Implementasi Teknik Sampling untuk Mengatasi *Imbalanced Data* pada Penentuan Status Gizi Balita dengan Menggunakan *Learning Vector Quantization*

Implementation of Sampling Techniques for Solving Imbalanced Data Problem in Determination of Toddler Nutritional Status using Learning Vector Quantization

Ade Yuni Triyanto[1], Retno Kusumaningrum[2]
Jurusan Ilmu Komputer/Informatika Universitas Diponegoro
Jalan Prof. Soedarto, SH Tembalang, Semarang, 50275

[1] ade.yuni.triyanto@if.undip.ac.id, [2]retno_ilkom@undip.ac.id

Naskah diterima: 12 Februari 2017, direvisi: 2 Mei 2017, disetujui: 17 Juli 2017

Abstrak

Balita memerlukan suatu pengawasan khusus karena pada masa-masa tersebut balita rentan terhadap serangan penyakit dan kekurangan gizi. Untuk itu penelitian ini bertujuan untuk menerapkan metode Learning Vector Quantization (LVQ) dalam proses klasifikasi status gizi balita ke dalam gizi lebih, gizi baik, gizi rentan, dan gizi kurang. Data yang digunakan dalam penelitian ini adalah data gizi balita sebanyak 612, terdiri dari 38 data gizi lebih, 491 data gizi baik, 63 gizi rentan, dan 20 data gizi kurang. Data tersebut disebut sebagai data tidak seimbang. Selanjutnya, penelitian ini menerapkan teknik undersampling dan oversampling untuk mengatasi permasalahan tersebut. Hasil penelitian menunjukkan bahwa penerapan metode LVQ terhadap data tak seimbang menghasilkan nilai akurasi sebesar 84.15 %, tetapi nilai overall accuracy sebesar 43.27%. Sedangkan penerapan metode LVQ terhadap data yang seimbang menghasilkan nilai akurasi dan overall accuracy yang sama yakni sebesar 74.38%.

Kata kunci: data tak seimbang, Learning Vector Quantization, SMOTE, undersampling, status gizi balita

Abstract

Toddlers require a special surveillance, since they are vulnerable to disease and malnutrition in these ages. Therefore, this study aims to apply Learning Vector Quantization (LVQ) as a model to classify nutritional status of the toddlers into four classes, i.e. over nutrition, good nutrition, vulnerable nutrition, and malnutrition. This study used data of toddlers nutritional status of about 612 which consist of 38 data is over nutrition, 491 data is good nutrition, 63 data is vulnerable nutrition, and 20 data is malnutrition. This condition is called as imbalanced data. Moreover, in this study we implement sampling techniques, i.e. undersampling and oversampling, to overcome the imbalanced data problem. The results show that a built LVQ model over imbalanced data produces the accuracy of 84.15%, but the overall accuracy value of 43.27%. Whereas a built LVQ model over balanced data produces same values of accuracy and overall accuracy of 74.38%.

Keywords: *imbalanced data, Learning Vector Quantization, SMOTE, undersampling, toddler nutritional status*

PENDAHULUAN

Anak balita adalah anak yang telah menginjak usia di atas satu tahun atau lebih popular dengan pengertian usia di bawah lima tahun (Muaris, 2006). Pengawasan khusus terhadap balita sangat diperlukan karena pada masa-masa tersebut balita rentan terhadap serangan penyakit dan kekurangan gizi. Atau dengan kata lain, status gizi balita merupakan salah satu indikator kesehatan seorang balita.

Berbagai penelitian mengenai penentuan gizi balita menggunakan Jaringan Syaraf Tiruan telah dilakukan di Indonesia, yaitu penerapan metode *Backpropagation* untuk mengetahui status gizi balita dengan nilai akurasi optimal sebesar 93,85% (Anggraeni & Indrarti, 2010), penerapan *Perceptron* untuk menentukan status gizi balita dengan rentang usia 7-60 bulan dengan nilai akurasi optimal sebesar 82,609% (Fitri, Setyawati, & S, 2013). Kedua penelitian tersebut hanya menerapkan teknik evaluasi kinerja berupa perhitungan nilai akurasi tanpa memperhatikan kondisi data dalam keadaan seimbang atau tidak. Kondisi seperti itu sering terjadi pada dataset medis, dimana data positif atau abnormal berjumlah sedikit (kelas minor). Sementara itu, data negatif atau normal berjumlah banyak (kelas mayor), sehingga menimbulkan permasalahan klasifikasi *imbalanced* (Wan, Liu, Cheung, & Tong, 2014). Permasalahan klasifikasi *imbalanced* merupakan permasalahan yang muncul, yaitu nilai kinerja klasifikasi menunjukkan nilai akurasi yang tinggi karena jumlah kelas mayor yang sangat banyak. Namun, hal itu sebenarnya memiliki kinerja klasifikasi yang sangat buruk pada saat mengklasifikasikan data dari kelas minor.

Apabila dilihat dari penggunaan metode klasifikasi pada penelitian sebelumnya (Anggraeni & Indrarti, 2010; Fitri et al., 2013), kedua penelitian tersebut menggunakan metode *Backpropagation* dan *Perceptron*, yaitu kedua metode tersebut memiliki berbagai kelemahan, yaitu metode *Backpropagation* memiliki kompleksitas waktu yang tinggi pada proses pembelajaran (Haykin, 2008; Khairani, 2014) dan metode *Perceptron* memiliki kelemahan berupa proses konvergensi yang lambat dan kondisi *local minima* yang mempengaruhi proses pembelajaran (Wankhede, 2014). Salah satu metode yang mampu mengatasi kelemahan-kelemahan tersebut adalah *Learning Vector Quantization* (LVQ), karena pada umumnya metode klasifikasi berbasis LVQ memiliki nilai sensitivitas, spesifisitas, dan akurasi yang tinggi (Dharmawan, 2014; Hatmojo, 2014; Wuryandari & Afrianto, 2012). Selain itu, LVQ juga memiliki kompleksitas waktu yang lebih rendah dibandingkan dengan *Backpropagation* (Wuryandari & Afrianto, 2012). Hal tersebut disebabkan arsitektur LVQ memiliki *linear layer* yang membuat metode ini mempunyai kemampuan pembelajaran yang cepat.

