PENGEMBANGAN NEURAL NETWORK UNTUK PREDIKSI KUALITAS AIR

Aretha Safira[1], L. M. Sarudi As.[1], Afifa Puspitasari[1], Nur Mayke Eka Normasari[1], Achmad P. Rifai[1]*
[1]Departemen Teknik Mesin dan Industri, Fakultas Teknik, Universitas Gadjah Mada
Jl. Grafika No. 2, Yogyakarta 55281
*E-mail:achmad.p.rifai@ugm.ac.id **

ABSTRACT

Research on artificial intelligence to determine water quality has been widely developed as a human endeavor to improve the quality of life. This study employs an artificial neural network (ANN) to determine the optimal classification model for determining the safety of water. This study uses existing Kaggle generic datasets. Numerous preprocesses were performed on the dataset starting from cleaning the data from missing values and outliers to equalizing the weights of each parameter with the min-max scaler. This study compares the accuracy of ANN model in various scenarios constructed with 10, 15, 20, and 30 neurons. Scaled Conjugate Gradient is implemented as the learning algorithm for developing the prediction model. The obtained results of the experiments vary between scenarios. Overall accuracy increases when the number of neurons is between 10 and 20, and decreases when the number of neurons is between 20 and 30.

Keyword: classification, water quality, artificial neural network, prediction accuracy, number of neurons

INTISARI

Penelitian kecerdasan buatan pada kualitas air telah banyak dikembangkan sebagai usaha manusia dalam meningkatkan kualitas hidupnya. Studi ini menerapkan jaringan saraf tiruan *atau artificial neural network (ANN)* untuk menemukan model klasifikasi paling optimal dalam penentuan aman tidaknya suatu air. Studi ini menggunakan *generic datasets* yang penulis dapat pada Kaggle. Sejumlah proses dilakukan pada dataset dimulai dari pembersihan data dari *missing values* & *outlier* hingga penyetaraan bobot tiap parameter dengan min-max scaler. Studi ini membandingkan hasil akurasi model ANN dari berbagai skenario dengan jumlah neuron yang berbeda. Pengembangan model untuk menyelesaikan masalah pada studi menggunakan metode *Scaled Conjugate Gradient* sebagai *learning algorithm.* Hasil studi yang diperoleh bervariasi antar skenarionya. Akurasi secara keseluruhan terdapat peningkatan keakuratan ketika jumlah neuron ditingkatkan dari neuron 10 ke 20 dan menurun dari neuron 20 ke 30.

Kata kunci: klasifikasi, kualitas air, jaringan saraf tiruan, akurasi prediksi, jumlah neuron

PENDAHULUAN (INTRODUCTION)

Air merupakan komponen penting dalam kehidupan, tanpa air makhluk hidup tidak dapat bertahan hidup. Di Indonesia sendiri, akses air bersih perlu diperhatikan apakah air tersebut layak untuk digunakan atau tidak dengan cara mengidentifikasi kandungan yang ada didalamnya. Berdasarkan RPJMN 2020-2024 (Bappenas, 2020), pencapaian kinerja akses pelayanan air bersih pada periode pembangunan sebelumnya di Indonesia belum cukup memuaskan. Hal ini ditunjukkan dari akses air minum perpipaan yang baru menjangkau 20,14% dari seluruh rumah tangga di Indonesia pada tahun 2018. Berdasarkan laporan Environmental Performance Index 2022 (Wolf *et al.*. 2022), Indonesia menempati peringkat ke-164 dari 180 negara yang diteliti berdasarkan keberlanjutan lingkungannya dengan skor 28,2 dari 100. Kemudian, salah satu indikator yang dinilai oleh EPI adalah pencemaran air yang terjadi di negara-negara tersebut. Makin tinggi skor yang dimiliki suatu negara maka keberlanjutan lingkungannya semakin baik. Hal ini menunjukkan bahwa Indonesia masih memiliki kualitas air yang masih rendah. Oleh karena itu, penentuan aman atau tidaknya air yang digunakan perlu diperhatikan untuk melakukan identifikasi secara tepat.

