Sites: https://www.jurnal-umbuton.ac.id/index.php/Pencerah
DOI: https://doi.org/10.35326/pencerah.v8i2.2251

Open Access

SANG PENCERAH
Jurnal Ilmiah Universitas Muhammadiyah Buton

E-ISSN: 2655-2906, P-ISSN: 2460-5697 Volume 8, No 2, Tahun 2022

Mengatasi Ketimpangan Data Deep Neural Network dengan Pelipatan Fitur Data Klasifikasi Spektroskopi Darah

Widi Hastomo[1]*, Adhitio Satyo Bayangkari Karno[2], Sutarno[3], Dodi Arif[2], Eka Sally Moreta[3], Sudjiran[3]

[1]Teknologi Informasi, ITB Ahmad Dahlan Jakarta, Indonesia

[2]Manajemen, Universitas Gunadarma, Indonesia

[3]Sistem Informasi, STMIK Jakarta STI&K, Indonesia

*Korespondensi: widie.has@gmail.com

Info Artikel

Diterima 30 *April 2022*

Disetujui 17 *Mei 2022*

Dipublikasikan 28 *Mei 2022*

Keywords:
Ketimpangan Data; Deep Neural Network; Spektroskopi Darah; Kecerdasan Buatan

Abstrak

Permasalahan utama dalam penelitian ini adalah ketimpangan data masukan menghasilkan dampak negatif yang signifikan terhadap hasil prediksi dari model Deep Neural Network (DNN). Kemampuan klasifikasi DNN sangat akurat hanya untuk dataset yang berimbang, namun DNN pada awalnya tidak di rancang untuk menangani ketimpangan data. Ketimpangan data merupakan hal yang sering dijumpai dalam dunia nyata, menjadikan ini sebagai tantangan besar dalam prediksi klasifikasi menggunakan model DNN. Penelitian ini berfokus untuk memprediksi tingkat kandungan kolesterol tinggi, kolesterol rendah dan hemoglobin, menggunakan data kasus di kompetisi Zindi Blood Spectroscopy Classification Challenge. Dengan melakukan analisa data, cleansing outlier, fine tunning, model neural network, jaringan pengelompokan data target dengan kategori sejenis, urutan pemrosesan, pemilihan nilai pelipatan (7 pelipatan) yang tepat terhadap data input train dan data test serta epoch 60, dapat meningkatkan hasil nilai score prediksi yang cukup tinggi sebesar 0.94594.

Abstract

The key issue in this study is the disparity in input data, which has a considerable detrimental impact on the Deep Neural Network (DNN) model's prediction outcomes. Only for balanced datasets is DNN's classification capabilities particularly accurate, however DNN was not built to manage data inequality. Inequality in data is a common occurrence in the actual world, which makes using the DNN model to predict categorization a difficult task. This study uses case data from the Zindi competition Blood Spectroscopy Classification Challenge to predict the levels of high cholesterol, low cholesterol, and hemoglobin. It is important to improve the result value by performing data analysis, cleaning outliers, fine tuning, neural network models, grouping target data networks with similar categories, processing sequences, and selecting the correct value for the multiples (7 folds) of the train input data and test data, as well as epoch 60. 0.94594 is a rather high prediction score.

1. Pendahuluan

Dunia medis telah banyak menggunakan teknologi *Artificial Intelligence (AI)*, selain akurasi yang tinggi (Hastomo, 2021b, 2021a; Hastomo, Bayangkari Karno, Kalbuana, Meiriki, & Sutarno, 2021; Karno, A. S. B., & Hastomo, 2020; Satyo, Karno, Hastomo, Efendi, & Irawati, 2021), juga dengan memanfaatkan AI diagnosa dan penanganan menjadi lebih mudah, cepat, berbiaya murah (Hastomo, 2021a, 2021b; Hastomo & Bayangkari, 2021; Satyo et al., 2021), mampu menjangkau daerah pelosok (daerah jauh dari pusat kesehatan) dan sangat membantu para medis terutama dalam era pandemi covid dimana kapasitas medis dan jumlah pasien sangat tidak berimbang. Agar AI dapat diterapkan dalam dunia medis tentu sebelumnya telah melalui proses *machine learning* (ML) dengan menggunakan data dalam jumlah yang cukup memadai untuk proses *training* (Karno, Hastomo, & Wardhana, 2020). Hasil dari proses *training* adalah suatu model dengan akurasi tinggi yang dapat digunakan untuk memprediksi penyakit tertentu, kandungan hemoglobin, tekanan darah, kolesterol dan diagnosa lainnya dalam tubuh manusia.

Eksperimen ini berdasarkan keikutsertaan dalam mengikuti salah satu kompetisi di Zindi ("Zindi Afica," n.d.) yaitu *Blood Spectrocopy Classification Challenge*. Oleh karena itu data yang dipergunakan berasal dari kompetisi tersebut. Zindi adalah komunitas *data scientist* terbesar di afrika selatan, komunitas ini dibentuk bertujuan untuk menjawab berbagai macam tantangan masalah di dunia dengan mengadakan berbagai macam kompetisi terbuka menggunakan *Machine Learning (ML)* dan *Artificial Intelligence (AI).* Dengan Zindi, afrika selatan berusaha menunjukkan bakat dan pengetahuan *data science* kepada dunia.

Penelitan tentang spektroskopi darah telah dilakukan oleh (Chen, F. et al., 2022) menggunakan *binary classifications* pada sampel kanker ovarium. Penelitian yang telah dilakukan oleh (Ciobanu, C. et al., 2022) pada spektroskopi darah untuk mendeteksi kanker prostat, kanker antigen, kanker ovarium. Sedangkan penelitian yang dilakukan oleh (Titus, J., Wu, A.H.B., & Biswal, 2022) pengembangan dan validasi awal pemantauan *transdermal non-invasif* dari *troponin-I* jantung untuk mendeteksi peningkatannya menggunakan spektroskopi darah. Penelitian yang dilakukan oleh (Fonseca, Pereira, Honorato, Bro, & Pimentel, 2022) menggunakan NIR pada spektroskopi darah dalam penyidikan tindak pidana pada jejak darah manusia di tempat kejadian perkara.