Terkait dengan permasalahan *imbalanced data*, salah satu strategi yang dapat diterapkan adalah teknik *sampling*, baik *oversampling* maupun *undersampling*. SMOTE *(Synthetic Minority Oversampling Technique)* merupakan salah satu teknik sampling yang mampu meningkatkan akurasi dari pengklasifikasi untuk kelas minor (Chawla, 2005; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Hal tersebut disebabkan oleh SMOTE mampu menyajikan sampel dari kelas minor untuk dipelajari dalam *learning process* sehingga memungkinkan model pembelajaran untuk menghasilkan area pembentukan model klasifikasi yang lebih luas cakupannya.

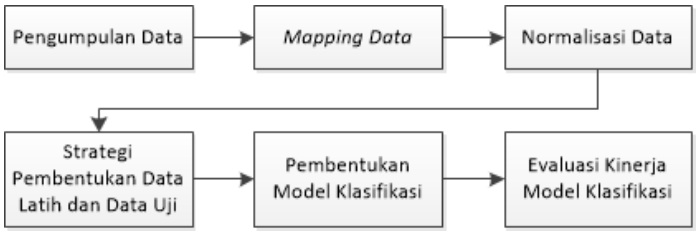
Berdasarkan penjelasan tersebut di atas, permasalahan utama yang dijumpai pada penerapan metode klasifikasi pada data medis adalah kondisi ketersediaan data yang

pada umumnya dalam keadaan *imbalanced*. Dengan demikian proses evaluasi terhadap kinerja model klasifikasi yang dihasilkan menjadi bias. Kondisi *imbalanced* yang terjadi dapat berupa munculnya kelas mayor yang sangat banyak, munculnya kelas minor yang sangat sedikit, ataupun munculnya kelas mayor dan kelas minor secara bersamaan. Dengan itu, ketepatan strategi menerapkan teknik sampling untuk mendapatkan dataset dalam kondisi seimbang untuk seluruh kelas merupakan permasalahan yang perlu dikaji lebih lanjut dalam permasalahan klasifikasi data medis untuk menentukan status gizi balita, khususnya untuk klasifikasi berbasis LVQ.

Pada penelitian ini diusulkan penerapan teknik *sampling* untuk mengatasi permasalahan *imbalanced data*, baik *oversampling* maupun *undersampling,* pada proses klasifikasi status gizi balita menggunakan LVQ. Selanjutnya, pembahasan pada bab kedua dari artikel ini adalah penjelasan mengenai metode penelitian. Sementara itu, penjelasan mengenai hasil penelitian dan pembahasannya dibahas pada bab ketiga. Bab keempat adalah bab penutup, meliputi kesimpulan dan saran.

METODE

Garis besar penyelesaian masalah dalam penerapan teknik sampling untuk mengatasi *imbalanced data* pada penelitian status gizi balita menggunakan *Learning Vector Quantization* (LVQ) dapat dilihat pada Gambar 1.



Gambar 1. Garis Besar Tahap Penyelesaian Masalah

Tahap pertama pada penelitian ini ialah tahap pengumpulan data balita dengan meminta langsung kepada Kader Posyandu Sekar Wangi Kelurahan Bandaharjo, Semarang Utara. Data yang didapat merupakan data status gizi balita pada tahun 2014 sebanyak 612 buah data. Sedangkan atribut-atribut pada data balita yang digunakan meliputi jenis kelamin, umur, berat badan, dan tinggi badan. Data yang diperoleh pada penelitian ini merupakan data tak seimbang *(imbalanced data)*, karena terdapat kelas yang mendominasi dari kelas lainnya. Komposisi data tersebut ialah kelas lebih, baik, dan rentan merupakan kelas mayoritas, sedangkan kelas kurang merupakan kelas minoritas. Detail mengenai dataset penelitian dapat Tabel 1.

Tabel 1. Dataset Penelitian

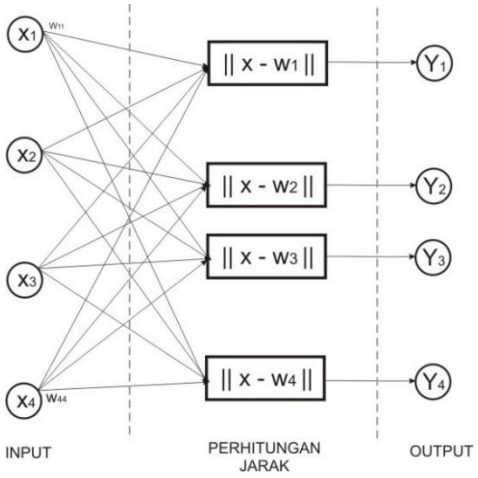
Status Gizi	Jenis Kelamin		Total
	Laki-laki	Perempuan	
Lebih	20	18	**38**
Baik	261	230	**491**
Rentan	27	36	**63**
Kurang	315	13	**20**
Jumlah	**315**	**297**	**612**

Tahap kedua ialah tahapan *mapping data*. Tahap ini dilakukan untuk mengidentifikasi data balita yang telah diperoleh ke dalam arsitektur LVQ, yaitu data yang diperoleh kemudian diidentifikasi untuk selanjutnya dipetakan mana yang akan menjadi *input neuron* dan mana yang akan menjadi *output neuron*. Pada proses ini nilai status gizi akan dipetakan sebagai *output neuron* sebagai berikut.