Pada penelitiannya, Prambudi dan Febrianti (2022) melakukan pengamatan terhadap parameter yang mempengaruhi kualitas air. Mereka kemudian dapat menentukan rentang kelas untuk tiap parameter kualitas air di Kota Balikpapan dengan *backpropagation artificial neural network*. Namun, penelitian ini tidak dapat menunjukkan seberapa akurat model yang terbentuk. Penelitian serupa dilakukan Singh et al (2008) pada kualitas air sungai Gomti (India) yang menunjukkan bahwa jaringan yang optimal mampu menangkap tren jangka panjang yang diamati untuk variabel kualitas air. Oleh karena itu, diperlukan jaringan optimal sehingga variabel kualitas air dapat diamati dalam jangka panjang. Selanjutnya untuk

menanggapi kekurangan dari penelitian Prambudi dan Febrianti (2022), dapat dilakukan klasifikasi menggunakan *Artificial Neural Network (ANN)* yang dapat dengan mudah mengklasifikasikan kualitas air dengan output yang dapat dipertanggungjawabkan (Sulaiman et al, 2019).

Beberapa jenis metode *machine learning* lainnya juga digunakan sebagai model prediksi kualitas air. Ahmed *et al.* (2019) membangun model *Adaptive Neuro-Fuzzy Inference System* (ANFIS), *Radial Basis Function Neural Networks* (RBF-ANN), dan *Multi-Layer Perceptron Neural Networks* (MLP-ANN) untuk prediksi kualitas air berdasarkan input parameter air berupa *ammoniacal nitrogen* (AN), *suspended solid* (SS) dan pH. Chen *et al.* (2020) mengusulnya *decision tree* (DT), *random forest* (RF) dan *deep cascade forest* (DCF) untuk prediksi kualitas air berdasarkan input parameter air berupa AN, *Dissolved oxygen* (DO), *Chemical oxygen demand* (COD), dan pH. *Decision tree* juga digunakan oleh Lu dan Ma (2020) yang mengembangkan *hybrid decision tree* untuk mengidentifikasi kualitas air berdasarkan DO, PH, *turbidity*, SS, dan suhu air. Bui et al. (2020) mengembangkan 16 algoritma *hybrid* untuk predksi kualitas air berdasarkan parameter COD, *biological oxygen demand* (BOD), SS, PH, DO, AN, dan suhu air.

Berdasarkan literatur-literatur sebelumnya, metode *machine learning* merupakan salah satu metode yang menjanjikan untuk prediksi kualitas air. Pada penelitian ini, penulis melakukan pengembangan dari penelitian Sarkar dan Pandey (2015) yang melakukan analisis penggunaan berbagai skenario dalam proses *training ANN* pada pembangunan model penentuan kualitas air. Data yang digunakan dalam penelitian tersebut merupakan data dari sungai yang berlokasi di hilir kota Mathura, India yang terletak di tepi Sungai Yamuna di negara bagian Uttar Pradesh, India. Penelitian tersebut menggunakan algoritma *feed forward error back propagation*. Data tersebut mencakup data bulanan tentang debit aliran, suhu, pH, kebutuhan oksigen biokimia , dan oksigen terlarut atau *Diluted Oxygen* di tiga lokasi, yaitu Mathura (hulu), Mathura (tengah), dan Mathura (hilir). Penentuan aman atau tidaknya air untuk dikonsumsi dapat dilihat dari konsentrasi oksigen terlarut yang telah digunakan sebagai indikator utama kualitas air sungai. Pada penelitian sebelumnya hasil penelitian dievaluasi menggunakan perhitungan *root mean square error*, koefisien korelasi dan koefisien determinan. Dan didapatkan hasil akurasi yang sangat baik dengan korelasi tinggi yaitu hingga 0.9 antara hasil perhitungan dan hasil prediksi. Sedangkan, pada penelitian ini penulis ingin mengetahui dampak dari penerapan perbedaan jumlah neuron pada perhitungan dan prediksi kualitas air dengan data yang sama pada penilitian sebelumnya. Jumlah neuron yang diterapkan adalah 10, 15, 20 dan 30 neuron.

