# Penggunaan Augmentasi Data pada Klasifikasi Jenis Kanker Payudara dengan Model Resnet-34

Lukas Hansel Ganda[#1], Hendra Bunyamin, S.Si., M.T.[*2]

[#]*Program studi S1 Teknik Informatika, Universitas Kristen Maranatha*
*Jl. Prof. Drg. Suria Sumantri No. 65 Bandung*
[1]lukaslynch98@gmail.com
[2]hendra.bunyamin@maranatha.ac.id

***Abstract*** **— A recent study from the Global Cancer Observatory (GLOBOCAN) revealed that in 2020 about 2.2 milion women worldwide have been diagnosed with breast cancer. Diagnostic tissue biopsy with hematoxylin and eosin stained images is used to make decisions on the final diagnosis. Computer-assisted diagnostic systems contribute to increasing the efficiency of this process. In this study using the dataset "BreAst Cancer Histology Images (BACH)". And a method was made to classify breast biopsy images stained with hematoxylin and eosin using convolutional neural networks. The images are classified into four classes, normal tissue, benign lesions, carcinoma in situ and invasive carcinoma. In this study, regularization techniques were also carried out in the convolutional neural networks model to achieve maximum accuracy.**

*Keywords*— **Breast Cancer, Convolutional Neural Networks, Data Augmentation, Multiclass Classification**

## I. PENDAHULUAN

### *A. Latar Belakang*

Kanker payudara merupakan salah satu penyebab kematian yang paling umum terjadi pada wanita dari segala usia, tetapi diagnosis dini dan pengobatan dini dapat dengan signifikan mencegah perkembangan penyakit dan mengurangi tingkat mortalitas [1].

Palpasi dan pemeriksaan rutin melalui ultrasound atau mamografi dapat dilakukan untuk diagnosis dini; jika dalam pemeriksaan ditemukan kelainan, biopsi jaringan payudara akan dilakukan. Umumnya, jaringan atau sampel yang terkumpul akan diberi cairan penanda hematoxylin dan eosin (H&E) yang dapat membedakan inti dari jaringan dasar dan diamati melalui mikroskop optik. Sampel-sampel ini juga dapat dipindai menjadi gambar dengan ukuran giga-pixel atau disebut *whole-slide image (WSI)*.

Untuk dapat membantu proses identifikasi jenis kanker yang lebih detil daripada identifikasi 2 jenis, yaitu jinak dan ganas, *Institute of Molecular Pathology and Immunology of the University of Porto* (IPATIMUP) dan *Institute for Research and Innovation in Health* (i3S) telah mengklasifikasi jenis kanker menjadi 4 kelas, yaitu normal, jinak, karsinoma in situ, dan karsinoma invasif; hasil klasifikasi ini menjadi *dataset* yang digunakan dalam *BACH challenge* [2]. Penelitian ini hendak mengeksplorasi *dataset BACH* dengan mengimplementasikan metode *deep learning* untuk dapat mengklasifikasi jenis kanker payudara. Adapun metode spesifik yang hendak digunakan adalah *Convolutional Neural Network (CNN).*

Terdapat 100 citra pada setiap kelas dengan total 4 kelas dan 400 citra, jumlah data yang sedikit ini menjadi alasan dilakukannya teknik augmentasi data pada penelitian ini, penelitian ini juga hendak membandingkan *CNN* sebelum dan sesudah diaplikasikan augmentasi data.

### *B. Rumusan Masalah*

Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana cara mengklasifikasi jenis kanker payudara dalam bentuk gambar histopatologis dengan menggunakan metode *Convolutional Neural Network*?
2. Bagaimana cara mengaplikasikan augmentasi data pada proses klasifikasi jenis kanker payudara pada *Convolutional Neural Network* dengan menggunakan *library* fast.ai?
3. Berapa hasil akurasi yang diperoleh dari proses klasifikasi jenis kanker payudara dengan menggunakan *Convolutional Neural Network* sebelum mengaplikasikan augmentasi data dan sesudah mengaplikasikan augmentasi data.

