Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning

Elly Pusporani[(1)] Siti Qomariyah[(2)] dan Irhamah[(3)]
[(1)(2) (3)]Statistika, Fakultas Matematika Komputasi dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111
E-mail: *[(1)]elly.pusporani@gmail.com, [(2)]sitidotkom@gmail.com, [(3)] irhamah@statistika.its.ac.id*

Abstrak—Liver atau hati adalah organ yang perannya sangat vital dalam tubuh manusia. Penyakit liver sering dianggap sebagai silent killer (pembunuh diam-diam) karena adanya kemungkinan tidak timbul gejala. Permasalahan yang terjadi adalah sulitnya mengenali penyakit liver sejak dini., bahkan saat penyakit ini sudah menyebar pun masih sulit untuk dideteksi. Padahal penderita perlu mengetahui adanya gejala penyakit liver sejak dini agar dapat segera melakukan pengobatan. Adanya diagnosa penyakit liver sejak dini mampu meningkatkan kelangsungan hidup pasien. Pada penelitian ini diterapkan metode untuk klasifikasi penyakit liver menggunakan machine learning dan dibandingkan hasilnya dengan metode klasik. Data yang digunakan adalah Indian liver patients dataset (ILPD)yang diambil dari UCI machine learning. Terdapat beberapa tahapan preprocessing yang dilakukan, antara lain pengecekan missing value, imputasi, feature selection, dan resampling untuk mengatasi data imbalance. Setelah dilakukan preprocessing, selanjutnya dilakukan analisis menggunakan metode regresi logistik, decision tree, naïvebayes, k-nearest neighbor, dan support vector machine. Berdasarkan nilai akurasi dan presisi, maka metode SVM memberikan hasil yang terbaik, tapi berdasarkan recall maka metode K-Nearest Neighbor memberikan hasil terbaik. Walaupun SVM memberikan hasil nilai akurasi dan presisi tertinggi tetapi terdapat ketimpangan yang besar antara nilai presisi dan recall yang dihasilkan, jika dibandingkan selisih nilai akurasi dan recall dari metode K-Nearest Neighbor.

Kata Kunci —Imbalance, Klasifikasi, Decision tree, Naïve Bayes, K-nearest Neighbor, Support Vector Machine, Penyakit Liver

I. PENDAHULUAN

LIVER atau hati adalah organ yang vital bagi manusia. Organ ini terletak di dalam rongga perut sebelah kanan, tepatnya di bawah diafragma. Terdapat beberapa fungsi kerja liver antara lain sebagai penawar dan penetralisir racun, mengatur sirkulasi hormon, mengatur komposisi darah yang mengandung lemak, gula, protein, dan zat lain. Liver juga berfungsi membuat empedu, zat yang membantu pencernaan lemak. Penyakit liver merupakan suatu gangguan pada setiap fungsi liver. Liver bertanggung jawab untuk fungsi-fungsi kritis dalam tubuh, dimana hilangnya fungsi-fungsi tersebut dapat menyebabkan kerusakan yang signifikan pada tubuh. Liver adalah satu-satunya organ dalam tubuh yang dapat dengan mudah mengganti sel-sel yang rusak, tetapi jika sel-sel itu hilang, maka liver tidak mungkin dapat memenuhi kebutuhan tubuh. Penyakit liver sering disebut sebagai pembunuh diam-diam karena kemungkinan tidak timbulnya gejala.

Permasalahan yang biasanya terjadi adalah sulitnya mengenali penyakit liver sejak dini, bahkan ketika penyakit tersebut sudah menyebar. Padahal mengetahui adanya gejala penyakit liver sejak dini ini sangat diperlukan, agar penderita dapat melakukan pengobatan dengan tepat. Dengan diagnosa adanya penyakit liver lebih awal dapat meningkatkan tingkat kelangsungan hidup pasien[1]. Olehkarena itu, pada penelitian ini peneliti ingin mengetahui apakah model yang paling sesuai untuk mengklasifikasikan data penyakit liver agar penyakit liver bisa dideteksi sejak dini.