BAHAN DAN METODE (MATERIALS AND METHODS)

Input data yang digunakan untuk menguji apakah kualitas air layak untuk digunakan pada kasus ini terdiri dari 20 variabel input yang merupakan senyawa kimia yang terkandung di dalam air. Dengan *output* data berupa pengklasifikasian, apakah air yang digunakan aman atau tidak. Data-data atribut yang didapatkan dari Kaggle (Mssmartypants, 2021) ditampilkan pada Tabel 1. Sumber data yang penulis gunakan merupakan data *generation* yang digunakan untuk tujuan akademik.

Tabel 1. Atribut input dan output

No.	Atribut	Tipe Data	Deskripsi
1	aluminium	Numerical	berbahaya apabila lebih dari 2.8
2	ammonia	Numerical	berbahaya apabila lebih dari 32.5
3	arsenic	Numerical	berbahaya apabila lebih dari 0.01 mg/l
4	barium	Numerical	berbahaya apabila lebih dari 2
5	cadmium	Numerical	berbahaya apabila lebih dari 0.005
6	chloramine	Numerical	berbahaya apabila lebih dari 4
7	chromium	Numerical	berbahaya apabila lebih dari 0.1
8	copper	Numerical	berbahaya apabila lebih dari 1.3
9	flouride	Numerical	berbahaya apabila lebih dari 1.5 mg/l
10	lead	Numerical	berbahaya apabila lebih dari 0.015
11	nitrates	Numerical	berbahaya apabila lebih dari 10
12	nitrites	Numerical	berbahaya apabila lebih dari 1
13	mercury	Numerical	berbahaya apabila lebih dari 0.002
14	perchlorate	Numerical	berbahaya apabila lebih dari 56 micrograms per liter

15	radium	Numerical	berbahaya apabila lebih dari 5 picoCuries per liter (pCi/L)
16	selenium	Numerical	berbahaya apabila lebih dari 0.5
17	silver	Numerical	berbahaya apabila lebih dari 0.1
18	uranium	Numerical	berbahaya apabila lebih dari 0.3
output	class attribute	Categorical	0 - not safe, 1 - safe

Sebelum data digunakan untuk pemodelan, perlu dilakukan *data preprocessing*. Dataset awal terdiri dari 7999 data. Kemudian setelah dilakukan *data cleaning* dan *data transformation*, dataset menjadi 7996. Langkah-langkah yang dilakukan adalah sebagai berikut:

1. *Data Cleaning*

a. Menghilangkan *missing values*

Apabila dalam 1 baris terdapat missing values maka baris tersebut dihilangkan.

b. Menguji ada tidaknya *outlier*

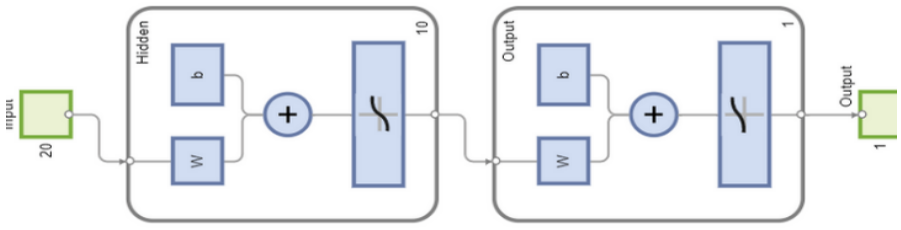
Dari hasil uji minitab menggunakan *Grubb test* didapatkan tidak ada outlier pada setiap variabel input.

2. *Data Transformation*

Karena masing-masing input variabel memiliki range yang berbeda maka *input variable* perlu dilakukan *scaling data* dengan metode *min-max scaler* sehingga semua data berada pada *range* antara 0 sampai 1. Rumus *min-max scaler* adalah sebagai berikut:

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

Model prediksi dibuat dengan bantuan *software Matlab*. Pada penelitian ini penulis ingin melakukan klasifikasi layak atau tidaknya air yang diuji. Metode analisis dilakukan dengan jumlah neuron yang berbeda dengan metode yang sama yaitu *Scaled conjugate gradient.* Selanjutnya dilakukan pengukuran evaluasi kinerja model menggunakan *objective cross-entropy.* Gambar 1 menunjukkan *network architecture* yang digunakan untuk mengklasifikasikan jenis air. *Network architecture* terdiri dari 20 *variable input* dengan jumlah neuron n dan *output* berupa 1 dan 0.