Hasil penelitian (Liu, Wang, Zhou, & Xiong, 2022) menggunakan spektroskopi raman pada sidik jari dapat digunakan untuk mengkarakterisasi berbagai sampel biologi dan kimia, dengan menggunakan tujuh model klasifikasi ML. Penelitian pada pemantauan kadar glukosa darah untuk deteksi diabetes dengan spektroskopi optik menggunakan *Ensemble* ML (Aloraynan, A., Rassel, S., Xu, C., & Ban, 2022). Penelitian yang dilakukan oleh (Guleken, Z. et al., 2022) untuk pengukuran level antibody pada penderita covid-19 menggunakan metode spektroskopi dan ML. Pemantauan glukosa dalam darah menggunakan NIR telah dilakukan oleh (Parab, J., Sequeira, M., Lanjewar, M., Pinto, C., & Naik, 2022) menggunakan algoritma *Partial Least Square Regression (PLSR)* dan *Backpropagation Artificial Neural Network (BP-ANN).*

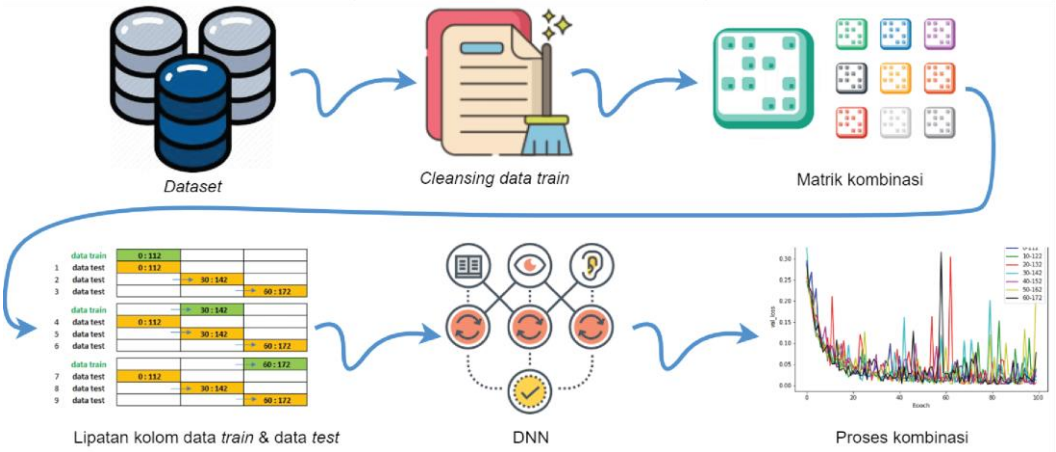
Tujuan dari eksperimen ini adalah membangun model *neural network* yang nantinya diharapkan dapat dipergunakan untuk mengklasifikasikan tingkat

kandungan kimia tertentu dari sampel data *spectroscopic*. Analisa spektrum dilakukan dengan memberikan seberkas cahaya dengan panjang gelombang tertentu kearah sampel darah, cahaya yang datang ke sampel sebagian akan dipantulkan *(reflected)* dan sebagian lagi akan di serap *(absorbed)* berdasarkan struktur molekul sampel darah. Banyaknya cahaya yang diserap bergantung pada besar panjang gelombang sumber cahaya yang digunakan. Proses pengambilan data ini menggunakan rentang panjang gelombang (frekuensi) yang dinamakan dengan *data spectrum*. *Spectral data* menggunakan rentang panjang gelombang *NIR (Near Infra-Red)* berkisar 350 nm-1350 nm. Berbeda dengan panjang gelombang lainnya, NIR mempunyai penetrasi tertinggi dan mampu menembus lebih dalam ke jaringan sample darah (Rizevsky, Zhaliazka, Dou, Matveyenka, & Kurouski, 2022).

Hasil dari eksperimen ini diharapkan dapat menciptakan tingkat kesehatan baru pada kebanyakan orang dengan menjadikan tes darah sebagai komoditi dan prosedur yang dilakukan tanpa usaha berulang-ulang dan akan sangat membantu ilmuwan medis untuk membuat kemajuan menuju analisa darah yang *non-invasive* (tanpa memasukkan alat medis kedalam tubuh). Model yang telah terbentuk akan dapat dipergunkan untuk menganalisa darah dalam waktu kurang dari 1 menit. Hanya dengan menyorotkan cahaya akan diperoleh informasi kandungan kolesterol, kadar vitamin dalam darah dan sebagainya. *Score* yang mendekati nilai 1 dijadikan ukuran capaian maksimum dari keberhasilan.

2. Metode Penelitian

Proses eksperimen ini diilustrasikan pada Gambar 1, diawali dengan proses unduh *dataset* dari Zindi, dilanjutkan dengan proses pembersihan data *train*, matrik kombinasi, lipatan kolom pada data *train* dan *testing*, implementasi model DNN dilanjutkan dengan proses kombinasi data *train* dan *testing*.



Gambar .1 Alur Eksperimen

Data *training* dan *testing* diperoleh dengan cara mengunduh data yang telah disediakan oleh Zindi. Spektrum data *absorbance* dipergunakan untuk mempredksi target dengan level hdl_cholesterol (kolesterol tinggi), hemoglobin dan Idl_cholesterol (kolesterol rendah) yang di kategorikan dalam bentuk *"low", "ok"* dan *"high"* (Aggarwal, 2018). Spektrum *absorbance* dilakukan sebanyak 170 jenis dengan rentang frekuensi NIR yang berbeda. Tiap jenis *absorbance* dilakukan pengambilan data sebanyak 60 kali untuk tiap sampel darah. Selain data

absorbance, ada 2 kolom data tambahan yaitu *temperature* dan *humidity*. Jadi jumlah kolom (label) untuk data *train* sebanyak 172 kolom dan target sebanyak 3 kolom *("low", "ok",* dan *"high"),* dengan masing berisi 29.160 baris data (Gambar 2).

```
train = pd.read_csv(traincsv)
print(len(train))
train.head()
train
```

