ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika

ISSN(p): 2338-8323 | ISSN(e): 2459-9638 | Vol. 12 | No. 3 | Halaman 556 - 568
DOI : http://dx.doi.org/10.26760/elkomika.v12i3.556 Juli 2024

Prediksi Kanker Paru menggunakan *Grid search* untuk Optimasi *Hyperparameter* pada Algoritma MLP dan *Logistic Regression*

NOR KUMALASARI CAECAR PRATIWI, NUR IBRAHIM, SOFIA SAIDAH

Fakultas Teknik Elektro, Universitas Telkom, Indonesia
Email: caecarnkcp@telkomuniversity.ac.id

Received 19 Januari 2024 **|** *Revised* 1 Maret 2024 **|** *Accepted* 5 April 2024

ABSTRAK

Kanker paru merupakan penyebab utama kematian akibat kanker di seluruh dunia. Prediksi dini kanker paru-paru telah banyak dilakukan, baik berbasis citra maupun data mentah. Prediksi kanker paru berbasis citra memberikan dampak positif dalam diagnosis dini, namun pendekatan berbasis data mentah juga penting dalam memahami faktor risiko dan kondisi yang dapat mempengaruhi perkembangan kanker. Penelitian ini mengusulkan sistem prediksi dini kanker paru dengan basis data klinis dan demografi, menggunakan Multi-Layer Perceptron (MLP) dan logistic regression dengan pemanfaatan grid search optimizer. Kedua model mencapai tingkat akurasi, presisi, recall, dan f1-score sebesar 1, optimal dalam melakukan prediksi data. Pada logistic regression, solver liblinear, penalty L1, dan nilai C yang lebih tinggi berkontribusi pada peningkatan akurasi. Sedangkan pada MLP, konfigurasi aktivasi tanh dan solver adam menghasilkan akurasi yang lebih baik. Hasil ini memberikan keyakinan implementasi MLP dan logistic regression, memiliki potensi dalam mendukung prediksi kanker paru-paru.

***Kata kunci**: kanker paru, multi-layer perceptron, logistic regression, grid search*

ABSTRACT

Lung cancer is a leading cause of cancer-related deaths worldwide. Early prediction of lung cancer has been widely conducted, both based on images and raw data. Image-based lung cancer prediction has a positive impact on early diagnosis, but a raw data-driven approach is also crucial for understanding risk factors and conditions that can influence cancer development. This research proposes an early lung cancer prediction system using clinical and demographic data, employing Multi-Layer Perceptron (MLP) and logistic regression with the utilization of grid search. Both models achieved an accuracy, precision, recall, and f1-score of 1, optimal in classifying data. In logistic regression, the liblinear solver, L1 penalty, and higher C values contributed to increased accuracy. Meanwhile, in MLP, the configuration of tanh activation and adam solver yielded better accuracy. These results instill confidence that the implementation of MLP and logistic regression has significant potential in supporting lung cancer prediction.

***Keywords**: lung cancer, multi-layer perceptron, logistic regression, grid search*

1. PENDAHULUAN

Kanker paru-paru adalah jenis kanker yang paling sering didiagnosis dan penyebab utama kematian akibat kanker secara global di dunia, menurut data dari GLOBOCAN2020, diperkirakan terdapat sekitar 1,8 juta kematian akibat kanker paru-paru (81%) pada tahun 2020 dari total 2,2 juta kasus **(Li, dkk, 2023)**. Polusi udara merupakan campuran kompleks dari berbagai zat, telah diidentifikasi berpotensi menjadi karsinogen penyebab utama kanker paru-paru **(Berg, dkk, 2023)**. Kondisi udara yang tidak sehat dapat menyebabkan kanker paru-paru bukan dengan merusak DNA, tetapi dengan menciptakan lingkungan yang mengakibatkan peradangan dan mendorong terjadinya proliferasi sel dengan mutasi pengendali kanker yang sudah ada **(Ledford, 2023)**. Faktor dominan lainnya yang bisa menjadi pendorong terjadinya kanker paru bisa berasal dari kebiasaan konsumsi alkohol, faktor genetik, dan obesitas **(Kanwal, dkk, 2017) (Troche, dkk, 2016) (Vedire, dkk, 2023)**. Sebuah studi mengemukakan bahwa penyintas kanker paru-paru yang tidak aktif bergerak tidak hanya bergejala batuk berkepanjangan saja, namun juga sering cepat kelelahan, merasa nyeri, dan sesak napas yang lebih parah dibandingkan dengan penyintas yang tetap aktif secara fisik **(Lehto, 2016)**. Dari beberapa fakta tersebut, penelitian ini akan melakukan prediksi ditemukannya atau terdiagnosanya seseorang dengan kanker paru-paru berdasarkan beberapa informasi terkait data lingkungan (tingkat polusi udara), gaya hidup (kebiasaan merokok, tingkat pasif merokok, konsumsi alkohol), serta kondisi kesehatan (level alergi, resiko genetik, nyeri dada, batuk, sesak nafas, penebalan kuku dan level mendengkur).

Prediksi dini terjadinya kanker paru-paru telah banyak dilakukan oleh para peneliti terdahulu, baik berbasi data citra maupun data mentah. Beberapa penelitian dan aplikasi klinis menggunakan data citra, seperti hasil *CT-scan* atau gambar radiologi untuk mendeteksi atau memprediksi kanker paru-paru. Teknik-teknik seperti analisis tekstur dan *deep learning* pada citra medis dapat memberikan informasi yang berharga. Penelitian yang dilakukan oleh Sharmila Nageswaran dan kolega melakukan klasifikasi dan prediksi terhadap kanker paru-paru berbasis citra *CT-scan*, menunjukan hasil bahwa model *Artificial Neural Network* (ANN) memiliki performansi yang lebih akurat untuk memprediksi kanker paru-paru dibandingkan *K-Nearest Neighbour* (KNN) dan *Random Forest* (RF) **(Nageswaran, dkk, 2022)**. Prediksi kanker paru dengan menggunakan masukan citra *CT-scan* dan *X-ray* juga dilakukan oleh **(Deepapriya, dkk, 2023)** dan **(Rajasekar, dkk, 2023)**. Dalam penelitian **(Shanbhag, dkk, 2022)** diusulkan algoritma klasifikasi *ensemble* untuk mendeteksi kanker paru-paru menggunakan *CT-scan*. Dalam klasifikasi *ensemble* disertai oleh lima model pembelajaran mesin yaitu SVM, LR, MLP, Decision-tree, dan KNN. Akurasi klasifikasi *ensemble* mencapai 85% dapat membedakan antara kanker Malignan dan Benign. Dengan menerapkan beberapa metode peningkatakan kualitas citra seperti *stretching* kontras dan koreksi nilai gamma, ternyata mampu dengan signifikan meningkatkan kinerja prediksi dengan akurasi mencapai 100% dan AUC 1.00 menggunakan kernel SVM RBF dan polynomial **c** Dari beberapa uraian penelitian tersebut, menunjukan bahwa prediksi kanker paru berbasis data citra telah menjadi bagian penting dalam dunia medis modern, memberikan dampak positif dalam diagnosis dini dan pengelolaan penyakit paru-paru.