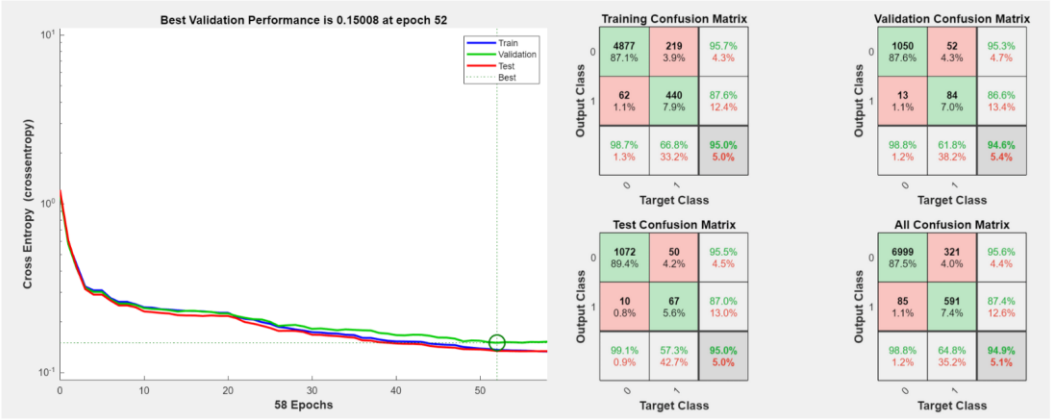


Gambar 1. *Network Architecture*

HASIL DAN PEMBAHASAN (RESULT AND DISCUSSIONS)

Masalah pada penelitian ini termasuk dalam kategori *classification* dikarenakan output hanya memiliki dua kemungkinan yaitu apakah air aman (1) atau tidak aman (0). Analisis dilakukan dengan jumlah neuron yang berbeda dengan *learning algorithm* yang sama yaitu *scaled conjugate gradient* dengan input 7996x20 dan respon 7996x1. Perbandingan *data training, testing,* dan *validation* yang digunakan dalam penelitian ini adalah 70:15:15 dengan 4 kali percobaan yaitu dengan jumlah neuron 10,15,20, dan 30.

Pada jumlah neuron 10 didapatkan bahwa *training model* berhenti di *epoch* 58. Dengan hasil *training* dengan jumlah *observation* 5598 didapat *Cross-entropy* 0,1362 dan *error* 0,0502. *Best validation* bernilai 0,14008 di *epoch* 52. Pengujian *performance* dari *network* dapat dilihat pada *test confusion matrix* yang menunjukkan 1.199 dataset pengujian. Model dengan jumlah neuron 10 dapat membaca data *water not safe* (0) sebanyak 1.082 dengan 1.072 data diklasifikasikan secara tepat atau memiliki akurasi sebesar 99,1%. Kemudian untuk *water safe*, seratus tujuh belas data diklasifikasikan, tetapi terdapat 50 data yang diklasifikasikan sebagai *water not safe* dengan akurasi sebesar 57,3%. Sehingga secara keseluruhan model ini dapat menghasilkan akurasi *overall* sebesar 95%.

