Prediksi Diagnosis dan Prognosis *Breast Cancer* menggunakan *Machine Learning*

Marcel Indra Yordanus[#1], Wenny Franciska Senjaya[#2]

[#]*Jurusan Teknik Informatika, Universitas Kristen Maranatha*
Jalan Surya Sumantri No 65, Kota Bandung
[1]1972025@maranatha.ac.id
[2]wenny.fs@it.maranatha.edu

Abstract — **Breast cancer is one of the most common health problems affecting women around the world. Breast cancer diagnostic involves the identification and assessment of tumors in the breast tissue to determine the malignant or benign nature of the cancer. Meanwhile, breast cancer prognostic aims to identify disease progression after treatment and predict the likelihood of recurrence. This research aims to analyze the latest developments for predicting breast cancer diagnosis and prognosis using machine learning with k-nearest neighbors and logistic regression models and deep learning using artificial neural network models with sequential models. in this research, things are done such as: conducting data exploration, preprocessing data, oversampling for unbalanced datasets and training models. The results show that deep learning and machine learning predictions are suitable for predicting breast cancer diagnosis while prediction for breast cancer prognosis is suitable using machine learning. All results were compared using the evaluation metrics used in this study such as accuracy, precision, recall and F1-Scores. The best-performing model for the diagnosis dataset is logistic regression, while for the prognosis dataset, the best-performing model is the deep learning model using oversampling. The best-performing model for the diagnosis dataset is logistic regression, while for the prognosis dataset, the best-performing model is the deep learning model using oversampling.**

Keywords— **accuracy, breast cancer, deep learning, diagnostic, machine learning, prognostic**

I. PENDAHULUAN

Kanker payudara saat ini masih menjadi salah satu masalah yang paling mengkhawatirkan di seluruh dunia sampai saat ini. Menurut Organisasi Kesehatan Dunia kanker payudara merupakan kanker yang paling umum terjadi pada perempuan di seluruh dunia. Terlebih lagi, jumlah kasus baru kanker payudara di seluruh dunia semakin meningkat. diagnosis dini dan pengobatan yang tepat terbukti dapat meningkatkan peluang kesembuhan dan menurunkan angka kematian akibat kanker payudara. Oleh karena itu penting untuk mengembangkan metode yang lebih canggih dan akurat untuk mendeteksi, mendiagnosis, dan memprediksi kanker payudara [1].

Diagnosis kanker payudara merupakan proses identifikasi dan penilaian tumor yang berkembang di dalam jaringan payudara dan diagnosis ini penting dilakukan dikarenakan untuk menentukan apakah kanker tersebut bersifat ganas atau jinak. Prognosis kanker payudara merupakan proses identifikasi perkembangan kanker payudara pada pasien yang telah melakukan pengobatan dan mengetahui apakah kanker tersebut kembali muncul atau tidak.

Salah satu cara untuk melakukan identifikasi terhadap diagnosis dan prognosis *breast cancer* adalah penggunaan teknologi *Deep Learning*. *Deep Learning* merupakan salah satu cabang kecerdasan buatan yang telah menunjukkan potensi besar dalam analisis data medis khususnya dalam interpretasi gambar medis seperti Mamografi, USG dan MRI. Dengan melakukan pemrosesan data yang kompleks dan berbasis pola, model deep learning dapat mendeteksi tanda-tanda kanker payudara pada manusia. Selain itu deep learning juga dapat digunakan untuk memprediksi perkembangan penyakit sehingga dapat membantu tenaga kesehatan dalam merencanakan pengobatan yang efektif [2].

Selain menggunakan *deep learning* penelitian ini juga menggunakan tradisional *machine learning* dengan model *K-Nearest Neighbors* dan *Logistic Regression* yang dimana *K-Nearest Neighbors* digunakan untuk tujuan mengklasifikasi objek baru berdasarkan atribut contoh latihannya. Algoritma K-Nearest Neighbors merupakan algoritma yang unik dikarenakan algoritma KNN merupakan algoritma yang diawasi serta KNN banyak digunakan dalam aplikasi pengembangan data, pengenalan pola, pemrosesan gambar dan lain-lain. Sedangkan *Logistic Regression* merupakan metode statistik yang kuat untuk memprediksi probabilistik, intrepretasi yang mudah dan efektif untuk masalah klasifikasi biner

