Perbandingan Metode *Multilayer Perceptron* (MLP) dan *Xtreme Gradient Boosting* (XGBoost) pada Data Ekspresi Gen *Hepatocelluler Carsinoma* Terinfeksi Hepatitis B

Oleh: Ervina Putri Indah Sari
Univeristas Muhammadiyah Semarang

Article history
Submission :
Revised :
Accepted :

Keyword:
HCC, Hepatitis, *Microarray*, MLP, XGBoost

Abstract

HCC is a malignant tumor and the most fatal type of cancer. One of the risk factors for HCC is chronic hepatitis B infection. Hepatitis B is an infectious disease that attacks the liver in humans and is caused by the Hepatitis B virus. The analysis that can be used to see the difference between two or more classes of a disease is a classification analysis. The aim of this study was to conduct a classification analysis on microarray data resulting from gene expression of hepatitis B infected HCC patients in adjacent normal and tumor tissues. In this study, the MLP and XGBoost classification methods were used to analyze the data. MLP is an algorithm in the neural network model with one or more hidden layers. Meanwhile, XGBoost is a regression and classification algorithm with the ensamble method and a variant of the tree gradient boosting algorithm. This study will classify HCC patients infected with Hepatitis B with tumor tissue and adjacent normal tissue. The results of the analysis that have been done, it is found that the classification method with MLP with three hidden layers is able to obtain higher accuracy and AUC values than XGBoost. The accuracy value obtained is 95.24% with an AUC value in 0.9643, the AUC value is said to be very good because it is close to 1.

PENDAHULUAN

Hepatocelluler Carsinoma (HCC) atau kanker hati merupakan salah satu tumor ganas dan heterogen serta jenis kanker yang paling fatal. Di wilayah dengan angka penderita HCC tinggi, rasio kasus laki-laki dan perempuan dapat mencapai 8:1. Hal tersebut dikarenakan laki-laki lebih banyak yang terpapar faktor-faktor resiko HCC, seperti virus Hepatitis B dan alkohol (Sulaiman, 2012). Hepatitis B merupakan suatu sindroma klinis atau patologis yang ditandai dengan berbagai tingkat peradangan dan nekrosis pada hepar yang disebabkan oleh virus Hepatitis B yang menyerang sel hati, dimana infeksi tersebut dapat berlangsung secara akut atau kronik.

Penyakit HCC di Amerika Serikat menjadi penyebab kematian akibat kanker terbanyak ke-9. HCC banyak dijumpai pada penderita sirosis, Hepatitis B kronis, NASH, dan obesitas. Distribusi global dari HCC erat kaitannya dengan prevalensi geografis VHB dan VHC kronik yang jumlahnya mencapai 400 juta penduduk di seluruh dunia. Secara epidemiologi, jumlah kasus VHB kronik adalah sebesar 52% kasus. Penderita VHB kronik memiliki resiko 102 kali lebih tinggi daripada yang bukan kronik untuk terjadinya HCC.

Banyak peneliti yang mempercayai bahwa kunci untuk menangani penyakit ini adalah dengan deteksi dini. Deteksi tersebut dapat diketahui dengan melihat ekspresi dari gen yang terkandung dalam DNA. Ilmu yang mempelajari tentang biologis dengan melibatkan suatu sistem yaitu bioinformatika. Fatchiyah (2009) menyatakan bahwa bioinformatika dapat dikatakan sebagai ilmu yang mempelajari tentang penerapan teknik kamputasi untuk menganalisis berbagai informasi biologi.

Salah satu teknologi yang berkembang dalam bidang bioinformatika adalah teknologi *microarray*, yang merupakan sebuah perangkat berupa chip yang didalamnya berisi ribuan gen. Pada data *microarray*, dapat digunakan untuk mendeteksi atau mengklasifikasikan jaringan suatu penyakit pada manusia. Data *microarray*

menghasilkan ekspresi gen yang berisi informasi gen-gen yang ada dan kemudian dicocokkan dengan suatu penyakit (Trevino, Falciani, & Barrera-Saldana, 2007).

Analisis yang dapat digunakan untuk melihat adanya perbedaan dari dua kelas pada penyakit tersebut adalah analisis klasifikasi. Klasifikasi merupakan proses menyatakan suatu objek data sebagai salah satu dari sebuah kategori (kelas) yang telah didefinisikan sebelumnya (Zaku., et al. 2013). Metode klasifikasi dapat dibangun menggunakan teknik pembelajaran pada bidang *Machine Learning*. Teknik komputasi *machine learning* dapat digunakan untuk menganalisis pemilihan gen ataupun protein yang memiliki sifat terkait dan mengklasifikasikan tipe sampel ekspresi gen pada data microarray (Yang, Yang, Zhou, & Zomaya, 2016).

Algoritma yang digunakan pada *machine learning* adalah *supervised learning*, dimana algoritma tersebut digunakan untuk pengamatan dengan hasil yang diperoleh telah diketahui kelasnya. Algoritma pemodelan pada supervised learning yaitu klasifikasi, dimana pada penelitian ini menggunakan metode *Multilayer Perceptron* (MLP) dan *Xtreme Gradient Boosting* (XGBoost).

Multilayer Perceptron (MLP) merupakan suatu algoritma pada pembuatan model neural network atau jaringan saraf tiruan. Jaringan syaraf tiruan memiliki kemampuan yang baik dalam menganalisa pola data. Kemampuan ini menjadi salah satu alasan mengapa jaringan syaraf tiruan banyak dipilih sebagai metode dalam melakukan prediksi. MLP merupakan *perceptron* dengan satu atau lebih hidden layer, dimana input sebuah perceptron adalah *output* dari *perceptron* sebelumnya. Tidak seperti *perceptron* yang hanya dapat memodelkan permasalahan linear, MLP juga dapat menyelesaikan permasalahan non-linear (Suyanto, Data Mining , 2017).

Xtreme Gradient Boosting (XGBoost) merupakan suatu metode pada *machine learning*, dimana XGBoost merupakan algoritma regresi dan klasifikasi dengan metode *ensemble* yang merupakan suatu varian dari algoritma *Tree Gradient Boosting* yang dikembangkan dengan optimasi 10 kali lebih cepat dibandingkan *Gradient Boosting* lainnya (Chen & Guestrin, 2016). Pembangunan model dilakukan dengan menggunakan metode *boosting*, yaitu dengan membuat model baru untuk memprediksi *error/residual* dari model sebelumnya. Model baru ditambahkan hingga tidak ada lagi perbaikan pada *error* yang dapat dilakukan.