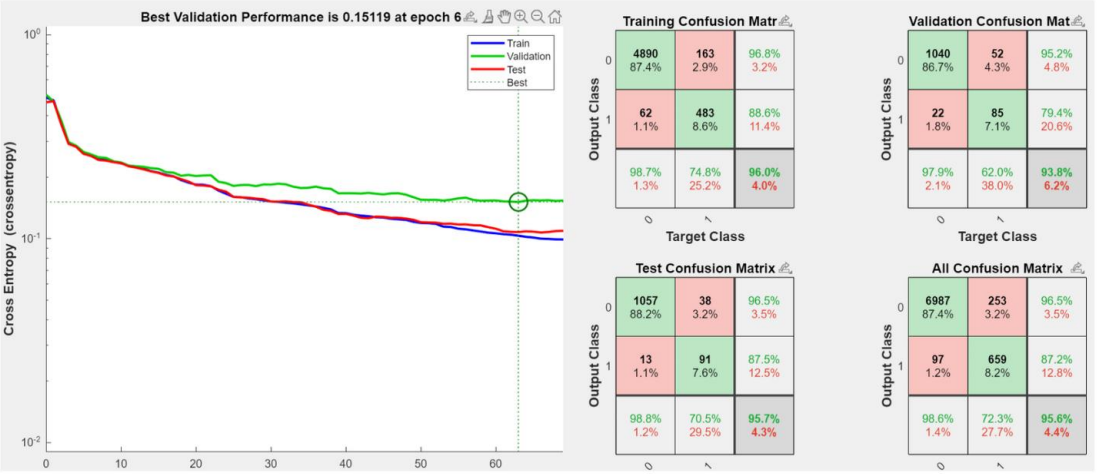


Gambar 2. Tampilan *Plotting* Hasil MATLAB dengan Neuron 10

Pada jumlah neuron 15 didapatkan *training model* berhenti di epoch 72. Hasil *training* dengan jumlah *observation* 5598 didapat *Cross-entropy* 0.098 dan *error* 0.0357. *Best validation* bernilai 0.093732 di epoch 66. Ilustrasi *confusion matrix* menunjukkan 1.199 dataset pengujian. Model ini dapat membaca data *water not safe* (0) sebanyak 1087 dengan 1066 data diklasifikasikan secara tepat atau memiliki akurasi sebesar 98,1%. Kemudian untuk *water safe* (1), seratus dua belas data diklasifikasikan namun terdapat 37 data yang diklasifikasikan sebagai *water not safe* dengan akurasi sebesar 67%. Sehingga secara keseluruhan model ini dapat menghasilkan akurasi *overall* sebesar 95,2%.

Gambar 3. Tampilan *Plotting* Hasil MATLAB dengan Neuron 15

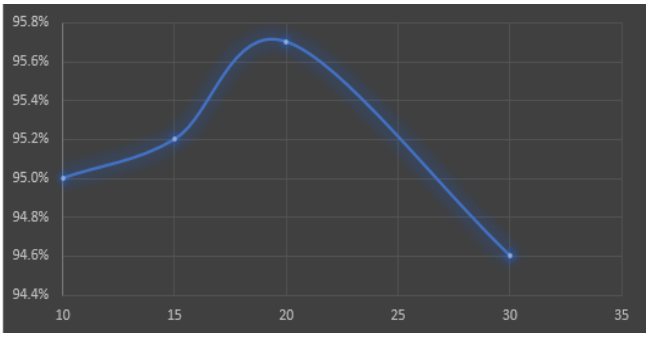
Pada jumlah neuron 20 *training model* berhenti di *epoch* 69. Hasil *training* dengan jumlah *observation* 5.598 didapat Cross-entropy 0,1033 dan error 0,0402. Best validation bernilai 0,15119 di epoch 63. Ilustrasi *confusion matrix* menunjukkan 1.199 dataset pengujian. Model ini dapat membaca data *water not safe* (0) sebanyak 1.070 dengan 1.057 data diklasifikasikan secara tepat atau memiliki akurasi sebesar 98,8%. Kemudian untuk *water safe*, seratus dua puluh sembilan data diklasifikasikan namun terdapat 38 data yang diklasifikasikan sebagai *water not safe* dengan akurasi sebesar 70,5%. Sehingga secara keseluruhan model ini dapat menghasilkan akurasi *overall* sebesar 95,7%.



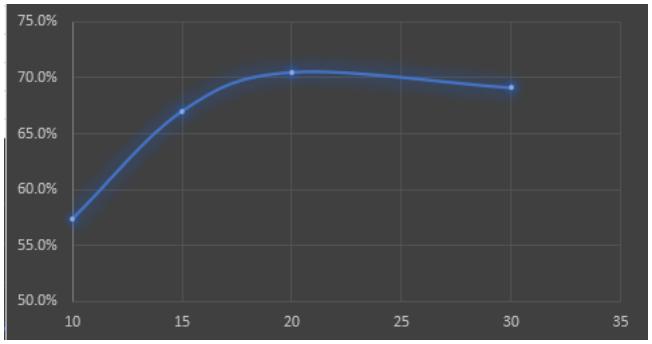
Gambar 4. Tampilan *Plotting* Hasil MATLAB dengan Neuron 20