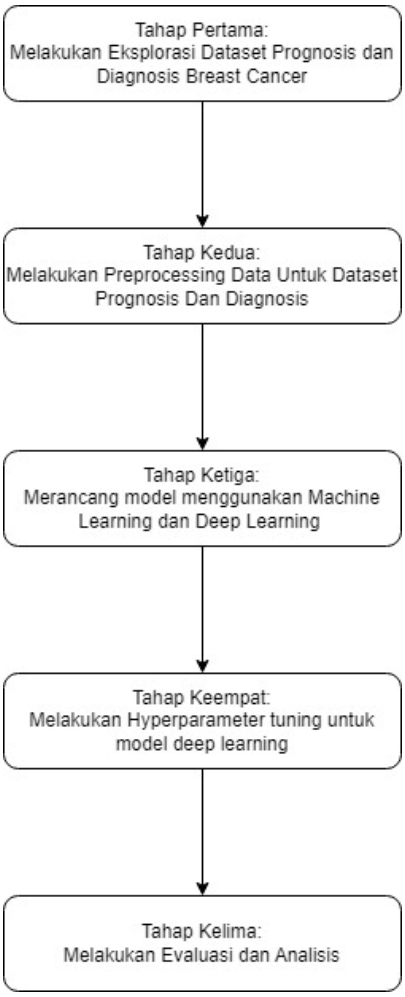
Penelitian ini dilakukan untuk mengetahui perkembangan terbaru dalam prediksi diagnosis dan prognosis kanker payudara menggunakan tradisional *machine learning* dan *deep learning*. Untuk mengetahui perkembangan terbaru dalam prediksi, model *deep learning* akan menggunakan *hyperparameter tuning* untuk mencari parameter-parameter terbaik serta

oversampling menggunakan *ADASYN* dan *k-fold cross validation* untuk *dataset prognostic*. Tahapan terakhir dari penelitian ini ialah membandingkan hasil *F1_Score* dari setiap model yang digunakan.

II. METODOLOGI PENELITIAN

A. Metodologi Penelitian

Gambar 1 merupakan metodologi penelitian yang akan dilakukan dalam penelitian ini



Gambar 1. Metodologi Penelitian

Tahapan pertama dari metodologi penelitian ini adalah melakukan eksplorasi *dataset* yang digunakan yaitu *Wisconsin Breast Cancer Dataset Diagnostic* dan *Prognostic*. Dalam *dataset Wisconsin Breast Cancer (Prognostic)*, terdapat 198 data pasien yang menderita kanker payudara. Setiap pasien memiliki 34 fitur yang digunakan untuk mengklasifikasikan perkembangan kanker payudara. Target yang digunakan adalah "*outcome*," yang memiliki dua nilai: *R (recur)* dan *N (nonrecur)*. 34 fitur yang digunakan untuk klasifikasi adalah*: id, outcome (R=recur, N=nonrecur), radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension, tumor size, dan lymph node status.*

Sedangkan dalam *dataset Wisconsin Breast Cancer (Diagnostic)* terdapat 569 data pasien yang terdiagnosis kanker payudara. Setiap pasien memiliki 30 fitur yang digunakan untuk melakukan pengklasifikasian apakah pasien tersebut terkena kanker atau tidak. Target yang digunakan adalah "*diagnosis*," yang memiliki dua nilai: *M (malignant) dan B (benign)*. 30 fitur yang digunakan untuk pengklasifikasian adalah: *id, diagnosis (M=malignant, B=benign), radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry,* dan *fractal dimension.*

Setelah melakukan eksplorasi *dataset* akan dilakukannya *preprocessing data* dengan melakukan *Data Cleaning*, Pemisahan Data, dan *Oversampling*.