Penelitian tentang *Multilayer Perceptron* (MLP) telah dilakukan beberapa kali, seperti penelitian yang dilakukan oleh (Naf'an & Arifin, 2017) yang berjudul "Identifikasi Tanda Tangan Berdasarkan Grid Entropy Menggunakan *Multi Layer Perceptron*" dengan hasil yang diperoleh menunjukkan bahwa pengujian terbaik adalah untuk pengujian ukuran grid 8x8 dan menggunakan citra *outline*, yaitu dengan tingkat akurasi sebesar 97,78%, nilai korelasi 0,981 dan nilai kappa 0,977. Penelitian tentang *Xtreme Gradient Boosting* (XGBoost) juga telah beberapa kali dilakukan, seperti penelitian yang dilakukan oleh (Prasetyo, Christianto, & Hartomo, 2019) berjudul "Analisis Data Citra Landsat 8 OLI Sebagai Indeks Prediksi Kekeringan Menggunakan *Machine Learning* di Wilayah Kabupaten Boyolali dan Purworejo" dengan hasil yang diperoleh menunjukkan bahwa metode XGBoost lebih baik digunakan dibandingkan dengan metode *Random Forest* karena mempunyai nilai akurasi dan nilai kappa yang lebih tinggi.

Berdasarkan permasalahan dan penjelasan yang telah diuraikan, maka penelitian ini bertujuan untuk melakukan klasifikasi terhadap penyakit *Hepatocelluler Carsinoma* terinfeksi Hepatitis B dengan membandingkan metode *Multilayer Perceptron* dan *Xtreme Gradient Boosting*, yang kemudian dipilih metode mana yang terbaik untuk melakukan klasifikasi pada data ekspresi gen HCC terinfeksi Hepatitis B tersebut.

LANDASAN TEORI

Hepatocelluler Carsinoma (HCC)

Penyakit kanker adalah salah satu penyebab kematian utama di seluruh dunia. Kanker menjadi penyebab kematian sekitar 8,2 juta orang. Menurut Data GLOBOCAN, *International Agency for Research on Cancer* (IARC) dikatakan bahwa pada tahun 2012 terdapat 14.067.894 kasus baru penyakit kanker dan 8.201.575 kematian akibat penyakit kanker di seluruh dunia. Setiap tahunnya kematian akibat kanker sebagian besar disebabkan oleh

kanker hati, paru, perut, kolorektal serta kanker payudara (KemenKes, 2015).

Hepatocelluler Carsinoma atau HCC merupakan suatu benjolan atau tumor ganas yang terjadi pada hati. Kanker hati primer yang terjadi berasal dari hepatosis (*Hepatocelluler Carsinoma*) ataupun berasal dari duktus empedu (*kongaliokarsinoma*). Sedangkan untuk kanker hati sekunder yang muncul terjadi akibat metastasis kanker yang berasal dari bagian tubuh lain yang mengalirkan darahnya ke hati melalui vena porta atau kanker lainnya (Corwin, 2008). Penyebab terjadinya HCC adalah hepatitis B , hepatitis C dan sirosis hati yang disebabkan oleh konsumsi alkohol, diet tinggi, *aflatoksin*, penderita diabetes, obesitas dan akibat dari seringnya terkena paparan bahan kimia (Maharani, 2015).

Hepatitis B

Hepatitis B adalah salah satu penyakit infeksi yang terjadi pada jaringan hati dan disebabkan oleh virus yang berasal dari *family hepadnavirus*. Hepatitis B kronik didefinisikan sebagai peradangan hati yang berlanjut yang lebih dari enam bulan sejak munculnya gejala penyakit dan keluhan dari penderita. Penularan VHB yang sering terjadi adalah adanya kontak seksual atau kontak rumah tangga dengan seseorang yang terkena VHB, penularan perianal yang terjadi dari ibu kepada banyinya, penggunaan alat suntik oleh pecandu obat-obatan terlarang serta melalui pajanan *nosocomial* di rumah sakit (Masriadi, 2014).

Hubungan HCC dengan Hepatitis B

Hubungan antara HCC dengan infeksi hepatitis B dapat dikatakan sangan kuat secara *epidemiologis*, klinis maupun eksperimental. Proses terjadinya HCC pada VHB ada tiga tahapan, yaitu inisiasi, promosi dan progresi. Tahapan inisiasi ini akan terjadi integrase antara genom VHB ke genom hepatosit. Pada tahapan promosi akan mulai terjadi ekspansi klonal dari sel-sel yang terangsang dalam tahapan inisiasi. Selanjutnya pada tahapan progresi, sel-sel yang mengalami transformasi keganasan akan mengalami replikasi lebih lanjut. Beberapa kasus yang terjadi pada penderita hepatitis B yang berkembang menjadi HCC bisa langsung terjadi tanpa adanya proses sirosis.

Bioinformatika

Bioinformatika merupakan ilmu yang menggabungkan beberapa disiplin ilmu yang saling terhubung, yaitu biologi, informatika, matematika dan statistika. Menurut (Luscombe, Greenbaum, & Gerstein, 2001), bioinformatika menyertakan diri dengan memanfaatkan komputer untuk menyimpan, mencari keterangan, manipulasi, serta distribusi terkait data biologi macromolekul seperti DNA, RNA, serta Protein. Program yang didukung oleh internet merupakan fitur utama untuk bioinformatika. Salah satu analisis yang digunakan pada bioinformatika adalah analisis pada data ekspresi gen, yaitu beberapa gen dapat ditentukan dengan mengukur level dari gen tersebut dengan menggunakan berbagai macam metode seperti *microarray* (Raza, 2012).

Microarray

Microarray merupakan salah satu dari kemajuan teknologi yang digunakan dalam penelitian dalam membantu pendekatan *farmakologis* untuk mengobati berbagai penyakit, serta dapat digunakan untuk pengukuran tingkat ekspresi dari gen yang terdapat pada sebuah jaringan atau sel tertentu. Data *Microarray* dapat digunakan dalam mendeteksi dan mengklasifikasikan suatu jaringan penyakit pada manusia. Microarray menghasilkan ekspresi gen yang berisi informasi-informasi gen, kemudian dicocokkan dengan penyakit tertentu. Teknologi yang digunakan pada Microarray biasanya untuk *genotype* dengan skala yang besar, profil pada ekspresi gen, hibridasi genomik serta penyeimbang dalam penggunaan aplikasi lainny. *Microarray* merupakan hasil dari kombinasi dari beberapa bidang teknolongi dan penelitian, seperti mekanik, pembuatan mikro, kimia, perilaku DNA, *mikrofluida*, enzim, optik serta bioinformatika (Dufva, 2009).

Ekspresi Gen

Ekspresi gen atau *Gene Expression* merupakan proses transkripsi pada DNA dalam sel menjadi RNA (Madigan, Martinko, V, & Clark, 2008). *Deoxyribonucleic Gen* merupakan urutan dari DNA yang memberikan kode protein, dimana protein merupakan pengendali dari sifat fisik sel seperti pada warna

mata dan rambut. *Acid* atau DNA adalah sebuah asam nukleat yang didalamnya menyimpan informasi-informasi tentang genetika (Bolstad, 2004).