Pada jumlah neuron 30 didapat hasil *training model* dengan jumlah *observation* 5.598 didapat *Cross-entropy* 0,0763 dan *error* 0,0277 dan *training* berhenti di *epoch* 92. *Best validation* bernilai 0,11705 di *epoch* 86. Ilustrasi *confusion* matrix menunjukkan 1.199 dataset pengujian. Model ini dapat membaca data *water not safe* (0) sebanyak 1.063 dengan 1.038 data diklasifikasikan secara tepat atau memiliki akurasi sebesar 97,6%. Kemudian untuk *water safe*, seratus tiga puluh enam data diklasifikasikan namun terdapat 42 data yang diklasifikasikan sebagai *water not safe* dengan akurasi sebesar 69,1%. Sehingga secara keseluruhan model ini dapat menghasilkan akurasi *overall* sebesar 94,4%.

Dari hasil keempat percobaan yang telah dilakukan maka didapatkan bahwa terdapat peningkatan akurasi *overall* ketika jumlah neuron ditingkatkan dan menurun dari neuron 20 ke 30. Peningkatan akurasi *overall* tidak terlalu signifikan yang secara lebih lanjut dapat dilihat di Gambar 5.



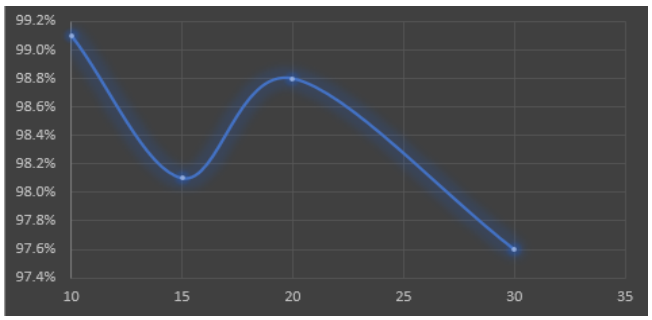
Gambar 5. Akurasi *Overall*



Gambar 6. Akurasi *Water Safe*

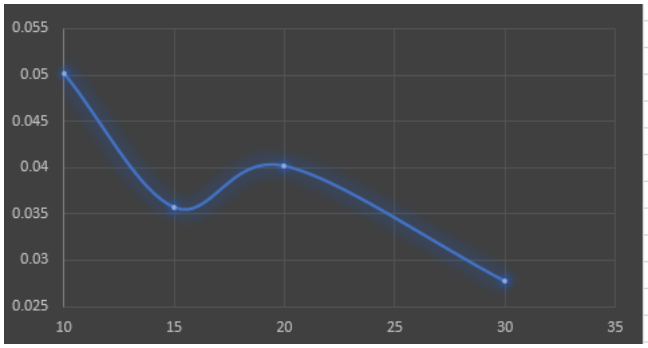
Akurasi dari sisi *water safe* dapat terlihat pada Gambar 6. Gambar tersebut menunjukkan peningkatan akurasi yang signifikan dari neuron 10 ke 20 dan terjadi penurunan akurasi *Water Safe* dari neuron 20 ke 30 walaupun penurunan akurasi tidak terlalu signifikan.

Sedangkan untuk akurasi dari sisi *water not safe*, dapat disimpulkan secara *general* peningkatan jumlah neuron akan mengurangi akurasi *water not safe*, seperti yang diperlihatkan pada Gambar 7. Penurunan akurasi tersebut tidak terlalu signifikan.

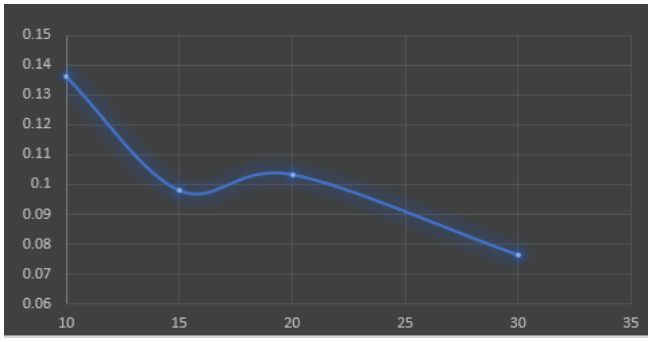


Gambar 7. Akurasi *Water Not Safe*

Dari sisi pengukuran error training dapat disimpulkan secara general peningkatan jumlah neuron akan mengurangi error training, penurunan error training tidak terlalu signifikan, seperti yang ditunjukkan oleh Gambar 8.



