Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 4 No. 2 (2023) 124 – 130 | ISSN: 2723-1453 (Media Online)

Penentuan Jalur Diagnostik Penyakit Berbasis Konsep Pembelajaran Mesin: Studi kasus Penyakit Hepatitis C

Jimmy Tjen[1], Valentino Pratama[2]
[1]Informatika, Fakultas Teknologi Informasi, Universitas Widya Dharma Pontianak
[2]Sekolah Menengah Atas Santo Paulus Pontianak
[1]jimmy.tjen@mathmods.eu, [2]pratamavalentino14@gmail.com

Abstract

Hepatitis is considered to be one of the most dangerous diseases, which often leads to death if not handled properly. Thus, early detection via precise diagnosis is needed in order to prevent the unfortunate event. This research aims to provide a novel hepatitis C diagnosis based on the machine learning algorithm, which is the classification tree from the decision tree learning and the distance correlation, which measures the Euclidean distance between 2 vectors. In particular, the goal is to develop a low computational cost yet precise algorithm for diagnosing the possibility of whether a person is being infected with Hepatitis C or not. Based on the experiment, the distance correlation-based classification tree algorithm outperforms the classical classification tree algorithm by around 3% while using only 7 features instead of 12 as in the classical algorithm. Furthermore, the algorithm identified albumin (ALB), Creatinine (CREA), Bilirubin (BIL), Aspartate Transaminase (AST) and Cholesterol (CHOL) as significant risk factors in determining whether someone is potentially infected with hepatitis C or not, with Creatinine is identified as the most important parameter among all 5 parameters mentioned above.

Keywords: Distance correlation, hepatitis C, diagnostic pathway, machine learning, classification tree

Abstrak

Penyakit hepatitis merupakan satu dari sekian banyak penyakit mematikan, yang jika tidak ditanggani dengan baik dapat menimbulkan kematian. Oleh karena itu, dirasa penting untuk dapat mengidentifikasi penyakit ini sedini mungkin guna mencegah hal yang buruk terjadi. Penelitian ini bertujuan untuk menghasilkan algoritma diagnosis penyakit hepatitis C berdasarkan pada konsep pembelajaran mesin, terutama pada metode pohon klasifikasi dari pembelajaran pohon keputusan dan korelasi jarak yang mengukur jarak *Euclidean* dari 2 vektor. Secara spesifik, penelitian bertujuan untuk menghasilkan algoritma baru yang memiliki kompleksitas rendah, namun memiliki presisi yang tinggi dalam menentukan apakah seseorang telah mengidap penyakit hepatitis C atau tidak. Berdasarkan pada penelitian yang telah dilakukan, metode pohon klasifikasi berbasis korelasi jarak memiliki performa yang lebih baik daripada metode klasik dengan peningkatan akurasi sebesar 3%, meskipun hanya menggunakan 7 dari 12 parameter yang ada. Lebih lanjut, algoritma yang digagas mengidentifikasi *albumin* (ALB), *Creatinine* (CREA), *Bilirubin* (BIL), *Aspartate Transaminase* (AST) dan *Cholesterol* (CHOL) sebagai faktor resiko signifikan dalam menentukan apakah seseorang berpotensi mengidap penyakit hepatitis C, dengan *creatinine* merupakan parameter terpenting dalam proses diagnosis, jika dibandingkan dengan 4 parameter lainnya.

Kata kunci: korelasi jarak, hepatitis C, jalur diagnostik, pembelajaran mesin, pohon klasifikasi

1. Pendahuluan

Penyakit hepatitis merupakan sebuah kondisi peradangan organ hati akibat infeksi virus [1]. Penyakit hepatitis digolongkan berdasarkan pada 5 golongan: A, B, C, D dan E berdasarkan pada jenis virus yang menginfeksi organ hati. Secara umum, hepatitis golongan A dan E bersifat lebih jinak dan dapat disembuhkan, sedangkan hepatitis C merupakan golongan penyakit hepatitis yang mengakibatkan gangguan fungsi hati secara kronis dan sering kali berdampak pada kematian [2, 3]. Sehingga, dibutuhkan pendektesian dini yang akurat agar dapat menyelamatkan lebih banyak pasien sebelum terlambat. Sebagai contoh dengan memadukan konsep kecerdasan buatan, secara khusus pembelajaran mesin (*machine learning*) dengan hasil pengukuran kondisi medis pasien guna menghasilkan diagnosis awal yang cepat dan presisi [4, 5]. Konsep ini secara spesifik dikenal sebagai *Intelligent Fault Diagnosis* (IFD).