- Jika status gizi bernilai lebih akan dikonversi menjadi 1 (y_1)
- Jika status gizi bernilai baik akan

dikonversi menjadi 2 (y_2)
- Jika status gizi bernilai rentan akan dikonversi menjadi 3 (y_3)
- Jika status gizi bernilai kurang akan dikonversi menjadi 4 (y_4)

Input neuron merepresentasikan atribut jenis kelamin (x_1), umur (x_2), berat badan (x_3), dan tinggi badan (x_4). Dengan itu, diperoleh ilustrasi arsitektur LVQ dari penelitian ini seperti pada Gambar 2.



Gambar 2. Arsitektur LVQ

Tahap selanjutnya ialah normalisasi yang bertujuan untuk mengubah nilai data ke dalam interval 0 – 1 sehingga diharapkan setiap atribut memiliki kontribusi yang sama dalam perhitungan jarak. Pada penelitian ini diterapkan dua proses normalisasi, yaitu:
- normalisasi untuk atribut jenis kelamin, yaitu data yang memiliki jenis kelamin laki-laki (L) akan dikonversi menjadi 1, sedangkan data yang memiliki jenis kelamin perempuan (P) akan dikonversi menjadi 0.
- normalisasi atribut umur, berat badan dan tinggi badan, yaitu proses normalisasi pada ketiga atribut ini menggunakan persamaan sebagai berikut.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

dimana :

x' : nilai hasil normalisasi

x : nilai atribut awal sebelum normalisasi

x_{max} : nilai terbesar dari atribut tersebut

x_{min} : nilai terkecil dari atribut tersebut

Tahap keempat ialah tahap pembentukan data latih dan data uji menggunakan teknik *10-folds cross validation*. Seperti telah dijelaskan sebelumnya, data balita yang akan digunakan berjumlah 612 yang terdiri dari 4 kelas dalam kondisi *imbalanced data* (lihat Tabel 1). Oleh karena itu, untuk mengatasi masalah tersebut digunakan suatu metode sampling yang dapat memberi distribusi data seimbang untuk setiap kelas. Metode sampling yang diterapkan ialah *oversampling* dan *undersampling*, dengan detail sebagai berikut:

1. *Oversampling*

   Oversampling merupakan metode pembangkitan data minoritas sebanyak data mayoritas. Pada penelitan ini metode SMOTE diterapkan sebagai teknik *oversampling*. Metode SMOTE berfungsi untuk membuat data buatan, yaitu suatu data replika atau data *syntetic* dari data minoritas. Adapun algoritma dari metode SMOTE adalah sebagai berikut (Chawla et al., 2002).

Algoritma *SMOTE (sum_minor, p_syn)*
Input : Jumlah data minoritas *sum_minor* dan persentase data syntetic *p_syn*;
Output : Data buatan Z, berukuran $num\_syn \times n$ dimana n adalah banyaknya atribut
1. Tetapkan nilai $k = 3$ dan $a = 1$
2. Jika $p\_syn > 100$, maka lanjut ke langkah 3, sedangkan jika $p\_syn < 100$, maka:
   - $num\_syn = (p\_syn/100) * sum\_minor$
   - $p\_syn = 100$
3. Ubah nilai $p\_syn = \lceil p\_syn/100 \rceil$ dalam bentuk bilangan bulat
4. Kerjakan dari $b = 1$ sampai num_syn
   a. Hitung *k-nearest neighbors* untuk b, dan simpan k terdekat ke dalam vektor v berukuran $1 \times k$
   b. Kerjakan, jika $p\_syn \neq 0$
      i. $r = random(0 - 1)$
      ii. Kerjakan $j = 1$ sampai n
         - Hitung $dif = y_{v_{rj}} - y_{bj}$
         - Hitung $gap = rand(0,1)$
         - Hitung $z_{aj} = y_{bj} + gap\ x\ dif$

iii. Lakukan proses *increment* nilai a dengan persamaan $a = a + 1$

iv. Lakukan proses *decrement* nilai p_syn sebagai berikut:
$p\_syn = p\_syn - 1$

2. *Undersampling*

Berkebalikan dengan *oversampling*, *undersampling* adalah metode untuk mengambil beberapa data mayoritas sehingga jumlah data mayoritas sama besar jumlahnya dengan jumlah data minoritas. Metode *undersampling* yang digunakan pada penelitian ini ialah metode *undersampling* yang berbasis *clustering*. Berikut ini merupakan algoritma dari metode *undersampling* tersebut (Yen & Lee, 2009) :

Algoritma *UNDERSAMPLING* (X, r)
Input : Dataset X berukuran $m \times n$ dimana m adalah banyaknya data dan n adalah banyaknya atribut, indeks kelas data minor r
Output : Data buatan
1. Tetapkan nilai k, untuk proses *k-means clustering*
2. Lakukan proses *k-means clustering*, dengan menghasilkan nilai centroid yang disimpan dalam D berukuran $k \times n$
3. Kerjakan $i = 1$ sampai m
   a. Kerjakan $c = 1$ sampai k
      i. Hitung jarak
      $$s(i,c) = \sqrt{\sum_{j=1}^{n} \left(x_{ij} - d_{cj}\right)^2}$$
      ii. Cari $s(i,c)$ minimum, kemudian simpan c
      iii. Periksa data ke-i termasuk dalam kelas mana
      - Jika data ke-i termasuk kelas lebih maka $w_{1c} = w_{1c} + 1$
      - Jika data ke-i termasuk kelas baik maka $w_{2c} = w_{2c} + 1$
      - Jika data ke-i termasuk rentan lebih maka $w_{3c} = w_{3c} + 1$
      - Selain itu $w_{rc} = w_{rc} + 1$
4. Kerjakan $g = 1$ sampai $sum\_g - 1$
   a. Cari rasio data mayoritas kelas ke-g dengan data minoritas
   $$v_{gc} = \frac{w_{gc}}{w_{rc}}$$
   b. Hitung sementara jumlah rasio data mayoritas kelas ke-g dengan data minoritas
   $$u_g = u_g + v_{gc}$$
5. Proses pengambilan seberapa banyak data mayoritas yang akan diambil pada setiap kluster.
   Kerjakan $g = 1$ sampai $sum\_g - 1$
   a. Kerjakan $c = 1$ sampai k
      Hitung:
      $$l_{gc} = m \times num\_minor \times v_{gc}/u_g$$