Gambar 8. *Error Training*



Gambar 9. *Cross Entropy Training*

Pada nilai cross-entropy didapatkan secara general peningkatan jumlah neuron akan mengurangi nilai Cross-entropy training, penurunan Cross-entropy training tidak terlalu signifikan, seperti yang ditunjukkan pada Gambar 9.

KESIMPULAN (CONCLUSION)

Pada penelitian ini, analisis terhadap kualitas air dilakukan pada data lingkungan perkotaan. Klasifikasi terhadap kualitas air berupa air yang aman atau tidaknya dilakukan berdasarkan kadar kimiawi yang sudah ditetapkan. Dataset total berjumlah 7999 data dan digunakan perbandingan jumlah data 70:15:15 untuk data *training*, *testing* dan *validation*. Dengan melakukan variasi pada jumlah neuron, penelitian ini dapat mengetahui dampak dari perubahan jumlah neuron terhadap analisis kualitas air menggunakan metode ANN. Dari percobaan yang telah dilakukan, peningkatan akurasi secara keseluruhan tidak terlalu signifikan antara jumlah neuron 10, 15, 20 dan 30. Kemudian, secara lebih detail didapatkan bahwa terdapat peningkatan akurasi pada neuron 10 ke 15 lalu ke 20 dilanjutkan dengan penurunan dari neuron 20 ke 30. Dari hasil pengukuran *error training* maupun *cross entropy*, secara *general* peningkatan jumlah neuron akan mengurangi nilai keduanya walaupun penurunan keduanya tidak terlalu signifikan. Setelah dianalisis lebih lanjut data yang digunakan pada penelitian ini mengalami *imbalanced dataset* yaitu 85% merupakan

data dengan kategori *water not safe* sehingga mempengaruhi hasil akurasi prediksi model. Pada hasil akurasi *water safe* didapatkan akurasi pada rentang 57%-71% sedangkan akurasi pada *water not safe* didapatkan akurasi pada rentang 97.6%-99.1%. Sehingga pada penelitian lanjutan dapat dilakukan analisis lebih lanjut menggunakan data yang lebih seimbang.

DAFTAR PUSTAKA

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., & Elshafie, A. 2019. Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578, 124084.

Bappenas. 2020. Rencana Pembangunan Jangka Menengah Nasional 2020-2024. Peraturan Presiden Republik Indonesia Nomor 18 Tahun 2020.

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 721, 137612.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., ... & Ren, H. 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, 171, 115454.

Mssmartypants. 2021. Water Quality. *Kaggle*. https://www.kaggle.com/datasets/mssmartypants/water-quality

Lu, H., & Ma, X. 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249, 126169.

Prambudi, D. A., & Febrianti, N. 2022. Penerapan Artificial Neural Network pada Prototyping Sistem Monitoring Kualitas Air di Kota Balikpapan untuk Mendukung Balikpapan sebagai Smart City. *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, 16(1), 30-38.

Singh, K. P., Basant, A., Malik, A., & Jain, G. 2009. Artificial neural network modeling of the river water quality—a case study. *Ecological modelling*, 220(6), 888-895.

Sulaiman, K., Ismail, L. H., Razi, M. A. M., Adnan, M. S., & Ghazali, R. 2019. Water quality classification using an Artificial Neural Network (ANN). In *IOP Conference Series: Materials Science and Engineering*, 601(1), 012005. IOP Publishing.

Sarkar, A., & Pandey, P. 2015. River Water Quality Modelling Using Artificial Neural Network Technique. *Aquatic Procedia*, 4, 1070-1077. doi: 10.1016/j.aqpro.2015.02.

Wolf, M. J, Emerson, J. W., Esty, D. C., de Sherbinin, A., Wendling, Z. A., et al. 2022. 2022 Environmental Performance Index. *New Haven, CT: Yale Center for Environmental Law & Policy*. epi.yale.edu