IFD merupakan konsep dari penerapan pembelajaran mesin untuk melakukan diagnosis secara otomatis, sehingga mempermudah pekerjaan manusia dibidang kesehatan dengan memanfaatkan mesin yang dirancang khusus untuk mendiagnosis penyakit. Konsep IFD dilakukan dengan cara mengajari mesin mengenali pola-pola atau membangun jalur diagnostik berdasarkan pada

masukan data penyakit pasien [6]. Terdapat beberapa jenis algoritma yang lazim digunakan untuk membangun IFD, salah satunya adalah metode pohon klasifikasi dan regresi dari pembelajaran pohon keputusan atau *decision tree learning.*

Metode pohon keputusan merupakan sebuah algoritma pemecahan biner yang memecah sekumpulan data berdasarkan pada logika jika dan maka. Melalui pemecahan biner ini, maka dapat diperoleh kumpulan data sejenis yang memudahkan interpretasi dari data. Pohon keputusan terbagi menjadi dua: pohon regresi yang mengolah keluaran data kontinyu dan pohon klasifikasi yang mengolah keluaran yang bersifat diskrit [7]. Terdapat beragam penelitian yang telah menunjukan kemampuan dari metode pohon klasifikasi dan regresi dalam memodelkan data yang memiliki kompleksitas tinggi, seperti pada [8, 9, 10, 11]

Penelitian terkait: pada [8], telah dibahas bagaimana metode pohon klasifikasi digunakan untuk mendiagnosis penyakit hepatitis C. Berdasarkan pada penelitian yang telah dilakukan, dapat disimpulkan bahwa metode pohon klasifikasi dapat mengkatagorikan apakah seseorang mengidap penyakit hepatitis C dengan akurasi di atas 80%. Pada [9] telah ditunjukan kemampuan dari metode pohon regresi berbasi entropi dalam memodelkan dinamika dari 3 bangunan berbeda, yang secara spesifik ditujukan untuk mengetahui apakah terdapat kerusakan di dalam struktur. Berdasarkan percobaan yang telah dilakukan, diketahui bahwa metode pohon regresi dapat digunakan untuk memodelkan kerusakan struktur bangunan dengan akurasi rerata di atas 90%. Pada [10] dan [11] telah dilakukan penelitian untuk menentukan tingkat kerusakan dari struktur bangunan dengan menggunakan metode pohon regresi, konsep entropi dalam teori informasi dan korelasi jarak atau *distance correlation*. Berdasarkan pada penelitian yang telah dilakukan, dapat disimpulkan bahwa metode pohon regresi berbasis korelasi jarak dapat menentukan kondisi dari sebuah bangunan dengan tingkat kesalahan di bawah 15%.

Berdasarkan pada pemaparan di atas, dapat terlihat bahwa metode pohon regresi dan klasifikasi dapat memodelkan dinamika dari sekumpulan data dengan presisi. Namun, masih terdapat kekurangan pada algoritma ini. Secara spesifik, algoritma pohon regresi dan klasifikasi memiliki kompleksitas waktu pada $O(m \times n^2)$ dengan m menyatakan jumlah sampel dan n menyatakan jumlah fitur atau parameter yang terkandung di dalam data, sehingga semakin besar data yang digunakan, semakin lama waktu yang dibutuhkan bagi algoritma ini untuk memproses keluaran [12].

Tujuan: meninjau permasalahan yang telah dipaparkan di atas, maka karya tulis ini bertujuan untuk menghasilkan algoritma diagnosis penyakit diabetes yang memiliki tingkat akurasi yang tinggi, namun tanpa membebani performa mesin dengan kompleksitas yang tinggi. Secara spesifik, yang menjadi tujuan pada karya tulis ini adalah untuk menghasilkan algoritma diagnosis penyakit hepatitis C berbasis metode pohon klasifikasi dan korelasi jarak dan untuk mengidentifikasi faktor resiko utama (*significant risk factor*) yang menentukan apakah seseorang terjangkit penyakit hepatitis C atau tidak berdasarkan pada metode pohon klasifikasi dan korelasi jarak.

Keterbaharuan topik: penelitian ini dilakukan dengan mengikuti alur penelitian serupa yang telah dilakukan pada [8]. Secara spesifik, penelitian ini akan meningkatkan kualitas diagnosis sesuai dengan menggabungkannya dengan konsep korelasi jarak. Tujuannya adalah untuk mengurangi jumlah parameter yang diduga berpengaruh pada diagnosis penyakit hepatitis C, sehingga dapat dihasilkan algoritma yang akurat dan tidak membutuhkan waktu kalkulasi yang lama. Lebih lanjut dengan menggunakan metode korelasi jarak, maka dapat dibentuk kumpulan dari himpunan bagian data yang berkorelasi baik (secara jarak *Euclidean*), sehingga dapat digunakan untuk mengidentifikasi faktor resiko yang paling akurat untuk menentukan apakah seseorang telah terjangkit hepatitis C atau tidak. Konsep ini merupakan gagasan baru terhadap penelitian yang telah dilakukan pada [8], yang dalam kasus ini, diharapkan modifikasi yang dilakukan akan meningkatkan performa algoritma yang telah digagas sebelumnya terutama dalam akurasi model prediktif. Terhadap [10] dan [11], metode korelasi jarak digunakan untuk mengidentifikasi data yang bersifat kontinyu. Pada karya tulis ini, konsep ini akan diekspansikan lebih lanjut, sehingga dapat mengidentifikasi pola untuk data yang bersifat diskrit.

