KLASIFIKASI HARAPAN HIDUP PASIEN KARSINOMA HEPATOSELULER MENGGUNAKAN *EXTREME LEARNING MACHINE* DENGAN PERBAIKAN DATA HILANG

Suci Permata Sari[1], Triando Hamonangan Saragih[2], Andi Farmadi[3], Radityo Adi Nugroho[4], Rudy Herteno[5]

[1,2,3,4,5]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat
[1]sucipermataa@gmail.com, [2]triando.saragih@ulm.ac.id, [3]andifarmadi@ulm.ac.id, [4]radityo.adi@ulm.ac.id, [5]rudy.herteno@ulm.ac.id

Abstrak

International Agency for Research on Cancer (IARC) mengestimasi bahwa pada tahun 2020 kanker hati primer berada di peringkat ke-6 sebagai kanker yang paling banyak didiagnosis dan peringkat ke-3 sebagai penyebab utama kematian akibat kanker di dunia. Mayoritas kanker hati primer muncul dari sel-sel hati dan disebut Karsinoma Hepatoseluler (KHS). Salah satu upaya yang dapat dilakukan untuk mengatasi permasalahan tersebut adalah dengan mengklasifikasikan harapan hidup pasien KHS. Terdapat banyak metode yang dapat digunakan dalam klasifikasi, salah satunya adalah menggunakan *Extreme Learning Machine* (ELM). Dataset yang digunakan pada penelitian ini adalah HCC Survival Data Set yang memiliki 49 fitur dengan rata-rata data hilang sebesar 10,22% secara keseluruhannya. ELM merupakan metode yang mengharuskan semua data pada datasetnya lengkap tanpa memiliki data hilang. Sehingga harus dilakukan penanganan data hilang terlebih dahulu sebelum dilakukan klasifikasi. Penanganan data hilang pada penelitian ini dilakukan dengan menggunakan teknik imputasi. Pada penelitian ini dilakukan perbandingan antara hasil klasifikasi dari data yang diimputasi menggunakan MissForest dengan hasil klasifkasi dari data yang diimputasi menggunakan *K-Nearest Neighbors Imputation* (KNNI). Perbandingan tersebut dilakukan untuk mengetahui metode imputasi mana yang menghasilkan data imputasi dengan kinerja terbaik pada klasifikasi kelangsungan hidup pasien KHS. Hasil menunjukkan bahwa data yang diimputasi menggunakan KNNI menghasilkan nilai akurasi rata-rata dan nilai rata-rata AUC yang lebih unggul dibandingkan dengan data yang diimputasi dengan MissForest, yaitu dengan nilai akurasi rata-rata sebesar 92,941% dan rata-rata AUC sebesar 0,9758.

Kata kunci : Karsinoma Hepatoseluler, *Extreme Learning Machine*, MissForest, *K-Nearest Neighbors Imputation*

1. Pendahuluan

International Agency for Research on Cancer (IARC) mengestimasi bahwa pada tahun 2020 kanker hati primer berada di peringkat ke-6 sebagai kanker yang paling banyak didiagnosis dan peringkat ke-3 sebagai penyebab utama kematian akibat kanker di dunia (Sung *et al*., 2021). Mayoritas dari kanker hati primer timbul dari sel-sel hati dan disebut Karsinoma Hepatoseluler (KHS). Berdasarkan data Riset Kesehatan Dasar 2013, sekitar 1.050.000 penduduk di Indonesia memiliki potensi untuk menderita kanker hati. Sehingga permasalahan ini akan berdampak besar terhadap kesehatan masyarakat, produktivitas, umur harapan hidup, dan dampak sosial ekonomi lainnya (Kementrian Kesehatan RI, 2013). Untuk membantu mengatasi permasalahan kanker hati, maka diperlukan suatu upaya penanganan. Salah satu upaya yang dapat dilakukan adalah dengan melakukan klasifikasi harapan hidup pasien KHS. Hasil klasifikasi dapat digunakan untuk membantu dokter atau tenaga kesehatan dalam menentukan bagaimana penanganan yang lebih tepat terhadap pasien. Terdapat banyak metode yang dapat digunakan dalam klasifikasi, salah satunya adalah menggunakan jaringan syaraf tiruan.

Extreme Learning Machine (ELM) merupakan jaringan syaraf tiruan feedforward dengan satu hidden layer *(single hidden layer feedforward neural network*). ELM memiliki kelebihan, yaitu dapat menghasilkan performa generalisasi yang lebih baik serta memiliki kecepatan pembelajaran lebih cepat dibandingkan metode *feedforward network* terdahulu seperti metode *Backpropagation* (Huang *et al*, 2006). Saragih *et al* (2018) juga telah melakukan penelitian dimana ELM menghasilkan akurasi yang lebih baik dibandingkan dengan metode *Fuzzy Neural Network*, *Fuzzy Neural Network-Simulated Annealing*, dan *Backpropogation* pada klasifikasi penyakit tanaman jarak pagar. Berbagai jenis penyakit pada manusia telah diklasifikasi menggunakan ELM. Nurdiansyah *et al* (2020) menggunakan ELM untuk melakukan klasifikasi penyakit Tuberkulosis, yang mana menghasilkan akurasi tertinggi sebesar 99,33%.