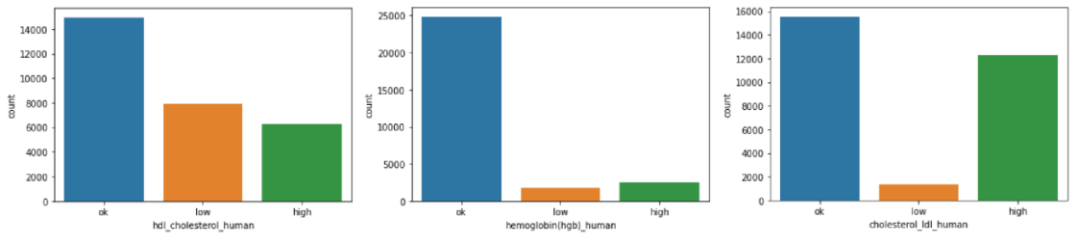
```
29160
```

	id	absorbance0	absorbance1	...	absorbance169	temperature	humidity	hdl_cholesterol_human	hemoglobin(hgb)_human	cholesterol_ldl_human
0	0	0.520883	0.528200		1.278293	39.27	44.17	low	ok	high
1	1	0.529109	0.534852	...	1.267465	39.39	44.09	low	ok	high
2	2	0.528434	0.532036		1.266464	39.50	44.00	low	ok	high
3	3	0.530528	0.531880	...	1.342224	39.60	43.89	low	ok	high
4	4	0.527530	0.536424		1.216146	39.70	43.83	low	ok	high
...	...	...	...	...	...	...	...	...	...	...
29155	36075	0.506799	0.509514		1.230291	45.56	37.06	low	ok	high
29156	36076	0.510311	0.506805	...	1.277773	45.58	37.06	low	ok	high
29157	36077	0.511655	0.513587		1.224521	45.58	37.01	low	ok	high
29158	36078	0.512617	0.514809	...	1.242985	45.60	36.97	low	ok	high
29159	36079	0.523990	0.522537		1.307333	45.61	36.98	low	ok	high

29160 rows × 178 columns

Gambar 2. Ekstraksi Data *Train*

Ukuran data yang tergolong besar ini cenderung mengakibatkan proses komputasi *training* secara keseluruhan menjadi berat. Untuk mengatasi ini, dalam proses *training* diperlukan pengaturan parameter dalam menentukan model *neural network*. Model yang tepat akan dapat menggunakan data tanpa *noise, history training* dengan *loss* rendah dan akurasi tinggi, dan terhindar dari *overfitting*. Selain data dalam jumlah besar juga terlihat sebaran data historgram target "*low", "ok"* dan *"high*" dari setiap data target sangat tidak seimbang (Gambar 3 dan 4). Untuk mengatasi ketimpangan data dilakukan teknik lipatan data *feature* untuk *training* dan data *testing* agar memperoleh nilai *score data test* yang tinggi (mendekati 1).



Gambar 3. Sebaran Data *Train* Target

Data *train* dan data *test* memiliki 170 kolom data *absorbance* dan 2 kolom untuk *temperature* dan *humidity* dengan jumlah data sebanyak 3.660 baris. Kolom pertama data *test* berlabel *"Reading_ID"* menyatakan identitas dari sampel darah yang akan diprediksi tingkat hdl_cholesterol, hemoglobin dan Idl_cholesterol (Gambar 4).

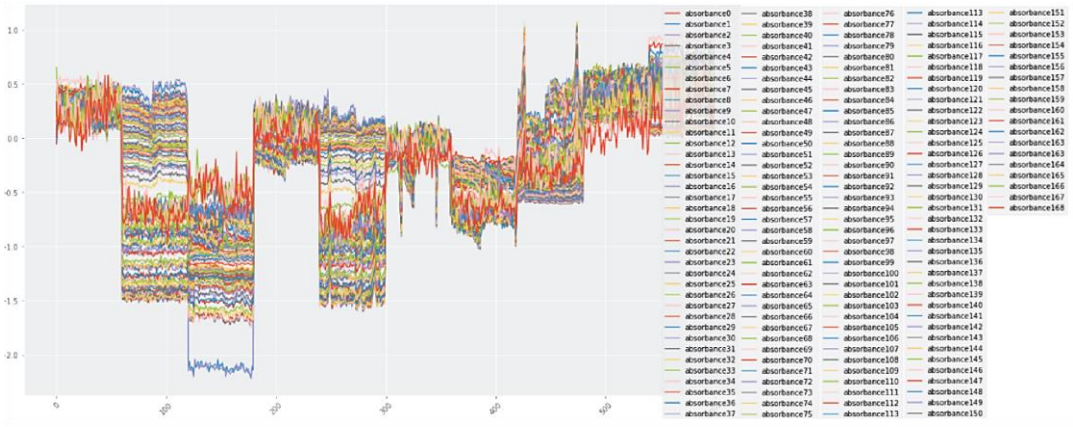
```
test_na=test.iloc[:,0:173].copy()
test_na
```

	Reading_ID	absorbance0	...	absorbance165	absorbance166	absorbance167	absorbance168	absorbance169	temperature	humidity
0	ID_2982	0.226909	...	-0.785705	-0.658222	-0.760531	-0.877281	-0.320136	37.09	27.22
1	ID_2982	0.224089	...	-0.723267	-0.890013	-0.739458	-0.752447	-0.369814	37.20	27.17
2	ID_2982	0.220644	...	-0.806759	-0.715664	-0.881390	-0.773883	-1.007555	37.30	27.07
3	ID_2982	0.404754	...	-1.052284	-0.836271	-0.546718	-1.068405	-0.791829	37.38	26.98
4	ID_2982	0.291962	...	-0.844797	-0.770166	-0.847213	-1.249808	-0.662884	37.48	26.89
...	...	...	...	...	...	...	...	...	...	...
3655	ID_3637	-0.299270	...	0.904249	0.955975	1.025743	1.262447	0.331985	40.63	36.33
3656	ID_3637	-0.282332	...	1.226686	0.884973	1.287536	0.620149	1.389364	40.66	36.33
3657	ID_3637	-0.117899	...	1.237441	0.965974	1.255336	1.021060	1.088025	40.70	36.23
3658	ID_3637	-0.021027	...	0.976432	1.025610	0.994409	1.049926	0.737960	40.72	36.23
3659	ID_3637	0.278158	...	1.012944	0.955143	0.851547	0.835106	0.800761	40.75	36.23

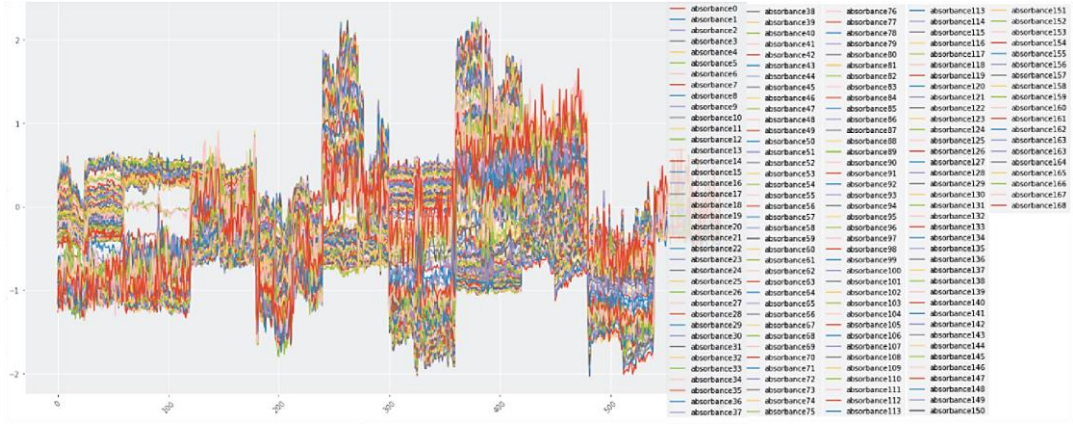
3660 rows × 173 columns

Gambar 4. Ekstraksi Data *Test*

Dengan *plotting* seluruh *absorbance train* dan *test* dari 600 data (Gambar 5 dan 6) terlihat kisaran data *train absorbance* berada diantara -2 dan 1, sedangkan untuk data *test absorbance* berada diantara rentang -2 dan 3.