Karya tulis ini terdiri atas 4 bagian: bagian pertama adalah pendahuluan yang menjelaskan tujuan, permasalahan dan solusi yang ditawarkan pada penelitian ini. Pada bagian kedua akan dijelaskan algoritma dari metode pohon regresi yang dipadukan dengan konsep korelasi jarak, untuk mendiagnosis apakah seseorang mengidap hepatitis C atau tidak. Bagian ketiga akan menjelaskan hasil simulasi sesuai dengan algoritma yang telah dijelaskan pada bagian kedua. Terahir, akan disajikan kesimpulan dari penelitian yang telah dilakukan serta arah penelitian mendatang yang dapat dilakukan.

2. Metode Penelitian

Pada bagian ini akan dibahas secara spesifik algoritma serta alur penelitian untuk membangun jalur diagnosis dari penyakit hepatitis C berdasarkan pada algoritma pohon klasifikasi dan pemilihan himpunan bagian berbasis korelasi jarak. Silahkan mengacu pada [9] terkait dengan konsep dasar dari metode pohon regresi dan [10, 11, 13] terkait dengan pemilihan himpunan bagian berbasis korelasi jarak.

DOI : https://doi.org/10.52158/jacost.v4i2.556

2.1. Algoritma Diagnosis Penyakit Hepatitis C Berbasis Korelasi Jarak

Dimisalkan terdapat sebuah himpunan data $H = [\boldsymbol{y}\ X\ D]; H \in \mathbb{R}^{m\times(n_1+n_2+1)}$ yang merupakan data hasil pengukuran laboratorium terkait dengan penyakit hepatitis C, dengan $y \in \{0,1\}^m$ menyatakan kondisi apakah seseorang mengidap hepatitis atau tidak (sebagai contoh $y = 0$ berarti pasien negatif hepatitis C dan begitupula sebaliknya), $X = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots \boldsymbol{x_{n+1}}]; X \in \mathbb{R}^{m\times n_1}$ merupakan kumpulan data yang berisikan parameter pengukuran yang bersifat numerik (sebagai contoh, tingkat *cholesterol*, dan seterusnya) dari pasien dan $D \in \mathbb{Z}^{m\times n_2}$ merupakan kumpulan data dari pasien yang bersifat diskrit, sebagai contoh: usia, jenis kelamin dan seterusnya. Tujuan dari penelitian ini adalah untuk merumuskan sebuah algoritma yang dapat memprediksi apakah seseorang mengidap hepatitis C atau tidak sesuai dengan informasi diatas. Secara spesifik, algoritma ini tersusun atas 3 bagian: penyusunan subhimpunan berbasis korelasi jarak, penyusunan model matematis dan tahap diagnosis.

Langkah pertama: berdasarkan pada data X sesuai dengan definisi diatas, maka langkah pertama yang dibutuhkan adalah mendefinisikan himpunan bagian data berbasis korelasi jarak sesuai dengan alur pada [13], untuk membangun himpunan bagian dengan vektor data yang memiliki jarak *Euclidean* [14] yang dekat satu sama lain. Namun, berbeda pada alur yang digunakan pada [13], pemilihan himpunan bagian pada penelitian ini tidak dapat dilakukan secara langsung akibat variabel terikat yang berupa data diskrit. Sehingga, pada tahapan ini, perlu untuk menentukan himpunan bagian bagi setiap parameter yang dirasa cocok untuk memodelkan variabel terikat.

Dimisalkan bahwa $\mathfrak{D}_i = [\boldsymbol{x_i}\ \boldsymbol{x_a}\ \boldsymbol{x_{a+1}}\ \dots \boldsymbol{x_{a^*}}];\ \mathfrak{D}_i \in \mathbb{R}^{m\times a^*}; i = 1,2,3,\dots,n_1$ merupakan himpunan data dari X yang memiliki korelasi jarak yang baik dengan x_i yang terbentuk sesuai dengan algoritma pemilihan himpunan bagian berbasis korelasi jarak pada [13], dengan $x_a = [x_a(1)\ x_a(2)\ \dots x_a(m)]^\intercal, x_a \in X$ merupakan parameter ke a dari data X. Pada tahap ini, akan diperoleh sebanyak n_1 himpunan bagian data yang memiliki korelasi jarak yang baik satu sama lain. Lebih lanjut, untuk setiap himpunan bagian data, misalkan bahwa $H_i = [\boldsymbol{y}\ D\ \mathfrak{D}_i]; H_i \in \mathbb{R}^{m\times(a^*+n_2+1)}$ adalah bentuk data teraugmentasi dari $\mathfrak{D}_i$ yang telah dilengkapi dengan parameter $\boldsymbol{y}$ dan D sehingga menjadi data utuh yang menjelaskan informasi pasien dengan data numerik yang berkorelasi (secara definisi korelasi jarak) satu sama lain. Pada tahap ini, data telah selesai dibentuk dan siap untuk digunakan untuk membentuk model prediktif berbasis algoritma pohon klasifikasi.