Multazam *et al* (2020) menerapkan ELM pada penyakit hepatitis yang mana menghasilkan akurasi rata-rata sebesar 80%. Penerapan ELM pada penyakit kanker serviks juga telah dilakukan oleh Hidayah *et al* (2019) yang mana menghasilkan akurasi sebesar 91,76%.

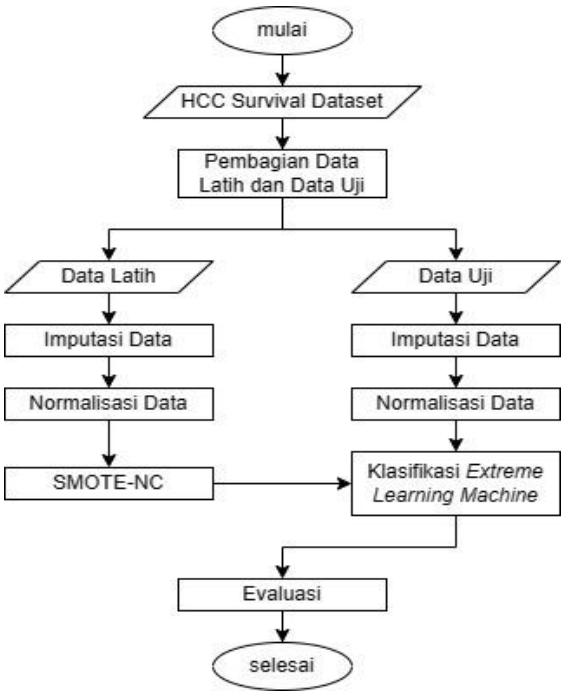
Di sisi lain, dataset yang akan digunakan pada penelitian ini memiliki permasalahan, yaitu adanya nilai hilang pada dataset. Nilai hilang (*missing value*) dapat diartikan sebagai data atau informasi yang tidak tersedia atau hilang atau tidak tersedia mengenai subjek penelitian pada variabel tertentu. Menurut Gao *et al* (2015), ELM merupakan metode yang mengharuskan semua data pada datasetnya lengkap tanpa memiliki nilai hilang. Sehingga pada penelitian ini nilai hilang pada dataset akan ditangani terlebih dahulu dengan menggunakan metode imputasi sebelum melakukan klasifikasi.

Santos *et al* (2015), telah melakukan imputasi nilai hilang menggunakan *K-Nearest Neighbor Imputation* (KNNI) pada dataset yang sama dengan penelitian ini. Namun penelitian tersebut berfokus pada penanganan data tidak seimbang pada dataset bukan penanganan nilai hilang pada data. Doni *et al* (2021) juga telah melakukan imputasi rata-rata pada dataset yang sama tetapi penelitian tersebut lebih berfokus untuk mengetahui akurasi klasifikasi menggunakan Naïve Bayes. Petrazzini *et al* (2021) telah melakukan perbandingan metode-metode imputasi pada data genomic. Dari penelitian tersebut KNNI dan MissForest menghasilkan rata-rata nilai error terkecil dibandingkan imputasi rata-rata, *Multivariate Imputation by Chained Equation*, dan *Multivariate Normal Distribution using EM*. Stekhoven dan Buhlmann (2012) juga telah melakukan perbandingan metode imputasi pada 11 dataset, yang mana MissForest berhasil menghasilkan performa dengan rata-rata NRMSE lebih kecil dibandingkan KNNI dan *Multiple Imputation by Chained Equations* (MICE) pada 10 dataset yang diuji. KNNI menghasilkan NRMSE lebih kecil dibandingkan MissForest hanya pada satu dataset, yaitu *Gene finding over prediction dataset*. Pada penelitian lainnya oleh Alsaber *et al* (2021), MissForest menghasilkan imputasi dengan nilai error terkecil diikuti oleh KNNI dibandingkan dengan *Bayesian Principal Component Analysis*, MICE dengan *Random Forest, Predictive Mean Matching*, dan *Expectation Maximization* dengan *Bootsraping*. Sedangkan pada penelitian dari Emmanuel *et al* (2021), MissForest dan KNNI masing-masing lebih unggul pada satu dataset.

Dari penelitian-penelitian sebelumnya tersebut dapat diketahui bahwa MissForest dan KNNI merupakan metode imputasi yang telah sering digunakan serta dapat menghasilkan akurasi ataupun nilai error lebih kecil dibandingkan metode imputasi lainnya. Oleh karena itu, pada penelitian ini dilakukan perbandingan metode imputasi MissForest dan KNNI untuk mengetahui metode imputasi mana yang akan menghasilkan data imputasi dengan kinerja terbaik pada klasifikasi harapan hidup pasien KHS. Kinerja pada penelitian ini diukur menggunakan nilai akurasi dan nilai AUC.

