

Klasifikasi Kematian Akibat Gagal Jantung Menggunakan Algoritma *Logistic Regression* Berbasis *Forward Selection*

Helmi Imaduddin¹, Brian Aditya Hermansyah², Muhammad Mutawadhi³ Alfajri³

^{1,2,3} Teknik Informatika, Fakultas Komunikasi dan Informatika, Universitas Muhammadiyah Surakarta, Indonesia

¹helmi.imaduddin@ums.ac.id

²brian.adityaherman@gmail.com

³alfajrimutawadhi@gmail.com

Received: 11-07-2022; Accepted: 27-08-2023; Published: 12-09-2023

Abstrak— *Gagal jantung adalah masalah kesehatan masyarakat utama yang beban penyakitnya meningkat seiring bertambahnya usia. Kondisi jantung dalam kasus ini menandakan bahwa jantung tidak mampu lagi untuk memompa darah secara optimal dan ketidakmampuan jantung dalam memenuhi kuota darah normal yang dibutuhkan oleh tubuh. Berdasarkan timbulnya gejala, gagal jantung dapat terjadi secara tiba-tiba atau lebih dikenal dengan gagal jantung akut, dan gagal jantung yang berkembang secara perlahan karena kondisi jantung yang melemah atau lebih dikenal dengan istilah gagal jantung kronis. Tujuan dari penelitian ini adalah mendapatkan model klasifikasi penyakit gagal jantung untuk membuat sistem penunjang keputusan sebagai deteksi dini penyakit gagal jantung. Setelah itu model yang sudah diperoleh akan dievaluasi untuk mengetahui performanya dengan akurasi, spesifisitas dan sensitivitas. Metode yang digunakan untuk melakukan klasifikasi menggunakan metode Support Vector Machine, Decision Tree, Logistic Regression dan Random Forest. Pengukuran performa klasifikasi menggunakan matrik akurasi, sensitivitas dan spesivisitas, hasil klasifikasi menunjukkan bahwa algoritma logistic regression memiliki performa paling baik dengan memperoleh akurasi sebesar 90% dan spesivisitas 80%.*

Kata kunci— *Gagal jantung, machine learning, forward selection*

Heart failure is a major public health problem whose disease burden increases with age. The heart condition, in this case, indicates that the heart is no longer able to pump blood optimally and the heart's inability to meet the normal blood quota needed by the body. Based on the onset of symptoms, heart failure can occur suddenly, or better known as acute heart failure, and heart failure that develops slowly due to a weakened heart condition, or better known as chronic heart failure. The purpose of this study was to obtain a classification model of heart failure to create a decision support system for the early detection of heart failure. After that, the model that has been obtained will be evaluated to determine its performance with accuracy, specificity, and sensitivity. The method used to classify uses the Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest methods. Measurement of classification performance using accuracy, sensitivity and specificity metrics, the classification results show that the logistic regression algorithm has the best performance by obtaining an accuracy of 90% and specificity of 80%.