Langkah kedua: pada tahap ini, untuk setiap himpunan bagian data akan dibangun sebuah model pohon klasifikasi yang berguna untuk melakukan diagnosis terhadap pasien yang diduga mengidap penyakit hepatitis C. Untuk setiap H_i, misalkan bahwa p_i adalah model pohon klasifikasi yang dibangun sesuai dengan parameter data yang terkandung pada H_i. Secara spesifik:

$$\begin{aligned} p_1 &: y = F_{PK}(D, \mathfrak{D}_1), \\ p_2 &: y = F_{PK}(D, \mathfrak{D}_2), \\ &\vdots \\ p_i &: y = F_{PK}(D, \mathfrak{D}_i) \end{aligned} \tag{1}$$

dengan F_{PK} menyatakan fungsi pohon klasifikasi [7, 9].

Langkah ketiga: setelah model prediktif telah berhasil dibangun, maka langkah terakhir adalah menentukan model prediktif terbaik untuk mendiagnosis apakah pasien mengidap hepatitis C atau tidak. Dimisalkan $H_v = [\boldsymbol{y_v}\ X_v\ D_v]; H_v \in \mathbb{R}^{m_2\times(n_1+n_2+1)}$ merupakan kumpulan data validasi model, yakni kumpulan data yang serupa dengan H, namun tidak digunakan sama sekali untuk membangun model sesuai pada langkah kedua (sebagai contoh, H_v adalah data uji yang berguna untuk memvalidasi kemampuan diagnosis dari model). Misalkan bahwa $\%A_i$ merupakan persentase akurasi model prediktif dengan

$$\%A_i = \frac{b_i}{m_2} \times 100\% \tag{2}$$

Dengan b_i menyatakan jumlah diagnosis benar yang dilakukan berdasarkan pada model prediktif ke-i dan m_2 menyatakan banyaknya sampel yang terkandung di dalam data uji, H_v. Pada kasus ini, model preditif terbaik yang digunakan untuk melakukan diagnosis, p^* dapat ditentukan sebagai

$$p^* \coloneqq \{p_i : \underset{i}{\operatorname{argmax}}\ \%A_i\} \tag{3}$$

atau dengan kata lain p_i dengan tingkat akurasi yang tertinggi. Dalam kasus ini, parameter yang terkandung di dalam $\mathfrak{D}^*$ (yang merupakan parameter data numerik dari p^*) akan dipilih sebagai faktor resiko yang paling signifikan dalam menentukan apakah seseorang mengidap penyakit hepatitis C atau tidak. Keseluruhan dari algoritma diagnosis hepatitis C berbasis metode pohon klasifikasi dan korelasi jarak dirumuskan pada algoritma 1.

2.2. Studi kasus: Data Hepatitis C

Data yang digunakan dalam penelitian ini merupakan data penelitian sekunder yang diperoleh dari repositori data milik *University of California Irvine* (UCI), Amerika Serikat [14], yang merupakan data yang disumbangkan oleh institut kimia klinis, *Medical University Hannover* (MHH) Jerman. Data tersebut mengandung 610 sampel dengan 14 parameter: ID pasien, katagori diagnosis (0=negatif, 1=positif hepatitis C), umur, jenis kelamin, *Albumin* (ALB), *Alkaline phosphatase* (ALP), *Alanine Transaminase* (ALT), *Aspartate Transaminase* (AST), *Bilirubin* (BIL),

DOI : https://doi.org/10.52158/jacost.v4i2.556

Acetylcholinesterase (CHE), *Cholesterol* (CHOL), *Creatinine* (CREA), *Gamma-Glutamyl Transferase* (GGT), dan *Protein* (PROT).

Pada penelitian ini, 50% dari sampel atau sebanyak 305 sampel akan digunakan untuk membangun model prediktif, sedangkan 305 sampel yang tersisa akan digunakan untuk memvalidasi kemampuan diagnosis dari model yang telah dibangun. Dari 14 parameter yang diberikan, parameter ID pasien akan dibuang dari himpunan data, karena tidak relevan. Parameter katagori akan digunakan sebagai variabel terikat (sebagai $\boldsymbol{y}$ dan $\boldsymbol{y_v}$ seperti pada algoritma 1) sedangkan parameter umur dan jenis kelamin akan dikatagorikan sebagai parameter diskrit (D dan D_v) dan selebihnya akan diklasifikasikan sebagai parameter numerik (X dan X_v).

Lebih lanjut, pada penelitian ini, digunakan kardinalitas himpunan bagian atau banyaknya parameter untuk setiap himpunan, a^* sebesar 5. Pemilihan dari a^* dilakukan dengan mempertimbangkan jumlah dari parameter numerik secara total berjumlah 10 buah. Tujuan dari pemilihan angka ini adalah untuk menunjukan bahwa algoritma yang dibangun dapat mendiagnosis kondisi pasien lebih baik daripada metode yang digunakan pada [8] meskipun dengan jumlah parameter yang lebih sedikit.