### *C. Tujuan Penelitian*

Tujuan penelitian ini adalah sebagai berikut:

1. Mempelajari proses klasifikasi jenis kanker payudara dengan menggunakan metode *Convolutional Neural Network*.
2. Mengetahui keakuratan metode *Convolutional Neural Network* dalam mengklasifikasi jenis kanker payudara.
3. Mengetahui apakah teknik *Augmentation Data* dapat meningkatkan keakuratan metode *CNN*.

### *D. Batasan Masalah*

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Citra gambar yang akan diolah menggunakan *dataset BACH (BreAst Cancer Histology).*
2. Sistem pengklasifikasian yang dibuat tidak diimplementasi ke platform seperti *website.*

## II. TINJAUAN PUSTAKA

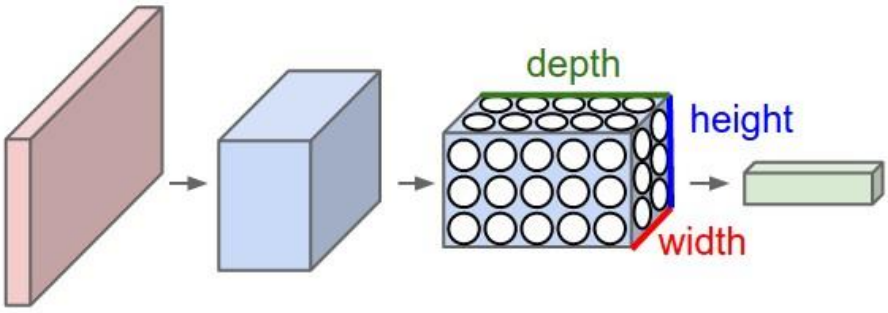
### *A. Deep Learning*

Definisi *deep learning* yang dapat berlaku bagi banyak orang agak sulit Ditentukan karena telah berubah definisi selama beberapa tahun terakhir. Salah satu definisi yang sederhana: "*Deep learning* adalah setiap jaringan saraf dengan lebih dari dua lapisan." [3].

### *B. Convolutional Neural Networks*

Jika dilihat dari arsitekturnya, *Convolutional Neural Network (CNN)* termasuk ke dalam kelas *feed-forward artificial neural network. CNN* adalah sebuah *Artificial Neural Network* yang memiliki satu atau lebih lapisan konvolusional. *CNN* banyak diaplikasikan pada analisis citra. Arsitektur *CNN* sangat sederhana, yaitu:satu lapis masukan *(input layer)*, sejumlah lapisan tersembunyi *(hidden layers)*, dan lapis hasil atau keluaran *(output layer)* [5].

*CNN* menggunakan arsitektur 3 dimensi: lebar *(width)*, tinggi *(height)*, dan dalam *(depth)*. Seperti yang diilustrasikan pada Gambar 1, setiap lapisan *CNN* mentransformasikan volume masukan tiga dimensi ke dalam volume keluaran tiga dimensi aktivasi-aktivasi sel saraf *(from 3D input volume to 3D output volume of neuron activations.)*. Di Gambar 1 masukkan berupa citra berwarna merah dengan lebar dan tingginya menyatakan dimensi citra tersebut sedangkan dalam *(depth)*-nya akan menjadi 3 yang menyatakan kanal *red, green, blue (RGB channels)* [6].

