PERBANDINGAN METODE ALGORITMA C4.5 DAN *EXTREME LEARNING MACHINE* UNTUK MENDIAGNOSIS PENYAKIT JANTUNG KORONER

Jefri Junifer Pangaribuan[1)], Cathlin Tedja[2)], Sentosa Wibowo[3)]

[1]Fakultas Ilmu Komputer, Universitas Pelita Harapan Medan
E-mail: jefrijuniferp@gmail.com[1)]
[2]Fakultas Ilmu Komputer, Universitas Pelita Harapan Medan
E-mail: cathlintedja09@gmail.com[2)]
[3]Fakultas Ilmu Komputer, Universitas Pelita Harapan Medan
E-mail: sentosa.huang1997@gmail.com[3)]

Abstract – According to the World Health Organization (WHO) an increase in cardiovascular disease uplift by 28% per year and will increase every year if the diagnosis of the disease is not done. In 2015, according to the World Health Organization (WHO) stated that there were 17.5 million people worldwide dying of cardiovascular disease or 31% of deaths worldwide, and in Indonesia the death rate caused by cardiovascular disease was 7.4 million (42.3%) Mostly are caused by coronary heart disease (CHD). Coronary heart disease is a disease caused by narrowing of the coronary arteries. There are various methods that can be used to diagnose disease, known as artificial neural networks. This research will implement a method known as Extreme Learning Machine (ELM). Extreme Learning Machine is an artificial neural network with one or more hidden layers known as the single hidden layer feed-forward neural. Algorithm C4.5 is an algorithm that can be used to create a decision tree. The result conducted from experiment using method algorithm C4.5 providing excellent diagnosis results. By using confusion matrix, the accuracy level represent that the algorithm C4.5 1.27 times better than extreme learning machine.

***Keywords**: Cardiovascular Disease, Coronary Heart Disease, Artificial Neural Network, Extreme Learning Machine, Algorithm C4.5*

Abstrak – Menurut World Health Organization (WHO) peningkatan penyakit kardiovaskuler meningkat sebanyak 28% per tahun dan akan semakin bertambah setiap tahunnya bila tidak di diagnosis penyakit tersebut. Pada tahun 2015, menurut World Health Organization (WHO) menyebutkan bahwa terdapat 17.5 juta orang di dunia meninggal akibat penyakit kardiovaskular atau 31% dari kematian di seluruh dunia, dan di Indonesia sendiri angka kematian yang disebabkan penyakit kardiovaskukar adalah 7,4 juta (42.3%) diantaranya disebabkan oleh penyakit jantung koroner (PJK). Penyakit Jantung koroner merupakan penyakit yang disebabkan karena penyempitan arteri koroner. Terdapat berbagai metode yang dapat digunakan untuk mendiagnosa apakah seseorang terkena penyakit jantung

koroner atau tidak yaitu dengan menggunakan metode jaringan saraf tiruan. Penelitian ini mengimplementasikan suatu metode yaitu Extreme Learning Machine (ELM). Extreme Learning Machine merupakan jaringan saraf tiruan feed-forward dengan satu atau lebih hidden layer yang dikenal dengan istilah single hidden layer feed-forward neural. Algortima C4.5 adalah salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan (decision tree). Setelah eksperimen dilakukan menggunakan metode algoritma C4.5 mampu memberikan hasil diagnosis yang sangat baik. Dengan menggunakan confusion matrix, didapatkan tingkat akurasi yang menunjukkan bahwa algoritma C4.5 memiliki nilai 1.27 kali lebih baik dibandingkan dengan extreme learning machine (ELM).

Kata Kunci: *Penyakit Kardiovaskular, Jantung Koroner, Jaringan Saraf Tiruan, Extreme Learning Machime, Algoritma C4.5*

PENDAHULUAN

Penyakit jantung koroner merupakan penyakit jantung yang disebabkan oleh penyempitan yang terjadi pada arteri koroner [1]. Arteri yang mengalirkan darah ke otot jantung mengalami gangguan, sehingga jantung tidak mampu untuk memompa darah untuk memenuhi organ-organ vital. Salah satu penyebab terjadinya penyakit jantung koroner adalah kadar kolestrol yang tinggi.