Preprocessing

Preprocessing merupakan tahapan yang penting dalam sebuah analisis. Hal tersebut dikarenakan pada sebagian besar data yang diambil atau diperoleh merupakan data mentah yang didalamnya terdapat data yang hilang, data noise (mengandung kesalahan), tidak konsisten (terdapat perbedaan kode atau nama) serta kualitas data. Salah satu tujuan pada tahap *preprocessing* adalah membersihkan data, normalisasi data, *integrase* data serta melakukan pengurangan data (Johan, 2017).

Filtering

Filtering pada data ekspresi gen merupakan tahapan yang penting, karena tahapan tersebut digunakan untuk penyaringan data dan mengurangi jumlah gen yang tidak diikutsertakan pada analisis atau tidak valid, sehingga akan meningkatkan kualitas model pada analisis ekspresi gen serta membuat proses pemodelan menjadi lebih efisien. Proses tersebut akan mengurangi ruang fitur dalam analisis prediksi yang tentunya akan menyebabkan suatu model yang dilatih akan lebih cepat dan variansinmya sama (Dozmorov, 2016).

Pemilihan Fitur

Pemilihan fitur atau *Feature Selection* merupakan proses yang dibuat agar pengklasifikasian lebih efektif dengan cara pengurangan data-data yang dianggap tidak relevan, sehingga dapat mempersingkat waktu pengklasifikasian dan dapat pula meningkatkan akurasi (Karabulut, Ozel, & Ibrikci, 2011). Pengujian yang digunakan pada penelitian ini adalah uji t-test karena memiliki dua sampel. Uji t merupakan suatu pengujian dengan satu individu dan dua perlakuan yang berbeda. Hal yang dapat dilakukan adalah dengan menghapus *row* yang mempunyai nilai p-value yang telah ditentukan dengan menggunakan package dplyr (Pollard, et al., 2019).

Machine Learning

Machine Learning atau bisa disebut dengan pembelajaran mesin adalah gabungan dari ilmu komputer dan statistik dan bisa juga disebut dengan komputer sains. Salah satu ciri yang khas pada *Machine Learning* adalah adanya proses pelatihan dan pembelajaran, oleh sebab itu dibutuhkan data untuk dipelajari atau data *training* dan data untuk di uji atau data *testing* (Ahmad, 2017). Program *machine learning* mampu mendeteksi pola yang ada pada data serta mampu menyelesaikan tindakan dari program yang sesuai. Sederhananya, machine learning merupakan pemrograman computer yang berguna untuk pencapaian kriteria/performa tertentu dengan sekumpulan data training atau pengalaman masa lalu (Primartha, 2018).

Klasifikasi

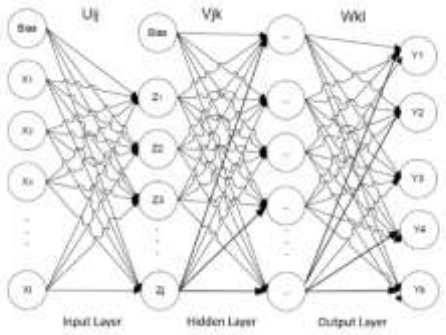
Teknik pada klasifikasi yang merupakan salah satu contoh dari algoritma *machine learning* adalah *supervised learning*. *Supervised Learning* adalah sebuah algoritma yang digunakan dalam *machine learning* dengan menggunakan data latih (*training*) yang sebelumnya telah diberi label dengan jawaban yang benar untuk memprediksi beberapa target dari beberapa kelas. Sedangkan algoritma merupakan suatu langkah-langkah atau urutan yang digunakan untuk perhitungan dalam menyelesaikan suatu masalah yang ditulis secara berurutan (Primartha, 2018).

Menurut (Gorunescu, 2011), klasifikasi adalah suatu proses dalam menempatkan suatu objek tertentu yang terdapat pada satu set kategori yang didasarkan pada empat komponen klasifikasi, yaitu kelas, prediktor, *training* dataset dan *testing* dataset.

Jaringan Syaraf Tiruan (JST)-*Multilayer Perceptron* (MLP)

Jaringan Syaraf Tiruan (JST) adalah suatu sistem komputasi yang di desain dengan menirukan cara kerja dari otak manusia untuk menyelesaikan dengan proses belajar melalui perubahan bobot sinapsisnya. Sekarwati (2005) mengatakan bahwa JST adalah sistem komputasi yang berdasarkan pemodelan pada sistem biologis *(neuron)* dengan pendekatan sifat-sifat biologis (*biological computation*). JST terdiri dari kumpulan grup *neuron* yang

tersusun dalam beberapa lapisan, yaitu *Input Layer*, *Hidden Layer* dan *Output Layer*. *Multilayer Perceptron* adalah sebuah arsitektur dari JST yang paling banyak digunakan dalam permasalahan klasifikasi yang berupa JST *feedforward* dengan satu atau lebih hidden layer. Berikut merupakan contoh dari arsitektur pada MLP:



Gambar 1. *Arsitektur* Multilayer Perceptron *(S.E Fahlaman, 1987)*

Berdasarkan gambar di atas, tiap koneksi antar *layer* dihubungkan dengan bobot Uij untuk koneksi antara *input layer* (xi) menuju *hidden layer* (Zj), kemudian untuk bobot Vjk untuk koneksi antara *hidden layer* (Zj) menuju *output layer* (Yk) berikutnya, serta bobot Wkl untuk koneksi antara *hidden layer* (Zj) menuju *output layer* (Yk). Proses pembelajaran yang terdapat pada MLP bertujuan untuk menemukan bobot sinaptik yang paling optimum untuk mengklasifikasian himpunan data latih dan data uji. Algoritma pembelajaran yang banyak digunakan untuk MLP yaitu *Backpropagation*.

Algoritma Backpropagation

Algoritma *Backpropagation* adalah algoritma pelatihan yang menggunakan banyak lapisan dalam mengubah bobot yang terhubung dengan neuron-neuron yang terdapat pada *hidden layer* (Haryati, Abdillah, & Hadiana, 2016). Algoritma *Backpropagation* melakukan pelatihan pada MLP dengan dua tahap, yaitu perhitungan maju (*Feedforward*) dan perhitungan mundur (*Backpropagation*). *Feedforward* digunakan untuk menghitung galat (*loss function*) antara prediksi target dan nilai aktual. Sedangkan *Backpropagation* digunakan untuk mempropagasikan balik galat untuk melakukan prediksi bobot sinaptik pada semua neuron yang ada. Dasar pada proses pembelajaran pada MLP yaitu menemukan bobot sinaptik yang menghasilkan *hyperplane* dengan kemiringan dan posisi yang tepat agar dapat mengklasifikasikan pola-pola yang ada pada himpunan data latih dengan kesalahan yang kecil (Suyanto, *Machine Learning* Tingkat Dasar dan Lanjut, 2018).