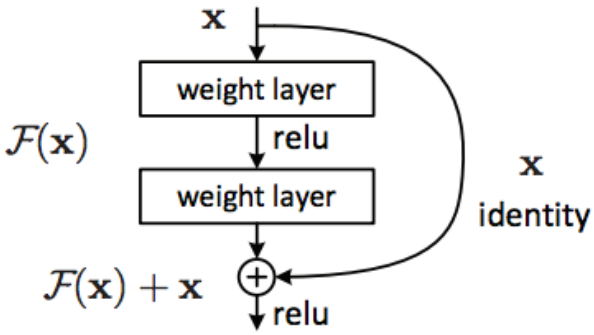



Gambar 1 Arsitektur Convolutinal Neural Networks

(Source: https://cs231n.github.io/convolutional-networks/)

### *C. Residual Network*

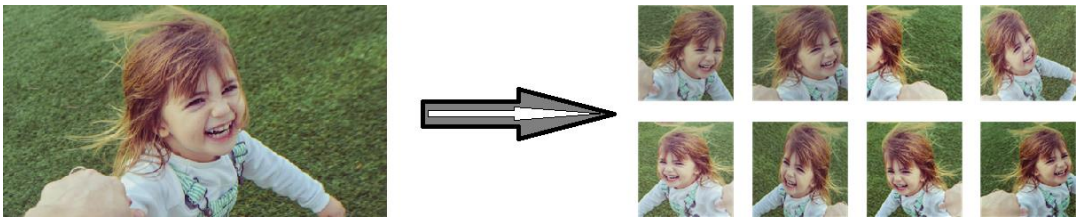
*Residual Network* atau yang biasa disebut *ResNet* adalah salah satu Arsitektur *CNN* yang dibangun oleh Kaiming He et al dan menjadi juara satu dalam kompetisi ILSVRC 2015. Arsitektur ini memiliki tiga ciri khusus yang berupa koneksi lompat (*skip connections*) menggunakan *batch normalization*, dan menghilangkan *fully connected layers* dibagian akhir. Pada sejumlah arsitektur *CNN* sebelumnya, saat jumlah *hidden layer* terlalu banyak, umumnya akurasi yang didapat mulai menurun dan dan berujung pada tingkat *error* yang lebih tinggi. *ResNet* berhasil memecahkan permasalahan tersebut dengan menambahkan sebuah cara untuk melompat atau melewati sejumlah *layer* yang dinamakan *Residual Learning*, seperti diilustrasikan pada Gambar 2. Cara ini berhasil menghilangkan *vanishing gradient problem* yang banyak terjadi pada arsitektur *CNN* lain. Arsitektur ResNet memiliki berbagai macam jenis arsitektur, mulai dari 18, 34, 50, 101, sampai 152 layer [8].




Gambar 2 *Residual Learning Block [8].*

*D. Data Augmentation*

*Data augmentation* atau augmentasi data adalah trik yang umum untuk mengurangi over-fitting. Pada augmentasi data, dihasilkan data baru dengan menggunakan transformasi pada data asli. Augmentasi data memungkinkan untuk meningkatkan generalisasi data [9].



Gambar 3 Ilustrasi Augmentasi Data

(Source : https://towardsdatascience.com/data-augmentations-in-fastai-84979bbcefaa)

Augmentasi data menggunakan transformasi untuk menghasilkan data baru [7].

I. Transformasi *Flip_lr*
   Transformasi *Flip_lr* adalah transformasi yang membalik citra secara horizontal dan mencerminkan citra.
II. Transformasi *Symmetric Warp*
   Transformasi *Symmetric Warp* melakukan perubahan pada sudut tampilan citra dalam besaran vektor.
III. Transformasi *Rotate*
   Transformasi ini melakukan rotasi pada citra dengan besaran sudut.
IV. Transformasi *Zoom*
   Transformasi ini melakukan *zoom* pada citra dengan besaran skala.
V. Transformasi *Brightness*
   Transformasi *brightness* mengubah pencahayaan *brightness* atau kecerahan pada citra dengan besaran skala.
VI. Transformasi *Contrast*
   Transformasi *contrast* mengubah pencahayaan kontras pada citra dengan besaran skala, semakin berbeda warna terang dan gelap maka semakin tinggi kontrasnya.