Data mining merupakan tahapan pengelolaan data yang menggunakan teknik statistika, kecerdasan buatan, *machine learning* untuk mengekstrak serta mengidentifikasi informasi yang bermanfaat beserta indikator penting dari berbagai *database* seperti *Kaggle*.

Extreme Learning Machine (ELM) adalah salah satu metode terbaru dari Jaringan Saraf Tiruan. ELM merupakan jaringan saraf tiruan *feedforward* dengan *single hidden layer* atau biasa disebut sebagai *single hidden layer feedforward neural networks* (SLFNs).

Algoritma C4.5 adalah algoritma yang pada umumnya digunakan guna untuk membentuk *decision tree*. *Decision tree* adalah salah satu metode klarifikasi yang paling popular karena mudah diinterpretasi oleh manusia. Kelebihan algoritma C4.5 adalah dapat menghasilkan pohon keputusan yang memiliki tingkat akurasi yang dapat diterima dan efisien dalam menangani atribut yang bertipe diskret atau numerik.

STUDI LITERATUR

Jantung Koroner

Terdapat berbagai jenis penyakit yang menjadi penyumbang angka mortalitas terbanyak pada kelompok penyakit bersifat tidak menular salah satunya adalah penyakit kardiovaskuler. Penyakit kardiovaskuler sendiri merupakan penyakit yang diakibatkan karena organ jantung atau pembuluh darah mengalami gangguan sehingga tidak berfungsi secara normal sehingga menyebabkan munculnya penyakit seperti penyakit jantung koroner, penyakit jantung rematik, dan lain-lain. Penyakit Jantung Koroner (PJK) merupakan salah satu penyakit degeneratif yang dialami penderita dikarenakan penyempitan pembuluh arteri yang mengalirkan darah ke otot jantung. Apabila penyempitan semakin parah, maka dapat mengakibatkan serangan jantung [1].

Data Mining

Data mining dapat dinyatakan juga sebagai prosedur mengekstrak pengetahuan dari sejumlah besar data yang tersedia, yaitu nama pasien, usia, jenis kelamin, dan lain-lain. Pengetahuan yang dihasilkan dari proses *data mining* harus mudah dimengerti dan bermanfaat [2].

Didasarkan atas apa yang dikemukakan [3], terdapat tujuh jenis aktivitas dasar *data mining*, yaitu: Karakterisasi dan Diskriminasi, Klasifikasi, Regresi, *Time Series*, Asosiasi, *Clustering*, Deskripsi dan Visualisasi.

Algoritma C4.5

Dari beberapa algoritma yang dapat digunakan untuk membuat pohon keputusan (*decision tree*) yaitu diantaranya algoritma C4.5.

Secara global algoritma C4.5 untuk menciptakan pohon keputusan sebagai berikut yaitu [4]:

a. Pemilihan atribut yang akan dijadikan sebagai akar
b. Pembuatan cabang untuk setiap nilai
c. Pembagian kasus dalam cabang
d. Repetisi proses sampai setiap cabang guna agar semua kasus pada cabang memiliki kelas yang sama

Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) yaitu jaringan saraf tiruan *feed-forward* yang didalamnya terdapat lapisan tunggal tersembunyi diketahui dengan istilah yaitu *single hidden layer feed-forward neural network* (SLFN).