Fungsi Aktivasi

Fungsi aktivasi merupakan fungsi yang berfungsi untuk melakukan tranformasi suatu input menjadi output. Fungsi aktivasi dapat dikatakan operasi matematika yang digunakan untuk perhitungan *output*. Terdapat beberapa fungsi aktivasi yang dapat digunakan pada arsitektur MLP, salah satunya adalah fungsi aktivasi *Rectified Linier Unit* (ReLu). Fungsi aktivasi tersebut termasuk fungsi yang sangat popular karena sangat mudah untuk dioptimalkan, serta tidak mudah jenuh karena tidak *asymptotic* (Jin, et al., 2015). Definisi fungsi ReLu adalah:

$$h(x_i) = \max(0, x_i)$$

Dimana, fungsi xi adalah input dan h(xi) adalah *output*.

***Xtreme Gradient Boosting* (XGBoost)**

Boosting adalah salah satu teknik ensemble yang biasanya digunakan untuk proses analisis klasifikasi maupun prediksi. Teknik *ensemble* adalah suatu metode pada algoritma pembelajaran yang dibangun oleh beberapa model klasifikasi maupun prediksi yang nantinya digunakan untuk melakukan klasifikasi pada data baru berdasarkan bobot prediksi dari hasil sebelumnya (Dietterich, 2000). *Xtreme Gradient Boosting* atau XGBoost adalah suatu metode kombinasi antara *boosting* dengan *gradient boosting*. Pertama kali metode tersebut diperkenalkan oleh Friedman, yaitu dalam penelitiannya tentang hubungan antara *boosting* dengan optimasi untuk membuat *Gradient Boosting Machine* (GBM). Model yang dibangun dengan metode boosting adalah dengan membuat model baru untuk melakukan prediksi dari error pada model sebelumnya. Algoritma yang semacam itu dinamakan *gradient boosting*, karena menggunakan *gradient descent* untuk memperkecil *error* pada saat pembentukan model baru. XGBoost adalah salah satu *tree ensemble algorithm* yang terdiri dari *classification and Regression trees* (CART).

Algoritma XGBoost ini dapat melakukan optimasi 10 kali lebih cepat dibandingkan GBM lainnya (Chen & Guestrin, 2016). Nilai akurasi dari hasil klasifikasi menggunakan metode XGBoost tergantung parameter-parameter yang akan digunakan. Parameter yang digunakan pada penelitian ini adalah Eta, Gamma, Max_depth, Min_child_weight, Subsample dan Colsample_bytree.

Confussion Matrix

Confussion matrix adalah salah satu alat ukur yang berbentuk matrix untuk memperoleh nilai ketepatan sistem klasifikasi terhadap kelas sesuai dengan algoritma yang digunakan. Tujuan digunakannya confussion matrix adalah untuk mempermudah dalam mencari ukuran performa pada hasil klasifikasi. Berikut merupakan tabel contoh dari confussion matrix dengan 2 klasifikasi benar dan salah (Sofi & Jajuli, 2015):

Tabel 1. *Confussion Matrix*

Confusion Matrix		**Nilai Sebenarnya**	
		TRUE	***FALSE***
Nilai rediksi	***TRUE***	*TP (True Positive) Correct result*	*FP (False Positive) Unexpected result*
	FALSE	*FN (False Negative) Missing result*	*TN (True Negative) Correct absence of result*

Confussion matrix mengandung nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) serta *False Negative* (FN). Untuk nilai TP dan TN memberikan informasi *classifer* dalam melakukan klasifikasi data yang bernilai benar. Sedangkan untuk FP dan FN memberikan informasi classifer salah dalam melakukan klasifikasi data. Berdasarkan hasil dari *confussion matrix* yang diperoleh dapat digunakan untuk mengukur nilai *Accuracy*, *Precision* serta *Recall*. Berikut merupakan rumus untuk mengitung nilai *Accuracy*, *Precision, Recall/Sensitivity, Specificity*, FPR dan AUC:

$$Precision = \frac{TP}{TP + FP}\ x\ 100\%$$

$$Sensitivity = \frac{TP}{TP+FN}\ x\ 100\%$$

$$Specificity = \frac{TN}{TN+FP}\ x\ 100\%$$

$$FPR = 1 - Specificity$$

$$AUC = \frac{1+sensitivity-FPR}{2}$$

Nilai *Area Under Curve* (AUC) berguna untuk pengukuran kinerja deskriminatif dengan memperkirakan propabilitas *output* sampel yang dipilih secara acak dari probabilitas positif dan negatif. AUC memiliki rentang nilai 0-1, semakin tinggi atau semakin besar nilai AUC makan klasifikasi dapat dikatakan baik. Tabel berikut merupakan penerapan klasifikasi AUC (Gorunescu, 2011):

Tabel 2. Nilai AUC

Nilai AUC	**Keterangan**
0,91 – 1	Klasifikasi sangat baik
0,81 – 0,90	Klasifikasi baik
0,71 – 0,80	Klasifikasi cukup
0,61 – 0,70	Klasifikasi buruk
≤ 0,60	Klasifikasi salah

METODOLOGI PENELITIAN

Sumber Data

Data yang digunakan untuk penelitian ini merupakan data sekunder yang diperoleh dari *website National Center of Biotechnology* (NCBI) *Gene Expression Omnibus* dengan url https://www.ncbi.nlm.nih.gov/. Objek pada penelitian ini adalah data microarray yang dimuat oleh server tersebut yang berisi tentang informasi bioteknologi seperti DNA, protein, senyawa aktif dan taksonomi serta merupakan hasil dari penelitian *National Cancer Centre Singapore*. *Platform* data yang digunakan adalah GPL570 dengan series GSE121248, dimana jumlah sampel pada data tersebut sebanyak 107 sampel dengan 54675 gen.

Variabel dan Definisi Operasional Variabel

Tabel 3. Variabel dan Definisi

Variabel	**Definisi Operasional**
X	Merupakan gen yang digunakan pada penelitian ini, dengan jumlah sebanyak 54675 gen.
Y	Merupakan *Tissue* (Jaringan), yaitu variabel yang menunjukkan kumpulan sel yang menyusun setiap tubuh manusia. Variabel ini memiliki 2 kelas, yaitu 0 untuk jaringan tumor dan 1 untuk jaringan normal berdekatan.