Tahapan selanjutnya ialah membuat model menggunakan *K-Nearest Neighbours* dan *Logistic Regression* untuk *Machine Learning* dan model untuk *Deep Learning*. Setelah merancang model yang ada untuk model *deep learning* akan dilakukan *hyperparameter tuning* dengan parameter sebagai berikut: jumlah unit pada layer pertama sebesar 32, 64, atau 128, jumlah unit pada layer kedua sebesar 16, 32, atau 64, batch_size sebesar 10, 20, atau 40, jumlah epoch sebanyak 50 atau 100, dan learning_rate sebesar 0.01, 0.001, atau 0.0001. Setiap angka yang dipilih dalam *hyperparameter tuning* merupakan pemilihan untuk mencapai kinerja optimal model. Pemilihan jumlah neuron pada model ini didasarkan pada aturan praktis yang umum digunakan. Untuk lapisan pertama jumlah neuron yang dipilih adalah 32, 64, 128 yang berada di dalam rentang ukuran lapisan *input* dan *output*. Sedangkan untuk lapisan kedua, jumlah *neuron* yang dipilih adalah 16, 32, atau 64 yang mengikuti prinsip bahwa jumlah *neuron* tersembunyi sebaiknya berada antara ukuran lapisan *input* dan *output* atau sekitar 2/3 ukuran lapisan *input* ditambah *output*. Pemilihan jumlah *neuron* ini bertujuan untuk mencapai keseimbangan optimal antara menghindari *overfitting* serta mempertimbangkan kompleksitas data dan kebutuhan pelatihan[3].

Tahapan terakhir dari metodologi penelitian ini ialah melakukan evaluasi dan analisis dengan melihat dan membandingkan nilai *F1_Score* dari setiap model untuk mendapatkan model terbaik dari setiap *dataset*

B. Breast Cancer

Breast Cancer merupakan kanker yang sering ditemukan pada wanita. Menurut *World Health Organization Breast Cancer* merupakan kanker yang sering terjadi pada wanita di seluruh dunia [1]. *Breast cancer* merupakan suatu penyakit keganasan oleh karena proliferasi tak terkontrol dari sel-sel di payudara[2].

C. Machine Learning

Machine Learning adalah mesin yang dikembangkan untuk belajar dengan sendirinya tanpa arahan dari penggunanya. *Machine Learning* merupakan cabang dari AI yang melakukan pencarian yang dirancang untuk memilih fungsi dari sekumpulan fungsi untuk menjelaskan hubungan antar fitur dalam sebuah kumpulan data [4].

D. K-Nearest Neighbors

K-Nearest Neighbors adalah algoritma metode pengklasifikasi pembelajaran non-parametrik yang menggunakan pendekatan untuk membuat klasifikasi atau prediksi tentang pengelompokan data individual. K-Nearest Neighbors merupakan salah satu algoritma pengklasifikasian yang paling sering digunakan dalam machine learning saat ini [5]. K-Nearest Neighbors digunakan untuk tujuan mengklasifikasi objek baru berdasarkan atribut contoh latihannya. Algoritma K-Nearest Neighbors merupakan algoritma yang unik dikarenakan algoritma KNN merupakan algoritma yang diawasi serta KNN banyak digunakan dalam aplikasi pengembangan data, pengenalan pola, pemrosesan gambar dan lain-lain. Cara kerja algoritma ini sebagai berikut [6]

1. Menentukan nilai k yang ada. penentuan nilai k yang digunakan dalam klasifikasi tidak memiliki aturan baku.
2. Menghitung jarak antar objek/data baru
3. Mengurutkan hasil perhitungan
4. Menentukan neighbors yang terdekat dengan nilai k
5. Menggunakan kategori sebagai klasifikasi

E. Logistic Regression

Logistic Regression merupakan teknik analisis prediktif di mana label keluaran yang akan diprediksi bersifat *dichotomous*, yang berarti *biner*. Semua model *Logistic Regression* lainnya digunakan untuk menggambarkan data antara dependen atau independen. Cara kerja algoritma *Logistic Regression* adalah dengan memprediksi data dengan memvisualisasikan letak data yang tidak terlihat pada garis lurus[7]. Variabel *biner* dalam *Logistic Regression* biasanya hanya terdiri dari dua nilai untuk mewakili ada atau tidak adanya suatu peristiwa dan biasanya diberi numerik 1 dan 0[8].

F. Oversampling

*Imbalanced d*apat ditemukan dalam berbagai *dataset* seperti klasifikasi *breast cancer* untuk mencegah *imbalance* yang terjadi maka digunakan nya *oversampling*. *Oversampling* merupakan metode untuk memperbanyak data minoritas sebanyak data mayoritas [9].

G. K-Folds Cross Validation

K-Fold Cross Validation merupakan sebuah teknik yang digunakan untuk membagi data menjadi *train* data dan *test* data. teknik ini banyak digunakan untuk mengurangi bias yang terjadi di dalam sebuah sampel. K-Fold Cross Validation bersifat terus menerus dalam membagi data *train* dan data *test* sehingga setiap data akan mendapatkan kesempatan untuk menjadi *test* data [10].