Extreme learning machine mempunyai karakteristik yang memukau dan substansial, namun berbeda dengan algoritma pembelajaran yang berdasarkan atas pada gradien yang populer untuk jaringan saraf *feed-forward*. Karakteristik yang dimaksudkan sebagai berikut [5]:

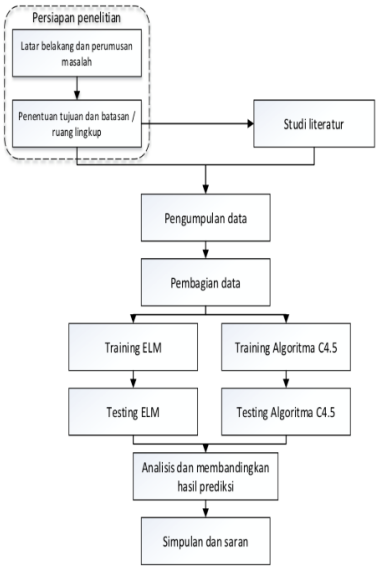
a. Kecekatan mengkaji pada *extreme learning machine* tergolong sangat cepat. Dalam simulasi yang diperoleh dari hasil laporan dalam literatur, fase pembelajaran pada *extreme learning machine* dapat dituntaskan cukup hitungan detik untuk aplikasi yang jumlahnya tergolong banyak.
b. *Extreme learning machine* mempunyai kinerja generalisasi yang jauh lebih baik dibandingkan dengan pembelajaran berbasiskan pada gradien, misalkan seperti dalam kebanyakan kasus yaitu *backpropagation* [6].
c. *Extreme learning machine* memiliki kecenderungan dalam pencapaian penyelesaian sederhana tanpa memiliki efek masalah yang sifatnya sepele. Algoritma pembelajaran *extreme learning machine* terlihat jauh lebih sederhana dibandingkan dengan algoritma pembelajaran jaringan saraf *feed-forward* pada kebanyakannya. Namun berbeda terhadap algoritma pembelajaran berbasiskan pada gradien yang cuma bekerja untuk fungsi aktivasi terdiferensiasi, algoritma *extreme learning machine* ini mampu diterapkan kedalam SLFNs guna berfungsi melatih SLFNs, dengan jumlah fungsi aktivasi yang banyak dan tidak mempunyai diferensiasi.

METODE PENELITIAN

Tahapan Penelitian

Tahap awal penelitian akan dilakukan penentuan latar belakang dan tujuan penelitian serta mendefinisikan batasan / ruang lingkup. Studi literatur mempunyai peran penting dalam memperdalam pemahaman akan cara kerja Algoritma C4.5 dan ELM serta yang tahapan-tahapan yang diperlukan agar dapat mendiagnosis penyakit jantung koroner dengan menggunakan Algoritma C4.5 dan ELM.

Tahap kedua dari penelitian ini yaitu pengumpulan data. Tahap ketiga yakni implementasi Algoritma C4.5 dan ELM melalui pembagian data, *training* dan *testing* ELM. Tahap keempat adalah membandingkan hasil analisis dari Algoritma C4.5 dan ELM. Tahap terakhir adalah menarik kesimpulan dan memberikan saran dari hasil penelitian yang diperoleh.



Gambar 1. Tahapan Penelitian

Metode Pengumpulan Data

Data set penyakit jantung koroner yang digunakan untuk melakukan penelitian ini didapatkan dari *Kaggle. Cleveland dataset (VA Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, MD, PhD) Irvine: The University of California Irvine; 1988.* Database yang digunakan dalam penelitian ini berjumlahkan 303 pasien yang dikumpulkan dari daerah *Cleveland* dengan menggunakan 14 atribut yang nantinya akan dipisahkan menjadi data *training* dan data *testing*.

Metode Diagnosis dengan Algoritma C4.5

Untuk mencari entropi pada algoritma C4.5, rumus daripada Entropi adalah sebagai berikut:

$$Entropi\ (S) = \sum_{i=1}^{k} -pi\ \log_2 pi \qquad (1)$$

Keterangan rumus:
S adalah himpunan (*dataset*) pada kasus, *k* adalah banyaknya partisi *S*, *Pi* adalah probabilitas yang didapat dari Sum (Ya) atau Sum (Tidak) dibagi total kasus.