Struktur Data

Tabel 4. Struktur Data

n	X_1	X_2	…	X_{54675}	Y
1	$X_{1,1}$	$X_{2,1}$	…	$X_{54675,1}$	0
2	$X_{1,2}$	$X_{2,2}$	…	$X_{54675,2}$	0
3	$X_{1,3}$	$X_{2,3}$	…	$X_{54675,3}$	0
⋮	⋮	⋮	…	⋮	⋮
37	$X_{1,37}$	$X_{2,37}$	…	$X_{54675,37}$	0
38	$X_{1,38}$	$X_{2,38}$	…	$X_{54675,38}$	1
⋮	⋮	⋮	…	⋮	⋮
107	$X_{1,107}$	$X_{2,107}$	…	$X_{54675,107}$	1

Keterangan :
0 : Kategori jaringan tumor
1 : Kategori jaringan normal berdekatan

Langkah-Langkah Penelitian

1. Mengambil data di *website* NCBI dengan menuliskan *keyword* atau kode akses GSE121248.
2. Melakukan eksplorasi data, yaitu untuk mengetahui gambaran data yang digunakan dalam penelitian.
3. *Preprocessing*, dilakukan untuk membersihkan data, normalisasi data, integrase data serta melakukan pengurangan data.
4. *Filtering*, yaitu untuk menghilangkan data-data yang tidak berguna dan tidak valid.
5. Dari data hasil *preprocessing* dan *filtering* dimasukkan kedalam data *expression set* yang baru, sehingga akan membentuk data baru berbentuk *frame* yang digunakan untuk analisis.
6. Membagi data kedalam 2 data baru, yaitu data latih *(training)* dan data uji *(testing)*. Dengan pembagiannya yaitu 80% untuk data *training* dan 20% untuk data *testing*.
7. Pengujian menggunakan metode MLP, yaitu dengan:
   a. Memasukkan data *training* kedalam pengujian klasifikasi dengan algoritma pembelajaran *Backpropagation*.
   b. Menentukan jumlah *hidden layer* dan jumlah *epoch* yang akan digunakan.
   c. Memasukkan fungsi aktivasi *Rectified Linier Unit* (ReLu) pada *input layer* dan *hidden layer* serta fungsi aktivasi *Sigmoid* pada *output layer*.
   d. Melakukan evaluasi model dengan optimasi *Adaptive Moment Estimation* (ADAM).
   e. Evalusi model yang diperoleh dan lakukan klasifikasi MLP terhadap data *training* menggunakan model yang terbaik.
8. Pengujian menggunakan metode XGBoost, yaitu dengan:
   a. Memasukkan data *training* kedalam pengujian klasifikasi.
   b. Uji coba/inisiasi parameter untuk memperoleh nilai akurasi terbaik. Parameter yang digunakan adalah Eta, Gamma, Max_depth, Min_child_weight, Subsample dan Colsample_bytree.
   c. Melakukan optimasi untuk nilai parameter dengan proses *tuning* untuk menentukan nilai parameter yang terbaik.
   d. Melakukan klasifikasi XGBoost terhadap data *training* menggunakan nilai parameter yang optimum agar menghasilkan model analisis terbaik
9. Selanjutnya dilakukan pengujian dengan menggunakan data *testing*, sehingga diperoleh nilai akurasi prediksi dari metode MLP dan XGBoost.
10. Melakukan perbandingan dari kedua metode untuk melihat metode mana yang memiliki nilai akurasi dan nilai AUC terbaik dalam melakukan prediksi atau klasifikasi data *testing*.
11. Melakukan interpretasi dari model dengan metode terbaik.

HASIL DAN PEMBAHASAN

Analisis Deskriptif

Data GSE121248 merupakan kumpulan sampel data ekspresi gen penyakit HCC terinfeksi Hepatitis B kronik dengan sampel jaringan tumor dan normal berdekatan. Informasi yang diperoleh dari data series GSE121248 berupa sejumlah variable yang digunakan untuk mendiagnosa penyakit pasien dan keterangan-keterangan lainnya.

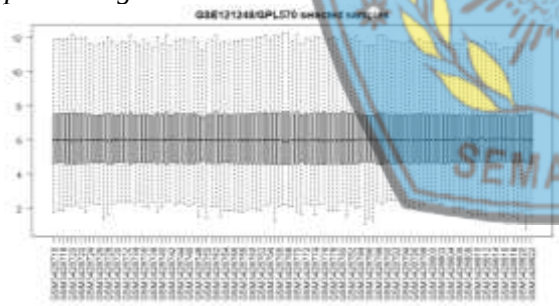
Tabel 4. Dataset

Informasi	Keterangan
Jumlah sampel	107
Jumlah gen	54675
Jumlah Kelas	2

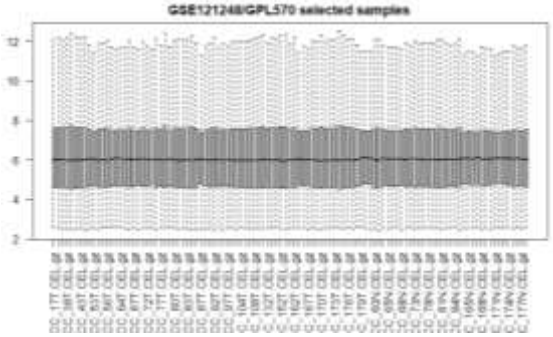
Berdasarkan tabel dapat diketahui bahwa terdapat 107 sampel dengan jumlah gen sebanyak 54675 gen yang terdiri dari 2 kelas, yaitu gen pada pasien dengan jaringan tumor dan jaringan normal berdekatan. Sampel tersebut diambil dari *National Cancer Centre Singapore*. Sebanyak 70 sampel mempunyai jaringan tumor pada HCC terfinfeksi Hepatitis B (65%) dan 37 sampel mempunyai jaringan normal berdekatan pada HCC terfinfeksi Hepatitis B (35%). Dimana HCC yang terinfeksi Hepatitis B ini banyak disebabkan oleh jaringan tumor dari pada jaringan normal berdekatan.

Preprocessing

Tahapan *preprocessing* digunakan untuk menghilangkan nilai *non biologis* dan *noise* yang terdapat pada data. Hasil yang diperoleh dari tahapan preprocessing yaitu berupa *boxplot*. Berikut merupakan *boxplot* sebelum dan setelah dilakukan *pre-processing* :



Gambar 1. *Boxplot* Sebelum *Pre-processing*



Gambar 2. *Boxplot* Setelah *Pre-processing*

Berdasarkan gambar terlihat perbedaan sebelum dilakukan dan setelah dilakukan *pre-processing*. Setelah dilakukan *pre-processing* dapat terlihat bahwa sebaran data menjadi lebih baik daripada sebelum dilakukan *pre-processing*. Tahapan *pre-processing* pada data *microarray* digunakan untuk menghapus nilai non biologis, sehingga yang terdapat atau tertera pada boxplot setelah dilakukannya *pre-processing* hanya terdapat data yang bersifat biologis saja dan menyebabkan *boxplot* tersebut mempunyai rata-rata yang sama.

Filtering

Tahapan *filtering* merupakan tahapan penentuan atau pemilihan gen yang akan siap untuk dianalisis. Tahapan ini digunakan fungsi *Non Speciefied Filtering*, yang mana fungsi tersebut menyediakan berbagai opsi pemfilteran (penghapusan fitur) dari *expression set*. Fungsi tersebut berguna untuk menghapus ataupun mengeluarkan variable yang mempunyai nilai *interquartile range* (IQR) tinggi. Setelah dilakukan tahapan filtering, hasil jumlah gen yang diperoleh sebesar 10080 dengan jumlah sampel sebesar 107. Dimensi data yang akan dilanjutkan untuk tahapan *feature selection* yaitu dengan dimensi 10080x107.