Keyword— *Heart failure, machine learning, forward selection*

I. PENDAHULUAN

Jantung merupakan organ berotot yang bekerja sebagai alat pemompa darah serta oksigen ke seluruh organ tubuh. Jantung manusia, memiliki rongga berupa kerucut serta mempunyai dimensi sebesar kepalan tangan [1]. Jantung berada di area rongga bagian dada, tepatnya paru-paru kiri bagian bawah, jantung terlindungi oleh beberapa tulang seperti tulang dada dan tulang rusuk. Jantung memiliki beberapa bagian antara lain *vena kava superior*, *katup trikuspidalis*, *arteri pulmonalis*, *katup aorta*, *aorta*, *ventrikel*, *vena pulmonalis*, *vena kava inferior*, *atrium*, *katup mitral*, *katup atrioventricular* dan dinding jantung. Setiap bagian dalam jantung menjalankan fungsinya masing-masing dalam memompa darah. Darah terdeoksigenasi dari *atrium* kanan dan *ventrikel* bertekanan rendah didorong ke bagian paru-paru. Kemudian darah yang sudah mengandung oksigen memasuki *atrium* kiri dan *ventrikel*, yang mendorong darah ke seluruh tubuh pada tekanan *sistemik*. Kamar dipisahkan oleh *septa interatrial* dan *interventrikular*, dan aliran searah dibentuk oleh katup *atrio-ventrikular* dan *ventrikel-arteri*. Sistem elektrofisiologi intrinsik dengan cepat menyebarkan impuls listrik dari *nodus sinoatrial* ke *nodus atrioventrikular*, dan sepanjang serat *Purkinje* ke puncak di mana kontraksi dimulai [2]. Seperti organ tubuh lainnya, jantung juga dapat rusak atau terkena suatu penyakit yang dapat menyebabkan kematian. Berdasarkan laman resmi *World Health Organization* (WHO), terdapat 17,9 juta orang yang meninggal setiap tahun di seluruh setiap tahunnya dikarenakan penyakit *kardiovaskular* yang menyerang bagian jantung. *Kardiovaskular* juga dapat menyebabkan penyakit gagal jantung.

Penelitian mengenai klasifikasi penyakit gagal jantung sudah pernah dilakukan, diantaranya Dhimas [3] melakukan penelitian dengan menggunakan metode ekstraksi fitur menggunakan *Multifractal Detrended Fluctuation Analysis* (MFDFA) dan algoritma klasifikasi menggunakan ANN, hasil penelitian tersebut memperoleh skor akurasi sebesar 67.24%.

Pada penelitian lain Duwi [4] menggunakan seleksi fitur algoritma genetika untuk membuat klasifikasi *naïve bayes* menjadi lebih optimal, hasilnya penelitian tersebut mendapatkan akurasi sebesar 96,67%, naik sangat signifikan daripada menggunakan *naïve bayes* saja yang hanya mendapatkan akurasi sebesar 69,6%.

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

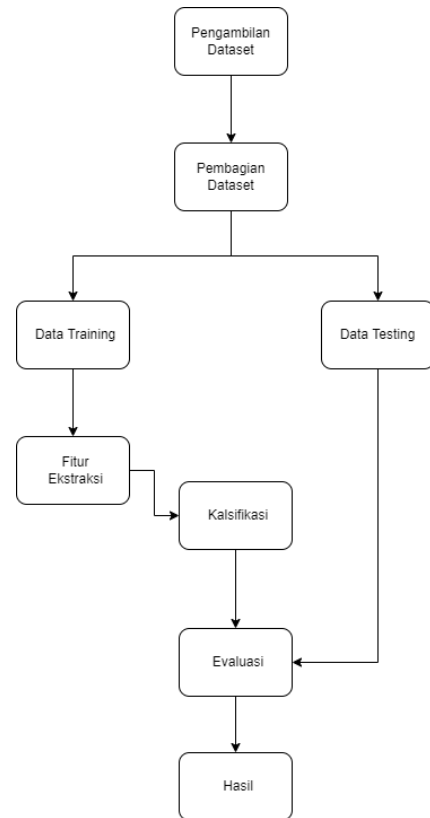
Selain itu penelitian lain yang dilakukan oleh Elin [5] menggunakan dataset yang berasal dari UCI repository, sama seperti penelitian yang sedang dilakukan, Elin menggunakan fitur seleksi *Forward Selection* yang dipadukan dengan algoritma C4.5 untuk mengklasifikasi penyakit gagal jantung, hasilnya akurasi naik dari 77,89% menjadi 84,29%.

Berdasarkan masalah tersebut, penulis melakukan penelitian untuk membuat sistem berbasis *machine learning* yang dapat mengklasifikasikan kematian pada gagal jantung. Algoritma yang digunakan untuk klasifikasi pada penelitian ini ada beberapa jenis yaitu *decision tree*, *logistic regression*, *support vector machine* dan *random forest*. Kemudian proses seleksi fitur menggunakan metode *forward selection*.