Namun, pendekatan berbasis data mentah juga penting dalam memahami faktor-faktor risiko dan kondisi yang dapat mempengaruhi perkembangan kanker paru-paru. Informasi seperti riwayat merokok, paparan asap rokok, kondisi lingkungan, dan riwayat kesehatan pribadi dapat memberikan konteks dan dukungan untuk analisis prediktif. Penelitian prediksi seorang berpotensi terkena kanker paru dilakukan kepada 504 pasien Rumah Sakit Universitas Karolinska Stockholm-Swedia, pasien terbagi kedalam beberapa kondisi yaitu aktif merokok, mantan perokok, dan yang tidak pernah merokok **(Nemlander, dkk, 2022)**. Hasil penelitian

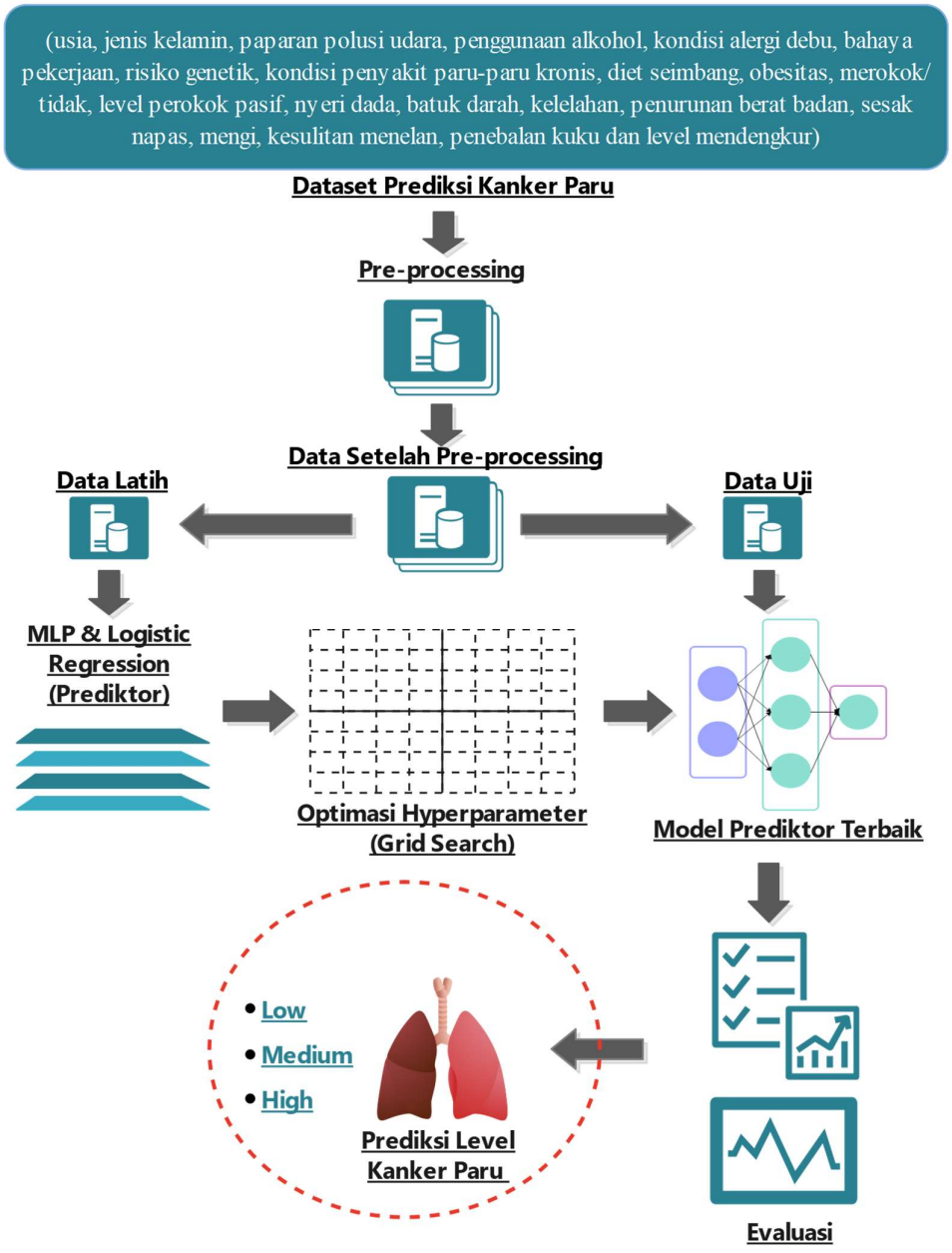
menyimpulkan bahwa alat penilaian (*tools assessing*) sangat dibutuhkan, hal ini terkait temuan bahwa ternyata seorang yang tidak pernah merokok sering terdeteksi pada tahap akhir kanker paru (tidak terdektsi dini). Menjawab tantangan besar untuk mengidentifikasi gejala pada tahap awal kanker paru-paru merupakan tujuan dari penelitian yang dilakukan oleh Y Ge dan rekan-rekan **(Ge, dkk, 2015)**. Sebanyak 345.600 orang terlibat sebagai responden penelitian, data yang dikumpulkan mencakup data demografis, CBC (*Complete Blood Count*), CMP (*Complete Metabolic Panel*), lipid, dan data urinalisis. Studi **(Munawar, dkk, 2022)** merancang metode yang efisien untuk diagnosis awal pasien kanker paru dengan gejala melalui data demografis dan riwayat klinis.