Feature Selection

Tahapan *feature selection* ini menggunakan fungsi multtest. Multtest merupakan fungsi yang berasal dari t-test yang digunakan untuk membandingkan dua sampel yang diambil dari dua populasi yang mempunyai variansi sama dengan data dan diasumsikan berdistribusi normal. Setelah data berdistribusi normal dan dilakukan pengujian maka hasil yang diperoleh dari feature selection yaitu dengan jumlah gen sebesar 82 dan jumlah sampel sebanyak 107. Jadi dimensi data yang terbentuk yaitu dengan dimensi 82x107.

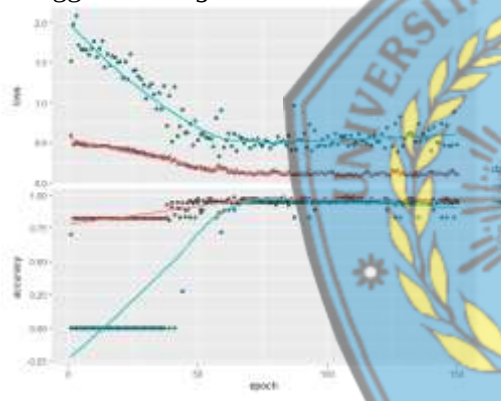
Data *Training* dan Data *Testing*

Data yang diperoleh dari tahapan *feature selection* yang telah siap dianalisis berdimensi 107x82, yang kemudian data tersebut dibagi menjadi data *training* dan data *testing* untuk dilakukan klasifikasi. Pembagian data training dan data testing menggunakan *split validation* dengan proporsi 80% untuk data training dan 20% untuk data testing dari total data yang telah

siap dianalisis untuk masing-masing kelas. Hasil yang diperoleh dari pembagian data *trining* dan data *testing* yaitu sebesar 86 sampel untuk data *training* dan 21 sampel untuk data *testing*.

Klasifikasi dengan MLP

Proses klasifikasi pada data ekspresi gen manusia dengan jaringan tumor dan jaringan normal berdekatan menggunakan Jaringan Syaraf Tiruan dengan arsitektur *Multilayer Perceptron* (MLP) digunakan data *input* sebesar 82 gen yang merupakan hasil dari tahapan *pre-processing* dan *filtering* dengan *output* 2 kelas. Dalam mendapatkan arsitektur MLP yang paling optimal dilakukan pengujian dengan menggunakan 3 *hidden layer*, penggunaan *optimizer* ADAM, dengan *epoch* 150, serta menggunakan *batch size* 10. Adapun fungsi aktivasi pada *input layer* dan *hidden layer* yaitu menggunakan *rectified linier unit* (ReLu), sedangkan fungsi aktivasi pada *output layer* menggunakan *Sigmoid.*



Gambar 3. Grafik Pergerakan Nilai Akurasi dan Nilai *Loss*

Gambar di atas merupakan hasil dari pelatihan data *training* menggunakan *epoch* sebanyak 150 yang ditunjukkan dengan nilai *loss* dan nilai akurasi. Grafik pergerakan nilai *loss* memiliki tren yang menurun yang menunjukkan bahwa semakin banyak *epoch* yang dilakukan maka *error* yang dihasilkan akan semakin kecil dan antara satu *epoch* dengan *epoch* berikutnya tidak mengalami banyak perubahan. Hal ini menunjukkan bahwa *epoch* yang dilakukan semakin konvergen. Sementara itu, grafik pergerakan nilai akurasi hasil *epoch* terus meningkat yang berarti bahwa nilai akurasi semakin besar seiring penambahan *epoch.* Dengan demikian, dapat diketahui bahwa untuk memperkecil nilai *loss*, maka dapat dilakukan dengan cara memperbanyak jumlah *epoch* sehingga menghasilkan nilai akurasi yang lebih tinggi. Dalam hal ini, nilai akurasi yang dihasilkan oleh data *training* adalah sebesar 95,5%. Hasil *Confusion matrix* yang digunakan untuk membandingkan hasil prediksi klasifikasi penderita HCC terinfeksi Hepatitis B dengan data sebenarnya ditunjukkan oleh tabel berikut:

Tabel 5. *Confussion Matrix* MLP

Prediksi	Aktual	
	0	1
0	13	0
1	1	7

Berdasarkan hasil klasifikasi dengan metode MLP dapat dilihat bahwa data aktual kategori kelas Jaringan Tumor, didapatkan ketepatan klasifikasi sesuai dengan kategori kelas sebanyak 13 data. Artinya terdapat 13 data yang dapat diprediksi oleh model secara tepat dan tidak ada terjadi *miss classification* pada kategori kelas Jaringan Tumor. Pada kategori Jaringan Normal Berdekatan, terdapat 7 data yang terklasifikasikan dengan tepat sesuai dengan data aslinya, namun dari total 8 data Jaringan Normal Berdekatan pada data uji terdapat 1 *miss clasification* data sehingga tidak sepenuhnya dengan data aktual. Total untuk data prediksi yang diperoleh pada kategori Jaringan Tumor didapatkan total sebesar 13, dan untuk kategori Jaringan Normal Berdekatan didapatkan sebanyak 8. Jika dibandingkan dengan dengan data asli aktualnya total kategori Jaringan Tumor sebanyak 14 dan Jaringan Normal Berdekatan sebanyak 7. Hasil prediksi dengan data asli terdapat perbedaan secara keseluruhan, namun untuk total keseluruhan data uji didapatkan sebesar 21. Berdasarkan tabel dapat dihitung nilai *Accuracy*, *Recall*, *Precision*, *Spesificity*, FPR dan AUC dengan hasil sebagai berikut:

Tabel 6. Evaluasi MLP

Kriteria Evaluasi	Nilai
Accuracy	95,24%
Precision	100%
Sensitivity	92,85%
Spesificity	100%
FPR	1
AUC	0,9643

Dari perhitungan nilai *accuracy*, dapat diketahui bahwa tingkat kedekatan antara nilai prediksi dan nilai aktual adalah sebesar 95,24%, yang mempunyai arti bahwa sistem dapat

melakukan klasifikasi antara kelas prediksi dan kelas aktual dengan sangat baik dimana dari 21 sampel terdapat 20 sampel yang tepat pengklasifikasiannya. Dari perhitungan nilai *precision*, dapat diketahui nilai presisi untuk pengklasifikasian mempunyai hasil prediksi yang tepat dengan nilai presisi sebesar 100%. Dari perhitungan nilai *sensitivity* yang diperoleh nilainya sebesar 92,85% yang berarti bahwa proporsi Jaringan Tumor yang diklasifikasikan dengan benar oleh sistem belum sempurna karena masih terdapat kesalahan klasifikasi pada sistem. Dari perhitungan *specificity* yang diperoleh nilainya sebesar 100% yang berarti bahwa proporsi Jaringan Normal Berdekatan yang diklasifikasikan dengan benar oleh sistem sudah mencapai nilai sempurna, dengan nilai FPR yang dihitung bernilai 0. Dari analisis dengan menggunakan metode MLP, diperoleh akurasi untuk model dengan menggunakan data uji sebesar 95,24%. Kemudian ditinjau dari nilai *AUC* untuk melihat nilai kinerja pengklasifikasian secara umum dimana nilai *AUC* berada pada rentang nilai 0 hingga 1, dimana jika nilai *AUC* mendekati 1 dapat dikatakan bahwa hasil prediksi semakin akurat. Dari perhitungan nilai *AUC*, diperoleh nilai sebesar 0,9643 yang berarti bahwa klasifikasi yang diperoleh sangat baik karena mendekati 1.