Algoritma 1: diagnosis hepatitis C berbasis metode pohon klasifikasi dan korelasi jarak

masukan: $H = [\boldsymbol{y}, D, X]$, $H_v = [\boldsymbol{y_v}, D_v, X_v]$, a^*
keluaran: $\hat{\boldsymbol{y}}_v, p^*$
Proses:
$[m, n] = size(X)$
$[m_2, n_2] = size(X_v)$
for $i = 1{:}n$
  $\mathfrak{D}_i = distcorr(X(:,i), X(:,:), a^*)$
  $p_i = F_{PK}(\boldsymbol{y}, D, \mathfrak{D}_i)$
  $\hat{\boldsymbol{y}} = predict(p_i, D_v, X_v)$
  $b = 0$
  for $j = 1{:}m_2$
    if $\hat{y}(j) == y_v(j)$
      $b = b + 1$
    Endif
  endfor
  $\%A_i = \frac{b}{m_2} \times 100\%$
endfor
$p^* := \{p_i : \underset{i}{\operatorname{argmax}}\, \%A_i\}$

3. Hasil dan Pembahasan

Pada bagian ini, akan ditunjukan hasil simulasi dari algoritma yang digagas, sesuai dengan data yang telah dibahas pada sub bagian 2.2. Untuk memperdalam analisis yang dihasilkan, maka selain menampilkan akurasi diagnosis dari setiap model, akan ditunjukan pula tingkat sensitivitas atau *True Positive Rate* (TPR), spesifisitas atau *True Negative Rate* (TNR), tingkat positif palsu atau *False Positive Rate* (FPR) dan tingkat negatif palsu atau *False Negative Rate* (FNR) yang dinyatakan dengan persamaan berikut:

$$TPR = \frac{TP}{TP + FN} \times 100\%,$$
$$TNR = \frac{TN}{TN + FP} \times 100\%,$$
$$FPR = \frac{FP}{TN + FP} \times 100\%,$$
$$FNR = \frac{FN}{FN + TP} \times 100\%. \quad (4)$$

Dengan FP, FN, TP dan TN secara berurutan menyatakan jumlah positif palsu, negatif palsu, positif sejati dan negatif sejati. Lebih lanjut, hasil dari penelitian yang dilakukan oleh [8] akan digunakan sebagai patokan untuk mengukur performa dari algoritma yang telah digagas.

3.1. Hasil Penelitian

Tabel 1. Akurasi model prediktif dari algoritma diagnosis penyakit hepatitis C berbasis metode pohon klasifikasi dan korelasi jarak

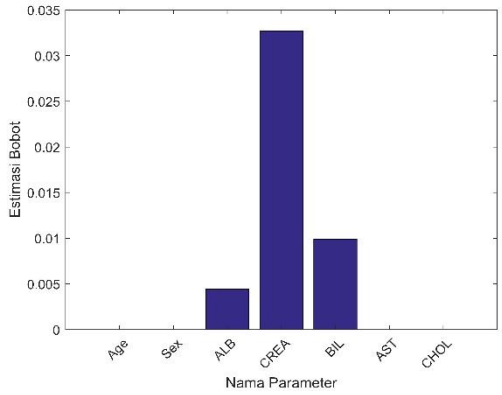
Model	%A	TPR	TNR	FPR	FNR
ALB-CREA-BIL-AST-CHOL*	98,68	94,87	99,25	0,75	5,13
ALP-GGT-CREA-CHOL-ALB	95,72	66,67	100,00	0,00	33,33
ALT-GGT-AST-CHE-ALB	94,08	69,23	97,74	2,26	30,77
AST-GGT-CREA-ALT-ALB	96,05	76,92	98,87	1,13	23,08
BIL-ALB-ALT-CREA-AST*	98,68	94,87	99,25	0,75	5,13
CHE-CHOL-ALB-AST-ALT	93,09	66,67	96,98	3,02	33,33
CHOL-CREA-CHE-ALB-ALP	95,72	66,67	100,00	0,00	33,33
CREA-AST-ALB-GGT-ALP	95,72	66,67	100,00	0,00	33,33
GGT-CREA-AST-ALB-ALT	96,05	76,92	98,87	1,13	23,08
PROT-ALB-AST-GGT-CHE	89,80	94,87	89,06	10,94	5,13
metode klasikal	95,72	66,67	100,00	0,00	33,33

*model terbaik secara keseluruhan

Tabel 1 menunjukan akurasi, TPR, TNR, FNR dan FPR dari algoritma yang digagas dalam mendiagnosis kemungkinan seseorang telah terjangkit hepatitis C. Berdasarkan pada Tabel 1, dapat diamati bahwa model yang mengandung parameter *Albumin* (ALB), *Creatinine* (CREA), *Bilirubin* (BIL), *Aspartate Transaminase* (AST), *Cholesterol* (CHOL) adalah model dengan akurasi terbaik, disusul oleh model *Bilirubin* (BIL), *Albumin* (ALB), *Alanine Transaminase* (ALT), *Creatinine* (CREA), dan *Aspartate Transaminase* (AST), dimana kedua model memiliki akurasi yang lebih tinggi dibandingkan dengan metode yang diusulkan pada [8], dengan akurasi sebesar 98,68%. Hal ini menunjukan bahwa algoritma yang digagas mampu meningkatkan performa diagnosis daripada algoritma yang digunakan sebelumnya.