2. Metode Penelitian

Adapun prosedur penelitian yang dilaksanakan dalam penelitian ini dapat dilihat pada Gambar 1 sebagai berikut:



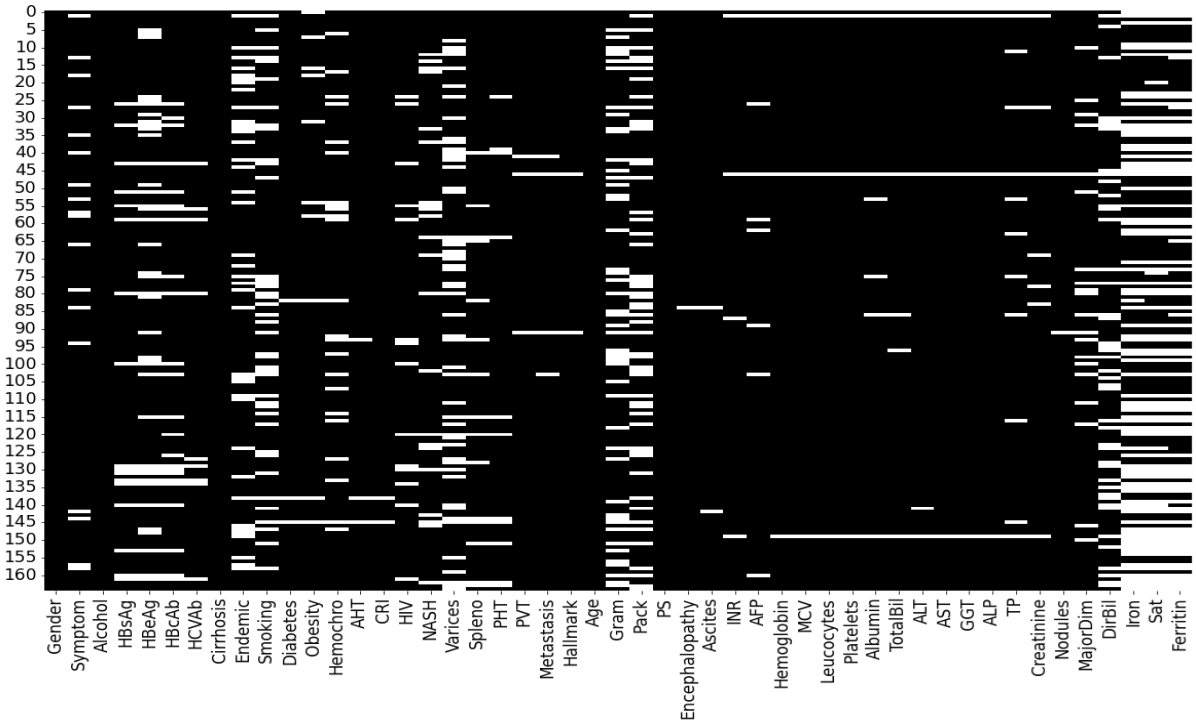
Gambar 1. Diagram Alur Prosedur Penelitian

2.1 Pengumpulan Data

Dataset digunakan pada penelitian ini adalah *HCC Survival Data Set* yang didapatkan dari *UCI Machine Learning Repository*. Dataset tersebut terdiri dari 49 fitur yang telah dipilih sesuai dengan pedoman klinis dari *European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer*, yang mana terdiri dari 23 variabel kuantitatif dan 26 variabel kualitatif. Kelas yang ada pada dataset ini adalah harapan hidup pasien KHS dalam kurun waktu 1 tahun, yang terbagi menjadi 2 kelas, yaitu kelas hidup yang terdiri dari 102 sampel dan kelas meninggal yang terdiri dari 63 sampel. Dataset ini memiliki rata-rata nilai hilang sebesar 10,22%, yang mana hanya terdapat 8 sampel yang memiliki data lengkap pada keseluruhan atributnya. Tabel 1 merupakan contoh data dari HCC Survival Dataset yang digunakan pada penelitian ini. Data kosong pada tabel merupakan data hilang yang akan diimputasi. Pola data hilang pada dataset ditunjukkan pada Gambar 2, yang mana data hilang pada dataset digambarkan dengan warna putih dan data yang tidak hilang atau terisi digambarkan dengan warna hitam.

Tabel 1. HCC *Survival Data Set*

No	*Gender*	*Symptom*	*Alcohol*	HBsAg	…	MajorDim	DirBil	*Iron*	Sat	*Ferritin*
1	1	0	1	0	…	3,5	0,5			
2	0		0	0	…	1,8				
3	1	0	1	1	…	13	0,1	28	6	16
4	1	1	1	0	…	15,7	0,2			
5	1	1	1	1	…	9		59	15	22
…	…	…	…	…	…	…	…	…	…	…
161	0	0	1		…	3				
162	0	1	0		…	2,2	2,3			
163	1	0	1	0	…	18,6				
164	1	0	1	1	…	18				
165	1	1	1	0	…	8,5	19,8			



Gambar 2. Pola Data Hilang pada *HCC Survival Data Set*

2.2 Pembagian Data

Pada tahap ini dilakukan pembagian data menjadi dua bagian, yaitu data latih dan data uji. Pembagian data dalam penelitian ini dilakukan menggunakan *stratified random sampling*. Pada penelitian ini pembagian data dilakukan sebelum tahap imputasi untuk menghindari kebocoran data (Marcinkevics *et al.*, 2021).

2.3 Imputasi Data

Imputasi adalah teknik untuk pengisian nilai hilang pada data dengan nilai yang mungkin berdasarkan informasi yang tersedia dari nilai-nilai yang diketahui (Susanti *et al.*, 2021). Pada penelitian ini digunakan dua jenis metode imputasi, yaitu MissForest dan *K-Nearest Neighbor Imputation* (KNNI). MissForest adalah metode imputasi non parametrik berbasis machine learning yang beroperasi dengan algoritma *Random Forest*. MissForest didasarkan pada pendekatan berulang, dimana pada setiap iterasi maka hasil imputasi yang dihasilkan semakin baik (Stekhoven dan Bühlmann, 2012). Sedangkan KNNI merupakan metode yang digunakan untuk mengisi nilai yang hilang menggunakan pendekatan *K-Nearest Neighbors*. Jika atribut data yang hilang bertipe data numerikal maka data hilang tersebut akan diganti oleh nilai rata-rata nilai k terdekat. Jika atribut data yang hilang bertipe data kategori maka data hilang tersebut akan diganti dengan nilai kategori yang

paling banyak muncul dari data lengkap sejumlah k terdekat (Sugiarta *et al*., 2019).