Klasifikasi dengan XGBoost

Proses klasifikasi pada data ekspresi gen manusia dengan jaringan tumor dan jaringan normal berdekatan menggunakan Jaringan Syaraf Tiruan dengan metode *Xtreme Gradient Boosting* (XGBoost) digunakan data *input* sebesar 82 gen yang merupakan hasil dari tahapan *pre-processing* dan *filtering* dengan *output* 2 kelas. Proses klasifikasi dengan menggunakan XGBoost dimulai dengan menentukan terlebih dahulu nilai parameter-parameter yang digunakan pada XGBoost. Parameter-parameter yang digunakan pada penelitian ini adalah *eta, max_depth, gamma, subsample, min_child_weight, dan colsample_by_tree.* Dalam menentukan model dengan parameter terbaik maka dilakukan proses *tuning parameter* agar memperoleh nilai terbaik dari mesing-masing parameter yang digunakan.

Hasil yang diperoleh dari proses *tuning parameter* pada XGBoost adalah 0,05 untuk nilai terbaik parameter *eta*, 1 untuk nilai terbaik parameter *max_depth*, 3 untuk nilai terbaik parameter *gamma*, 1 untuk nilai terbaik parameter *colsample_bytree*, 3 untuk nilai terbaik parameter *min_child_weight* dan 0,25 untuk nilai terbaik parameter *subsample*. Sedangkan untuk jumlah iterasi yang digunakan proses XGBoost pada penelitian ini adalah 150, dimana semakin banyak jumlah iterasi yang digunakan maka nilai error akan semakin kecil sehingga nilai akurasi akan semakin tinggi. Dalam hal ini, nilai akurasi yang dihasilkan oleh data *training* adalah sebesar 93,75%. Hasil *Confusion matrix* yang digunakan untuk membandingkan hasil prediksi klasifikasi penderita HCC terinfeksi Hepatitis B dengan data sebenarnya ditunjukkan oleh tabel berikut:

Tabel 7. *Confussion Matrix* XGBoost

Prediksi	Aktual	
	0	1
0	13	1
1	1	6

Confusion matrix merupakan alat untuk melakukan evaluasi model yang telah dibentuk, maka dilakukan suatu prediksi pada data *testing* setelah dilakukan *training* pada data latih. Berdasarkan hasil klasifikasi dengan metode XGboost pada tabel dapat dilihat bahwa data aktual kategori kelas Jaringan Tumor, terdapat 13 data yang terklasifikasikan dengan tepat sesuai dengan data aslinya, namun dari total 14 data Jaringan Tumor pada data uji terdapat 1 *miss clasification* data sehingga tidak sepenuhnya dengan data aktual. Pada kategori Jaringan Normal Berdekatan, terdapat 6 data yang terklasifikasikan dengan tepat sesuai dengan data aslinya, namun dari total 7 data Jaringan Normal Berdekatan pada data uji terdapat 1 *miss clasification* data sehingga tidak sepenuhnya dengan data aktual. Total untuk data prediksi yang diperoleh pada kategori Jaringan Tumor didapatkan total sebesar 14, dan untuk kategori Jaringan Normal Berdekatan didapatkan sebanyak 7. Jika dibandingkan dengan dengan data asli aktualnya total kategori Jaringan Tumor sebanyak 14 dan Jaringan Normal Berdekatan sebanyak 7. Hasil prediksi dengan data asli terdapat perbedaan secara keseluruhan, namun untuk total keseluruhan data uji didapatkan sebesar 21. Berdasarkan tabel dapat dihitung nilai *Accuracy, Recall, Precision, Spesificity,* FPR dan AUC dengan hasil sebagai berikut:

Tabel 8. Evaluasi XGBoost

Kriteria Evaluasi	Nilai
Accuracy	90,48%
Precision	92,85%
Sensitivity	92,85%
Spesificity	85,7%
FPR	0,143
AUC	0,866

Dari perhitungan nilai *accuracy*, dapat diketahui bahwa tingkat kedekatan antara nilai prediksi dan nilai aktual adalah sebesar 90,48%, yang mempunyai arti bahwa sistem dapat melakukan klasifikasi antara kelas prediksi dan kelas aktual dengan sangat baik dimana dari 21 sampel terdapat 19 sampel yang tepat pengklasifikasiannya. Dari perhitungan nilai *precision*, dapat diketahui nilai presisi untuk pengklasifikasian mempunyai hasil prediksi yang tepat dengan nilai presisi sebesar 92,85%. Dari perhitungan nilai *sensitivity* yang diperoleh nilainya sebesar 92,85% yang berarti bahwa proporsi Jaringan Tumor yang diklasifikasikan dengan benar oleh sistem belum sempurna karena masih terdapat kesalahan klasifikasi pada sistem. Dari perhitungan *specificity* yang diperoleh nilainya sebesar 85,7% yang berarti bahwa proporsi Jaringan Normal Berdekatan yang diklasifikasikan dengan benar oleh sistem sudah mencapai nilai sempurna, dengan nilai FPR yang dihitung bernilai 0,143. Dari analisis dengan menggunakan metode XGBoost, diperoleh akurasi untuk model dengan menggunakan data uji sebesar 90,48%. Kemudian ditinjau dari nilai *AUC* untuk melihat nilai kinerja pengklasifikasian secara umum dimana nilai *AUC* berada pada rentang nilai 0 hingga 1, dimana jika nilai *AUC* mendekati 1 dapat dikatakan bahwa hasil prediksi semakin akurat. Dari perhitungan nilai *AUC*, diperoleh nilai sebesar 0,866 yang berarti bahwa klasifikasi yang diperoleh dapat dikatakan karena nilainya antara 0,81 sampai 0,9.