Penelitian yang diuraikan di atas, menyimpulkan bahwa pendekatan berbasis data mentah sangat penting dalam pemahaman faktor-faktor risiko dan dapat mengoptimalkan upaya pencegahan dan deteksi dini kanker paru. Tujuan penelitian ini adalah untuk mengembangkan sebuah model prediktif yang dapat memprediksi kemungkinan seseorang terkena kanker paru-paru berdasarkan beberapa faktor, termasuk data lingkungan (tingkat polusi udara), gaya hidup (kebiasaan merokok, tingkat pasif merokok, konsumsi alkohol), serta kondisi kesehatan (level alergi, resiko genetik, nyeri dada, batuk, sesak nafas, penebalan kuku, dan level mendengkur). Nilai yang ingin dicapai dalam penelitian ini adalah model prediktif yang memiliki tingkat akurasi yang signifikan (> 95%) dalam memprediksi kemungkinan seseorang terkena kanker paru-paru.

2. METODE PENELITIAN

Sistem prediksi dini kanker paru-paru ini diusulkan dengan membandingkan kinerja dua algoritma *predictor* yang berbeda, yaitu *Logistic Regression* dan *Multi-Layer Perceptron* (MLP). Setiap algoritma diuji dengan *hyperparameter* yang berbeda-beda untuk menemukan hasil kinerja system terbaik. Sistem dirancang dengan menggunakan *grid search*, sebuah teknik pencarian atau pengaturan konfigurasi *hyperparameter* yang digunakan dalam pembelajaran mesin untuk menemukan kombinasi optimal yang menghasilkan kinerja model terbaik. Gambar 1 di bawah ini mengilustrasikan sistem yang diusulkan secara keseluruhan. Dataset yang digunakan ialah dataset publik **(Kaggle, n.d.)**, berisi informasi tentang pasien dengan kondisi kanker paru-paru level *Low*, *Medium* dan *High*. Informasi pendukung berupa data usia pasien, jenis kelamin, paparan polusi udara, konsumsi alkohol, alergi debu, risiko pekerjaan, risiko genetik, penyakit paru-paru kronis, pola makan seimbang, obesitas, merokok, perokok pasif, nyeri dada, batuk darah, kelelahan, kehilangan berat badan, sesak napas, mendengus, kesulitan menelan, penebalan kuku jari, dan mendengkur. Data dikumpulkan dari 1000 pasien, Dimana sebanyak 303 pasien berada dalam kondisi kanker paru *low*, 332 kondisi kanker paru *medium* dan 365 kondisi kanker paru *high*. Dari total 1000 *raw data*, pada tahap *pre-processing* akan dibagi menjadi data latih dan data uji, dengan komposisi sebesar 80% untuk data latih dan 20% untuk data uji. Data paparan polusi udara, kebiasaan merokok, konsumsi alkohol, dan faktor genetik merupakan data untuk identifikasi faktor risiko, memungkinkan analisis untuk menilai hubungan antara faktor-faktor ini dan risiko pengembangan kanker paru-paru. Penelitian telah menunjukkan bahwa paparan polusi udara dan merokok aktif secara signifikan meningkatkan risiko kanker paru-paru **(Buana & Harahap, 2022).** Informasi tentang gejala awal seperti nyeri dada, batuk darah, kelelahan, penurunan berat badan, dan sesak napas dapat digunakan untuk mengidentifikasi individu yang mungkin membutuhkan evaluasi lebih lanjut untuk deteksi dini kanker paru-paru. Gejala-gejala ini sering kali menjadi tanda-tanda awal penyakit dan dapat mempercepat diagnosis **(Prado, dkk, 2023)**. Dengan memanfaatkan data tersebut di atas, penelitian dapat mengembangkan model prediktif yang menggunakan faktor-faktor risiko dan gejala sebagai prediktor untuk

kemungkinan seseorang terkena kanker paru-paru. Penggunaan teknik analisis statistik dan *machine learning* dapat membantu mengidentifikasi pola-pola kompleks dalam data yang dapat digunakan untuk prediksi dini. Data ini dapat digunakan untuk memvalidasi model prediktif melalui studi klinis dan epidemiologi. Validasi medis penting untuk menilai kinerja model dalam dunia nyata dan memastikan bahwa model tersebut dapat diterapkan secara efektif dalam praktik klinis.



Gambar 1. Sistem Prediksi Kanker Paru

2.1 *Logistic Regression*

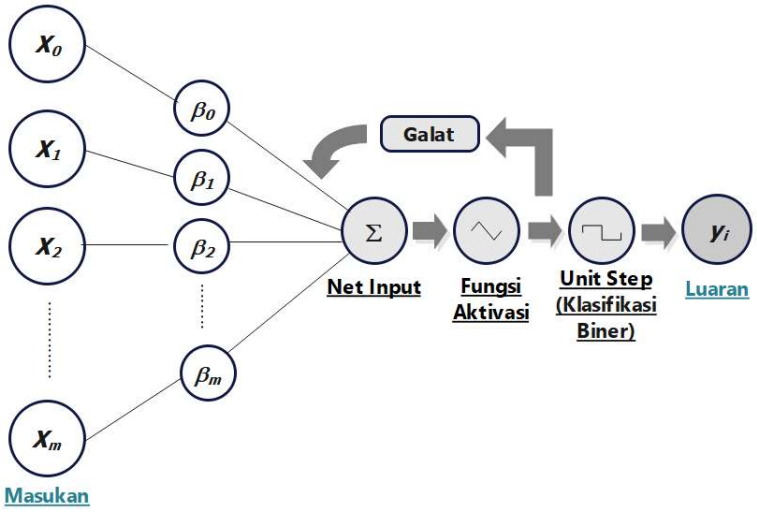
Pada bagian ini akan dijelaskan secara singkat cara kerja dari *logistic regression*. Misal kita mempunyai sekumpulan data X dan Y, dengan dua kemungkinan luaran, yaitu $y_i = 0$ (kelas negatif) atau $y_i = 1$ (kelas positif) untuk setiap kondisi pada sekumpulan data X. Proses selanjutnya ialah menentukan fungsi yang dapat membedakan apakah suatu kejadian x_i termasuk dalam kelas yang positif atau negatif. Fungsi ini ditunjukkan dalam persamaan berikut **(Biswas, dkk, 2023)**:

$$P(y = 1|X, \beta) = \frac{e^{\beta X}}{1 + e^{\beta X}} \tag{1}$$

Persamaan (1) digunakan untuk prediksi besarnya peluang luaran $y = 1$ (positif) berdasarkan (syarat) variabel *predictor* X dan parameter model β. Lebih lanjut, β menunjukkan seberapa besar pengaruh setiap variabel prediktor terhadap variabel target. Nilai β ditentukan selama proses pelatihan model, menggunakan fungsi:

$$\beta^* = \arg\left\{\prod_{i=1}^{n} P(y_i|x_i, \beta)\right\} = arg\left\{\sum_{i=1}^{n} y_i \log\log\left(\frac{1}{1 + e^{-\beta X}}\right) + (1 - y_i) \log\log\left(\frac{1}{1 + e^{-\beta X}}\right)\right\} \tag{2}$$

Adapun e ialah konstanta euler yang bernilai sekitar 2.71828. Sementara Persamaan (2) ialah fungsi *likelihood* model *logistic regression*, yang digunakan untuk mencari parameter model β yang memberikan probabilitas paling tinggi (*likelihood* tertinggi) terhadap data yang diamati. Fungsi log digunakan untuk mengonversi probabilitas ke dalam bentuk logaritmik, yang dapat mempermudah proses optimisasi parameter model. Gambar 2 di bawah ini menunjukan arsitektur dasar *logistic regression* dengan menggunakan array data sebagai masukan ke bagian *net-input*, *activation* dan *unit step* **(Biswas, dkk, 2023)**.

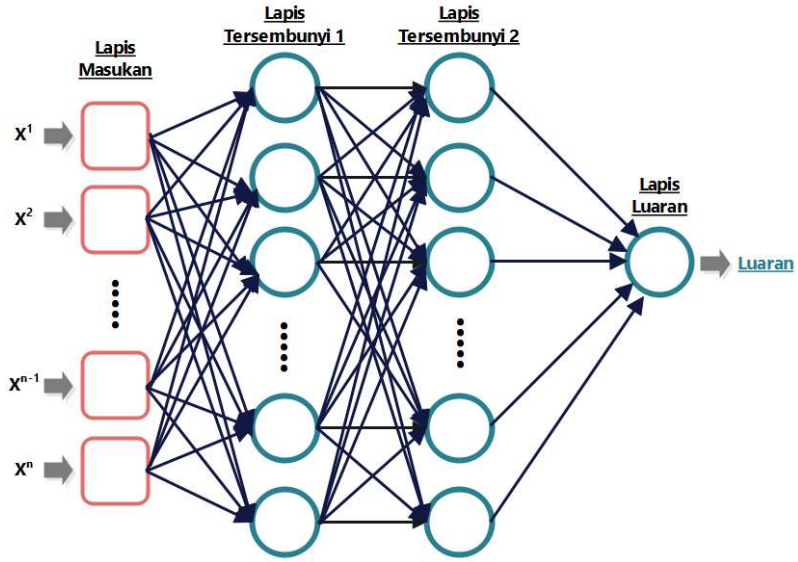


Gambar 2. Arsitektur Umum *Logistic Regression*

2.2 *Multi-Layer Perceptron* (MLP)

Artificial Neural Network (ANN) terdiri dari sejumlah besar elemen pemrosesan yang saling terhubung, disebut sebagai neuron atau sel, yang terhubung bersama dengan koneksi berbobot **(Panjaitan, dkk, 2018)**. ANN yang paling banyak diterapkan untuk tujuan regresi dan klasifikasi salah satunya ialah *Multi-Layer Perceptron* (MLP), yang terdiri dari lapisan masukan, lapisan keluaran, dan satu atau lebih lapisan tersembunyi (*hidden layer*) di mana

setiap lapisan memiliki matriks bobot W dan vektor bias b **(Shirazi & Frigaard, 2021)**. Gambar 3 mengilustrasikan arsitektur jaringan MLP secara umum. Setiap *node* atau neuron di setiap lapisan MLP akan terhubung sepenuhnya ke semua *node*/neuron pada lapisan berikutnya.



Gambar 3. Arsitektur Umum *Multi-Layer Perceptron* (MLP)

Dalam penelitian ini kami menggunakan konfigurasi satu atau dua lapisan tersembunyi, dengan masing-masing berukuran 50 neuron. Dari dua kondisi ini, akan dipilih konfigurasi jumlah lapisan tersembunyi terbaik oleh GridSearchCV. Saat melakukan GridSearchCV, model akan mencoba berbagai kombinasi dari *hyperparameter* (termasuk jumlah lapisan/arsitektur) dan menilai kinerja setiap model berdasarkan metrik evaluasi yang ditentukan, dalam kasus ini adalah akurasi. Tidak ada batasan pada jumlah lapisan tersembunyi yang dapat digunakan, pada penelitian ini kami mengusulkan maksimal dua lapisan tersembunyi. Faktanya menunjukan bahwa model dengan dua lapisan tersembunyi sudah cukup untuk menangani masalah dengan akurasi 100%. Namun, dalam beberapa kasus, menambahkan lebih banyak lapisan tersembunyi atau meningkatkan jumlah neuron dalam setiap lapisan tersembunyi dapat membantu dalam menangani masalah yang lebih kompleks atau meningkatkan generalisasi model. Hal ini tergantung pada kompleksitas masalah dan karakteristik data yang dihadapi.