Setelah melalui proses sampling, data yang digunakan hanya 80 data dari 612 data, yang terdiri dari 4 kelas dan dibagi secara proporsional, yaitu 20 data setiap kelasnya. Tahapan ini terdiri dari 3 strategi dalam pembentukan data latih dan data uji, yaitu:

1. Strategi 1

Strategi 1 akan berisi data latih dan data uji dengan jenis kelamin tidak terpisah. Teknik *sampling* yang akan digunakan pada strategi ini adalah *undersampling* karena proses yang diperlukan ialah mengurangi jumlah data pada kelas mayoritas sehingga berjumlah sama dengan kelas minoritas. Dalam penelitian ini kelas mayoritas (kelas lebih, baik, dan rentan) dan kelas minoritas (kelas kurang) berjumlah 20 data.

2. Strategi 2

Strategi 2 akan berisi data latih dan data uji dengan jenis kelamin laki-laki saja. Teknik *sampling* yang akan digunakan pada strategi ini adalah *oversampling* dan *undersampling*. Proses *oversampling* disini berfungsi untuk membuat data buatan sebanyak yang diperlukan. Dalam penelitian ini terdapat kelas yang belum memenuhi kriteria, yaitu setiap kelas berjumlah 20 data, kelas kurang yang hanya berjumlah 7 data. Sementara itu, proses *undersampling* berfungsi seperti yang sudah dijelaskan sebelumnya. Proses *undersampling* akan dilakukan pada kelas mayoritas yaitu kelas lebih, baik, dan rentan.

3. Strategi 3

   Strategi 3 akan berisi data latih dan data uji dengan jenis kelamin perempuan saja. Teknik *sampling* yang akan digunakan pada strategi ini adalah *oversampling* dan *undersampling*. Proses *oversampling* yang dilakukan pada strategi ini dilakukan pada kelas minoritas yaitu kelas lebih dan kurang, karena jumlah data pada kelas tersebut kurang dari 20 data. Sedangkan proses *undersampling* dilakukan pada kelas baik dan rentan yang merupakan kelas mayoritas.

Jumlah data dalam dataset hasil proses sampling untuk masing-masing strategi adalah 80 buah data. Seperti telah dijelaskan sebelumnya, teknik pembentukan data latih dan data uji adalah *10-Fold Cross Validation*, yaitu dataset akan dibagi menjadi 10 partisi dengan tiap partisi akan terdiri dari jumlah kelas yang sama. Pada penelitian ini jumlah dataset yang digunakan pada setiap strategi sebanyak 80 data, kemudian data tersebut akan dibagi menjadi 10 partisi, dengan setiap partisi berjumlah 8 data, 8 data tersebut terdiri dari 4 kelas, dimana tiap kelas terdiri dari dua data.

Setelah diperoleh data latih dan data uji dalam kondisi data yang seimbang untuk masing-masing kelasnya (kelas gizi lebih, gizi baik, gizi rentan, dan gizi kurang), tahap selanjutnya adalah proses pembentukan model klasifikasi. Dalam penelitian ini berupa proses pelatihan LVQ yang bertujuan untuk melatih metode LVQ sehingga mampu mengenali suatu pola pada data gizi balita berupa nilai bobot (w) yang memetakan nilai *input* neuron dalam menghasilkan *output* neuron. Parameter yang ditetapkan dalam proses pelatihan LVQ meliputi *learning rate* (laju perubahan), *epsilon* (*eps*), dan teknik penetapan bobot awal. Nilai laju perubahan yang ditetapkan bernilai antara 0.1 sampai 0.9, eps berkisar antara 10^{-1} sampai 10^{-5}, sedangkan teknik penetapan bobot awal secara random, nilai tengah dari *range data*, serta centroid dari *cluster K-Means clustering*.

Berdasarkan berbagai model klasifikasi (dengan berbagai parameter uji) yang telah diperoleh, maka tahap selanjutnya adalah melakukan evaluasi kinerja terhadap model klasifikasi tersebut. Tujuan dari tahap evaluasi ini adalah menemukan arsitektur terbaik dari model klasifikasi, yakni menemukan penggunaan kombinasi laju perubahan, nilai eps dan teknik penetapan nilai bobot awal yang menghasilkan nilai akurasi tertinggi. Selain itu, pada penelitian ini juga akan dilakukan proses evaluasi untuk mengetahui bagaimana perbandingan kinerja dari model klasifikasi yang menerapkan maupun tidak menerapkan teknik *sampling* dalam mengatasi permasalahan *imbalance data*. Detail mengenai berbagai skenario dari eksperimen yang dikerjakan dalam penelitian ini dapat dilihat pada bab selanjutnya mengenai hasil eksperimen dan analisa.

HASIL DAN PEMBAHASAN

Penelitian ini menerapkan beberapa skenario eksperimen, seperti berikut.