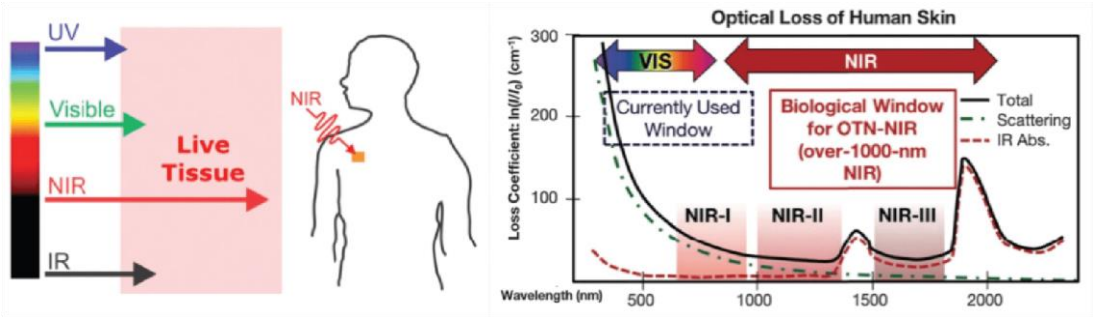
Gambar 5. *Plotting* 170 *absorbance* untuk 600 data *train*



Gambar 6. *Plotting* 170 *absorbance* untuk 600 data *test*

3. Hasil dan Pembahasan

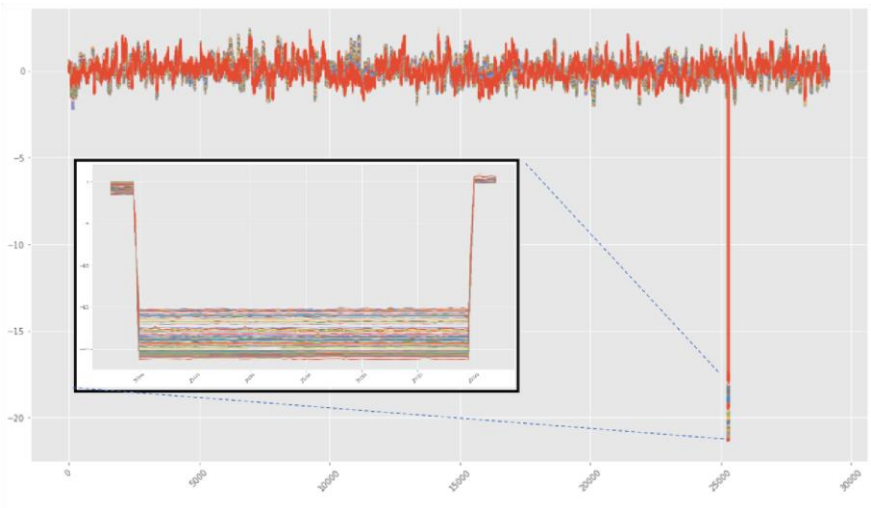
Penyerapan minimal NIR oleh air dan hemoglobin, dikombinasikan dengan hamburan marginal serta dibandingkan dengan cahaya visual atau ultraviolet, memungkinkan cahaya ditransmisikan secara efektif ke seluruh tubuh tanpa gangguan, menjadikannya media yang sempurna untuk pencitraan jaringan dalam (Hemmer, E., Benayas, A., Légaré, F., & Vetrone, 2016). Kemampuan NIR yang penting ini, digambarkan secara visual pada Gambar 7, membuat kasus yang kuat untuk efektivitas aplikasi NIR dalam menjelajahi jendela biologis (Wang, L. V. & Wu, 2012). Banyak ahli belum menggunakan teknik ML untuk analisa *spectral* karena risiko *overfitting* yang tinggi. Hasil dari eksperimen ini adalah suatu model yang mampu mengeneralisir dengan baik tanpa *overfitting*.



Gambar 7. *NIR Biological Window*

3.1 *Cleansing data train*

Proses *plotting data train* secara keseluruhan diketahui adanya sekelompok data yang berada di luar *range data absorbance* umumnya (Gambar 8). Karena data *outlier* ini dapat menurunkan nilai akurasi *training*, maka kelompok data ini tidak akan digunakan dalam proses *training deep learning*. Dengan melakukan perubahan skala sehingga *plotting* menjadi lebih besar di *range data outlier*, maka dapat diketahui kelompok data *outlier* berada dalam rentang baris diantara 25.260-25.320. Dengan tidak mengunakan data *outlier* maka data yang digunakan untuk *training* berkurang semula 29.160 menjadi 29.090 baris (Gambar 9).



Gambar 8. Plotting keseluruhan data train menunjukan adanya *outlier*

	absorbance0	absorbance1	...	absorbance169	temperature	humidity	hdl_cholesterol_human	hemoglobin(hgb)_human	cholesterol_ldl_human
0	0.127368	0.199785	...	0.203795	39.27	44.17	low	ok	high
1	0.180694	0.257928	...	0.127738	39.39	44.09	low	ok	high
2	0.176315	0.233318	...	0.120704	39.50	44.00	low	ok	high
3	0.189888	0.231955	...	0.652837	39.60	43.89	low	ok	high
4	0.170459	0.271669	...	-0.232726	39.70	43.83	low	ok	high
...	...	...	...	...	...	...	...	...	...
29155	0.036079	0.036459	...	-0.133368	45.56	37.06	low	ok	high
29156	0.058844	0.012776	...	0.200140	45.58	37.06	low	ok	high
29157	0.067553	0.072058	...	-0.173896	45.58	37.01	low	ok	high
29158	0.073790	0.082743	...	-0.044205	45.60	36.97	low	ok	high
29159	0.147512	0.150287	...	0.407766	45.61	36.98	low	ok	high

29090 rows × 177 columns

Gambar 9. Data *train* setelah proses

3.2 Matrik kombinasi data *train* target

Jumlah data target 3 yaitu, (hdl_cholesterol (A), hemoglobin_hgb (B) dan ldl_cholesterol (C)) berisi kombinasi dari "*low", "high",* dan *"ok*". Berdasarkan matrik yang dibuat untuk mengetahui kombinasi data target yang terbentuk dari data *train* adalah sebanyak 24 kombinasi (Gambar 10). Dari matrik juga dapat diketahui kuantitas jenis kombinasi data *train* target, dimana data 3 target berisi *"ok","ok","ok"* adalah yang mendominasi (6.300 baris data) dari seluruh data yang ada.