2.3 Parameter Kinerja Sistem

Evaluasi dilakukan dengan menggunakan parameter akurasi, presisi, *recall*, dan F1-*score*. Akurasi dilakukan untuk mengukur sejauh mana model klasifikasi benar-benar dapat mengidentifikasi dan memisahkan setiap kelasnya dengan tepat. Akurasi cenderung lebih sederhana dan mudah untuk diinterpretasi namun tidak cocok untuk dataset yang tidak seimbang (*imbalance dataset*). Presisi dalam model klasifikasi adalah ukuran perbandingan jumlah *instance* yang benar-benar positif (*True Positive*) dengan jumlah *instance* yang diprediksi sebagai positif oleh model (*True Positive + False Positive*). *Recall* juga dikenal sebagai sensitivitas atau *true positive rate*, adalah metrik evaluasi dalam model klasifikasi yang membandingkan jumlah instance yang benar-benar positif (*True Positive*) dengan jumlah instance positif secara keseluruhan (*True Positive + False Negative*). F1-*Score* adalah metrik evaluasi dalam model klasifikasi yang menyatukan presisi dan *recall* menjadi satu skor, hal ini

berguna ketika kita ingin mencari keseimbangan antara kedua metrik tersebut. Parameter tersebut dapat diselesaikan denga menggunakan persamaan berikut **(Pratiwi, dkk, 2020)**:

$$Akurasi = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (3)$$

$$Presisi = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (6)$$

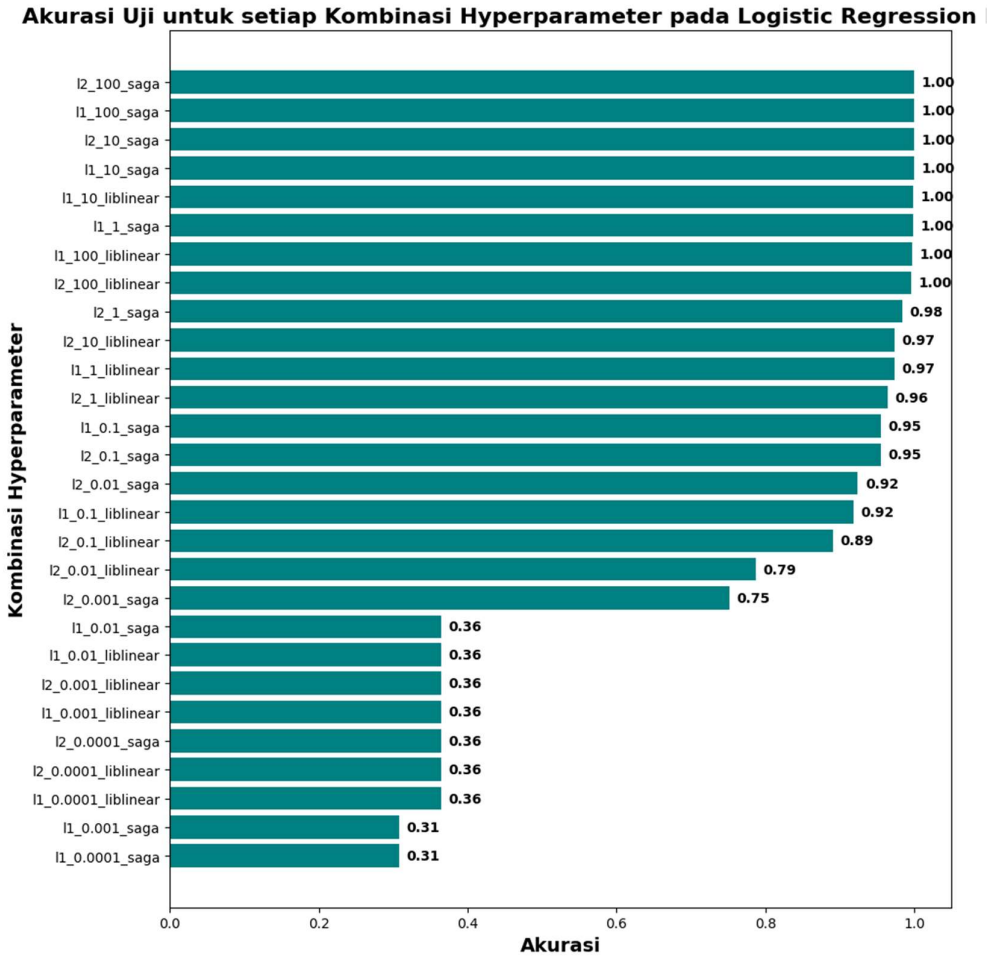
3. HASIL DAN DISKUSI

Penelitian ini menghadirkan prediksi kanker paru-paru dengan memfokuskan perbandingan kinerja dua *predictor* utama, yaitu *Multi-Layer Perceptron* (MLP) dan Logistic Regression. Pemilihan dua *predictor* ini didasarkan pada karakteristik unggul masing-masing, dengan harapan mendapatkan wawasan yang lebih mendalam mengenai kemampuan keduanya dalam memprediksi kemungkinan kanker paru-paru. MLP, sebagai model jaringan saraf tiruan, memiliki keunggulan memproses informasi dari beberapa lapisan atau tingkatan, membuat MLP sangat sesuai untuk masalah yang melibatkan hubungan yang kompleks di antara variabel. Di sisi lain*, logistic regression* sebagai model linier, menawarkan kemampuan untuk memberikan bobot atau koefisien untuk setiap variabel *input*, sehingga memudahkan identifikasi faktor-faktor yang signifikan dalam kaitannya dengan risiko kanker paru-paru. Penelitian ini menjadikan *grid search* sebagai elemen kunci dalam penentuan parameter yang optimal untuk model prediksi kanker paru-paru. Pendekatan ini memungkinkan peneliti untuk mencari kombinasi parameter terbaik dengan menguji berbagai nilai yang mungkin, sehingga memastikan bahwa model yang dikembangkan memiliki kinerja optimal. Dengan mengintegrasikan *grid search* ke dalam metodologi penelitian, diharapkan dapat ditemukan konfigurasi parameter yang paling sesuai untuk meningkatkan ketepatan prediksi kanker paru-paru, memperkaya kontribusi penelitian ini dalam pengembangan metode diagnostik yang lebih canggih. Pada *predictor logistic regression*, tiga parameter kunci yang dioptimalkan melalui *grid search* adalah *penalty*, nilai C, dan jenis *solver*. Parameter *penalty* menentukan jenis regularisasi yang diterapkan pada model. Regularisasi adalah teknik untuk mencegah *overfitting* dengan menambahkan sejumlah pembobotan tambahan pada model. Nilai C mewakili *invers* dari kekuatan regularisasi. Parameter *solver* menentukan algoritma yang digunakan untuk menyelesaikan masalah optimasi. *Liblinear* cocok untuk dataset yang relatif kecil, sementara *saga* lebih cocok untuk dataset yang lebih besar dan menangani regularisasi *lasso*. Tabel 1 di bawah ini memberikan informasi *hyperparameter* yang digunakan pada *logistic regression*.