Salah satu metode statistika yang dapat melakukan pengkategorian adalah klasifikasi. Terdapat beberapa metode yang dapat digunakan untuk kasus klasifikasi, antara lain regresi logistik, *naïve bayes, k-nearest neighbor* (KNN), *support vector machine* (SVM). Regresi Logistik adalah salah satu metode klasik yang biasanya digunakan dalam klasifikasi. Metode selanjutnya adalah *naïve bayes*. Kelebihan dari metode *naïve bayes* adalah algoritmanya sederhana namun mampu menghasilkan akurasi yang tinggi [2]. Selain itu, metode lain yang dapat digunakan yaitu SVM. SVM mampu bekerja dengan sangat baik pada data dengan banyak dimensi dan menghindarkan dari permasalahan dimensionalitas[3]. Pada penelitian sebelumnya, Ramana dkk melakukan penelitian mengenai diagnosis penyakit liver dengan menggunakan *naïve bayes classifier*, *decision tree*, *back propagation neural network algorithm*, *k-nearest neighbor dan* SVM[4]. Shambel dan Pooja pada penelitiannya juga menggunakan metode SVM, *decision tree, dan naïve bayes* untuk melakukan prediksi pada data penyakit liver[5]. Berdasarkan dua penelitian sebelumnya, peneliti hanya menggunakan metode bebas asumsi saja pada penelitiannya. Pada penelitian ini, peneliti ingin membandingkan bagaimana hasil klasifikasi menggunakan metode bebas asumsi dan metode konvensional. Oleh karena itu, pada penelitian ini beberapa metode yang akan digunakan peneliti untuk melakukan klasifikasi pada *Indian liver dataset* adalah regresi logistik, *decision tree, naïve bayes, k-nearest neighbor, dan support vector machine*.

II. TINJAUAN PUSTAKA

A. Regresi Logistik

Regresi logistik biner merupakan salah satu teknik analisis statistika dengan satu atau lebih variabel bebas dan satu variabel respon. Variabel bebas dapat berupa data kategorik maupun kontinu, sedangkan untuk variabel respon harus berskala kategorik. Peubah bebas dalam hal ini biasanya ditunjukan oleh vektor $\mathbf{X} = \left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_p\right)$ dan variabel respon Y. Dimana variabel Y adalah biner. Artinya variabel Y hanya mempunyai

dua kemungkinan yaitu 0 dan 1. Variabel Y sendiri mengikuti distribusi bernouli. Regresi Logistik adalah model yang paling penting digunakan untuk data dengan respon kategori [6]. Tujuan yang ingin dicapai dari analisis dengan menggunakan metode ini adalah untuk menemukan model yang paling tepat dan paling persimoni.

$$f(Y = y) = \pi^y(1-\pi)^{1-y} \quad (1)$$

Jika peubah respon Y berjumlah n, peluang setiap kejadian sama dan setiap kejadian saling bebas dengan kejadian lainnya maka peubah respon Y akan mengikuti sebaran Binomial. Hosmer dan Lemeshow (2000) menjelaskan bahwa model regresi logistik yang dibentuk $E(Y = 1|X = x)$ dengan $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ model regresi logistiknya adalah sebagai berikut [8].

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (2)$$

Dalam model regresi logistik diperlukan suatu fungsi penghubung yang sesuai dengan model regresi logistik yaitu fungsi logit. Transformasi logit sebagai fungsi dari $\pi(x)$. Persamaan model regresi logistik didapatkan

$$logit[\pi(x)] = log\left(\frac{\pi(x)}{1-\pi(x)}\right) \quad (3)$$

Dengan :

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (4)$$

Dimana p merupakan jumlah variabel bebas. Sebagai catatan jika variabel bebas berbentuk data kartegorik maka diperlukan variabel dummy, sehingga

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_u^{k_j-1} \beta_{ju} D_{ju} + \beta_p x_p \quad (5)$$

B. *Decision Tree*

Decision tree adalah suatu bentuk pohon yang menyerupai diagram alir, dimana internal node menunjukkan variabel prediktor yang digunakan sebagai pemisah yang dihubungkan oleh cabang, dan setiap *leaf node* merupakan kelas hasil klasifikasi [7]. Algoritma ini pertama kali dikembangkan oleh J Ross Quinlan pada awal tahun 1980, yang mengembangkan jenis Decision Tree ID3 (Iterative Dichot-omiser).