Metode Terbaik

Berdasarkan hasil analisis klasifikasi yang diperoleh dari metode *Multilayer Perceptron* dan *Xtreme Gradient* Boosting dapat dilakukan perbandingan antar kedua metode yang telah dianalisis untuk menentukan metode terbaik dalam data ekspresi gen HCC terinfeksi Hepatitis B. Penentuan metode terbaik dapat dilihat berdasrkan nilai *accuracy* dan *AUC* yang didapatkan dari data *testing*. Hasil perbandingan kedua metode dipaparkan pada tabel berikut :

Tabel 9. Perbandingan MLP dan XGBoost

Metode	*Accuracy*	*AUC*
MLP	95,24%	0,9643
XGBoost	90,48%	0,866

Berdasarkan tabel di atas dapat diketahui bahwa dari kedua metode yang digunakan untuk mengklasifikasikan data ekspresi gen HCC terinfeksi Hepatitis B pada pasien dengan Jaringan Tumor dan Jaringan Normal Berdekatan diperoleh hasil klasifikasi dengan menggunakan metode *Multilayer Perceptron* mampu mengklasifikasikan data ekspresi gen dengan nilai akurasi sebesar 95,24% dan nilai AUC sebesar 0,9643. Sedangkan metode klasifikasi *Xtreme Gradient Boosting* mampu mengklasifikasikan data ekspresi gen dengan nilai akurasi sebesar 90,48% dan nilai AUC sebesar 0,866. Berdasarkan hasil tersebut dapat diketahui bahwa dalam analisis ini metode klasifikasi terbaik untuk mengklasifikasikan data ekspresi gen HCC terinfeksi Hepatitis B adalah metode klasifikasi *Multilayer Perceptron* (MLP) karena memiliki nilai akurasi yang tinggi dan nilai AUC yang mendekati 1.

SIMPULAN DAN SARAN

Simpulan

Berdasarkan hasil analisis klasifikasi dengan menggunakan metode *Multilayer Perceptron* dengan 3 *hidden layer* diperoleh bahwa sistem dapat melakukan klasifikasi dengan sangat baik dimana dari 21 sampel terdapat 20 sampel yang tepat pengklasifikasiannya. Nilai akurasi yang diperoleh yaitu sebesar 95,24% dengan nilai AUC sebesar 0,9643. Hasil analisis klasifikasi dengan menggunakan metode *Xtreme Gradient Boosting* diperoleh bahwa system dapat melakukan klasifikasi dengan baik dimana dari 21 sampel terdapat 19 sampel tepat pengklasifikasiannya. Nilai akurasi yang diperoleh yaitu sebesar 90,48% dengan nilai AUC sebesar 0,866. Berdasarkan hasil klasifikasi dengan menggunakan metode *Multilayer Perceptron* dan *Xtreme Gradient Boosting*, metode terbaik dalam melakukan klasifikasi data ekspresi gen HCC terinfeksi Hepatitis B pada pasien dengan Jaringan Tumor

dan Jaringan Normal adalah metode *Multilayer Perceptron* (MLP).

Saran

Berdasarkan hasil analisis dan kesimpulan yang diperoleh, peneliti dapat memberikan saran sebagai upaya perbaikan dan pengembangan penelitian selanjutnya adalah pada penelitian selanjutnya diharapkan dapat menggunakan metode-metode klasifikasi lain sebagai bentuk perbandingan dalam melihat tingkat akurasi klasifikasi yang paling baik dan diharapkan dapat menerapkan metode *Cross Validation* untuk mengatasi nilai akurasi yang acak.

DAFTAR PUSTAKA

Ahmad, A. 2017. *Mengenal Artificial Intelligence, Machine learning, Neural Network, dan Deep Learning*. Yayasan Cahaya Islam Jurnal Teknologi Indonesia.

Bolstad, B. M. 2014. Low level Analysis of High density Oligonucleotide Array Data : Background, Normalization and Summarization. *University Of California*.

Chen, T., & Guestrin, C. 2016. XGBoost : A Scalable Tree Boosting System. *Vol.42, no. 8*, 665.

Corwin, E. J. 2008. *Handbook of Pathopysiology*. USA: Penerbit Buku Kedokteran EGC.

Dietterich, T. G. 2000. Multiple Classifier Systems. *International Workshop on Multiple Classifier Systems* (pp. 1-15). Berlin: Springer.

Dozmorov,M. 2016, *Filtering*. Diambil kembali dari https://mdozmorov.github.io/BIOS567/assets/presentation_Bioconductor/*Filtering*.pdf. [Diakses 3 Desember 2020].

Dufva, M. 2009. Introduction to microarray technology, *PubMed NCBI, PMID*: 19381982.

Gorunescu, F. 2011. *Data Mining Concepts, Models and Techniques*. Verlag Berlin Heidelberg: Springer.

Jin, X., Xu, C., Feng, J., Wei, Y., Xiong, J., & Yan, S. 2015. Deep Learning with S-shaped Rectified Linear Activation Units. *arXiv*.

Johan,Y. 2017. Diambil dari http://rosyid.lecturer.pens.ac.id/dataMining/Data%20Preprocessing.pdf. [Diakses 2 Desember 2020].

Karabulut, E. M., Ozel, S. A., & Ibrikci, T. 2011. *A comparative study on the effect of feature selection on classification accuracy*. Procedia Technology, 323-327.

Kementerian Kesehatan RI. 2015. Pusat Data dan Informasi Kementerian Kesehatan RI. [Online]. http://www.depkes.go.id. [Diakses 29 November 2020].

Luscombe, M. N., Greenbaum, D., & Gerstein, M. 2001. *What is Bioinformatics ? A Proposed Definition and Overview of the Field*. Method Inform Med, 346-58.

Maharani, S. 2015. *Mengenal 13 Jenis Kanker dan Pengobatan*. Yogjakarta: KATA HATI.

Masriadi, H. 2014. *Epidemiologi Penyakit Menular*. Depok: PT. Rajagrafindo Persada.

Pollard, et al. 2019. *Package 'multtest'*.

Primartha, R. 2018. *Belajar Machine Learning Teori Dan Praktik*. Bandung: Informatika Bandung.

Raza, K.2012. *Application of Data Mining in Bioinformatics*. Indian Journal of Computer Science and Engineering, 114-118.

Sofi, D., & Jajuli. 2015. Integrasi Metode Klasifikasi dan Clustering dalam Data Mining. *Jurnal Teknik Informatika Fakultas Ilmu Komputer*, Universitas Singaperbangsa Karawang.

Sulaiman, H. A. 2012. *Buku Ajar Ilmu Penyakit Hati*. Jakarta: CV. Agung Seto.

Suyanto. 2017. *Data Mining* . Bandung: Informatika Bandung.

Suyanto. 2018. *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung.

Trevino, V., Falciani, F., & Barrera-Saldana, H. A. 2007. DNA Microarrays : a Powerful Genomic Tools for Biomedical and Clinical Research. *Molecular Medicine*, 527-541.

Yang, P., Yang, Y. H., Zhou, B. B., & Zomaya, A. Y. 2016. A Review of ensemble methods in Bioinformatics. 1.