ABC	Jumlah
ok,ok,ok	6300
ok,ok,high	5760
low,ok,ok	4560
high,ok,high	3000
high,ok,ok	2160
low,ok,high	1980
ok,high,ok	1020
ok,high,high	600

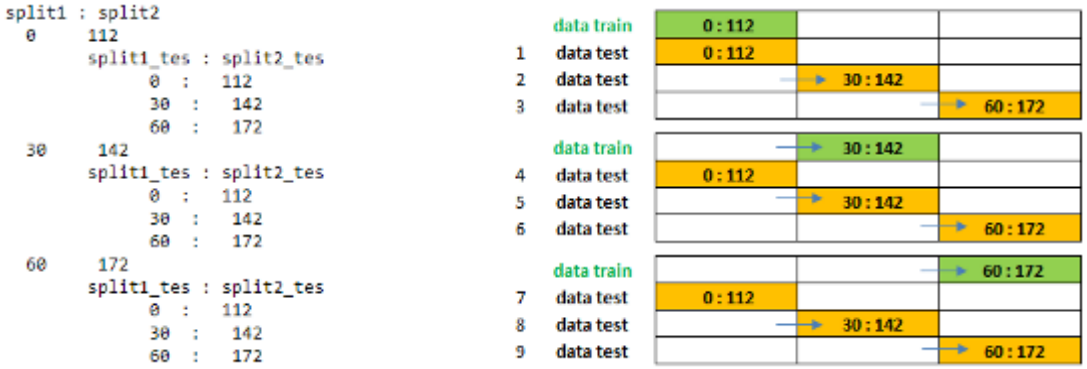
ABC	Jumlah
ok,low,ok	540
ok,ok,low	420
low,high,ok	420
high,low,high	420
low,ok,low	360
high,ok,low	300
low,low,ok	240
high,low,ok	180

ABC	Jumlah
ok,low,high	180
low,low,high	180
high,high,ok	120
low,high,low	120
high,high,high	120
low,high,high	60
ok,high,low	60
ok,low,low	60

Gambar 10. Matrik 24 kombinasi data *train* target

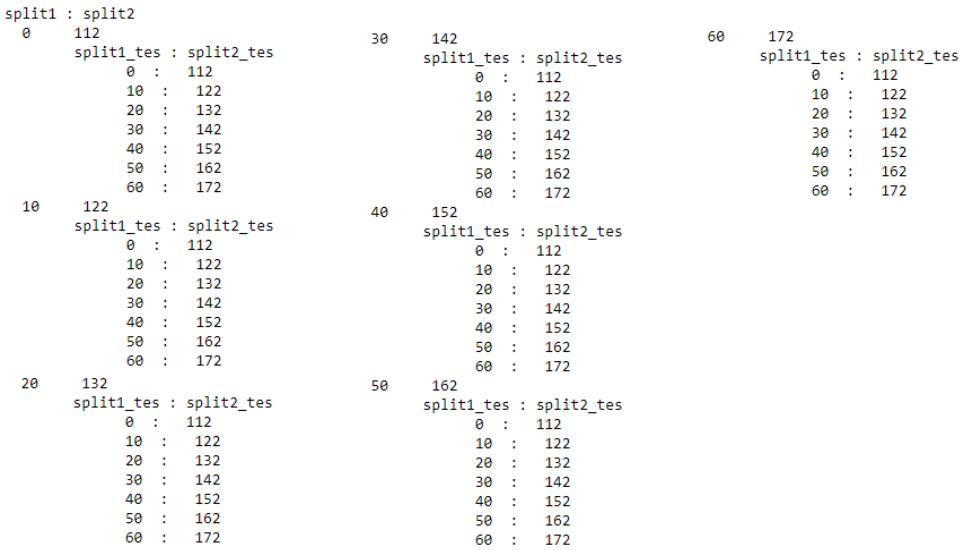
3.3 Membuat lipatan kolom data *train* dan kolom data *test*

Eksperimen ini memiliki 172 kolom data *absorbance* di bagi menjadi beberapa lipatan. Sebagai ilustrasi misal untuk membentuk 3 lipatan dengan *range* lipatan kolom 0-112 akan diperlukan variabel geser sebesar 30 *step*. Dari program yang telah di buat (Gambar 11) dengan mengganti parameter lebar_fitur 112 dan geser 30 menghasilkan 3 lipatan dengan lipatan pertama adalah kolom data *train* 0-112 dan 3 variasi data tes yang bergeser sebesar 30 *step*, untuk lipatan ke dua kolom data *train* bergeser 30 *step* kekanan (30-142) dengan 3 variasi data untuk *testing*, dan lipatan terakhir kolom 60-172 dengan 3 variasi data *testing*. Variasi data tes yang terbentuk untuk setiap lipatan adalah (0-112, 30-142, 60-172).



Gambar 11. Bentuk 3 lipatan kolom data *train* dan *test*

Proses selanjutnya dalam eksperimen ini dilakukan sebanyak 7 lipatan dengan mengubah parameter geser menjadi 10 (Gambar 12), perlu juga untuk dipertimbangkan bahwa semakin banyak lipatan akan memerlukan proses iterasi *training* lebih banyak dan waktu yang diperlukan untuk proses *trainning* akan menjadi lebih lama.



Gambar 12. Bentuk 7 lipatan kolom data *train* dan *test*

3.4 Model *Deep Neural Network*

Proses untuk memperoleh hasil skor yang baik diperlukan penyesuaian *(fine tuning)* baik dari arsitektur model juga dari pengaturan *hyperparameter* nya. Dalam merancang suatu arsitektur, model DNN dapat mengalami *underfitting* atau *overfitting*. Khusus untuk *overfitting* hasil prediksi dapat berada jauh di luar jangkauan data *training* (Feng, N., Wang, F., & Qiu, 2016). Jika jumlah *neuron* yang terlalu sedikit tidak akan mampu menangkap pola komplek diantara variabel *input* dan target, namun jika *neuron* berlebihan didalam *hidden layer* akan mengalami *over parameter,* menuju ke *overfitting* (Okut et al., 2013). Metode *Early stopping* digunakan untuk mencegah *overfitting*. Metode ini akan memberhentikan proses iterasi pada saat *training* jika nilai akurasi sudah tidak mengalami perubahan (Wang, H. J., Ji, F., Leung, C. S., & Sum, 2019).