Lakukan analisa pada setiap atribut beserta nilai-nilainya dan juga perhitungan entropinya setelah mendapatkan hasil entropi menggunakan rumus sebelumnya. Langkah selanjutnya adalah dengan menghitung *Gain*. Rumus perhitungan *Gain* sebagai berikut:

$$Gain(A)\ = Entropi\ (S) - \sum_{i=1}^{k} \frac{|Si|}{|S|}\ x\ Entropi\ (Si.) \qquad (2)$$

Keterangan rumus:
S adalah himpunan kasus, *A* adalah atribut yang tersedia, *k* adalah jumlah partisi pada atribut A, |*Si*| adalah jumlah kasus yang ada pada partisi ke-i, |*S*| adalah jumlah kasus yang ada dalam S.

Metode Diagnosis dengan ELM

Tool yang diterapkan pada penelitian ini adalah untuk mendiagnosis penyakit jantung koroner adalah MATLAB.

Langkah 1: Pembagian data *training* dan data *testing*.

Proses *training* dan *testing* bersifat absolut pada prosedur diagnosis menggunakan ELM. Proses *training* digunakan untuk menjabarkan secara luas proses ELM berbeda dengan proses *testing* yang digunakan untuk menilai kapabilitas ELM sebagai alat diagnosis. Menurut [7] pembagian komposisi data

training dan *testing* yang efektif diantaranya:

1. Data *training* yang digunakan berjumlahkan 80% dari total data
2. Data *testing* yang digunakan berjumlahkan 20% dari total data.

Langkah 2: Training

a. Normalisasi Data *Training*
Data yang akan dimasukkan ke dalam ELM dinormalisasi sehingga mempunyai nilai dengan rentang tertentu. Hal ini diperlukan agar fungsi aktivasi yang digunakan dapat memberikan hasil *output* yang mempunyai nilai rentang data 0,1 atau -1,1. Berikut merupakan rumus yang diterapkan dalam melakukan perhitungan normalisasi

$$x = 2 \times \frac{(x_p - min\ x_p)}{(\max x_p - \min x_p)} - 1$$

Dimana:
x = nilai hasil normalisasi dengan rentang antara -1 sampai 1
x_p = nilai data aktual yang belum dinormalisasi
Min x_p = nilai minimum pada *data set*
Max x_p = nilai maksimum pada *data set*

b. Menentukan fungsi aktivasi dan jumlah *hidden neuron*
Pada proses *training*, jumlah *hidden neuron* beserta peranan aktivasi dari ELM ditentukan terlebih dahulu. Tujuan dari penelitian ini untuk menghasilkan fungsi aktivasi sigmoid. Menurut [8], ELM menghasilkan *output* peramalan yang stabil dengan jumlah *hidden neuron* 0-30. Tetapi jika *output* yang dihasilkan ELM kurang optimal, maka jumlah *hidden neuron*-nya akan diubah

c. Menghitung bobot *input*, bias dari *hidden neuron*, dan bobot *output*
Output yang dihasilkan dari proses *training* ELM yaitu: bobot *input*, bobot *ouput*, dan bias dari *hidden neuron* dengan tingkat *error* yang rendah dan diukur menggunakan *Mean Square Error* (MSE). Bobot *input* ditentukan secara acak, sedangkan bobot *output* merupakan *invers* dari matriks *hidden layer* dan *output*. Secara matematis dapat ditulis sebagai berikut:

$$\beta = H^T T \quad (1)$$

$$H = (W_{i,\dots,} W_N, b_{i,\dots,} b_n, x_{i,\dots,} x_N) \quad (2)$$

$$= \begin{matrix} g(w_i.x_i + b_i) & \cdots & g(w_N.x_i + b_N) \\ \vdots & \vdots & \vdots \\ g(w_i.x_N + b_i) & \cdots & g(w_N.x_N + b_N) \end{matrix}$$