H. Deep Learning

Deep Learning merupakan cabang dari Machine Learning yang menggunakan metodologi artificial neural network. Deep Learning sangat baik untuk diterapkan untuk supervised learning, unsupervised learning, maupun semi-supervised learning [11]. Model pada Deep Learning dibangun berdasarkan jaringan neural network.

Deep Learning dapat memecahkan masalah utama dalam representation learning dengan cara memperkenalkan representation yang lebih sederhana. Deep Learning juga memungkinkan komputer untuk konsep yang kompleks dari konsep yang lebih sederhana.

I. Artificial Neural Network

Artificial Neural Network merupakan metode yang berasal dari *Deep Learning*. ANN dapat digunakan untuk melakukan regresi dan klasifikasi. Hal ini memanfaatkan untuk melakukan *deep learning* dengan memproses data pelatihan dengan cara meniru otak kerja manusia yang melewati node node [12]. ANN disusun dari *neural network* yang ketika *neural network* diberi *input* sebanyak jumlahnya, maka *neural network* akan melakukan penjumlahan semua input yang diberikan. Sedangkan jika *output* melebihi ambang batas yang diberikan, maka node akan diaktifkan dan node akan meneruskan data ke *layer* berikutnya di dalam *network* [13].

III. HASIL DAN PEMBAHASAN

Tahapan pengujian yang dilakukan pada penelitian ini merupakan hasil prediksi dari setiap model yang digunakan

A. Hasil Prediksi Breast Cancer

Tabel 1 merupakan hasil pengujian dari model *Logistic Regression, K-Nearest Neighbour, Deep Learning* menggunakan *oversampling* dan *Deep Learning* tidak menggunakan *oversampling* dengan membandingkan nilai *F1_Score*. Semua model *deep learning* yang digunakan dalam penelitian ini menggunakan *hyperparameter tuning* untuk mengetahui parameter-parameter terbaik untuk model *deep learning*.

TABEL I
HASIL PENGUJIAN YANG DILAKUKAN

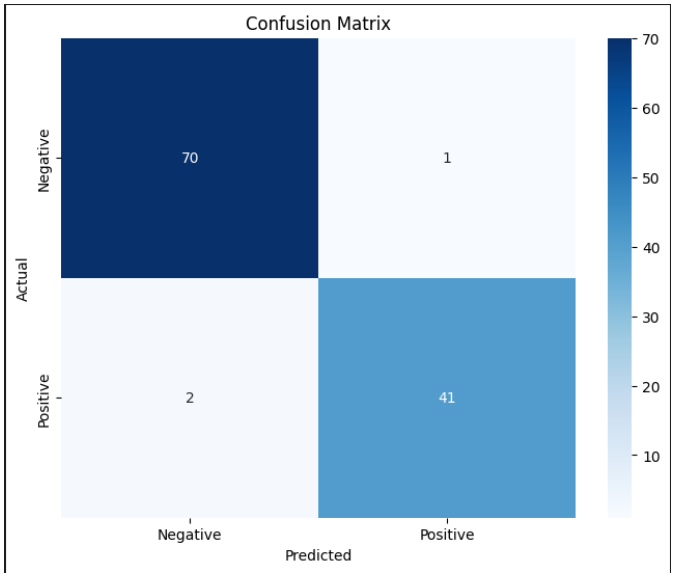
Model	**F1-Score**
Logistic Regression Prognostic	0.545
Logistic Regression Diagnostic	0.976
K-Nearest Neighbors Prognostic	0.615
K-Nearest Neighbors Diagnostic	0.953
Deep Learning Prognostic menggunakan oversampling Folds 1	0.950
Deep Learning Prognostic tanpa Oversampling Folds 2	0.777
Deep Learning Diagnostic	0.964

Berdasarkan hasil Tabel I didapatkan bahwa model *Logistic Regression* dan *K-Nearest Neighbors* menunjukkan kinerja yang baik pada *dataset diagnostic* tetapi tidak optimal untuk *dataset prognostic*. Sedangkan model yang menggunakan *deep learning* menunjukkan kinerja yang baik pada *dataset diagnostic* dan *prognostic* menggunakan *oversampling*, tetapi ketika tidak menggunakan *oversampling* kinerja pada model *deep learning prognostic* menurun. Hasil *hyperparameter tuning* untuk setiap model *deep learning* sebagai berikut:

a. Deep Learning Prognostic Menggunakan Oversampling:
   Batchsize:10, epochs:100, learning rate:0.01, units1:128, units2:64
b. Deep Learning Prognostic Tanpa Oversampling
   Batchsize:10, epochs:50, learning rate:0.001, units1:64, units2:64
c. Deep Learning Diagnostic
   Batchsize:40, epochs:50, learning rate:0.001, units1:128, units2:64