Model *neural network* dalam penelitian ini menggunakan 15 *layer* mulai dari 300 *node* di *hidden layer* pertama dan seterusnya hingga 1 *node output* di *layer* terakhir (Gambar 13). Aktivasi *ReLU (Rectifier Linier Unit)* disisipkan disetiap *hidden layer*, dan aktivasi *sigmoid* di bagian akhir *hidden layer*, pengukuran *accuracy* dan *loss* menggunakan *binary crossentropy*. Untuk parameter *batch size* adalah 32, berfungsi membagi *dataset* menjadi bagian kecil. *Batch size* mampu meminimalisir *gradient* agar tidak masuk dalam *local optimum*.

```
def nn(x,y):
    x_train, x_test, y_train, y_test = train_test_split(
                        x, y, test_size=0.20, random_state=42)
    model = Sequential()
    model.add(Dense(300, input_dim=x.shape[1], activation='relu',
                    kernel_initializer='random_normal'))
    model.add(Dense(225,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(200,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(175,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(150,activation='relu',kernel_initializer='random_normal'))
    #model.add(Dense(135,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(120,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(100,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(75,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(50,activation='relu',kernel_initializer='random_normal'))
    #model.add(Dense(25,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(20,activation='relu',kernel_initializer='random_normal'))
    #model.add(Dense(10,activation='relu',kernel_initializer='random_normal'))
    model.add(Dense(1,activation='sigmoid',kernel_initializer='random_normal'))
    model.compile(loss='binary_crossentropy',
                  optimizer=tensorflow.keras.optimizers.Adam(),
                  metrics =['accuracy'])
    monitor = EarlyStopping(monitor='val_loss', min_delta=delta,
        patience=pati, verbose=1, mode='auto', restore_best_weights=True)
    model.fit(x_train,y_train,validation_data=(x_test,y_test),
              callbacks=[monitor],batch_size=bs,verbose=2,epochs=e)

    return(model)
```

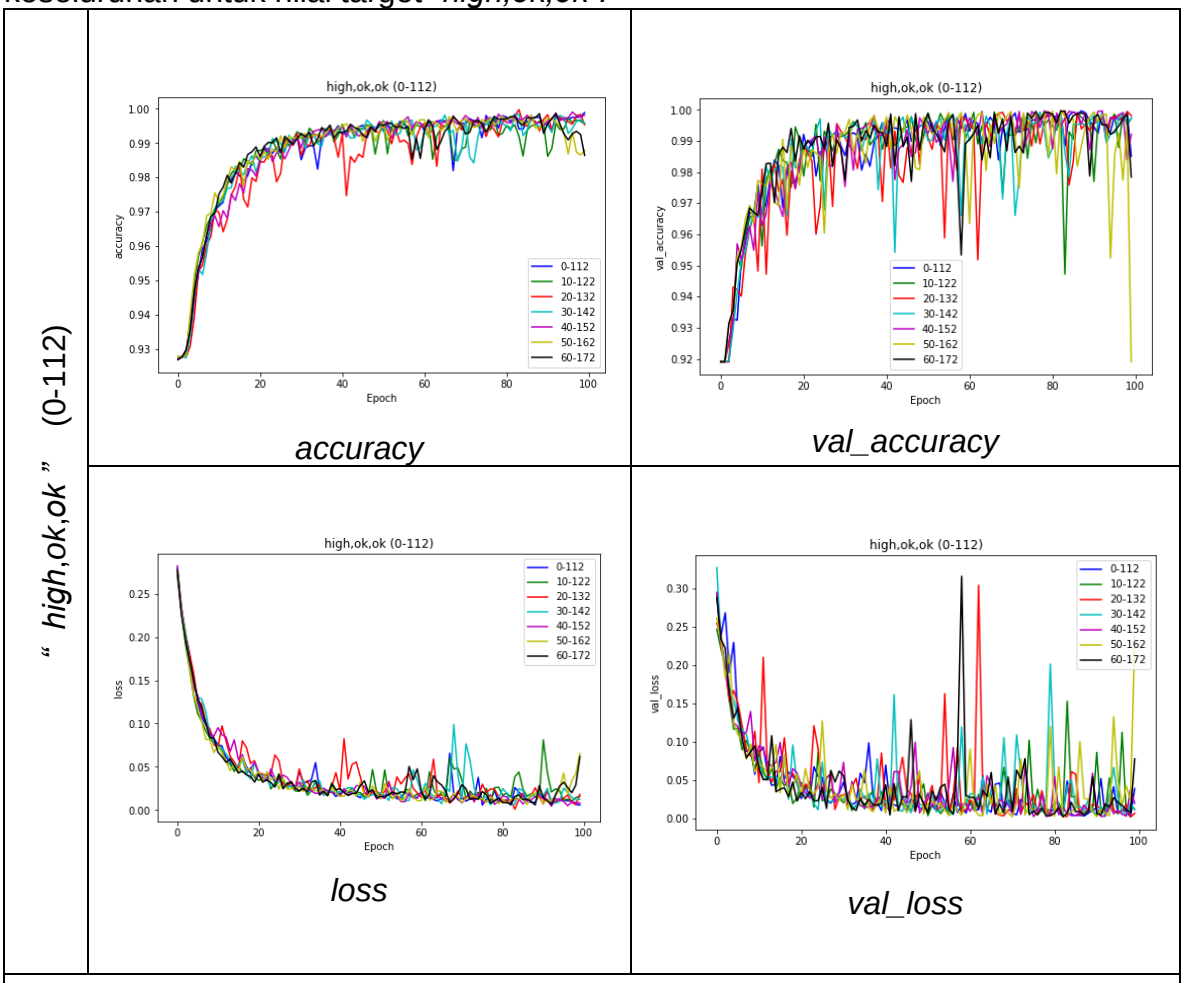
Gambar 13. Arsitektur *Deep Neural Network*