Parameter yang digunakan pada imputasi MissForest adalah parameter *default* pada program. Sedangkan untuk metode KNNI, dilakukan pengujian parameter jumlah tetangga atau k berjumlah ganjil pada rentang 1 sampai 20. Pengujian k tersebut bertujuan untuk mengetahui jumlah k yang dapat menghasilkan performa terbaik setelah diklasifikasi. Perhitungan jarak KNNI pada penelitian ini menggunakan *Heterogeneous Euclidean-Overlap Metric* (HEOM). Penggunaan HEOM merujuk kepada penelitian terdahulu dari Santos *et al* (2015), yang mana HEOM dapat mengatasi perhitungan jarak dari dataset yang memiliki data diskrit dan data kontinu.

2.4 Normalisasi Data

Metode normalisasi yang digunakan pada penelitian ini adalah *Min Max Normalization*, sebagaimana penelitian terdahulu dari Nurdiansyah *et al*. (2020). Pada penelitian tersebut normalisasi dilakukan untuk mempersiapkan data sebelum memasuki proses klasifikasi menggunakan *Extreme Learning Machine*. Normalisasi dilakukan untuk mengurangi perbedaan yang besar pada rentang nilai antar data. Data yang digunakan diubah menjadi nilai sedemikian rupa sehingga data bernilai pada rentang antara 0 hingga 1. Berikut ini merupakan persamaan normalisasi data menggunakan metode *Min-Max Normalization* (Nishom, 2019).

$$x' = \frac{x - nilai_{min}}{nilai_{max} - nilai_{min}} \quad (1)$$

Dengan x merupakan data yang akan dinormalisasi, $nilai_{min}$ adalah nilai terkecil atau minimum pada data, dan $nilai_{max}$ adalah nilai terbesar atau maksimum pada data.

2.5 SMOTE-NC

Dataset yang digunakan pada penelitian ini memiliki distribusi data yang tidak seimbang. Karena adanya ketidakseimbangan data tersebut, maka sebelum memasuki proses klasifikasi perlu dilakukan penyeimbangan data untuk membuat jumlah data minoritas menjadi sama dengan data mayoritas. Penyeimbangan data dilakukan untuk membuat pelatihan data saat klasifikasi dapat menjadi lebih baik. Pada penelitian ini data akan diseimbangkan SMOTE-NC (*Synthetic Minority Over-sampling Technique-Nominal Continuous*). SMOTE-NC merupakan jenis SMOTE yang dapat menangani data campuran numerik dan kategori (Wijaya *et al.*, 2018).

2.6 Klasifikasi

Pada penelitian ini metode klasifikasi yang digunakan adalah *Extreme Learning Machine* (ELM). Fungsi aktivasi yang digunakan pada penelitian ini adalah *Sigmoid Biner*. Penggunaan Sigmoid Biner pada penelitian ini didasarkan pada penelitian dari Nurdiansyah *et al* (2020) yang menyebutkan bahwa fungsi aktivasi tersebut menghasilkan hasil akurasi tertinggi dibandingkan fungsi aktivasi lainnya pada klasifikasi menggunakan ELM. Jumlah neuron pada *hidden layer* juga merupakaan salah satu faktor penting dalam perfoma generalisasi metode ELM. Sehingga pada penelitian ini dilakukan pengujian jumlah neuron pada *hidden layer* dari 5-70 pada neuron dengan kelipatan 5, yang mana pada setiap neuronnya dilakukan pengujian masing-masing sebanyak 10 kali.

2.7 Evaluasi

Proses evaluasi dilakukan untuk membandingkan hasil keluaran dari metode yang diterapkan dengan keadaan sebenarnya sesuai dengan data yang telah tersedia. Evaluasi yang digunakan pada penelitian ini adalah nilai AUC dan akurasi dari hasil klasifikasi. Akurasi dihitung dengan membagi jumlah keseluruhan klasifikasi yang benar dengan jumlah semua instance pada data awal (Han *et al*., 2011). Sedangkan AUC merupakan suatu daerah di bawah kurva receiver operating characteristic (ROC). ROC adalah kurva hasil dari tarik ulur antara spesifisitas dan sensitivitas pada berbagai titik potong. Nilai AUC secara teoritis berada di antara 0 dan 1. Semakin besar nilai AUC maka berarti semakin baik pula kinerja klasifikasi (Chohan *et al*., 2020). Selanjutnya, hasil pegujian pada MissForest-ELM dan KNNI-ELM tersebut akan dibandingkan.

