penyakit stroke yang terdiri dari 12 attribut. Attribut stroke dijadikan Label untuk kelompok hasil prediksi.

Tabel 2. Deskripsi attribut data

No.	**Attribut**	**Tipe Data**
1.	Gender	Polinomial
2.	Age	Integer
3.	Hypertension	Binomial
4.	Heart_disease	Binomial
5.	Ever_married	Binomial
6.	Work_type	Polinomial
7.	Residence_type	Binomial
8.	Avg_glucose_level	Real
9.	bmi	Real
10.	Smoking_status	Polinomial

11	Stroke	Integer Binomial (Label)

3.2 Preprocessing Data

Dataset penyakit stroke memiliki 11 attribut. Terdapat *missing value* pada attribut bmi sejumlah 201 data. Untuk itu data *missing value* diisi dengan rata-rata bmi, hal ini dilakukan untuk meningkatkan akurasi algoritma K-NN.

3.3 Split Data

Untuk pembagian data training dan testing penulis menggunakan 3 skenario yaitu: 90%:10%, 80%:20% , 70%:30%.

3.4 Evaluasi Algoritma K-NN

Percobaan menggunakan nilai k= 3 dan k=5. Berikut adalah hasil yang didapatkan.

3.4.1 Hasil Pengujian Algoritma KNN Untuk Nilai k=3

Berikut adalah akurasi yang diperoleh dari hasil pengujian algoritma KNN dengan menggunakan k=3, dengan membagi data testing dan training: 90%:10%, 80%:20% , 70%:30%.

A. 90% : 10%

Pada Gambar 2 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 3 dan perbandingan data training dan testing 90% : 10%. Akurasi yang didapat adalah 88.45%.

accuracy: 88.45%

	true 1	true 0	class precision
pred. 1	2	36	5.26%
pred. 0	23	450	95.14%
class recall	8.00%	92.59%	

Gambar 2. *Confusion matrix* dan akurasi untuk k= 3 dengan perbandingan data 90% : 10%

B. 80% : 20%

Pada Gambar 3 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 3 dan perbandingan data training dan testing 80% : 20%. Akurasi yang didapat adalah 88.55%.

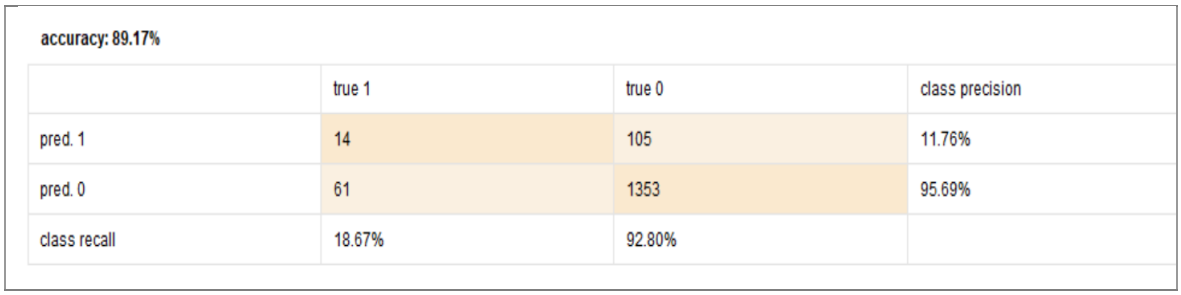
accuracy: 88.55%

	true 1	true 0	class precision
pred. 1	8	75	9.64%
pred. 0	42	897	95.53%
class recall	16.00%	92.28%	

Gambar 3. *Confusion matrix* dan akurasi untuk k= 3 dengan perbandingan data 80% : 20%

C. 70% : 30%

Pada Gambar 4 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 3 dan perbandingan data training dan testing 70% : 30%. Akurasi yang didapat adalah 89.17%.



accuracy: 89.17%

	true 1	true 0	class precision
pred. 1	14	105	11.76%
pred. 0	61	1353	95.69%
class recall	18.67%	92.80%	

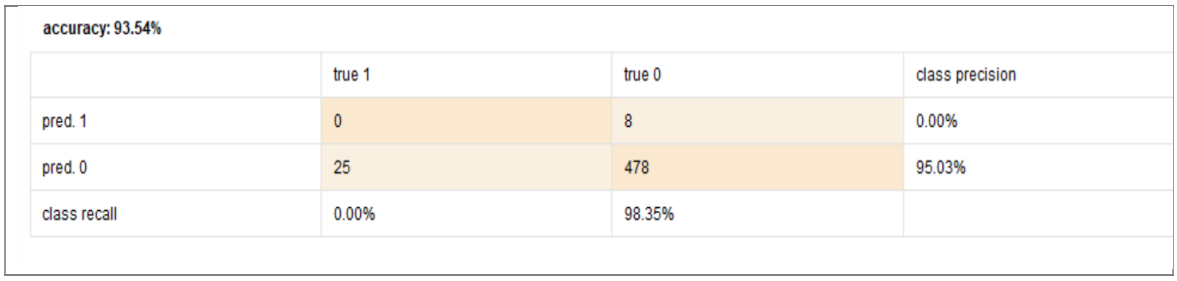
Gambar 4. *Confusion Matrix* dan Akurasi Untuk k= 3 Dengan Perbandingan Data 70% : 30%

3.4.2 Hasil Pengujian Algoritma KNN Untuk Nilai k=5

Berikut adalah *confusion matrix* dan akurasi yang diperoleh dari hasil pengujian algoritma KNN dengan menggunakan k=5, dengan membagi data training dan testing: 90%:10%, 80%:20% , 70%:30%.