DOI : https://doi.org/10.52158/jacost.v4i2.556

Lebih lanjut, dari Tabel 1, diperlihatkan bahwa algoritma yang digagas membutuhkan lebih sedikit informasi, yakni hanya 7 parameter dan bukan 12 seperti pada metode klasik untuk mendiagnosis apakah seseorang mengidap hepatitis C atau tidak. Ini menunjukan bahwa algoritma yang digagas lebih efisien dalam penggunaan informasi dalam data untuk menghasilkan diagnosis yang diinginkan. Meskipun, terdapat model yang tidak menghasilkan akurasi yang lebih baik dari metode klasik, namun harus diperhatikan bahwa, selisih akurasi maksimumnya berada pada sekitar 6%, namun dengan 42% lebih sedikit informasi yang dibutuhkan (hanya membutuhkan 7 dari 12 parameter) relatif terhadap metode klasik. Sehingga



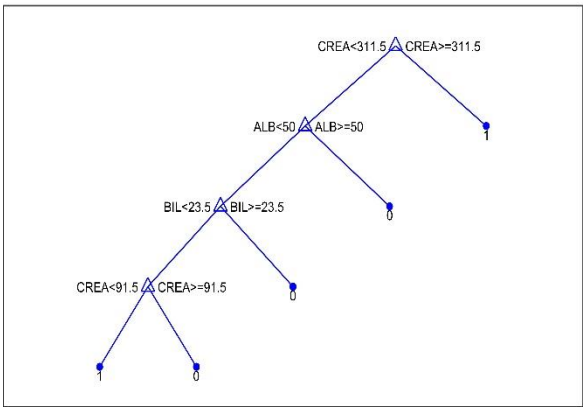
Gambar 1. Ukuran tingkat kepentingan dari parameter untuk model ALB-CREA-BIL-AST-CHOL

dalam kasus ini, dapat disimpulkan bahwa metode yang digagas memiliki performa yang lebih baik jika dibandingkan dengan metode pendahulunya, seperti yang dijelaskan pada [8].

Gambar 1 menunjukan tingkat kepentingan dari setiap parameter atau *parameter importance* (lihat: [15]) untuk model ALB-CREA-BIL-AST-CHOL. Tingkat kepentingan dari parameter diukur berdasarkan pada perubahan galat rerata kuadrat atau *root mean square error* yang berubah akibat pembagian data oleh parameter tertentu. Dalam kasus ini, semakin penting sebuah parameter dalam struktur pohon, maka semakin tinggi estimasi bobot yang dimiliki oleh parameter tersebut. Berdasarkan pada Gambar 1, terlihat bahwa parameter *Creatinine* atau CREA adalah parameter yang paling signifkan untuk menentukan apakah seseorang terjangkit hepatitis C atau tidak, diikuti oleh *bilirubin* dan kemudian *albumin*. Dalam kasus ini, terlihat dari gambar 1 bahwa umur dan jenis kelamin tidak berpengaruh terhadap diagnosis apakah seseorang mengidap diabetes atau tidak.

3.2. Pembahasan

Berdasarkan pada pemaparan diatas, terlihat bahwa algoritma metode pohon klasifikasi dan korelasi jarak seperti yang telah diulas pada Bab 2 mampu meningkatkan performa diagnosis penyakit hepatitis C relatif terhadap metode yang dipaparkan pada [8]. Lebih lanjut, terlihat bahwa algoritma yang digagas membuthkan lebih sedikit informasi jika dibandingkan dengan metode klasikal. Hal ini diakibatkan karena algoritma yang digagas pada penelitian ini berfokuskan membangun data berdasarkan kemiripan antar parameter yang dalam kasus ini, diukur dalam jarak *Euclidean*. Dengan melakukan tahap ini, maka sebenarnya data telah diklasterifikasi terlebih dahulu



Gambar 2. Jalur diagnostik penyakit hepatitis C yang terbentuk oleh model ALB-CREA-BIL-AST-CHOL. Angka 0 menandakan negatif hepatitis C, sedangkan 1 positif hepatitis C

sebelum dimodelkan, sehingga menggurangi kemungkinan terjadinya permasalahan multikolinearitas yang mengakibatkan model menjadi tidak presisi untuk mendiagnosa. Hal ini menjadi keunggulan terhadap metode yang digunakan pada [8], dimana pada penelitian tersebut, optimasi performa diagnosis tidak mendapat sorotan utama.