3. Hasil dan Pembahasan

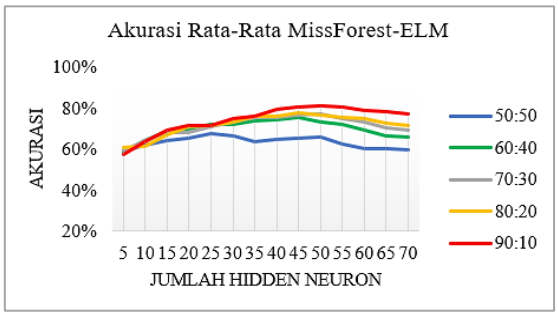
Pada penelitian ini, dilakukan beberapa skenario percobaan untuk mendapatkan hasil akurasi dan AUC tertinggi. Skenario tersebut dilakukan terhadap pembagian data, dataset hasil imputasi, dan parameter *k* pada SMOTE-NC.

3.1 Skenario Pemilihan Rasio Pembagian Data

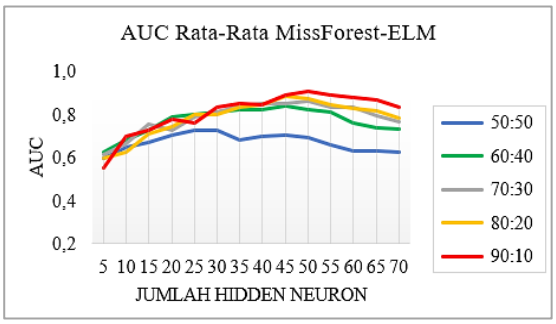
Pengujian rasio pembagian data dengan proporsi 90:10, 80:20, 70:30, 60:40, dan 50:50. Tahapan ini dilakukan untuk mengetahui proporsi pembagian data yang menghasilkan performa terbaik pada klasifikasi menggunakan data yang telah dimputasi. Pemilihan rasio tersebut didasarkan dari penelitian oleh Fadilla *et al* (2018) yang menggunakan rasio serupa, yang mana perubahan jumlah rasio pada data latih dan data uji tersebut memberikan pengaruh pada hasil akurasi dari klasifikasi. Pada penelitian ini, rasio pembagian data

yang menghasilkan akurasi tertinggi digunakan untuk skenario percobaan berikutnya. Pada skenario ini parameter MissForest yang digunakan, yaitu parameter *default* dan parameter KNNI yang digunakan adalah k=5. Parameter SMOTE-NC yang digunakan pada pengujian ini adalah parameter *default* k=5.

Gambar 3 dan Gambar 4 menunjukkan klasifikasi MissForest-ELM menghasilkan performa terbaik pada proporsi data 90:10 dengan jumlah *hidden neuron* sebanyak 50. Akurasi rata-rata tertinggi yang didapatkan, yaitu sebesar 81,176% dan AUC rata-rata sebesar 0,9061.

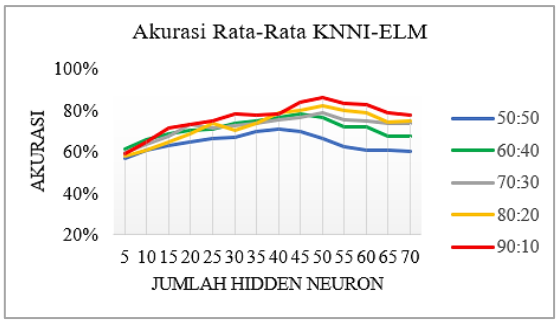


Gambar 3. Akurasi Pengujian Rasio Pembagian Data MissForest-ELM

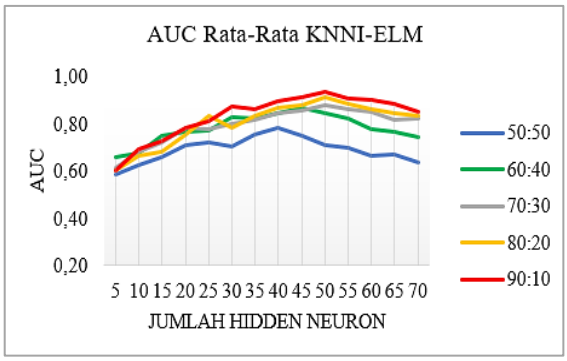


Gambar 4. AUC Pengujian Rasio Pembagian Data MissForest-ELM

Gambar 5 dan Gambar 6 menunjukkan klasifikasi MissForest-ELM menghasilkan performa terbaik pada proporsi data 90:10 dengan jumlah *hidden neuron* sebanyak 50. Akurasi rata-rata yang dihasilkan tersebut, yaitu sebesar 85,882% dengan diikuti AUC rata-rata sebesar 0,9318 pada jumlah *hidden neuron* sebanyak 50.



Gambar 5. Akurasi Pengujian Rasio Pembagian Data KNNI-ELM



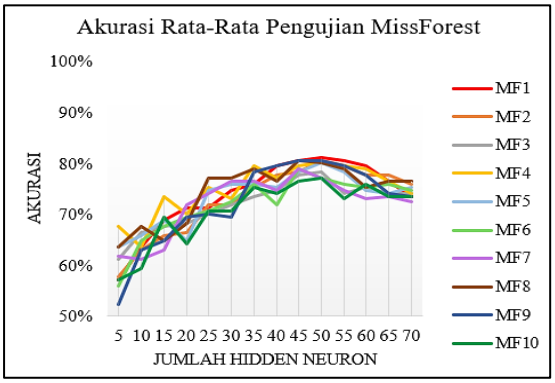
Gambar 6. AUC Pengujian Rasio Pembagian Data KNNI-ELM

Dari hasil yang telah didapatkan, maka dapat diketahui bahwa pada penelitian ini rasio pembagian data latih dan data uji 90:10 menghasilkan akurasi terbaik pada data yang diimputasi MissForest maupun data yang diimputasi KNNI dibandingkan akurasi dari rasio pembagian data lainnya.