A. 90% : 10%

Pada Gambar 5 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 5 dan perbandingan data training dan testing 90% : 10%. Akurasi yang didapat adalah 93.54%.



accuracy: 93.54%

	true 1	true 0	class precision
pred. 1	0	8	0.00%
pred. 0	25	478	95.03%
class recall	0.00%	98.35%	

Gambar 5. *Confusion matrix* dan akurasi untuk k= 5 dengan perbandingan data 90% : 10%

B. 80% : 20%

Pada Gambar 6 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 5 dan perbandingan data training dan testing 80% : 20%. Akurasi yang didapat adalah 93.25%.

accuracy: 93.25%

	true 1	true 0	class precision
pred. 1	3	22	12.00%
pred. 0	47	950	95.29%
class recall	6.00%	97.74%	

Gambar 6. *Confusion matrix* dan akurasi untuk k= 5 dengan perbandingan data 80% : 20%

C. 70% : 30%

Pada Gambar 7 disajikan *confusion matrix* dan akurasi yang dihasilkan oleh algoritma KNN dengan k= 5 dan perbandingan data training dan testing 70% : 30%. Akurasi yang didapat adalah 93.28%.

accuracy: 93.28%

	true 1	true 0	class precision
pred. 1	6	34	15.00%
pred. 0	69	1424	95.38%
class recall	8.00%	97.67%	

Gambar 7. *Confusion matrix* dan akurasi untuk k= 5 dengan perbandingan data 70% : 30%

Pada Tabel 3 disajikan rangkuman nilai akurasi algoritma KNN untuk nilai k=3 dan k=5 dengan komposisi perbandingan data: 90% : 10%, 80% : 20%, dan 70% : 30%.

Tabel 3. Rangkuman nilai akurasi

Nilai k	Persentase perbandingan data training dan data testing		
	90% : 10%	80% : 20%	70% : 30%
k = 3	88.45%	88.55%	89.17%
k = 5	93.54%	93.25%	93.28%

Nilai akurasi tertinggi didapatkan pada nilai k = 5 dengan komposisi perbandingan data training dan testing 90% : 10% dengan nilai akurasi 93.54%.

4. KESIMPULAN

Berdasarkan hasil percobaan pada penelitian ini disimpulkan bahwa klasifikasi menggunakan algoritma KNN untuk dataset penyakit stroke telah berhasil dilakukan, nilai akurasi tertinggi didapatkan pada nilai k = 5 dengan komposisi perbandingan data training dan testing 90% : 10% dengan nilai akurasi 93.54%.

2. DAFTAR PUSTAKA

[1] World Health Organization, "a Vital Investment," *World Health*, p. 202, 2005.
[2] L. Ghani, L. K. Mihardja, and D. Delima, "Faktor Risiko Dominan Penderita Stroke di Indonesia," *Bul. Penelit. Kesehat.*, vol. 44, no. 1, pp. 49–58, 2016.
[3] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, no. August, pp. 986–996, 2003.
[4] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020.
[5] A. K. Tiwari, "Introduction to machine learning," *Ubiquitous Mach. Learn. Its Appl.*, pp. 1–14, 2017.
[6] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.
[7] A. Puspitawuri, E. Santoso, and C. Dewi, "Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, pp. 3319–3324, 2019.
[8] D. Ulfatul, M. Rachmad, H. Oktavianto, and M. Rahman, "Jurnal Smart Teknologi Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke Comparison Of K-Nearest Neighbor And Gaussian Naive Bayes Methods For Stroke Disease Classification Jurnal Smart Teknologi bidang peng," vol. 3, no. 4, pp. 405–412, 2022.
[9] S. Sanjaya, M. L. Pura, S. K. Gusti, F. Yanto, and F. Syafria, "K-Nearest Neighbor for Classification of Tomato Maturity Level Based on Hue, Saturation, and Value Colors," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 2, p. 101, 2019.
[10] S. Sugriyono and M. U. Siregar, "Preprocessing kNN algorithm classification using K-means and distance matrix with students' academic performance dataset," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, 2020.
[11] S. Sumarlin, "Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM," *J. Sist. Inf. Bisnis*, vol. 5, no. 1, pp. 52–62, 2015.
[12] D. Prasad, S. Kumar Goyal, A. Sharma, A. Bindal, and V. Singh Kushwah, "System model for prediction analytics using /(-nearest neighbors algorithm," *J. Comput. Theor. Nanosci.*, vol. 16, no. 10, pp. 4425–4430, 2019.
[13] E. Tsalera, A. Papadakis, and M. Samarakou, "Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm," *Energy Reports*, vol. 6, no. June, pp. 223–230, 2020.
[14] K. Hemachandran, P. M. George, R. V. Rodriguez, R. M. Kulkarni, and S. Roy, "Performance analysis of k-nearest neighbor classification algorithms for bank loan sectors," *Adv. Parallel Comput.*, vol. 38, pp. 9–13, 2021.
[15] S. Arif, P. Aji, H. Oktavianto, and Q. A'yun, "KLASIFIKASI PENERIMA BANTUAN DANA DESA MENGGUNAKAN METODE KNN (K-NEAREST NEIGHBOR) (Studi Kasus : Desa Andongsari Kecamatan Ambulu Kabupaten Jember)," *J. TECHNO Nusa Mandiri*, 2019.
[16] I. A. A. Angreni, S. A. Adisasmita, and M. I. Ramli, "Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," vol. 7, no. 2, pp. 63–70, 2018.