Tabel 1. *Hyperparameter* yang digunakan pada *Logistic Regression*

Hyperparameter	**Nilai**
Penalty	*Lasso* (L1), *Ridge* (L2)
C	0.001, 0.01, 0.1, 1, 10, 100
Solver	*Liblinear*, *Saga*

Gambar 4 menunjukkan akurasi uji dari *predictor logistic regression* pada setiap kombinasi *hyperparameter*. Jika meninjau dari segi jenis *solver*, untuk *saga*, akurasi uji tetap sama sebesar 0.31 baik untuk *penalty* L1 (*lasso*) maupun L2 (*ridge*) dengan berbagai nilai C. Sedangkan untuk *liblinear*, akurasi uji lebih tinggi sebesar 0.36 untuk berbagai kombinasi, dan akurasi tertinggi dicapai adalah sebesar 1.0 baik untuk *penalty* L1 maupun L2 dengan nilai C sebesar 10, 100, dan 1. Dari segi *penalty*, L1 dengan *solver liblinear* menunjukkan akurasi uji yang konsisten sebesar 0.36 untuk berbagai nilai C. *Penalty* L2 dengan *solver liblinear* menunjukkan peningkatan akurasi, mencapai 0.964 untuk $C = 1$ dan mencapai akurasi sempurna (1.0) untuk nilai C sebesar 10, 100, dan *solver Saga* dengan $C = 10$. Untuk nilai C, secara umum nilai C yang lebih tinggi (10, 100) umumnya menghasilkan peningkatan akurasi uji, terutama untuk *solver liblinear* dengan *penalty* L1 maupun L2. Beberapa kombinasi, terutama yang menggunakan *solver liblinear* dan *penalty* L1, menunjukkan akurasi uji yang tinggi dan konsisten untuk berbagai nilai C. Performa model sensitif terhadap pilihan *hyperparameter*, dan beberapa kombinasi menghasilkan akurasi sempurna pada set uji. Secara keseluruhan, analisis menunjukkan bahwa *solver liblinear*, *penalty* L1, dan nilai *C* yang lebih tinggi berkontribusi pada peningkatan akurasi, dengan beberapa kombinasi mencapai akurasi sempurna pada set uji.



Gambar 4. Akurasi Uji pada setiap Kombinasi *Hyperparameter* pada *Logistic Regression*

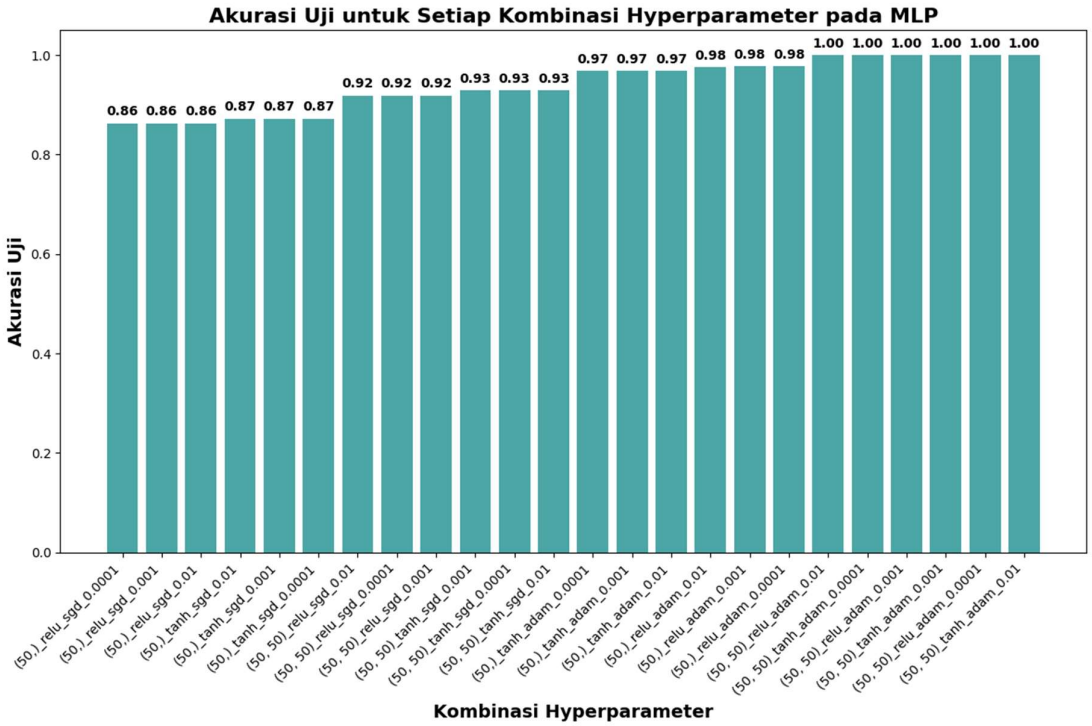
Pada model MLP, parameter-parameter yang digunakan pada penyetelan *hyperparameter* mencakup konfigurasi ukuran *hidden layer*, pemilihan fungsi aktivasi, pemilihan *solver*, dan

parameter *alpha*. Penyesuaian parameter ini bertujuan untuk menemukan konfigurasi optimal yang dapat menghasilkan kinerja model MLP yang unggul untuk tugas klasifikasi yang diberikan. Ukuran *hidden layer* akan menentukan arsitektur jaringan dengan menyediakan tuple yang berisi jumlah unit di setiap *hidden layer*. Misalnya (50) akan memiliki satu *hidden layer* dengan 50 unit, sementara (50, 50) akan memiliki *dua hidden layer*, masing-masing dengan 50 unit. Fungsi aktivasi digunakan oleh neuron di dalam jaringan. Pilihan umum melibatkan *relu* (*Rectified Linear Unit*), fungsi aktivasi yang umum digunakan untuk memperkenalkan non-linearitas, tanh (tangen hiperbolik), fungsi aktivasi yang menghasilkan nilai antara -1 dan 1 dan *logistic*, yang menghasilkan nilai antara 0 dan 1. *Solver* digunakan untuk menentukan algoritma yang digunakan untuk menemukan bobot optimal. Nilai *alpha* ialah parameter regularisasi yang mengendalikan kekuatan regularisasi pada bobot. Nilai yang lebih tinggi dapat membantu mencegah *overfitting.* Tabel 2 di bawah ini memberikan informasi *hyperparameter* yang digunakan pada MLP.