Selain itu, pada Tabel 1, dapat dilihat pula bahwa model dari algoritma yang digagas memiliki tingkat negatif positif palsu yang lebih rendah jika dibandingkan dengan metode yang diperkenalkan pada [8]. Rendahnya tingkat negatif palsu merupakan salah satu keunggulan utama dari algoritma yang digagas, karena sebagaimana telah diketahui bahwa tingkat negatif palsu berhubungan dengan kesalahan klasifikasi, dimana sampel yang positif digolongkan sebagai negatif. Dibidang medis, hal ini merupakan hal yang harus dihindari, karena menunjukan bahwa seseorang didiagnosa sebagai sehat, padahal telah ada indikasi bahwa orang tersebut sedang sakit.

Identifikasi faktor resiko. Pada Gambar 1, terlihat bahwa parameter *creatinine*, *bilirubin* dan *albumin* adalah 3 parameter yang paling signifikan dalam menentukan apakah seseorang mengidap penyakit hepatitis C atau tidak. Gambar 2 secara khusus

DOI : https://doi.org/10.52158/jacost.v4i2.556

menunjukan jalur diagnostik yang terbentuk oleh model yang menggunakan parameter estimasi sesuai dengan pada Gambar 1.

Berdasarkan pada jalur diagnostik yang terbentuk, dapat diidentifikasi bahwa seseorang diduga mengidap hepatitis C apabila memenuhi salah satu dari kriteria berikut:

1. *Creatinine* di bawah 91,5; *bilirubin* di bawah 23,5 dan *albumin* di bawah 50.
2. *Creatinine* di atas 311,5.

Sedangkan seseorang dinyatakan negative hepatitis C apabila:

1. *Creatinine* di bawah 311,5; dan *albumin* di atas 50.
2. *Creatinine* di bawah 311,5; dan *albumin* di bawah 50 namun, *bilirubin* di atas 23,5
3. *Bilirubin* di bawah 23,5; *albumin* di bawah 50, namun *creatinine* berada diantara 91,5 hingga 311,5.

Hasil temuan ini selaras dengan beberapa penelitian dibidang kesehatan [17, 18, 19, 20] yang menjelaskan bahwa kadar *creatinine* yang terlalu tinggi atau rendah dalam tubuh mengidenfikasikan adanya kerusakan fungsi organ hati, yang dalam kasus ini disebabkan oleh virus hepatitis C. *Creatinine* merupakan hasil sisa metabolisme tubuh yang timbul akibat aktivitas tubuh manusia. Tingkat *creatinine* yang tinggi diasosiasikan dengan diasosiasikan dengan adanya indikasi kerusakan ginjal yang berdasarkan temuan pada [20] dapat mengisyaratkan adanya kerusakan organ hati pula. Lebih lanjut, *creatinine* yang terlalu rendah mengindikasikan bahwa organ hati tidak bekerja dengan optimal yang dalam kasus ini diduga pasien gangguan fungsi organ hati [17].

Berdasarkan pada ulasan di atas, terlihat bahwa algoritma yang dibangun telah mampu membangun jalur diagnostik yang seusuai dengan hasil riset dibidang medis. Hal ini menunjukan adanya potensi dari algoritma yang dibangun untuk disusun menjadi sebuah perangkat lunak yang dapat membantu untuk menghasilkan alur diagnostik yang presisi. Secara khusus data informasi kesehatan dari berbagai pasien dapat dikumpulkan sehingga dapat dibangun sebuah sistem informasi yang dapat mempermudah tim medis dalam mendiagnosa penyakit dari seorang pasien. Secara spesifik, arah penelitian mendatang yang dapat diusulkan adalah untuk menguji algoritma yang telah digagas pada berbagai penyakit lainnya, seperti diabetes ataupun gagal jantung dan penggunaan parameter dengan jumlah yang lebih besar untuk menguji kemampuan dari algoritma pada data yang tergolong ke dalam kelas mahadata (*big data*). Sehingga, dapat dihasilkan sebuah perangkat lunak penentu jalur diagnosis penyakit yang lengkap dan dapat diterapkan di berbagai fasilitas kesehatan, guna mendukung alur tranformasi digital.

4. Kesimpulan

Pada penelitian ini, telah digagas sebuah algoritma baru untuk mendiagnosis pasien hepatitis C berdasarkan pada konsep pembelajaran mesin. Algoritma diagnosis ini dibangun dengan menggunakan metode pohon klasifikasi dan pemilihan himpunan bagian berbasis korelasi jarak. Berdasarkan pada hasil simulasi yang telah dilakukan, terlihat bahwa algoritma yang digagas miliki akurasi sebesar 98,68% dalam mendiagnosis seseorang mengidap penyakit hepatitis C atau tidak, dengan tingkat positif palsu sebesar 5%. Hal ini menunjukan potensi dari algoritma yang digagas untuk diterapkan dalam dunia medis guna menghasilkan diagnosis penyakit yang cepat dan presisi.