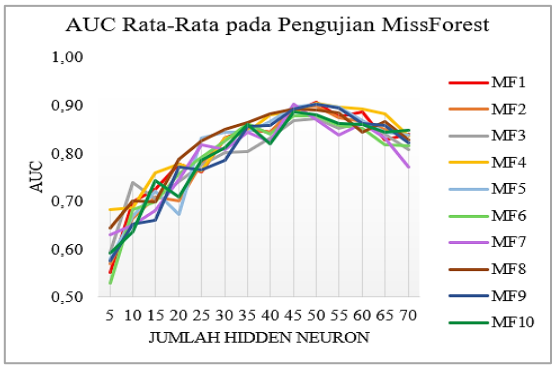
3.2 Skenario Pemilihan Dataset pada Imputasi

Pada skenario ini, metode imputasi MissForest dan KNNI diuji untuk mendapatkan dataset hasil imputasi yang menghasilkan performa terbaik pada klasifikasi. Pembagian data uji dan data latih yang digunakan adalah 90:10. Pengujian pada MissForest dilakukan dengan mengimputasi data sebanyak 10 kali menggunakan parameter default. Sedangkan pengujian pada KNNI, dilakukan pengujian parameter k bernilai ganjil dengan rentang 1 sampai dengan 20.

Pada skenario ini, hasil klasifikasi ELM pada data yang diimputasi menggunakan MissForest mendapatkan hasil akurasi rata-rata dan AUC tertinggi pada pengujian dataset pertama. Akurasi rata-rata tertinggi yang didapatkan, yaitu sebesar 81,176% dan AUC rata-rata sebesar 0,9061, pada jumlah *hidden neuron* sebanyak 50. Dengan demikian dataset hasil imputasi MissForest ke-1 akan digunakan pada skenario selanjutnya. Hasil kinerja dari pengujian tiap dataset yang dihasilkan MissForest dapat dilihat pada Gambar 7 dan Gambar 8.

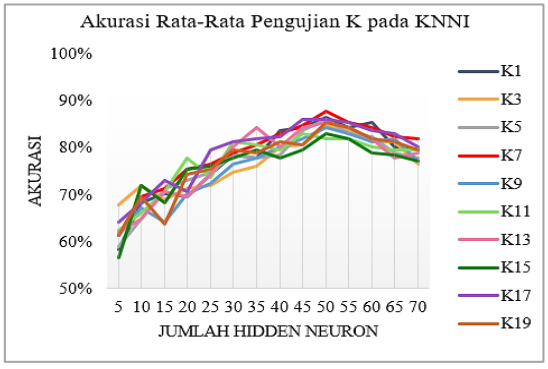


Gambar 7. Akurasi Rata-Rata Pengujian Imputasi MissForest

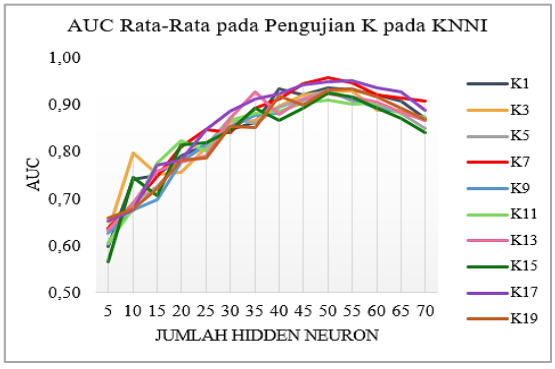


Gambar 8. AUC Rata-Rata Pengujian Imputasi MissForest

Klasifikasi KNNI-ELM menghasilkan performa terbaik pada k=7 dengan jumlah *hidden neuron* sebanyak 50. Akurasi rata-rata tertinggi yang didapatkan adalah sebesar 87,647% dengan AUC rata-rata yang didapatkan adalah sebesar 0,9576. Grafik hasil pengujian k pada KNNI dapat dilihat pada Gambar 9 dan Gambar 10. Pada pengujian ini, terlihat bahwa nilai k pada KNNI memiliki pengaruh terhadap akurasi dan AUC yang dihasilkan pada klasifikasi ELM.