$$\beta = \begin{matrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{matrix} \qquad T = \begin{matrix} t_1^T \\ \vdots \\ t_N^T \end{matrix} \quad (3)$$

d. Denorminalisasi *output*
Output yang dihasilkan dari proses *training* selanjutnya didenormalisasi. Berikut rumus denormalisasi yang digunakan:

$$x = 0.5 \times (x_p + 1) \times (\max x_p - \min x_p) + \min x_p$$

Dimana:
x = nilai data setelah denormalisasi
x_p = nilai pada data aktual yang belum denormalisasi
Min x_p = nilai minimum pada *data set* sebelum denormalisasi
Max x_p = nilai maksimum pada *data set* sebelum denormalisasi

Langkah 3: *Testing* ELM

Berdasarkan bobot *input* dan bobot *output* yang didapatkan dari proses *training*, maka tahap selanjutnya adalah melakukan diagnosis dengan menggunakan ELM. Pada tahap ini, data *input* dinormalisasi dan didenormalisasi

dengan rentang dan rumus yang sama dengan data *training*.

Metode Analisis Kinerja

Evaluasi keakuratan untuk memperkuat hasil prediksi ELM adalah *classification accuracy*, *sensitivity*, dan *specificity*. *Classification accuracy* merupakan sebuah ketepatan klasifikasi yang diperoleh. Sedangkan *sensitivity* merupakan ukuran ketepatan dari suatu kejadian yang diinginkan. Dan *specificity* merupakan suatu paramater yang menyatakan persentase kejadian-kejadian yang tidak diinginkan. *Classification accuracy*, *sensitivity*, dan *specificity* dapat ditentukan dengan menggunakan nilai yang terdapat dalam *confusion matrix*. Rumus untuk menghitung *classification accuracy*, *sensitivity*, dan *specificity* pada *Confusion Matrix* adalah:

$$Classification\ Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{a + d}{a + b + c + d}$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{a}{a + b}$$

$$Specificity = \frac{TN}{FP + TN} = \frac{d}{c + d}$$

Dimana:

TP = total *true positives*; total data positif yang terdeteksi benar

TN = total *true negatives*; total data negatif yang terdeteksi benar

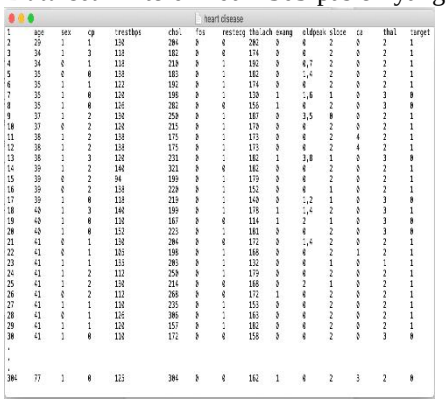
FP = total *false positives*; total data positif yang terdeteksi salah

FN = total *false negatives*; total data negatif yang terdeteksi salah

HASIL DAN PEMBAHASAN

Hasil Pengumpulan Data

Dengan melakukan pencarian pada situs *Kaggle*, diperoleh *data set* penyakit jantung koroner yang bersifat klasifikasi. *Data set* ini terdiri dari 303 pasien yang



dikumpulkan dari daerah *Cleveland*. Informasi nilai *data set* disediakan dalam *file* teks dengan format yang ditampilkan di bawah ini.