Tabel 2. *Hyperparameter* yang digunakan pada *Logistic Regression*

Hyperparameter	**Nilai**
Hidden Layer	50 dan (50,50)
Activation Function	Relu, tanh
Solver	*Adam, SGD*
Alpha value	*0.0001, 0.001, 0.01*

Gambar 5 menunjukkan akurasi uji dari prediktor MLP pada setiap kombinasi *hyperparameter.* Nilai akurasi berkisar dari 0.23 hingga 1, menunjukkan variasi kinerja model yang luas di berbagai kombinasi hyperparameter.



Gambar 5. Akurasi Uji pada setiap Kombinasi *Hyperparameter* pada MLP Prediktor

Meninjau pengaruh ukuran *hidden layer* dan fungsi aktivasi, model dengan ukuran *hidden layer* (50,50) umumnya lebih baik daripada yang memiliki satu *layer* saja (50). Konfigurasi

seperti (50,50)_relu_sgd_0.01, (50,50)_relu_sgd_0.0001, dan (50,50)_relu_sgd_0.001, menunjukkan peningkatan akurasi hingga mencapai 0.92. Pilihan fungsi aktivasi memengaruhi kinerja, dengan *tanh* seringkali lebih baik *relu*. Sebagai contoh, (50,)_tanh_sgd_0.01, (50,)_tanh_sgd_0.001, dan (50,)_tanh_sgd_0.0001 semua memiliki akurasi sebesar 0.87. Secara keseluruhan, pemilihan *hyperparameter* secara signifikan memengaruhi kinerja model MLP. Hasilnya menunjukkan bahwa konfigurasi dengan aktivasi 'tanh', *solver* 'adam', dan beberapa tingkat pembelajaran tertentu menghasilkan akurasi yang lebih baik. Selain itu, memiliki lapisan tersembunyi tambahan tampaknya meningkatkan performa dalam beberapa kasus. Secara keseluruhan, pemilihan *hyperparameter* secara signifikan memengaruhi kinerja model MLP. Hasilnya menunjukkan bahwa konfigurasi dengan aktivasi *tanh*, *solver adam*, dan beberapa tingkat pembelajaran tertentu menghasilkan akurasi yang lebih baik.

	precision	recall	f1-score	support
High	1.00	1.00	1.00	92
Low	1.00	1.00	1.00	82
Medium	1.00	1.00	1.00	76
accuracy			1.00	250
macro avg	1.00	1.00	1.00	250
weighted avg	1.00	1.00	1.00	250

Logistic Regression

	precision	recall	f1-score	support
High	1.00	1.00	1.00	92
Low	1.00	1.00	1.00	82
Medium	1.00	1.00	1.00	76
accuracy			1.00	250
macro avg	1.00	1.00	1.00	250
weighted avg	1.00	1.00	1.00	250

MLP

Gambar 6. *Classification Report* pada *Logistic Regression* dan MLP

Gambar 6 menunjukkan *classification report* untuk data uji pada *predictor logistic regression* dan MLP. Kedua model telah mencapai performa sempurna pada set uji, menunjukkan kemampuan yang sangat baik dalam mengklasifikasikan data dengan benar. Akurasi, presisi, *recall*, dan f1-*score* yang mencapai nilai sempurna menunjukkan bahwa *predictor* dapat mengidentifikasi dan memisahkan dengan sempurna antara setiap kelas set uji. Konsistensi performa yang tinggi menunjukkan bahwa *predictor* ini mampu menjaga tingkat keakuratan yang tinggi di berbagai skenario. Hasil ini juga menunjukkan bahwa pengaturan *hyperparameter* yang tepat dapat berkontribusi pada pencapaian performa sempurna.

4. KESIMPULAN

Dengan menggunakan model *Multi-Layer Perceptron* (MLP) dan *Logistic Regression*, prediksi kanker paru-paru berdasarkan data *raw* telah berhasil dilakukan dengan kinerja yang sangat baik. Kedua model tersebut mencapai tingkat akurasi, presisi, *recall*, dan f1-*score* sebesar 100%, menunjukkan kemampuan optimal dalam mengklasifikasikan *instance-instance* pada dataset. Proses pencarian kombinasi *hyperparameter* terbaik menggunakan *Grid search* juga memberikan kontribusi besar terhadap kesuksesan ini, memungkinkan identifikasi konfigurasi *hyperparameter* yang optimal untuk mencapai kinerja model yang sempurna. Pada *logistic regression*, secara keseluruhan, analisis menunjukkan bahwa *solver liblinear*, *penalty* L1, dan nilai *C* yang lebih tinggi berkontribusi pada peningkatan akurasi, dengan beberapa kombinasi mencapai akurasi sempurna pada set uji. Sedangkan pada MLP, hasilnya menunjukkan bahwa konfigurasi dengan aktivasi *tanh*, *solver adam*, dan beberapa tingkat pembelajaran tertentu menghasilkan akurasi yang lebih baik. Hasil ini memberikan keyakinan bahwa implementasi model-*machine learning*, khususnya MLP dan *logistic regression*, memiliki potensi besar dalam mendukung prediksi kanker paru-paru dengan tingkat akurasi yang sangat tinggi.