3.5 Urutan proses kombinasi *data train* dan *data test*

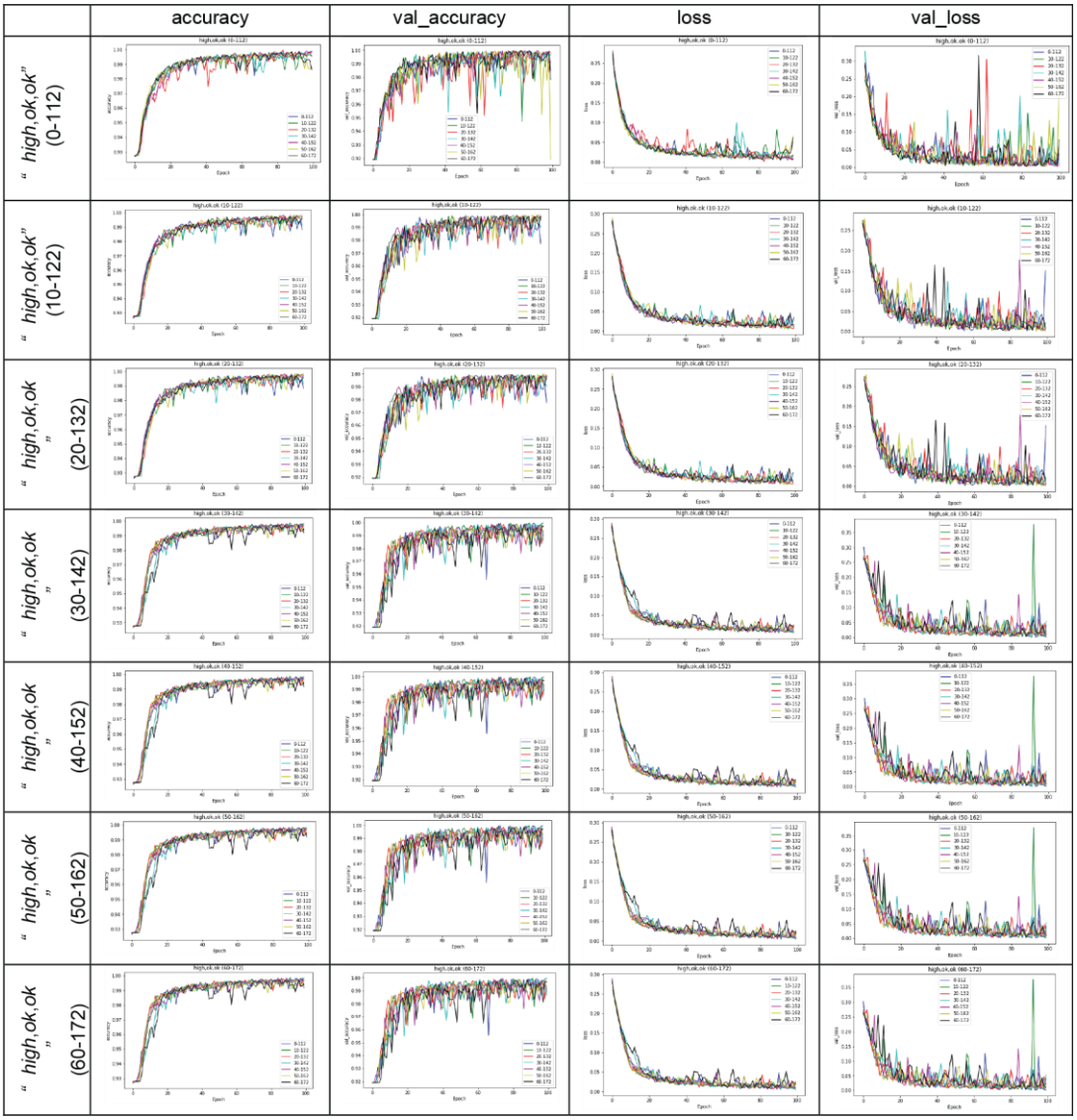
Banyak lipatan pada proses *training* dan *testing* akan memerlukan waktu proses lama untuk memperoleh hasil akhir prediksi. Agar waktu proses lebih efisien maka penentuan urutan jenis kombinasi target untuk proses *training* akan mengurangi waktu proses *trainning* dan sangat berpengaruh pada nilai *score* hasil *testing* yang diperoleh. Berdasarkan matrik kombinasi target *"ok,ok,ok"* memiliki jumlah 6.300 baris data tertinggi dari kombinasi lainnya (Gambar 10) dan berdasarkan pengamatan *trial* jika hasil *testing* semua berisi kombinasi "ok,ok,ok" maka akan memperoleh *score* dari Zindi sekitar 0.70 (70%). Artinya 70% prediksi data testing berisi kombinasi "ok,ok,ok". Selanjutnya proses *training* dilakukan untuk kombinasi nilai target tertentu saja, yaitu : *"ok,ok,high"* , *"low,ok,ok"* , *" high,ok,high"* , *"high,ok,ok"* , *"low,ok,high"* , *"ok,high,high"* , *"high,high,high"* , *"low,high,ok"* , *"low,ok,ok".* Dimana dalam setiap kombinasi target yang telah dipilih tersebut akan dipergunakan dalam proses *training* menggunakan 7 lipatan *input feature* dengan variasi kombinasi data *test* (Gambar 12).

Matrik "*accuracy"* dan *"loss"* dipergunakan dalam proses *training*, dimana model yang baik akan menghasilkan nilai akurasi bertambah mendekati nilai 1 dan *loss* menurun mendekati nilai 0 sesuai dengan pertambahan nilai *epoch* selama proses *trainning*. Karena banyaknya hasil iterari proses *training* yang dihasilkan,

sebagai visualisasi diambil satu contoh hasil grafik yang telah dilakukan. Gambar 14 menunjukan hasil grafik *metric accuracy* dan *loss* untuk nilai target "*high,ok,ok*" *input data train* untuk kolom *feature* 0-112 dengan data *test* yang bergeser 10 langkah, Gambar 15 memperlihatkan hasil grafik *metric accuracy* dan *loss* secara keseluruhan untuk nilai target *"high,ok,ok".*



Gambar 14. Grafik *metric "high,ok,ok "* kolom *input feature* 0 sampai dengan 112



Gambar 15. Grafik *metric "high,ok,ok"* secara keseluruhan

Hasil dari proses *training* adalah sebuah model yang selanjutnya dipergunakan untuk memprediksi data *test* tanpa target terhadap tingkat kandungan hdl_cholesterol, hemoglobin_hgb dan ldl_cholesterol dalam kategori *"low", "ok"* atau *"high".* Dengan menggunakan metode yang telah diuraikan sebelumnya, dengan mengatur nilai *epoch* > 60, *batch size* = 20, serta membuat kolom *train* dan *test* dengan lipatan sebanyak 7 (*range* kolom awal 0-112 dan variabel geser sebesar 30 *step*), diperoleh *score* hasil prediksi dari Zindi adalah 0.94594