II. METODOLOGI PENELITIAN

Penelitian dilakukan dengan menggunakan bahasa pemrograman *python* untuk menyusun kode programnya. Selain itu penulis juga menggunakan beberapa *library* untuk menjalankan *machine learning* yang sudah disediakan oleh *python*. Kode program dijalankan pada *cloud computing* yang dimiliki oleh google, yaitu *google colabatory* yang memiliki GPU untuk mempercepat proses training *machine learning*. Selain itu fasilitas ini bersifat gratis yang disediakan oleh *google* dan bisa diakses pada laman <https://colab.research.google.com/>.

Alur penelitian dimulai dari mencari dataset pada UCI *machine learning repository*, setelah dataset diperoleh kemudian dataset tersebut dibagi menjadi dua macam data yang berbeda, yaitu data *training* yang berfungsi dalam proses pembuatan model dan data *testing* berfungsi dalam proses evaluasi model. Setelah itu proses selanjutnya adalah melakukan seleksi fitur untuk membuat dataset menjadi lebih kecil tanpa harus mengurangi nilai penting yang digunakan untuk klasifikasi, selain itu fitur seleksi ini juga akan mempercepat proses klasifikasi. Proses selanjutnya melakukan klasifikasi dengan beberapa lagoritma dan melakukan evaluasi terhadap model yang sudah dibuat. Alur penelitian dapat dilihat selengkapnya pada Gambar 1.



Gambar 1. Alur Penelitian

A. Dataset

Dataset pada penelitian ini berasal dari *Government College University*, Faisalabad, Pakistan, dan dapat diakses secara gratis pada laman UCI *machine learning repository*. Dataset yang dipakai memiliki 299 records dan 13 atribut yang digunakan dalam pengukuran, yaitu *age*, *creatinine phosphokinase*, *anemia*, *serum creatinine*, *sex*, *diabetes*, *ejection fraction*, *platelets*, *serum sodium*, *high blood pressure*, *smoking* dan *time*. Secara lengkap dataset beserta keterangan atribut bisa dilihat pada Tabel 1.

TABEL I
 KETERANGAN DATASET

No	Atribut	Keterangan
1	<i>Age</i>	Usia
2	<i>Anemia</i>	Penurunan sel darah merah atau hemoglobin
3	<i>High blood pressure</i>	Riwayat hipertensi
4	<i>Creatinine phosphokinase</i>	Tingkat enzim CPK pada darah
5	<i>Diabetes</i>	Riwayat diabetes
6	<i>Ejection fraction</i>	Presentase darah meninggalkan jantung pada setiap kontraksi
7	<i>Platelets</i>	Kadar trombosit dalam darah
8	<i>Sex</i>	Jenis kelamin
9	<i>Serum creatinine</i>	Tingkat serum kreatinin pada darah

Terakreditasi Peringkat Sintia 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

10	<i>Serum sodium</i>	Kadar serum natrium pada darah
11	<i>Smoking</i>	Perokok atau tidak
12	<i>Time</i>	Periode tindak lanjut
13	<i>Death event</i>	Apakah pasien meninggal ketika sedang dalam masa tindak lanjut

B. Pembagian Dataset

Proses pembagian dataset atau *split dataset* akan membagi dataset yang diperoleh menjadi dua macam, bagian pertama sebagai data *training*, bagian sisanya sebagai data *testing*. Proses *training* merupakan proses yang berfungsi untuk pembuatan model *machine learning*, model yang sudah dibuat akan dipakai untuk klasifikasi data *testing*.

Pada proses pembuatan model *machine learning* data yang digunakan adalah data *training*, sedangkan data *testing* yang diperoleh dari hasil *split* dataset dipakai untuk melakukan proses *testing* pada model yang sudah jadi. Hal tersebut diperlukan untuk mengetahui performa klasifikasi yang dilakukan diperoleh. Data yang digunakan untuk *testing* akan dipisahkan dengan data *training*, data tersebut tidak boleh digunakan pada proses pembelajaran, supaya pada saat *testing*, model benar benar akan belajar dari data baru. Pembagian besarnya dataset yang dipakai adalah 8:2, dimana sebanyak 80% digunakan untuk data *training* dan 20% untuk data *testing*.