*E. Over-fitting*

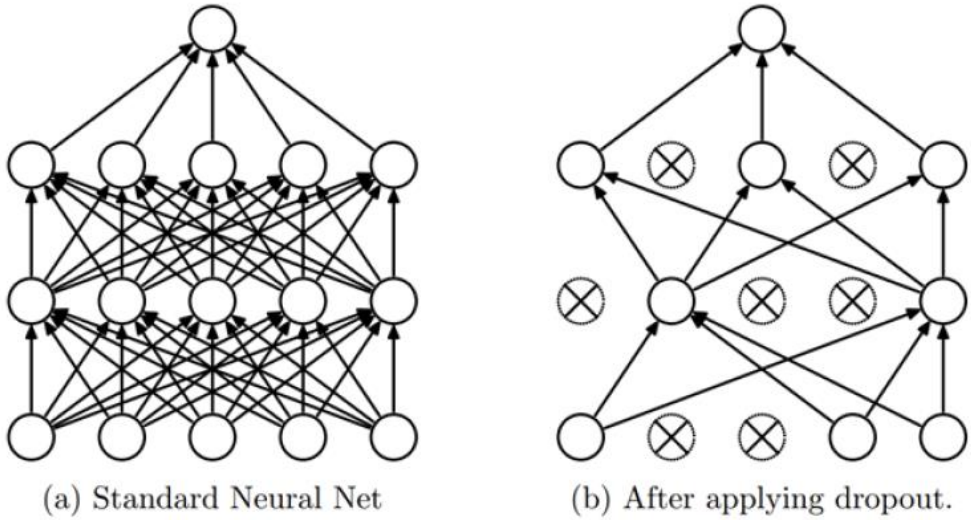
*Over-fitting* adalah kondisi yang terjadi ketika model memiliki *error* yang rendah selama pelatihan tetapi berfungsi dengan buruk saat memprediksi data baru. *Over-fitting* disebabkan oleh pembuatan model yang lebih kompleks dari yang diperlukan [4].

*F. Under-fitting*

*Under-fitting* adalah kebalikan dari *Over-fitting*, *Under-fitting* terjadi ketika sebuah model tidak mampu menangkap variabilitas dari data. *Under-fitting* dapat dicegah dengan meningkatkan kompleksitas model dan data [5].

*G. Dropout*

*Dropout* adalah salah satu teknik regularisasi yang paling efektif dan paling umum digunakan pada *neural networks*, *dropout* diperkenalkan oleh Geoffrey Hinton et al. pada *paper* yang berjudul "*Improving Neural Networks by Preventing Co-Adaption of Feature Detectors*". Ide dasarnya adalah mengubah beberapa fungsi aktivasi secara acak menjadi nol pada waktu pelatihan.




Gambar 4 Contoh aplikasi *dropout* pada *neural networks* [10].

Seperti yang terlihat pada gambar 4. Sebelah kiri adalah ilustrasi *neural network* standar dan sebelah kanan adalah ilustrasi *neural networks* yang telah mengaplikasikan *dropout* sehingga beberapa neuron di non-aktifkan. Menggunakan *dropout* membantu mengurangi *overfitting* [10].

### *H. Dataset BACH*

*Dataset BACH (BreAst Cancer Histology)* [2] adalah *dataset* berupa citra / gambar yang diklasifikasikan ke dalam 4 kelas yaitu normal, jinak, karsinoma in situ dan karsinoma invasif. Kanker payudara karsinoma in situ adalah pertumubuhan sel yang tidak terkendali di dalam saluran payudara, karsinoma ini bersifat noninvasif, artinya belum tumbuh ke dalam jaringan payudara di luar saluran. Biasanya karsinoma in situ tidak menimbulkan gejala dan baru bisa dideteksi dengan pemeriksaan mammogram. Peluang kesembuhan akan lebih besar jika kondisi ini terdeteksi sejak dini. Jika pengobatan telat dilakukan, kanker payudara ini dapat berubah menjadi ganas. Penelitian menunjukan bahwa resiko terkena kanker invasif rendah jika pernah dirawat karena karsinoma in situ. Jika tidak dirawat, 30% hingga 50% wanita dengan kanker payudara karsinoma in situ akan terkena kanker invasif. Kanker payudara karsinoma invasif adalah kanker payudara yang bersifat ganas, kanker ini bermula dari mengganasnya sel-sel di saluran payudara yang kemudian menerobos dinding saluran tersebut dan menyerang jaringan payudara lain di dekatnya. Sel kanker juga bisa menyebar ke jaringan tubuh lainnya [11].