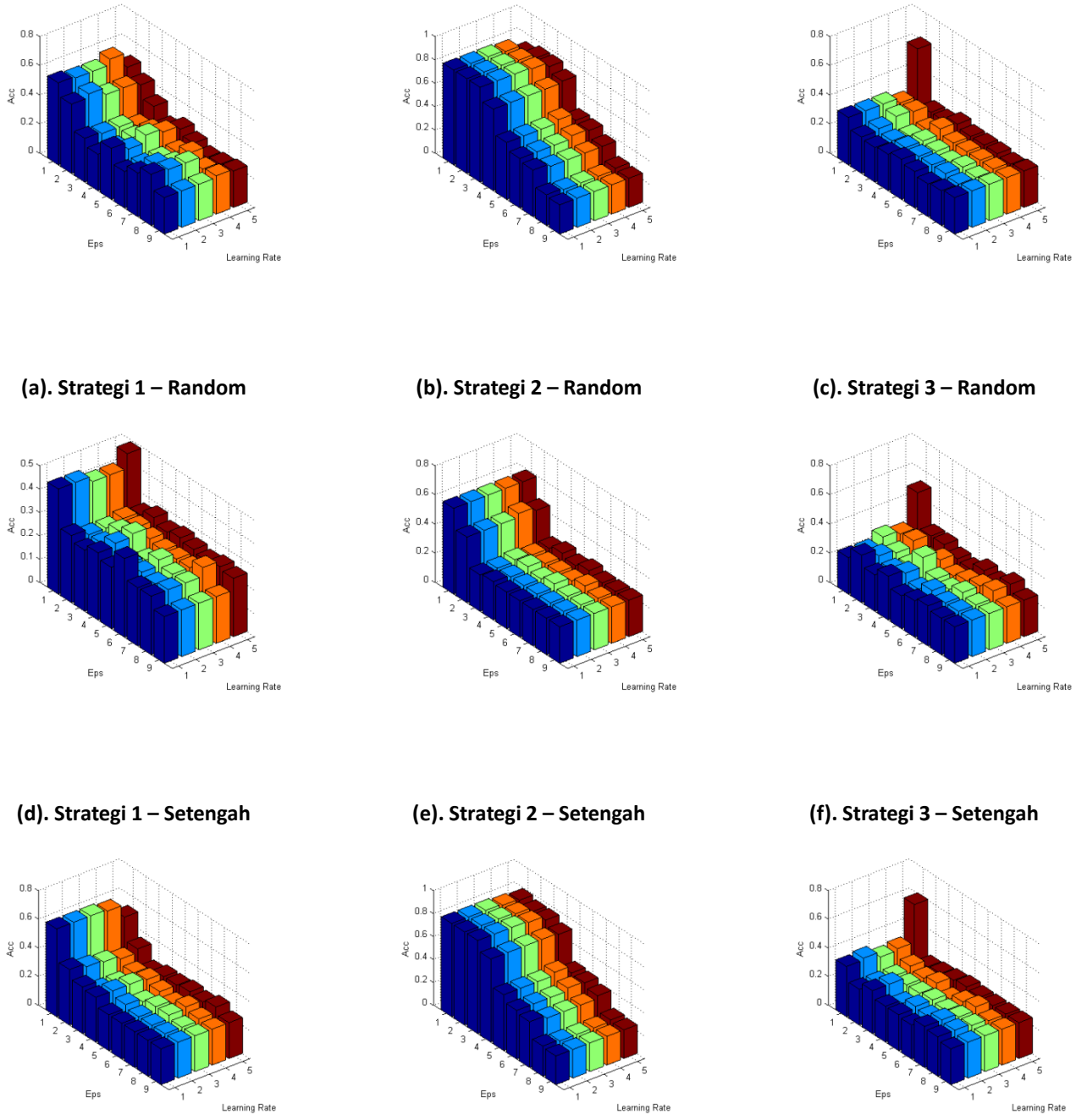
1. Skenario Eksperimen 1, bertujuan untuk mengetahui nilai laju perubahan (0,1 sampai dengan 0,9), epsilon (10^{-1} sampai 10^{-5}) dan teknik penetapan bobot awal (random, nilai setengah, dan nilai centroid dari *k-means clustering*) terbaik pada proses pembentukan model LVQ
2. Skenario Eksperimen 2, bertujuan untuk mengetahui bagaimana kinerja dari model klasifikasi penentuan status gizi balita **tanpa menerapkan** teknik penanganan *imbalance data*.
3. Skenario Eksperimen 3, bertujuan untuk mengetahui bagaimana kinerja dari model klasifikasi penentuan status gizi balita **dengan menerapkan** teknik penanganan *imbalance data*.
4. Skenario Eksperimen 4, bertujuan untuk mengetahui perbandingan kinerja dari model klasifikasi penentuan status gizi balita **dengan** dan **tanpa menerapkan** teknik penanganan *imbalance data*.

Penjelasan mengenai hasil dan analisa dari masing-masing eksperimen dijelaskan pada sub bab berikut ini.

A. Hasil dan Analisa dari Skenario

Eksperimen 1

Seperti telah dijelaskan sebelumnya, tujuan dari eksperimen pertama ini adalah untuk mendapatkan parameter terbaik dari model LVQ yang digunakan untuk menentukan status gizi balita, meliputi laju perubahan, epsilon, dan teknik penetapan bobot awal terbaik pada proses pembentukan model LVQ. Gambar 3 menjelaskan perbandingan kinerja dari berbagai model.



(a). Strategi 1 – Random **(b). Strategi 2 – Random** **(c). Strategi 3 – Random**

(d). Strategi 1 – Setengah **(e). Strategi 2 – Setengah** **(f). Strategi 3 – Setengah**

(g). Strategi 1 – Centroid Kmeans **(h). Strategi 2 – Centroid Kmeans** **(i). Strategi 3 – Centroid Kmeans**

Gambar 3. Perbandingan Kinerja LVQ untuk berbagai Parameter Eksperimen

Berdasarkan Gambar 3 maka kinerja model klasifikasi terbaik untuk strategi 1 adalah teknik penetapan bobot awal kmeans, laju perubahan 0,1, dan epsilon 10^{-2} dengan akurasi sebesar 61,25%. Kinerja strategi 2 memperoleh nilai akurasi yang sama baik menggunakan teknik penetapan bobot awal random maupun kmeans sebesar 81,25% untuk laju perubahan 0,1 dan dan epsilon 10^{-2}. Sedangkan kinerja terbaik untuk strategi 3 diperoleh dengan teknik penetapan bobot awal random, laju perubahan 0,1 dan dan epsilon 10^{-1} dengan nilai akurasi sebesar 61,25%.