DAFTAR RUJUKAN

Berg, C. D., Schiller, J. H., Boffetta, P., Cai, J., Connolly, C., Kerpel-Fronius, Kitts, A., Lam, A. B., C.L., D., Mohan, A., Myers, R., Suri, T., Tammemagi, M. C., Yang, D., & Lam, S. (2023). Air Pollution and Lung Cancer: A Review by International Association for the Study of Lung Cancer Early Detection and Screening Committee. *Journal of Thoracic Oncology*, *18*(10), 1277–1289.

Biswas, S., Ghosh, S., Roy, S., Bose, R., & Soni, S. (2023). A Study of Stock Market Prediction through Sentiment Analysis. *Mapana Journal of Sciences*, *22*(1), 89–120.

Buana, I., & Harahap, D. A. (2022). Asbestos, Radon Dan Polusi Udara Sebagai Faktor Resiko Kanker Paru Pada Perempuan Bukan Perokok. *AVERROUS: Jurnal Kedokteran dan Kesehatan Malikussaleh*, *8*(1), 1–16. https://doi.org/10.29103/averrous.v8i1.7088

Deepapriya, B. S., Kumar, P., Nandakumar, G., Gnanavel, S., Padmanaban, R., Anbarasan, A. K., & Meena, K. (2023). Performance evaluation of deep learning techniques for lung cancer prediction. *Soft Computing*, *27*(13), 9191–9198. https://doi.org/10.1007/s00500-023-08313-7

Ge, Y., Ma, L., Tao, L. W., Han, M. F., & Ma, L. M. (2015). Predicting Early Lung Cancer Using Big Data. *Annals of Oncology*, *26*(1), 6–9. https://doi.org/10.1093/annonc/mdv044.04

Kaggle. (n.d.). *Lung Cancer Prediction Dataset*. Diambil 6 Januari 2023, dari https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?select=cancer+patient+data+sets.csv

Kanwal, M., Ding, X. J., & Cao, Y. (2017). Familial risk for lung cancer. *Oncology Letters*, *13*(2), 535–542. https://doi.org/10.3892/ol.2016.5518

Ledford, H. (2023). How air pollution causes lung cancer - without harming DNA. *Nature*, *616*(7957), 419–420. https://doi.org/10.1038/d41586-023-00989-z

Lehto, R. H. (2016). Symptom burden in lung cancer: management updates. *Lung Cancer Management*, *5*(2), 61–78. https://doi.org/10.2217/lmt-2016-0001

Li, C., Lei, S., Ding, L., Xu, Y., Wu, X., Wang, H., Zhang, Z., Gao, T., Zhang, Y., & Li, L. (2023). Global burden and trends of lung cancer incidence and mortality. *Chinese Medical Journal*, *136*(13), 1583–1590. https://doi.org/10.1097/CM9.0000000000002529

Munawar, Z., Ahmad, F., Awadh Alanazi, S., Nisar, K. S., Khalid, M., Anwar, M., & Murtaza, K. (2022). Predicting the prevalence of lung cancer using feature transformation techniques. *Egyptian Informatics Journal*, *23*(4), 109–120. https://doi.org/10.1016/j.eij.2022.08.002

Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R. S., Jawarneh, M.,

& Asenso, E. (2022). Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. *BioMed Research International*, *2022*, 1–8. https://doi.org/10.1155/2022/1755460

Nemlander, E., Rosenblad, A., Abedi, E., Ekman, S., Hasselström, J., Eriksson, L. E., & Carlsson, A. C. (2022). Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. *PLoS ONE*, *17*(10), 1–11. https://doi.org/10.1371/journal.pone.0276703

Panjaitan, C., Silaban, A., Napitupulu, M., & Simatupang, J. W. (2018). Comparison K-nearest neighbors (K-NN) and artificial neural network (ANN) in real time entrants recognition. *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, 1–4. https://doi.org/10.1109/ISRITI.2018.8864366

Prado, M. G., Kessler, L. G., Au, M. A., Burkhardt, H. A., Zigman Suchsland, M., Kowalski, L., Stephens, K. A., Yetisgen, M., Walter, F. M., Neal, R. D., Lybarger, K., Thompson, C. A., Al Achkar, M., Sarma, E. A., Turner, G., Farjah, F., & Thompson, M. J. (2023). Symptoms and signs of lung cancer prior to diagnosis: Case-control study using electronic health records from ambulatory care within a large US-based tertiary care centre. *BMJ Open*, *13*(4), 1–10. https://doi.org/10.1136/bmjopen-2022-068832

Pratiwi, N. C., Ibrahim, N., Fu'adah, Y. N., & Masykuroh, K. (2020). Computer-Aided Detection (CAD) for COVID-19 based on Chest *X-ray* Images using Convolutional Neural Network. *International Conference on Engineering, Technology and Innovative Researches*, *982*(1), 1–10. https://doi.org/10.1088/1757-899X/982/1/012004

Rajasekar, V., Vaishnnave, M. P., Premkumar, S., Sarveshwaran, V., & Rangaraaj, V. (2023). Lung cancer disease prediction with *CT-scan* and histopathological images feature analysis using deep learning techniques. *Results in Engineering*, *18*, 1–9. https://doi.org/10.1016/j.rineng.2023.101111

Shanbhag, G. A., Prabhu, K. A., Reddy, N. V. S., & Rao, B. A. (2022). Prediction of Lung Cancer using Ensemble Classifiers. *Journal of Physics: Conference Series*, *2161*, 1–11. https://doi.org/10.1088/1742-6596/2161/1/012007

Shirazi, A. S., & Frigaard, I. (2021). Slurrynet: Predicting critical velocities and frictional pressure drops in oilfield suspension flows. *Energies*, *14*(5), 1–21. https://doi.org/10.3390/en14051263

Troche, J. R., Mayne, S. T., Freedman, N. D., Shebl, F. M., & Abnet, C. C. (2016). The Association between Alcohol Consumption and Lung Carcinoma by Histological Subtype. *American Journal of Epidemiology*, *183*(2), 110–121.

https://doi.org/10.1093/aje/kwv170

Vedire, Y., Kalvapudi, S., & Yendamuri, S. (2023). Obesity and lung cancer—a narrative review. *Journal of Thoracic Disease*, *15*(5), 2806–2823. https://doi.org/10.21037/jtd-22-1835