Gambar 9. Akurasi Rata-Rata Pengujian k pada KNNI

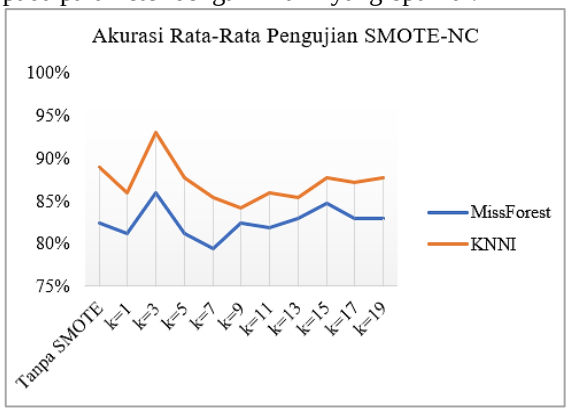


Gambar 10. AUC Rata-Rata Pengujian k pada KNNI

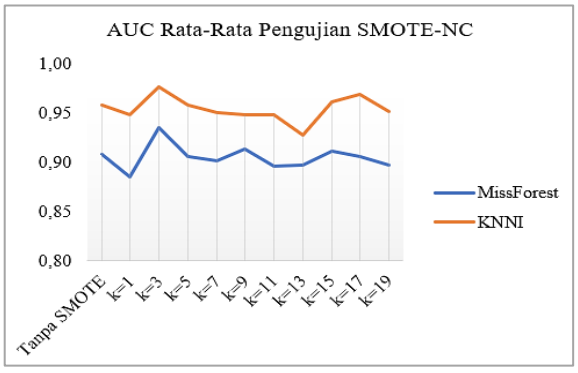
3.3 Skenario Pemilihan K pada SMOTE-NC

Pada skenario ini, data sintetis untuk data minoritas pada data latih akan dibentuk menggunakan SMOTE-NC pada k berjumlah ganjil pada rentang 1 sampai 20. Pada pengujian k untuk SMOTE-NC ini jumlah hidden neuron yang akan digunakan adalah sebanyak 50. Hal tersebut dikarenakan jumlah *hidden neuron* sebanyak 50 menghasilkan performa dengan akurasi dan AUC tertinggi dibandingkan yang lainnya pada skenario pengujian ke-2. Rasio pembagian data yang digunakan adalah 90:10. Dataset hasil imputasi MissForest yang digunakan adalah dataset pertama dan parameter k pada imputasi KNNI yang digunakan adalah k=7.

Gambar 11 dan Gambar 12 menunjukkan bahwa klasifikasi MissForest-ELM dan KNNI-ELM menghasilkan kinerja terbaik pada SMOTE-NC dengan jumlah k=3. Akurasi rata-rata tertinggi yang dihasilkan pada MissForest-ELM, yaitu sebesar 85,882% dan AUC rata-rata sebesar 0,9348. Sedangkan akurasi rata-rata tertinggi yang dihasilkan pada klasifikasi KNNI-ELM, yaitu sebesar 92,941% dengan AUC rata-rata sebesar 0,9758. Hasil pengujian menunjukkan bahwa pada penelitian ini SMOTE-NC dapat meningkatkan akurasi dan AUC pada klasifikasi jika dilakukan pada parameter dengan nilai k yang optimal.



Gambar 11. Akurasi Pengujian K pada SMOTE-NC

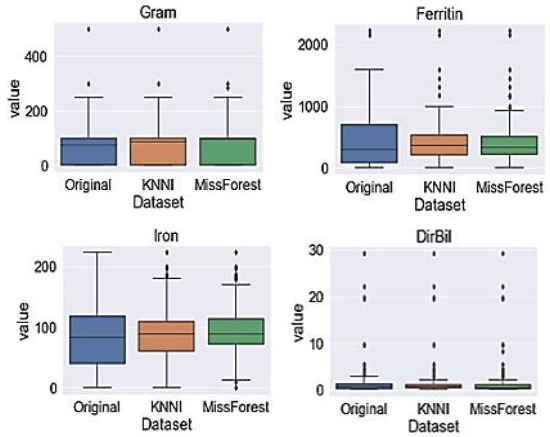


Gambar 12. AUC Pengujian K pada SMOTE-NC

Dari semua skenario pengujian yang telah dilakukan, didapatkan hasil akurasi rata-rata tertinggi pada MissForest-ELM adalah sebesar 85,882% dengan AUC rata-rata sebesar 0,9348. Sedangkan hasil akurasi rata-rata tertinggi pada KNNI-ELM adalah sebesar 92,941% dengan AUC rata-rata sebesar 0,9758. Dengan hasil pengujian

tersebut dapat disimpulkan bahwa pada penelitian ini data yang diimputasi menggunakan KNNI menghasilkan kinerja yang lebih baik dibandingkan dengan data yang diimputasi menggunakan MissForest pada klasifikasi harapan hidup pasien Karsinoma Hepatoseluler menggunakan *Extreme Learning Machine.*

Pada penelitian ini, data yang diimputasi menggunakan MissForest menghasilkan lebih banyak data outlier dibandingkan dengan hasil data imputasi KNNI. Data *outlier* adalah data pengamatan yang berada jauh dari pengamatan-pengamatan lainnya. Gambaran outlier pada beberapa fitur data sebelum dan sesudah diimputasi ditunjukkan pada Gambar 13. Adanya data *outlier* ini membuat analisis terhadap serangkaian data dapat menjadi bias, atau tidak mencerminkan fenomena yang sebenarnya. *Outlier* tersebut mempengaruhi kinerja klasikasi, sehingga akurasi dan AUC dari hasil klasifikasi ELM menggunakan data imputasi MissForest menjadi lebih rendah dibandingkan akurasi dan AUC dari hasil klasifikasi menggunakan data imputasi KNNI.