TABEL 1

JUMLAH GAMBAR DALAM SETIAP KELAS.

| **Klasifikasi** | **Jumlah Gambar** |
|---|---|
| Normal | 100 |
| Jinak | 100 |
| Karsinoma in situ | 100 |
| Karsinoma invasif | 100 |

Citra dalam *dataset* ini disediakan dalam format RGB dengan format .tiff yang memiliki resolusi 2048 × 1536 piksel dan skala piksel 0.42 μm × 0.42 μm. Gambar / citra dalam *dataset* ini diperoleh pada tahun 2014, 2015 dan 2017 menggunakan mikroskop Leica DM 2000 LED dan kamera Leica ICC50 HD dan semua pasien berasal dari distrik Porto dan dan distrik Castelo Branco (Portugal). Kasus yang didapatkan dari *Ipatimup Diagnostic* yang berasal dari 3 rumah sakit (*Hospital CUF Porto, Centro Hospitalar do Tâmega e Sousa* dan *Centro Hospitalar Cova da Beira)* [2]

## III. ANALISIS DAN RANCANGAN SISTEM

### *A. Tahapan Penelitian*

Pada penelitian ini akan dilakukan dengan beberapa tahap. Tahapan penelitian ini antara lain adalah studi literatur, generalisasi dataset, rancang model, implementasi, perbandingan, prediksi, analisis dan pembahasan, dan terakhir penarikan kesimpulan.

*I. Studi Literatur*
Studi literatur dilakukan dengan cara mencari referensi yang bersumber dari buku, jurnal, dan artikel yang memiliki keterkaitan dengan permasalahan yang telah direncanakan untuk diteliti untuk mencari solusinya.

*II. Generalisasi Dataset*
Generalisasi *dataset* dilakukan untuk mempersiapkan data yang dibutuhkan pada sistem klasifikasi jenis kanker payudara. Resolusi citra *original dataset* ini adalah 2048 × 1536. Citra diubah resolusinya menjadi 512 × 384 pixel untuk memenuhi kebutuhan sistem [12].

*III. Rancang Model*
Pada tahapan ini peneliti melakukan pembelajaran lebih lanjut pada arsitektur *CNN* yang hendak dipakai untuk diimplementasikan pada sistem klasifikasi jenis kanker payudara.

*IV. Implementasi*
Setelah melakukan augmentasi data, peneliti melakukan *training* pada *dataset* dengan arsitektur *CNN ResNet*.

*V. Prediksi*
Pada tahap ini model akan diuji coba untuk memprediksi sebuah gambar setelah kedua model tersebut selesai melakukan proses *training.*

*VI. Perbandingan*
Pada tahap ini model akan diduplikat sehingga ada 2 model, satu model tanpa augmentasi data dan satunya memanfaatkan augmentasi data dan dilakukan proses *training*.

*VII. Analisis dan Pembahasan*
Pada tahap ini akan dilakukan analisis dan evaluasi untuk menguji performa dari model *ResNet* sebelum dan sesudah diaplikasikan augmentasi data. Untuk mengetahui performa dari model *CNN* dapat dievaluasi dengan menghitung akurasinya dan tingkat *loss*-nya.