Daftar Pustaka

[1] D. Bradshaw, J. L. Mbisa, A. M. Geretti, B. J. Healy, G. S. Cooke, G. R. Foster, E. C. Thomson, J. McLauchlan, K. Agarwal and C. a. o. Sabin, "Consensus recommendations for resistance testing in the management of chronic hepatitis C virus infection: Public Health England HCV Resistance Group," *Journal of Infection*, vol. 76, no. 9, pp. 503-512, 2019.

[2] M. G. Ghany, T. R. Morgan and h. C. g. p. AASLD-IDSA, "Hepatitis C guidance 2019 update: American Association for the Study of Liver Diseases--Infectious Diseases Society of America recommendations for testing, managing, and treating hepatitis C virus infection," *Hepatology*, vol. 71, no. 2, pp. 686-721, 2020.

[3] S. Blach, N. A. Terrault, F. Tacke, I. Gamkrelidze, A. Craxi, J. Tanaka, I. Waked, G. J. Dore, Z. Abbas and A. R. a. o. Abdallah, "Global change in hepatitis C virus prevalence and cascade of care between 2015 and 2020: a modelling study," *The Lancet Gastroenterology & Hepatology*, vol. 7, no. 5, pp. 396-415, 2022.

[4] A. Souri, M. Y. Ghafour, A. M. Ahmed, F. Safara, A. Yamini and M. Hoseyninezhad, "A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment," *Soft Computing*, vol. 24, no. 22, pp. 17111-17121, 2020.

[5] A. B. Shatte, D. M. Hutchinson and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological medicine*, vol. 49, no. 9, pp. 1426-1448, 2019.

[6] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.

[7] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and regression trees, Routledge, 2017.

[8] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—a case study," *J Lab Precis Med*, vol. 3, no. 6, 2018.

[9] F. Smarra, J. Tjen and A. D'Innocenzo, "Learning methods for structural damage detection via entropy-based sensors selection," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 10, pp. 6035-6067, 2022.

[10] J. Tjen, G. Hoendarto and T. Darmanto, "Ensemble of the Distance Correlation-Based and Entropy-Based Sensor Selection for Damage Detection," in *2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, Online Virtual meeting, 2022.

[11] J. Tjen, G. Hoendarto, T. Darmanto and T. Willay, "Distance Correlation-Based Regression Tree Algorithm For Structural

DOI : https://doi.org/10.52158/jacost.v4i2.556

Damage Detection," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 10, no. 2, pp. 440-455, 2023.

[12] J. Tjen, F. Smarra and A. D'Innocenzo, "An entropy-based sensor selection algorithm for structural damage detection," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, Online virtual meeting, 2020.

[13] J. Tjen, "Algoritma Pendeteksi Kerusakan Struktur Bangunan Berbasis Korelasi Jarak dan Metode Kuadrat Terkecil Parsial," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 8, no. 3, pp. 459-469, 2022.

[14] R. Huang, C. Cui, W. Sun and D. Towey, "Is euclidean distance the best distance measurement for adaptive random testing?," in *IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*, Porto, 2020.

[15] University of California Irvine, "HCV data," 2020. [Online]. Available: https://archive.ics.uci.edu/dataset/571/hcv+data. [Accessed 13 June 2023].

[16] G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor, "Tree-based methods," in *An Introduction to Statistical Learning: with Applications in Python*, Springer, 2023, pp. 331-366.

[17] B. Cerbu, M. L. Grigoras, F. Bratosin, I. Bogdan, C. Citu, A. V. Bota, M. Timircan, M. L. Bratu, M. C. Levai and I. Marincu, "Laboratory Profile of COVID-19 Patients with Hepatitis C-Related Liver Cirrhosis," *Journal of Clinical Medicine*, vol. 11, no. 3, p. 652, 2020.

[18] M. F. Alsaffar, "Elevation of some biochemical and immunological parameters in hemodialysis patients suffering from hepatitis C virus infection in Babylon Province," *Indian Journal of Forensic Medicine & Toxicology*, vol. 15, no. 3, pp. 2354-2362, 2021.

[19] S. Y. Han, H. Y. Woo, J. Heo, S. G. Park, S. I. Pyeon, Y. J. Park, D. U. Kim, G. H. Kim, H. H. Kim and G. Am Song, "The predictors of sustained virological response with sofosbuvir and ribavirin in patients with chronic hepatitis C genotype 2," *The Korean Journal of Internal Medicine*, vol. 36, no. 3, p. 544, 2021.

[20] S. Pol, L. Parlati and M. Jadoul, "Hepatitis C virus and the kidney," *Nature Reviews Nephrology*, vol. 15, no. 2, pp. 73-86, 2019.

DOI : https://doi.org/10.52158/jacost.v4i2.556