Gambar 13. Box-Plot Data Sebelum dan Sesudah Imputasi

4. Kesimpulan dan Saran

Penanganan nilai hilang pada klasifikasi harapan hidup menggunakan ELM menggunakan data imputasi KNNI menghasilkan akurasi dan nilai AUC yang lebih baik dibandingkan dengan klasifikasi menggunakan data imputasi MissForest. Klasifikasi dengan penanganan nilai hilang menggunakan KNNI menghasilkan akurasi rata-rata yang didapatkan adalah sebesar 92,941% diiringi dengan AUC rata-rata sebesar 0,9758. Sedangkan klasifikasi dengan penanganan nilai hilang menggunakan MissForest Akurasi rata-rata yang didapatkan adalah sebesar 85,882% dan AUC rata-rata sebesar 0,9348. Penelitian berikutnya sebaiknya menggunakan dataset yang memiliki data lengkap, yang mana kemudian dataset tersebut dibuat menjadi memiliki nilai hilang. Sehingga dampak dari imputasi terhadap klasifikasi ELM bisa dianalisis lebih lanjut. Selain itu penelitian selanjutnya bisa mencoba untuk menggunakan metode imputasi lainnya.

Daftar Pustaka:

Alsaber, A., Pan, J., Al-Hurban, A. (2021). Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018). *International Journal of Environmental Research and Public Health*, 18(3), 1333. https://doi.org/10.3390/ijerph18031333

Chohan, S., Nugroho, A. S., Aji, A. M. B., & Gata, W. (2020). Analisis Sentimen Pengguna Aplikasi Duolingo Menggunakan Metode Naïve Bayes dan Synthetic Minority Over Sampling Technique. Paradigma (Jakarta), 22(2), 139–144. https://doi.org/10.31294/p.v22i2.8251

Doni, R., Tri, B., Susanti, S., & Mubarok, A., 2021. Penerapan Data Mining Untuk Klasifikasi Penyakit Hepatocellular Carcinoma Menggunakan Algoritma Naive Bayes. *Jurnal Responsif*, 3(1), 12-19.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of big data*, Volume 8, 1-37. https://doi.org/10.1186/s40537-021-00516-9

Fadilla, I., Adikara, P. P., & Setya Perdana, R. (2018). Klasifikasi Penyakit Chronic Kidney Disease (CKD) Dengan Menggunakan Metode Extreme Learning Machine (ELM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(10), 3397–3405.

Gao, H., Liu, X.-W., Peng, Y.-X., & Jian, S.-L. (2015). Sample-Based Extreme Learning Machine With Missing Data. *Mathematical Problems in Engineering*, Volume 2015, 1–11. https://doi.org/10.1155/2015/145156

Han, J., Kamber, M., & Pei, J, *Data Mining: Concepts and Techniques*, Amsterdam, Morgan Kauffman Publishers, 2011.

Hidayah, U. R., Cholissodin, I., & Adikara, P. P. (2019). Klasifikasi Penyakit Kanker Serviks dengan Extreme Learning Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(7), 6575–6582.

Huang, G. bin, Zhu, Q. Y., & Siew, C. K. (2006). Extreme Learning Machine: Theory And Applications. *Neurocomputing*, Volume 70, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Kementrian Kesehatan RI. (2013). *Riset Kesehatan Dasar*. Balitbang Kemenkes RI: Jakarta.

Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., & Vogt, J. E. (2021). Using

Machine Learning to Predict the Diagnosis, Management and Severity of Pediatric Appendicitis. *Frontiers in pediatrics*, Volume. 9, 1–12. https://doi.org/10.3389/fped.2021.662183

Multazam, S., Cholissodin, I., and Adinugroho, S. (2020). Implementasi Metode Extreme Learning Machine pada Klasifikasi Jenis Penyakit Hepatitis berdasarkan Faktor Gejala. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 4(3), 789-797.

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Nurdiansyah, V. V., Cholissodin, I., & Adikara, P. P. (2020). Klasifikasi Penyakit Tuberkulosis (TB) menggunakan Metode Extreme Learning Machine (ELM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 4(5), 1387– 1393.

Petrazzini, B.O., Naya, H., Lopez-Bello, F. Vazquez, G., & Spangenberg, L. (2021). Evaluation Of Different Approaches For Missing Data Imputation On Features Associated To Genomic Data. *BioData Mining*, 14(44), 1– 13. https://doi.org/10.1186/s13040-021-00274-7

Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A. (2015). A New Cluster-Based Oversampling Method For Improving Survival Prediction Of Hepatocellular Carcinoma Patients. *Journal of Biomedical Informatics*, Volume 58, 49–59. https://doi.org/10.1016/j.jbi.2015.09.012

Saragih, T. H., Fajri, D. M. N., Mahmudy, W. F., Abadi, A. L., & Anggodo, Y. P. (2018). Jatropha Curcas Disease Identification With Extreme Learning Machine. *Indonesian Journal of Electrical Engineering and Computer Science*, 12(2), 883-888. https://doi.org/10.11591/ijeecs.v12.i2.pp883-888

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-Parametric Missing Value Imputation For Mixed-Type Data. *Bioinformatics*, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sugiarta, K.A., Cholissodin, I., Santoso, E. Optimasi K-Nearest Neighbor Menggunakan Bat Algorithm Untuk Klasifikasi Penyakit Ginjal Kronis. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(10), 10301–10308.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Susanti, Shantika Martha, Evy Sulistianingsih. (2018). K Nearest Neighbor dalam Imputasi Missing Data. *Buletin Ilmiah Math. Stat. dan Terapannya*, 7(1), 9 -14. http://dx.doi.org/10.26418/bbimst.v7i1.23498

Wijaya, J., Soleh, A. M., & Rizki, A. (2018). Penanganan Data Tidak Seimbang pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB. Xplore, 2(2), 32–40. https://doi.org/10.29244/xplore.v2i2.99