C. Fitur Ekstraksi

Fitur ekstraksi merupakan proses yang digunakan untuk mengurangi fitur yang ada pada dataset. Proses pengurangan ini berguna untuk membuat data menjadi lebih efisien, mempercepat waktu komputasi dan meningkatkan performa klasifikasi. Proses pemilihan fitur paling optimal dilakukan dengan cara mencari hubungan paling dekat antara fitur dengan target klasifikasi. Ada dua macam pendekatan yang digunakan pada fitur ekstraksi, yaitu metode *filter* dan metode *wrapper*. Metode *filter* akan memberikan waktu yang lebih cepat, sedangkan metode *wrapper* akan memberikan hasil yang lebih baik [6].

Pada penelitian ini penulis menggunakan pendekatan *forward selection* yang merupakan salah satu pendekatan di metode *wrapper* yang dapat menghilangkan atribut yang tidak relevan atau pengaruhnya terhadap klasifikasi sedikit. *Forward selection* akan memilih variabel *independent* yang memiliki korelasi paling besar kedalam variabel *dependent*, proses itu akan berlanjut ke variabel berikutnya sampai tidak ada lagi variabel *independent* yang potensial.

D. Klasifikasi Data

Klasifikasi adalah suatu proses analisis pada dataset yang berfungsi untuk mengelompokkan data menjadi beberapa bagian sesuai dengan jenisnya. Model klasifikasi yang telah dibuat dapat digunakan untuk memastikan kelas yang tercantum pada suatu data yang telah diberikan label. Model untuk klasifikasi tersebut diberi nama *classifier*, jadi dengan menggunakan *classifier* kita bisa menentukan beberapa kelas yang terdapat dalam dataset [7]. Pada riset yang dilakukan model klasifikasi digunakan

sebagai alat untuk mengelompokkan dataset menjadi dua polaritas utama, yaitu polaritas 1 serta 0.

Proses Klasifikasi dilakukan dengan menggunakan beberapa algoritma yaitu *decision tree*, *logistic regression*, *support vector machine* dan *random forest* untuk menentukan kelas dari *death event*. Setelah diperoleh hasil klasifikasi, kemudian penulis membandingkan seluruh hasil tersebut untuk mencari performa klasifikasi yang terbaik diantara beberapa algoritma tersebut.

E. Gagal Jantung

Gagal jantung merupakan istilah umum untuk menyebutkan gagal jantung kongestif atau *Congestive Heart Failure* (CHF). Gagal jantung adalah masalah kesehatan masyarakat utama yang beban penyakitnya meningkat seiring bertambahnya usia [8]. Gagal jantung adalah sindrom kompleks dengan tampilan gejala khas seperti sesak saat aktivitas, kelelahan, serta tanda retensi cairan seperti kongesti paru [9]. Kondisi jantung dalam kasus ini menandakan bahwa jantung tidak mampu lagi untuk memompa darah secara optimal dan ketidakmampuan jantung dalam memenuhi kuota darah normal yang dibutuhkan oleh tubuh. Berdasarkan timbulnya gejala, gagal jantung dapat terjadi secara tiba-tiba atau lebih dikenal dengan gagal jantung akut, dan gagal jantung yang berkembang secara perlahan karena kondisi jantung yang melemah atau lebih dikenal dengan istilah gagal jantung kronis.