*VIII. Penarikan Kesimpulan*
Pada bagian ini akan dibuat suatu kesimpulan yang berasal dari hasil analisis dan pembahasan dari data yang sudah diuji menurut rumusan masalah. Dengan cara ini bisa ditarik kesimpulan dalam bentuk akurasi.

### *B. Alat Penelitian*

Penelitian ini menggunakan platform yang telah disediakan oleh *Google*, yaitu *Google Colaboratory* dengan menggunakan pengaturan *runtime GPU* yang terhubung dengan *Google Drive* sebagai media penyimpanannya. Untuk menggunakan *Google Colaboratory*, koneksi yang stabil dibutuhkan agar saat *runtime*, proses *training* tidak terputus. Penulis menggunakan *Google Drive* sebagai media penyimpanan *dataset* dan bahasa pemrograman yang digunakan pada penelitian ini adalah bahasa pemrograman Python dengan *library fastai.*

### *C. Transformasi*

Transformasi yang dilakukan adalah; transformasi *flip_lr*, transformasi *symmetric_warp,* transformasi *rotate,* transformasi *zoom,* transformasi *brightness,* transformasi *contrast.*

## IV. IMPLEMENTASI

### *A. Implementasi Tanpa Augmentasi Data*

Dilakukan pelatihan dengan total 150 iterasi tanpa augmentasi data dan didapatkan hasil terbaik yaitu *train loss* sekitar 20%, *valid loss* sekitar 51,7% dan akurasi sekitar 86,25%.

### *B. Implementasi dengan Augmentasi Data*

Dilakukan pelatihan dengan total 150 iterasi dengan memperhitungkan augmentasi data dan didapatkan hasil terbaik yaitu *train loss* sekitar 32,6%, *valid loss* sekitar 32,7% dan akurasi sekitar 93,75%.

### *C. Perbandingan Transformasi*

Pada penelitian ini dilakukan juga perbandingan terhadap transformasi yang sudah dilakukan untuk mengetahui transformasi mana yang memiliki dampak paling signifikan terhadap dataset yang digunakan pada penelitian ini. Dilakukan pelatihan sebanyak 150 *epoch* terhadap masing-masing transformasi dengan *batch size* 8 dan resolusi citra 512 x 384 pixel.

TABEL 2

HASIL PERBANDINGAN TRANSFORMASI

| Nama Transformasi | *Train Loss* | *Valid Loss* | *Accuracy* |
|---|---|---|---|
| *Flip_lr* | 26,5% | 43,2% | 88,75% |
| *Symmetric_warp* | 29% | 39,7% | 91,25% |
| *Rotate* | 41,7% | 47,1% | 92,50% |
| *Zoom* | 24,5% | 42,7% | 88,75% |
| *Brightness* | 22,6% | 44,2% | 87,50% |
| *Contrast* | 21,3% | 47,1% | 87,50% |

Setelah dilakukannya pelatihan terhadap masing – masing transformasi dapat diketahui bahwa transformasi rotasi merupakan transformasi yang memiliki dampak paling besar karena transformasi ini memiliki dampak akurasi yang paling tinggi yaitu 92,50% sedangkan transformasi pencahayaan seperti *brightness* dan *contrast* merupakan transformasi yang memiliki dampak paling tidak berpengaruh, keduanya memiliki hasil akurasi yang sama yaitu sebesar 87,50%/

## V. KESIMPULAN

Dalam penelitian ini telah dilakukan *multi-class classification* pada dataset *BACH*, dan dilakukan perbandingan hasil sebelum dan sesudah *augmentation data* dilakukan. Berdasarkan penilitian yang dilakukan, akurasi yang didapat dengan menggunakan teknik *augmentation data* lebih besar dibandingkan akurasi yang didapat tanpa menggunakan teknik *augmentation data*.