Selanjutnya, Tabel 2 menunjukkan ringkasan perbandingan kinerja model LVQ terbaik untuk berbagai teknik penentuan bobot awal pada proses pelatihan LVQ.

Tabel 2. Perbandingan Kinerja Model Klasifikasi LVQ untuk Berbagai Teknik Penentuan Bobot Awal

Strategi	Bobot Awal	Laju Perubahan	Epsilon	Akurasi
1	Random	0.1	0,01	58,75%
	Setengah	0.1	0,1	48,75%
	Kmeans	**0.1**	**0,001**	**61,25%**
2	**Random**	**0.1**	**0,001**	**81,25%**
	Setengah	0.1	0,1	58,75%
	Kmeans	**0.1**	**0,001**	**81,25%**
3	**Random**	**0.1**	**0,1**	**61,25%**
	Setengah	0.1	0,1	51,25%
	Kmeans	0.1	0,1	58,75%

Berdasarkan Tabel 2, teknik penentuan nilai bobot awal secara random maupun menggunakan nilai *centroid* dari *Kmeans clustering* selalu menunjukkan kinerja yang lebih baik dibandingkan penggunaan nilai bobot awal skalar 0,5. Hal tersebut dikarenakan penggunaan kedua teknik tersebut menunjukkan nilai bobot awal yang mewakili data-data pelatihan.

B. Hasil dan Analisa dari Skenario Eksperimen 2

Model klasifikasi pada skenario eksperimen kedua ini dilakukan dengan menerapkan metode LVQ dengan menggunakan keseluruhan dataset tanpa menerapkan teknik penangan *imbalance data,* selanjutnya disebut sebagai MK1. Hasil eksperimen 2 dapat dilihat pada *confusion matrix* yang tersaji pada Tabel 3.

Tabel 3. Confusion Matrix dari Hasil Klasifikasi MK1

Aktual	Prediksi			
	Lebih	Baik	Rentan	Kurang
Lebih	19	19	0	0
Baik	1	485	4	1
Rentan	0	54	9	2
Kurang	0	12	6	2

Berdasarkan *confusión matrix* tersebut maka dapat diketahui nilai *accuracy (acc), accuracy* untuk masing-masing kelas (*acc_lebih, acc_baik, acc_rentan, acc_kurang*), serta *overall accuracy (oa)* dari MK1 sebagai berikut:

$$acc = \frac{19+485+9+2}{612} x100\% = 84{,}15\%$$

$$acc\_lebih = \frac{19}{19+19+0+0} \; x\; 100\% = 50{,}00\%$$

$$acc\_baik = \frac{485}{1+485+4+1} \; x\; 100\% = 98{,}78\%$$

$$acc\_rentan = \frac{9}{0+54+9+2} \; x\; 100\% = 13{,}85\%$$

$$acc\_kurang = \frac{2}{0+12+6+2} \; x\; 100\% = 10{,}00\%$$

$$oa = \frac{50{,}00\% + 98{,}78\% + 13{,}85\% + 10{,}00\%}{4} = 43{,}16\%$$

Apabila dilihat dari keenam nilai tersebut, maka dapat disimpulkan bahwa MK1 mampu menghasilkan nilai *acc* yang tinggi, yakni mencapai 84,15%, tetapi nilai *oa* sangat kecil, yaitu 43,16%, yang disebabkan oleh MK1 hanya mampu mengklasifikasikan data dengan baik untuk kelas gizi baik yang memiliki sampel data mayoritas, yaitu 491 sampel data dari keseluruhan data sebanyak 612. Sedangkan pada data minoritas seperti gizi lebih, gizi rentan, dan gizi kurang memiliki nilai akurasi yang rendah, yakni sebesar ≤ 50%. Hal

tersebut dikarenakan proses pembelajaran dilakukan pada data yang terbatas dan dengan variasi data yang sangat heterogen sehingga sulit untuk menemukan bobot yang mampu memetakan setiap *input neuron* kepada *output neuron* secara tepat.

C. *Hasil dan Analisa dari Skenario Eksperimen 3*

Seperti telah dijelaskan sebelumnya, model klasifikasi pada skenario ketiga ini dilakukan dengan menerapkan metode LVQ, menggunakan dataset yang seimbang untuk masing-masing kelas, yaitu sebanyak 160 data (40 data per kelas). Model itu selanjutnya disebut sebagai MK2 dengan hasil eksperimen dapat dilihat pada *confusion matrix* yang tersaji pada Tabel 4.

Tabel 4. Confusion Matrix dari Hasil Klasifikasi MK2

Aktual	Prediksi			
	Lebih	Baik	Rentan	Kurang
Lebih	38	1	0	1
Baik	4	19	8	9
Rentan	1	8	28	3
Kurang	1	1	4	34

Berdasarkan *confusión matrix* tersebut, dapat diketahui nilai *accuracy (acc), accuracy* untuk masing-masing kelas (*acc_lebih, acc_baik, acc_rentan, acc_kurang*), serta *overall accuracy (oa)* dari MK2 sebagai berikut:

$$acc = \frac{38+19+28+34}{160} x100\% = 74{,}38\%$$

$$acc\_lebih = \frac{38}{38+1+0+1} \; x \; 100\% = 95{,}00\%$$

$$acc\_baik = \frac{19}{4+19+8+9} \; x \; 100\% = 47{,}50\%$$

$$acc\_rentan = \frac{28}{1+8+28+3} \; x \; 100\% = 70{,}00\%$$

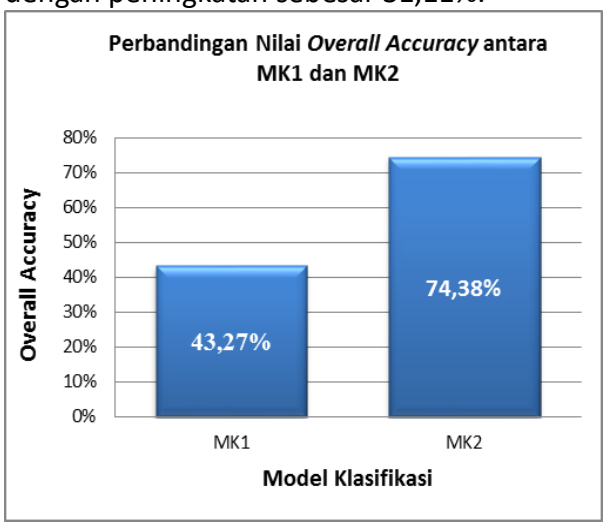
$$acc\_kurang = \frac{34}{1+1+4+34} \; x \; 100\% = 85{,}00\%$$