Penyakit gagal jantung dapat dipicu dikarenakan beberapa faktor. Menurut Hou [10] menyatakan bahwa usia merupakan faktor utama penyebab seseorang mengalami gagal jantung. Pada penelitian lainnya Ardini [11] menyebutkan bahwa CHF atau CHF kongestif terjadi paling banyak pada orang yang berusia kurang dari 60 tahun atau pada kelompok usia dewasa dibandingkan dengan kelompok usia lanjut atau lebih dari 60 tahun. Hal lain yang menjadi faktor penyebab gagal jantung adalah rokok, hipertensi, kadar gula darah dan obesitas. Berdasarkan hasil penelitian da Lam [12] menunjukkan bahwa peningkatan pasien gagal jantung di Asia Tenggara lebih banyak dibandingkan dengan negara barat.

F. Evaluasi

Evaluasi adalah proses yang digunakan untuk mengukur seberapa baik model klasifikasi, proses evaluasi akan memakai data *testing* yang sudah disiapkan. Evaluasi pada penelitian ini dilakukan menggunakan metode *confusion matrix*, nilai yang diperoleh dari *confusion matrix* akan digunakan untuk mencari skor akurasi, sensitifitas spesifisitas. Metode *confusion matrix* akan menghasilkan nilai *True Positive*, *False Negative*, *True Negative* dan *False Positive*. Gambar 2 menunjukkan ilustrasi dari *Confusion matrix*.

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

		NILAI AKTUAL	
		1	0
NILAI PREDIKSI	1	TP	FP
	0	FN	TN

Gambar 2. Confusion Matrix

Penelitian ini akan menggunakan matriks akurasi untuk mengukur

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Spesificity = \frac{TN}{TN+FP} \quad (3)$$

III. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan menggunakan data rekam medis gagal jantung yang berasal dari UCI *machine learning repository*. Proses seleksi fitur menggunakan metode *forward selection* untuk memilih atribut yang memiliki pengaruh paling besar terhadap kelas target klasifikasi, dan algoritma klasifikasi menggunakan *decision tree*, *logistic regression*, *support vector machine* dan *random forest*.

Pada awalnya dataset memiliki 13 atribut, kemudian setelah melalui proses *forward selection* atribut dipilih menjadi lima saja yang memiliki pengaruh kuat terhadap kelas klasifikasi yaitu kelas *death event*. Lima atribut tersebut adalah *creatinine_phosphokinase*, *diabetes*, *ejection_fraction*, *serum_creatinine* dan *time*. Kemudian fitur yang sudah dipilih tersebut dijadikan input untuk beberapa algoritma klasifikasi. Hasil klasifikasi dari penelitian yang dilakukan diperlihatkan pada Tabel 2.

TABEL 2
HASIL KLASIFIKASI

Algoritma	Akurasi	Sensitivitas	Spesifisitas
<i>Support Vector Machine</i>	86%	93%	70%
<i>Decision Tree</i>	81%	92%	60%
<i>Logistic Regression</i>	90%	93%	80%
<i>Random Forest</i>	88%	95%	72%

Pada Tabel 2 menunjukkan bahwa nilai akurasi paling baik diperoleh oleh algoritma *logistic regression* dengan akurasi 90%, sedangkan nilai akurasi paling rendah didapat oleh algoritma *decision tree* dengan 81%. Nilai sensitivitas paling tinggi diperoleh oleh algoritma *random forest*

sebanyak 95%, nilai sensitivitas paling rendah 92% dengan algoritma *decision tree*. Dan nilai spesifisitas paling besar adalah 80% diperoleh algoritma *logistic regression* dan paling rendah 60% oleh algoritma *decision tree*.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dijelaskan sebelumnya, dapat disimpulkan bahwa algoritma *logistic regression* memiliki performa yang sangat baik ketika digunakan untuk klasifikasi penentuan kematian akibat gagal jantung, karena bisa menghasilkan akurasi sebesar 90% dan spesifisitas sebesar 80%. Selain itu penambahan metode seleksi fitur *forward selection* dapat membuat dataset yang ada menjadi lebih kecil tanpa menghilangkan nilai penting didalamnya. Hasil penelitian ini memperoleh nilai akurasi lebih baik daripada penelitian terdahulu yang dilakukan oleh Elin menggunakan algoritma C4.5 dan fitur seleksi *forward selection*.