Variabel yang dipilih sebagai pemilah adalah variabel yang memiliki nilai goodness of split terbesar, karena variabel tersebut mampu mereduksi heterogenitas lebih besar. Jika variabel prediktor yang digunakan merupakan data kategorik, maka pemisah simpul sebelumnya menjadi simpul kanan dan simpul kiri dapat menggunakan nilai kategori pada variabel tersebut. Namun jika variabel predikor yang digunakan berskala rasio atau merupakan data numerik, maka digunakan berbagai kemungkinan nilai tengah antar setiap data yang telah diurutkan sebagai pemisah simpul, kemudian akan dipillih nilai tengah yang menghasilkan nilai *goodness of fit* terbesar.

C. *Naïve Bayes*

Naïve bayes adalah salah satu metode klasifikasi probabilistik paling sederhana yang didasarkan pada teorema *Bayes*. *Naïve Bayes* merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam data mining [8]. Diketahui $A_1, A_2, \dots, A_p$ merupakan kejadian independen di dalam ruang sampel Ω sehingga $\sum_{k=1}^{p} A_k = \Omega$. Kejadian $A_1, A_2, \dots, A_p$ merupakan partisi dalam $\boldsymbol{\Omega}$, dan B merupakan suatu kejadian acak dari peristiwa $A_1 \cap B, A_2 \cap B, \dots, A_p \cap B$ yang membentuk partisi dalam B.

$$P(B) = \sum_{k=1}^{p} P(A_k \cap B) \quad (6)$$

Jika $P(A_i) > 0$ pada $k = 1,2, \dots, p$ maka $P(A_k \cap B) = P(B|A_k)P(A_k)$

$$P\{B\} = \sum_{k=1}^{p} P\{B|A_k\}P\{A_k\} \quad (7)$$

Kejadian acak B dan $A_1, A_2, \dots, A_p$merupakan partisi dari ruang sampel Ω. Apabila $P\{B\} > 0$ dan $P\{A_k\} > 0$ untuk $k = 1,2, \dots, p$ maka

$$P\{A_k|B\} = \frac{P\{B|A_k\}P\{A_k\}}{\sum_{k=1}^{p} P\{B|A_k\}P\{A_k\}} \quad (8)$$

$P\{A_k|B\}$ disebut *posterior probability* karena nilainya bergantung pada nilai B, $P\{A_k\}$ merupakan *prior probability* karena nilainya tidak bergantung pada B, dan $P\{B|A_k\}$ merupakan fungsi *likelihood* dan $P\{B\}$ merupakan keterangan (*evidence*).

Diberikan $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}\}$ merupakan variabel yang digunakan untuk menentukan kelas $\boldsymbol{y}$. Perhitungan *posterior probability* untuk setiap kelas $\boldsymbol{y_j}$ menggunakan teorema *Bayes* adalah sebagai berikut.

$$P(\boldsymbol{y_j}|\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}) = \frac{P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}|\boldsymbol{y_j}).P(y_j)}{P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p})} \quad (9)$$

Kelas yang terpilih adalah kelas yang memaksimalkan nilai $P(\boldsymbol{y_j}|\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p})$ atau memaksimalkan nilai dari peluang $P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}|\boldsymbol{y_j}).P(\boldsymbol{y_j})$. Berdasarkan Persamaan 2.20 maka diperlukan perhitungan $P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}|\boldsymbol{y_j})$. Setiap variabel diasumsikan saling bebas untuk kelas $\boldsymbol{y}$.

$$P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}|\boldsymbol{y_j}) = P(\boldsymbol{x_1}|\boldsymbol{y_j}).P(\boldsymbol{x_2}|\boldsymbol{y_j}) \dots P(\boldsymbol{x_p}|\boldsymbol{y_j}) \quad (10)$$

Jika terdapat variabel yang bersifat kuantitatif atau kontinu, maka $P(\boldsymbol{x_k}|\boldsymbol{y_j})$ dihitung menggunakan pendekatan distribusi normal.

$$P(\boldsymbol{x_k}|\boldsymbol{y_j}) = \frac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(-\frac{(\boldsymbol{x_k} - \mu_{kj})^2}{2{\sigma_{kj}}^2}\right) \quad (11)$$