B. Confusion Matrix Diagnostic

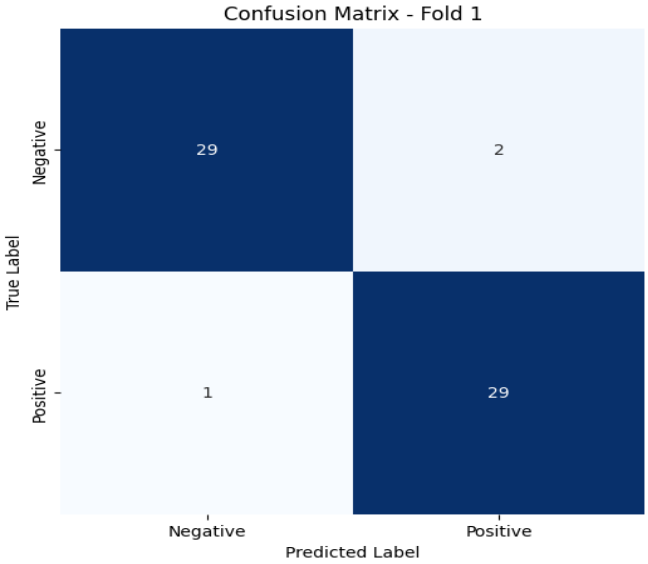
Gambar 2 menunjukkan *confusion matrix* dari hasil *deep learning* bahwa 70 data berhasil diprediksi dengan benar sebagai *benign* atau *true negative* dan 1 data *benign* yang salah diidentifikasi sebagai *malignant* atau *false positive* dan 41 data berhasil diprediksi sebagai *malignant* atau *true positive* dan 2 data *malignant* yang salah diidentifikasi sebagai *benign* atau *false negative.*



Gambar 2. Confusion Matrix Diagnostic

C. Confusion Matrix Prognostic menggunakan Oversampling

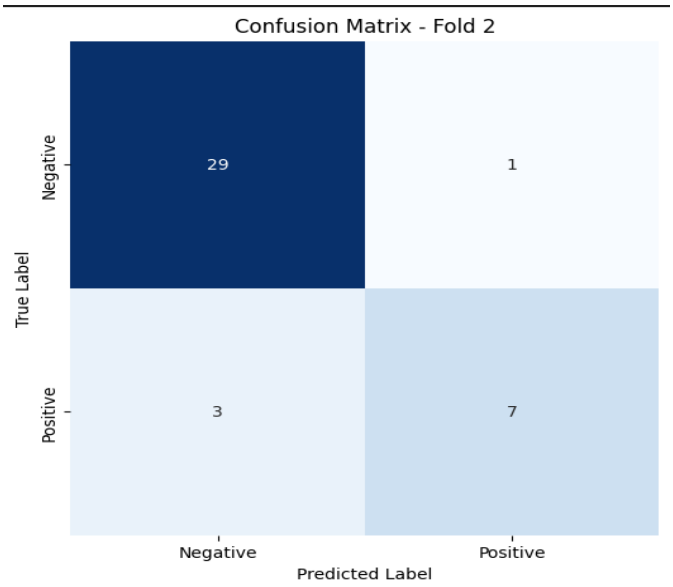
Gambar 3 menunjukkan *confusion matrix* dari hasil model *deep learning* menggunakan *oversampling fold* 1 bahwa 29 data berhasil diprediksi dengan benar sebagai *non-recur* atau *true negative* dan 2 data *non-recur* yang salah diidentifikasi sebagai *recur* atau *false positive* dan 29 data berhasil diprediksi sebagai *recur* atau *true positive* dan 2 data *non-recur* yang salah diidentifikasi sebagai *recur* atau *false negative.*