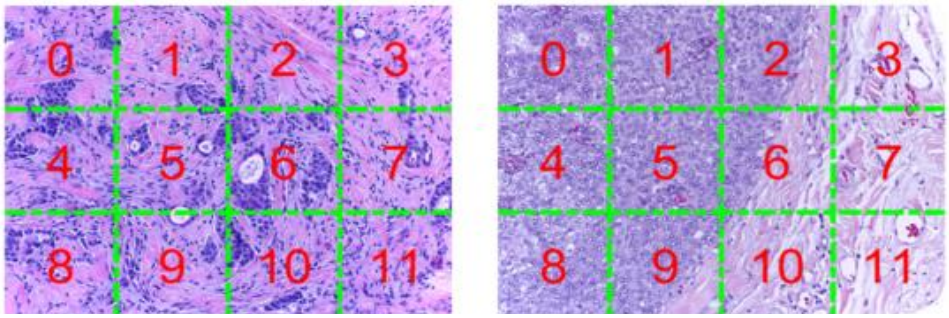
TABEL 3

HASIL PERBANDINGAN PENGGUNAAN TEKNIK AUGMENTASI DATA

| | **Tanpa *Augmentation Data*** | **Dengan *Augmentation Data*** |
|---|---|---|
| *Train Loss* | 20% | 32,6% |
| *Valid Loss* | 51,7% | 32,7% |
| *Accuracy* | 86,25% | 93,75% |

Dapat ditarik kesimpulan bahwa *augmentation data* merupakan teknik yang tepat untuk meningkatkan performa model. Selisih performa akurasi yaitu 7,5% dan teknik data augmentatition yang memiliki dampak paling besar yaitu transformasi rotasi.

Pada paper *"breast cancer histopathological image classification using a hybrid deep neural network"*, *image preprocessing* menggunakan teknik yang berbeda, tidak mengubah resolusi citra melainkan membaginya ke dalam 12 *patch* [13].




Gambar 5 Sampel *image patch [13].*

## DAFTAR PUSTAKA

[1] A. Migowski, "Early detection of breast cancer and the interpretation of results of survival studies/A deteccao precoce do cancer de mama e a interpretacao dos resultados de estudos de sobrevida," *Ciência & Saúde Coletiva*, vol. 20, no. 4, 2015.

[2] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. Nhat To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia and P. Aguiar, "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122-139, 2019.

[3] K. H. Mahmud, A. and S. A. Faraby, "Klasifikasi Citra Multi-Kelas Menggunakan Convolutional Neural Network," *eProceedings of Engineering*, vol. 6.1, 2019.

[4] "Generalisasi: Bahaya Overfitting," Google Developers, [Online]. Available: https://developers.google.com/machine-learning/crash-course/generalization/peril-of-overfitting?hl=id. [Accessed 8 May 2020].

[5] H. K. Jabbar and R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study).," *Computer Science, Communication and Instrumentation Devices,* 2015.

[6] C. C. Aggarwal, "Neural Networks and Deep Learning," New York, Springer, 2018.

[7] F. AI, "Vision Transform Fast AI," [Online]. Available: https://fastai1.fast.ai/vision.transform.html. [Accessed 20 11 2020].

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," vol. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[9] G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2017.

[10] J. Howard and S. Gugger, Deep Learning for Coders with fastai and PyTorch, Sebastopol: O'Reilly Media, 2020.

[11] L. J. Martin, "Invasive Ductal Carcinoma (IDC) & Ductal Carcinoma In Situ (DCIS)," WebMD Medical Reference, 27 February 2019. [Online]. Available: https://www.webmd.com/breast-cancer/ductal-carcinoma-invasive-in-situ. [Accessed 23 December 2020].

[12] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks.," *PloS one,* vol. 12, no. 6, 2017.

[13] R. Yan, F. Ren, Z. Wang, L. Wang , T. Zhang, Y. Liu, X. Rao, C. Zheng and F. Zhang, "Breast cancer histopathological image classification using a hybrid deep neural network," *Elsevier,* vol. 173, pp. 52-60, 2020.