Selanjutnya hasil penelitian ini bisa digunakan untuk menjadi acuan bagi peneliti selanjutnya dalam melakukan analisis klasifikasi, baik untuk data medis ataupun non-medis. Penulis menyarankan penelitian selanjutnya bisa menggunakan algoritma *deep learning* supaya dapat memperoleh hasil yang lebih baik.

UCAPAN TERIMA KASIH / ACKNOWLEDGMENT

Penulis mengucapkan terima kasih kepada Universitas Muhammadiyah Surakarta atas semua dukungan yang telah diberikan kepada penulis, sehingga dengannya penulis dapat menyelesaikan penelitian ini.

REFERENSI

- [1] Wahyudi, A. K., Kairupan, Y. J., & Masengi, Y. C. Alat Peraga Jantung Manusia Berbasis Augmented Reality dengan Menggunakan Teknik 3D Object Tracking. *CogITo Smart Journal*, 2018, 4(1), 46–59.
- [2] Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C. L., Lindberg, E. L., Kanda, M., Polanski, K., Heinig, M., Lee, M., Nadelmann, E. R., Roberts, K., Tuck, L., Fasouli, E. S., DeLaughter, D. M., McDonough, B., Wakimoto, H., Gorham, J. M., Samari, S., ... Teichmann, S. A. (2020). Cells of the adult human heart. *Nature* 2020 588:7838, 588(7838), 466–472. <https://doi.org/10.1038/s41586-020-2797-4>.
- [3] Dhimas, Y. A. Klasifikasi Penyakit Gagal Jantung Kongestif Menggunakan Artificial Neural Network (Ann) Berdasarkan Ekstraksi Fitur Multifractional Detrended Fluctuation Analysis (Mfdfa) Pada Variabilitas Denyut Jantung, 2017.
- [4] Duwi, C. P. B. Penerapan Algoritma Naïve Bayes dengan Seleksi Fitur Algoritma Genetika untuk Prediksi Gagal Jantung. *Evolusi: Jurnal Sains dan Manajemen*, 2021, Vol 9 No. 2.
- [5] Elin, N., Ultach, E. Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5. *JUTIM (Jurnal Teknik Informatika Musirawas)*, 2021, Vol. 6, No. 1.
- [6] Matthias Reif, Faisal Shahait, "Efficient Feature Size Reduction Via Predictive Forward Selection". *Pattern Recognition* ,2014, 1664–1673.
- [7] Sutoyo, I. Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik. *Jurnal PILAR Nusa Mandiri*, 2018, Vol. 14, No.2.
- [8] Wong, C. M., Hawkins, N. M., Ezekowitz, J. A., Jhund, P. S., Savu, A., MacDonald, M. R., Kristensen, S. L., Petrie, M. C., McMurray, J. J. V., McAlister, F. A., & Kaul, P. Heart Failure in Young Adults Is Associated With High Mortality: A Contemporary Population-Level Analysis. *Canadian Journal of Cardiology*, 2017, 33(11), 1472–1477. <https://doi.org/10.1016/j.cjca.2017.05.009>

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

- [9] Imaligy, E. U. Gagal Jantung pada Geriatri. *Cermin Dunia Kedokteran*, 2014, 41(1), 19–24. <https://doi.org/10.55175/CDK.V41I1.1167>.
- [10] Hou, N. Relationship of age and sex to health-related quality of life in patients with heart failure. *Am J Crit Care* 2004 March, 2004, 13(20:153-61).
- [11] Ardini, D. N. E. Perbedaan Etiologi Gagal Jantung Kongestif Pada Usia Lanjut Dengan Usia Dewasa Di Rumah Sakit Dr. Kariadi Januari- Desember 2006. 2007.
- [12] Lam, C. S. P. Heart failure in Southeast Asia: facts and numbers. *ESC Heart Failure*, 2015, 2(2), 46–49. <https://doi.org/10.1002/EHF2.12036>.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