Gambar 3. Confusion Matrix Prognostic menggunakan *oversampling*

D. Confusion Matrix Prognostic tanpa menggunakan Oversampling

Gambar 4 menunjukan confusion matrix dari hasil model *deep learning* tanpa menggunakan *oversampling fold* 2 bahwa 29 data berhasil diprediksi dengan benar sebagai *non-recur* atau *true negative* dan 1 data *non-recur* yang salah diindentifikasi sebagai *recur* atau *false positive* dan 7 data berhasil diprediksi sebagai *recur* atau *true positive* dan 3 data *non-recur* yang salah diindentifikasi sebagai *recur* atau *false negative.*



Gambar 4. *Confusion Matrix Prognostic* Tanpa *Oversampling*

IV. Simpulan Dan Saran

Berdasarkan dari hasil penelitian yang dilakukan dengan menggunakan metode-metode seperti *oversampling* yang digunakan untuk bertujuan mengatasi ketidakseimbangan kelas dalam *dataset* dengan memperbanyak jumlah contoh dari kelas minoritas, *hyperparameter tuning* yang digunakan untuk menemukan parameter yang memberikan kinerja yang terbaik dalam setiap model serta menggunakan *k-fold cross validation* yang digunakan untuk memastikan bahwa model tidak *overfitting* dan memberikan kinerja yang akurat. Analisis pada penelitian ini dilakukan dengan melihat nilai *F1-Score* dari setiap model yang digunakan dapat disimpulkan bahwa model *Logistic Regression* merupakan model yang paling cocok untuk melakukan analisis dalam *dataset diagnostic* dibandingkan dengan model *deep learning* dan *K-Nearest Neighbors* sedangkan untuk *dataset prognostic* model yang paling cocok merupakan model *Deep Learning* menggunakan *oversampling.*

Saran yang dapat digunakan untuk penelitian selanjutnya dapat melakukan eksplorasi dengan menggunakan metode yang berbeda, menggunakan *random search* pada *hyperparameter tuning* agar dapat membandingkan hasil dengan *grid-search* serta melakukan *feature selection* untuk mencari fitur terbaik yang dapat digunakan agar model yang telah dibuat dapat dikembangkan menjadi aplikasi.

Daftar Pustaka

[1] A. I. S. Azis, I. S. K. Idris, B. Santoso, and Y. A. Mustofa, “Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara,” 2019.
[2] A. L. Munawir, S. A. Srigati, and P. Wulandari, “POTENSI KECERDASAN BUATAN DALAM PENINGKATAN AKURASI PEMBACAAN HASIL MAMOGRAFI: TINJAUAN SISTEMATIS DAN METAANALISIS,” 2023.
[3] S. KRISHNA, “How to decide Number of Neurons In Input and Output Layers and how many Hidden Layers and how many neurons are required in Hidden Layers?” Accessed: Jul. 01, 2024. [Online]. Available: https://www.kaggle.com/discussions/general/321114
[4] A. A. Soebroto, “Buku Ajar AI, Machine Learning & Deep Learning,” 2019. [Online]. Available: https://www.researchgate.net/publication/348003841
[5] S. A. Medjahed, T. A. Saadi, and A. Benyettou, “Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules,” 2013.
[6] R. Sitepu, “Implementasi Algoritma K-Nearest Neigbor Untuk Klasifikasi Pengajuan Kredit.”
[7] A. Sharma, S. Kulshrestha, and S. Daniel, “Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis.”
[8] D. Y. Utami, E. Nurlelah, and F. N. Hasan, “Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes,” *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 5, no. 1, pp. 53–64, Jul. 2021, doi: 10.31289/jite.v5i1.5201.
[9] J. Brandt and E. Lanzén, “A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification.”

[10] T. Ridwansyah, "KLIK: Kajian Ilmiah Informatika dan Komputer Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *Media Online*, vol. 2, no. 5, pp. 178–185, 2022, [Online]. Available: https://djournals.com/klik
[11] M. Elgendy, O'Reilly for Higher Education (Firm), and an O. M. Company. Safari, *Deep Learning for Vision Systems*.
[12] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Comput Appl*, vol. 29, no. 12, pp. 1545–1554, Jun. 2018, doi: 10.1007/s00521-016-2701-1.
[13] K. Wadkar, P. Pathak, and N. Wagh, "Article ID: IJCET_10_03_009 Detection using Ann Network and Performance Analysis with SVM," *International Journal of Computer Engineering & Technology (IJCET)*, vol. 10, no. 3, pp. 75–86, [Online]. Available: http://www.iaeme.com/IJCET/index.asp75http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=10&IType=3JournalImpactFactor