4. Kesimpulan

Model jaringan *neural network* yang baik (terhindar dari *underfitting* atau *overfitting*) dan *fine tunning hyperparameter* saja ternyata tidak cukup untuk memperoleh nilai *score* maksimal dalam kompetisi ini (*score* hanya berkisar 0.7xx). Tapi analisa sebaran, *cleansing data*, dan pengaturan data *input* dan uji untuk proses *training* dan *testing* dalam penelitian ini menjadi sangat berpengaruh (terutama untuk data *input* tak berimbang) untuk optimasi nilai *score*. Hasil dari eksperimen ini mencapai *score* > 0.94 menyiratkan bahwa model prediksi yang

digunakan dalam eksperimen ini dalam kategori baik (mendekati nilai 1). Capaian ini diharapkan dapat dijadikan pertimbangan dalam klasifikasi spektroskopi darah pada penelitian berikutnya. Eksperimen selanjutnya dapat menggunakan metode kombinasi pada algoritma ML.

5. Ucapan Terima Kasih

Ucapan terimakasih ditujukan kepada LP3M ITB Ahmad Dahlan Jakarta yang telah mendukung penelitian ini sehingga berjalan dengan baik.

Daftar Pustaka

Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer- Verlag New York*, *10*(3), 978.

Aloraynan, A., Rassel, S., Xu, C., & Ban, D. (2022). A Single Wavelength Mid-Infrared Photoacoustic Spectroscopy for Noninvasive Glucose Detection Using Machine Learning. *Biosensors*, *12*(3), 166.

Chen, F., Sun, C., Yue, Z., Zhang, Y., Xu, W., Shabbir, S., & Yu, J. (2022). Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *265*, 120355.

Ciobanu, C., Ember, K. J., Nyíri, B. J., Rajan, S., Chauhan, V., Leblond, F., & Murugkar, S. (2022). Potential of Raman Spectroscopy for Blood-Based Biopsy. *IEEE Instrumentation & Measurement Magazine*, *25*(1), 62–68.

Feng, N., Wang, F., & Qiu, Y. (2016). Novel approach for promoting the generalization ability of neural networks. *International Journal of Signal Processing*, *2*(2).

Fonseca, A. C., Pereira, J. F., Honorato, R. S., Bro, R., & Pimentel, M. F. (2022). Hierarchical classification models and Handheld NIR spectrometer to human blood stains identification on different floor tiles. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *267*, 120533.

Guleken, Z., Tok, Y. T., Jakubczyk, P., Paja, W., Pancerz, K., Shpotyuk, Y., & Depciuch, J. (2022). Development of novel spectroscopic and machine learning methods for the measurement of periodic changes in COVID-19 antibody level. *Measurement*, 111258.

Hastomo, W. (2021a). *Convolution Neural Network Arsitektur Mobilenet-V2 Untuk Mendeteksi Tumor Otak*. *5*(Gambar 1).

Hastomo, W. (2021b). *Klasifikasi Covid-19 Chest X-Ray Dengan Tiga Arsitektur Cnn (Resnet-152, Inceptionresnet-V2, Mobilenet-V2)*. *5*(Dl).

Hastomo, W., Bayangkari Karno, A. S., Kalbuana, N., Meiriki, A., & Sutarno. (2021). Characteristic Parameters of Epoch Deep Learning to Predict Covid-19 Data in Indonesia. *Journal of Physics: Conference Series*, *1933*(1). https://doi.org/10.1088/1742-6596/1933/1/012050

Hastomo, W., & Bayangkari, S. (2021). *Diagnosa Covid-19 Chest X-Ray Dengan Convolution Neural Network Arsitektur Resnet-152*. *2*(1), 26–33.

Hemmer, E., Benayas, A., Légaré, F., & Vetrone, F. (2016). Exploiting the biological windows: current perspectives on fluorescent bioprobes emitting

above 1000 nm. *Nanoscale Horizons*, *1*(3), 168–184.

Karno, A. S. B., & Hastomo, W. (2020). Optimalisasi Data Terbatas Prediksi Jangka Panjang Covid-19 Dengan Kombinasi LSTM Dan GRU. *Prosiding SeNTIK*, *4*(September), 181–191.

Karno, A. S. B., Hastomo, W., & Wardhana, I. S. K. (2020). Prediksi Jangka Panjang Covid-19 Indonesia Menggunakan Deep Learning Long-Term. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, 483–490.

Liu, Y., Wang, Z., Zhou, Z., & Xiong, T. (2022). Analysis and comparison of machine learning methods for blood identification using single-cell laser tweezer Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 121274.

Okut, H., Wu, X. L., Rosa, G. J., Bauck, S., Woodward, B. W., Schnabel, R. D., & Gianola, D. (2013). Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. *Genetics Selection Evolution*, *45*(1), 1–13.

Parab, J., Sequeira, M., Lanjewar, M., Pinto, C., & Naik, G. M. (2022). *Blood Glucose Prediction Using Machine Learning on Jetson Nanoplatform. Handbook of Intelligent Computing and Optimization for Sustainable Development, 835-848.*

Rizevsky, S., Zhaliazka, K., Dou, T., Matveyenka, M., & Kurouski, D. (2022). Characterization of Substrates and Surface-Enhancement in Atomic Force Microscopy Infrared Analysis of Amyloid Aggregates. *The Journal of Physical Chemistry C*, *126*(8), 4157–4162.

Satyo, A., Karno, B., Hastomo, W., Efendi, Y., & Irawati, R. (2021). *Arsitektur Alexnet Convolution Neural Network (CNN) Untuk Mendeteksi Covid-19 Image Chest-Xray*. 482–485.

Titus, J., Wu, A.H.B., & Biswal, S. (2022). Development and preliminary validation of infrared spectroscopic device for transdermal assessment of elevated cardiac troponin. *Commun Med 2*, *42*. https://doi.org/https://doi.org/10.1038/s43856-022-00104-9

Wang, H. J., Ji, F., Leung, C. S., & Sum, P. F. (2019). Regularization parameter selection for faulty neural networks. *International Journal of Intelligent Systems and Technologies*, *4*(1), 45–48.

Wang, L. V., & Wu, H. I. (2012). *Biomedical optics: principles and imaging*. John Wiley & Sons.

Zindi Afica. (n.d.). Retrieved from https://zindi.africa/