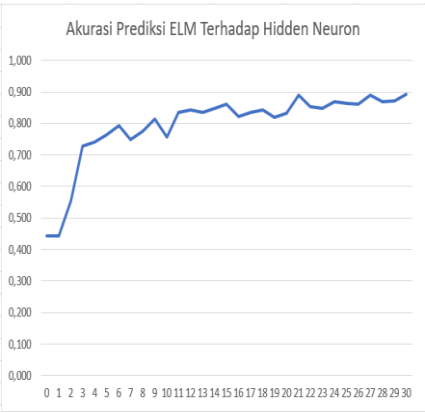
Gambar 2. *Data Set Heart Disease*

Hasil Prediksi

Hal yang perlu ditentukan dalam prediksi menggunakan ELM adalah jumlah *input*. Penelitian ini akan menggunakan 13 buah variabel *input*, diantaranya umur, jenis kelamin, *chest pain type*, *resting blood pressure*, *serum cholestrol*, *fasting blood sugar*, *resting electrocardiographic results*, *maximum heart rate achieved*, *exercise induced angina*, *ST depression induced by exercise relative to rest*, *the slope of the peak exercise ST segment*, *number of major vessels*, dan *thal*.

Penentuan jumlah *hidden neuron* akan berpengaruh pada prediksi ELM yang optimal. Prediksi ELM sendiri menghasilkan *output* peramalan yang stabil dengan rentang jumlah *hidden neuron* 0 sampai dengan 30 [8]. Data hasil tingkat akurasi yang dihasilkan dari prediksi menggunakan metode ELM

didasarkan atas nilai jumlah *hidden neuron* dan dapat ditinjau pada Gambar 3 di bawah ini.



Gambar 3. Grafik Akurasi Prediksi ELM Terhadap Perubahan Jumlah *Hidden Neuron*

Tabel 1 dibawah menunjukkan perbandingan tingkat keakuratan kinerja ELM dan Algoritma C4.5 pada *data testing*.

Tabel 1. Perbandingan tingkat akurasi *data testing* ELM dan Algoritma C4.5

Metode	Classification Accuracy	Sensitivity	Specificity
ELM	73,33%	56,67%	90%
Algoritma C4.5	93,33%	93,33%	93,33%

Berdasarkan tabel 1 diatas, dapat dilihat bahwa tingkat akurasi Algoritma C4.5 1.27 kali lebih tinggi dibandingkan dengan metode ELM.

KESIMPULAN

Hasil yang diperoleh dari percobaan dan analisis diagnosis penyakit jantung koroner dengan menggunakan metode *Extreme Learning Machine* (ELM) dan Algoritma C4.5 adalah Algoritma C4.5 mempunyai tingkat akurasi yang lebih tinggi dibandingkan dengan ELM dikarenakan Algoritma C4.5 dijabarkan berdasarkan nilai *gain* tertinggi yang dijadikan *nodes* dan hasil akhirnya dalam bentuk *leaves* yang berisikan indikator penentu seseorang penyakit jantung koroner atau tidak.

DAFTAR PUSTAKA

[Perhatikan cara penulisan Daftar Pustaka, dibawah ini]

[1] I. Soeharto, "Kolestrol dan Lemak Jahat, Kolestrol dan Lemak Baik serta Proses Terjadinya Serangan Jantung dan Stroke," 2001.

[2] Gambbeta and Windy, "Pohon Keputusan (Decision Tree)," *Departemen Teknik Informatika Institute Teknologi Bandung*, 2012.

[3] Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making," 2009.

[4] B. Max, "Principles of Data Mining," *Springer Science*, 2007.

[5] Huang, Zhu and Siew, "Extreme Learning Machine: Theory and Applications," *Neurocomputing Vol. 70*, pp. 489-501, 2006.

[6] J. J. Pangaribuan, "Mendiagnosis Penyakit Diabetes Melitus Dengan Menggunakan Metode Extreme Learning Machine," *Isd Vol. 2 No. 2*, pp. 2528-5114, 2016.

[7] G. Zhang, B. Pattuwo and M. Hu, "Forecasting with Artificial Neural Networks: The State of the Art," *Elsevier International Journal of Forecasting 14 (1998)*, pp. 35-62, 1997.

[8] C. A. Sun and Yu, "Sales Forecasting using Extreme Learning Machine with Application in Fashion Retailing," *Elsevier Decision Support System 46*, pp. 411-419, 2008.