$$oa = \frac{95{,}00\% + 47{,}50\% + 70{,}00\% + 85{,}00\%}{4} = 74{,}38\%$$

Hasil tersebut menunjukkan bahwa nilai *acc* dan nilai *oa* menunjukkan nilai yang sama, yakni 74,38% . Atau dengan kata lain nilai *acc* yang diperoleh telah mampu merepresentasikan kinerja dari model klasifikasi untuk masing-masing kelas. Permasalahan yang muncul adalah nilai *accuracy* untuk kelas baik menjadi turun hanya mencapai nilai 47,50%. Hal tersebut disebabkan proses *undersampling* yang diterapkan untuk kelas baik menggunakan algoritma *K-means clustering*, dimana pada proses tersebut menggunakan inisialisasi *centroid* awal secara random. Dengan demikian, hasil akhir dari *proses K-means clustering* selalu tidak sama untuk setiap prosesnya atau dikenal juga sebagai *uncertainty final result*, yang berpengaruh terhadap hasil dari proses *undersampling*.

D. *Hasil dan Analisa dari Skenario Eksperimen 4*

Perbandingan kinerja dari kedua model tersebut (MK1 dan MK2) dilakukan berdasarkan nilai *overall accuracy*. Pemilihan matrik ini dikarenakan *oa* merupakan matrik yang representatif untuk menggambarkan *accuracy* untuk masing-masing kelas dan *accuracy* dari model klasifikasi secara utuh. Berdasarkan Gambar 3 mengenai perbandingan nilai *oa* MK1 dan MK2 dapat disimpulkan bahwa kinerja dari model klasifikasi yang dibentuk pada *balance dataset* memberikan kinerja yang lebih baik dengan peningkatan sebesar 31,11%.



Gambar 3. Grafik Perbandingan Nilai *Overall Accuracy* antara Model Klasifikasi untuk Data Tidak Seimbang (MK1) vs Model Klasifikasi Data Seimbang (MK2)

Berdasarkan keseluruhan eksperimen yang dilakukan model terbaik untuk penentuan status gizi balita adalah dengan memisahkan data antara data laki-laki dan data perempuan. Oleh karena itu, model yang memberikan kinerja klasifikasi terbaik adalah model *hybrid* yaitu gabungan antara *rule base* dan model klasifikasi berbasis LVQ, yaitu :

$$f(\boldsymbol{x}) = \begin{cases} 1, & jika\ D_e(\boldsymbol{w_1}, \boldsymbol{x}) = minimal \\ 2, & jika\ D_e(\boldsymbol{w_2}, \boldsymbol{x}) = minimal \\ 3, & jika\ D_e(\boldsymbol{w_3}, \boldsymbol{x}) = minimal \\ 4, & jika\ D_e(\boldsymbol{w_4}, \boldsymbol{x}) = minimal \end{cases}$$

dimana,

$D_e(\boldsymbol{w_g}, \boldsymbol{x}) = \sqrt{\sum_{j=1}^{n} \left(\boldsymbol{x_j} - w_{gj}\right)^2}$, untuk $g = 1,2,3,4$

$$W_{gj} = \begin{cases} A, & jika\ x_{j=1} = 1 \\ B, & jika\ x_{j=1} = 0 \end{cases}$$

$$A = \begin{bmatrix} \boldsymbol{a_1} \\ \boldsymbol{a_2} \\ \boldsymbol{a_3} \\ \boldsymbol{a_4} \end{bmatrix} = \begin{bmatrix} 1,0.5817,0.7441,0.7673 \\ 1,0.6104,0.4652,0.7248 \\ 1,0.1760,0.1557,0.3893 \\ 1,0.7231,0.2137,0.5918 \end{bmatrix}$$

$$B = \begin{bmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \\ \boldsymbol{b_3} \\ \boldsymbol{b_4} \end{bmatrix} = \begin{bmatrix} 0,0.3570,0.2701,0.4652 \\ 0,0.4396,0.1601,0.4707 \\ 0,0.6633,0.7745,0.8204 \\ 0,0.6967,0.3324,0.5697 \end{bmatrix}$$

Hal tersebut berarti apabila balita berjenis kelamin laki-laki, akan mengambil bobot akhir pada arsitektur terbaik laki-laki. Sebaliknya, jika balita tersebut berjenis kelamin perempuan maka akan mengambil bobot akhir pada arsitektur terbaik perempuan. Berikut merupakan algoritma dari penentuan status gizi balita pada aplikasi:

1. Masukkan data balita ($\boldsymbol{x}$)
2. Lakukan proses normalisasi data balita, jika diketahui nilai maksimal dan minimal data balita sebagai berikut:
   - Untuk atribut jenis kelamin ($\boldsymbol{x_1}$)
     Jika $\boldsymbol{x_1}$ bernilai L(Laki-Laki) maka $\boldsymbol{x_1}$ bernilai 1, dan jika $\boldsymbol{x_1}$ bernilai P(Perempuan) maka $\boldsymbol{x_1}$bernilai 0.
   - Untuk atribut umur ($\boldsymbol{x_2}$), berat badan ($\boldsymbol{x_3}$) dan tinggi badan $\boldsymbol{x_4}$ menggunakan persamaan berikut:
     $$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$
     Dimana nilai berdasarkan nilai berikut ini

$\boldsymbol{x}$	Nilai Minimal	Nilai Maksimal
$\boldsymbol{x_2}$	1	60
$\boldsymbol{x_3}$	3.5	28
$\boldsymbol{x_4}$	50	112

3. Jika $\boldsymbol{x_1} = 1$, maka $W = A$; selain itu jika $\boldsymbol{x_1} = 0$, maka $W = B$
   Dimana A merupakan bobot akhir dari laki-laki dan B merupakan bobot akhir dari perempuan
4. Kerjakan $g = 1$ sampai dengan sum_g
   - Hitung $D_e(w_g) = \sqrt{\sum_{j=1}^{n} \left(\boldsymbol{x_j} - w_{gj}\right)^2}$
5. Cari nilai minimum dari $D_e(w_g)$, kemudian simpan nilai g
6. Tentukan data tersebut termasuk ke dalam kelas apa menurut nilai g yang disimpan.

Keterangan :

n : jumlah variabel penentu gizi balita
j : indeks untuk variabel penentu status gizi balita $(1,2,3,\dots,n)$
sum_g : jumlah kelas status gizi balita
g : indeks untuk kelas status gizi balita $(1,2,3,\dots,sum\_g)$
$\boldsymbol{x}$: data balita berukuran $1xn$
W : bobot akhir yang dipakai saat penentuan status gizi balita yang berukuran $sum\_g \ \ x\ n$
w_{gj} : elemen dari matriks W yang menunjukkan bobot kelas ke-g dan variabel ke-j
A : bobot akhir dari jenis kelamin laki-laki yang berukuran $sum\_g \ \ x\ n$
B : bobot akhir dari jenis kelamin perempuan yang berukuran $sum\_g \ \ x\ n$
$D_e(w_g)$: nilai *euclidean distance* kelas ke-g

PENUTUP

Kesimpulan yang dapat diambil ialah penerapan metode LVQ terhadap data yang tak seimbang (*imbalanced data*) menghasilkan nilai akurasi sebesar 84.15 %, tetapi nilai *overall* akurasinya hanya 43.16%. Sedangkan penerapan metode LVQ terhadap data yang seimbang menghasilkan nilai akurasi sebesar 74.38% dan nilai *overall* akurasinya juga sama sebesar 74.38%. Jadi, jika suatu data dalam keadaan tidak seimbang (*imbalanced data*) sangat berpengaruh terhadap nilai akurasi yang akan dihasilkan.

Berdasarkan penjelasan-penjelasan di atas, hal utama yang perlu dipersiapkan pada penyelesaian permasalahan klasifikasi adalah ketersedian dataset dalam kondisi seimbang untuk seluruh kelas. Hal tersebut lebih lanjut dapat memberikan manfaat berupa diperolehnya informasi mengenai kinerja metode klasifikasi yang konsisten atau tidak bias. Adapun proses penyeimbangan data tersebut dapat dilakukan menggunakan teknik *sampling*, baik *undersampling* maupun *oversampling*.

Adapun kelemahan utama dari penelitian ini ialah pemilihan metode *undersampling* yang diusulkan berupa proses *undersampling* berbasis *Kmeans clustering,* dimana metode *clustering* tersebut memiliki permasalahan terkait penentuan *centroid* awal secara random yang mengakibatkan munculnya permasalahan *uncertainty final result*. Hal tersebut dapat diatasi dengan menerapkan metode *Minimum Description Length (MDL)* untuk mendapatkan jumlah *cluster* yang optimal beserta nilai *centroid*-nya. Selanjutnya nilai *centroid* tersebut dapat diterapkan sebagai *centroid* awal pada proses *undersampling* berbasis *Kmeans clustering*.

DAFTAR PUSTAKA

Anggraeni, R., & Indrarti, A. (2010). Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Jaringan Syaraf Tiruan. In *Prosiding Seminar Nasional Teknologi Informasi 2010 (SNASTI 2010)* (p. ICCS-14 s/d ICCS 18). Surabaya.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853–867).

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Dharmawan, D. A. (2014). Deteksi kanker serviks otomatis berbasis Jaringan Saraf Tiruan LVQ dan DCT. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *3*(4), 269–272.

Fitri, Setyawati, O., & S, D. R. (2013). Aplikasi jaringan syaraf tiruan untuk penentuan status gizi balita dan rekomendasi menu makanan yang dibutuhkan. *Jurnal EECCIS*, *7*(2), 119–124.

Hatmojo, Y. I. (2014). Implementasi Wavelet Haar dan Jaringan Tiruan Pada Pengenalan Pola Selaput Pelangi Mata. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *3*(1), 58–62.

Haykin, S. (2008). *Neural Networks and Learning Machines*. New Jersey: Prentice Hall.

Khairani, M. (2014). Improvisasi Backpropagation menggunakan penerapan adaptive learning rate dan parallel training. *TECHSI - Jurnal Penelitian Teknik Informatika*, *4*(1), 157–172.

Muaris, H. (2006). *Sarapan Sehat untuk Anak Balita* (PT Gramedi). Jakarta.

Wan, X., Liu, J., Cheung, W. K., & Tong, T.

(2014). Learning to improve medical decision making from imbalanced data without a priori cost. *BMC Medical Informatics and Decision Making*, *14*(111), 1–9. https://doi.org/10.1186/s12911-014-0111-9

Wankhede, S. B. (2014). Analytical Study of Neural Network Techniques : SOM , MLP and Classifier-A Survey. *IOSR Journal of Computer Engineering*, *16*(3), 86–92.

Wuryandari, M. D., & Afrianto, I. (2012). Perbandingan Metode Jaringan Syaraf Tiruan Backpropagation dan Learning Vector Quantization pada Pengenalan Wajah. *Jurnal Komputer Dan Informatika*, *1*, 45–51.

Yen, S.-J., & Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108