Estimasi peluang $P(\boldsymbol{x_k}|\boldsymbol{y_j})$ dapat dihitung untuk setiap variabel $\boldsymbol{x_k}$ dan kelas $\boldsymbol{y_j}$ sehingga data baru akan dapat diklasifikasikan ke dalam kelas $\boldsymbol{y_j}$ apabila peluangnya lebih besar dibandingkan yang lainnya. Pada *Naïve Bayes* digunakan *Hypothesis Maximum A Posterior* (HMAP) untuk memaksimalkan nilai probabilitas dari masing-masing kelas dengan rumus sebagai berikut [9].

$$H_{MAP} = \arg max \frac{P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}|\boldsymbol{y_j}).P(\boldsymbol{y_j})}{P(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p})} \quad (12)$$

D. *K-Nearest Neighbor*

KNN merupakan salah satu pendekatan yang sederhana untuk diimplementasikan dan merupakan metode lama yang digunakan dalam pengklasifikasian. Menurut Y. Hamamoto, dkk dan E.Alpaydin pada tahun 1997 menyebutkan bahwa KNN memiliki tingkat efisiensi yang tinggi dan dalam beberapa kasus memberikan tingkat akurasi yang tinggi dalam hal pengklasifikasian.

Dalam istilah lain, K-*Nearest Neighbor* merupakan salahsatu metode yang digunakan dalam pengklasifikasian. Prinsip kerja K-*Nearest Neighbor* (KNN) adalah melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain [10]. Dekat atau jauhnya lokasi (jarak) bisa dihitung melalui salah satu dari besaran jarak yang telah ditentukan yakni jarak *Euclidean*, jarak *Minkowski*, dan jarak *Mahalanobis*. Konsep jarak *Minkowski* ini memperlakukan semua peubah adalah bebas (tidak berkorelasi). Transformasi baku yang dilakukan berarti menghilangkan pengaruh keragaman data atau dengan kata lain semua peubah akan memberikan kontribusi yang sama untuk jarak. Rumus jarak *Minkowski* adalah sebagai berikut.

$$d(x_i, x_j) = \left(\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r} \quad (13)$$

dengan:

x_{ik} : data *testing* ke-*i* pada variabel ke-*k*

x_{jk} : data *training* ke-*j* pada variabel ke-*k*

$d(x_i, x_j)$: jarak

k : dimensi data variabel bebas

Jika nilai r pada jarak *Minkowski* adalah 1 maka jarak ini sama dengan jarak *Manhattan*. Jika nilai r adalah 2 maka jarak Minkowski sama dengan jarak *Euclidean.*

E. SVM

Support Vector Machine merupakan sistem pembelajaran yang menggunakan hipotesis fungsi linear dalam ruang berdimensi tinggi dan dilatih dengan algoritma berdasarkan teori optimasi dengan menerapkan *learning* bias yang berasal dari teori statistik. [11]. Tujuan utama dari metode ini adalah untuk membangun OSH (Optimal Separating Hyperplane), yang membuat fungsi pemisahan optimum yang dapat digunakan untuk klasifikasi.

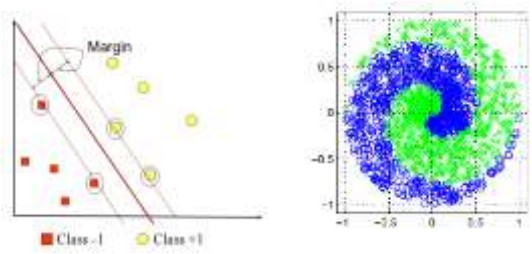
Salah satu contoh kasus dasa dalam klasifikasi adalah data yang bisa dipisah secara linier oleh suatu garis lurus. Persamaan *hyperplane* untuk kasus ini adalah:

$$\mathbf{WX} + b = 0 \quad (14)$$

Dengan $\mathbf{W} = \{w_1, w_1, \ldots, w_p\}$ adalah vektor pembobot, p adalah banyak variabel X, dan b adalah suatu konstanta atau biasa disebut dengan bias. Saat data dapat dipisahkan dengan *hyperplane* yang linier, maka fungsi 14 dapat berubah menjadi

$$f(x) = w^T x + b \quad (15)$$

Jika $f(x) \geq 0$ untuk $y_i = +1$ dan jika $f(x) < 0$ untuk $y_i = -1$ [12]. Untuk kasus data yang tidak bisa dipisahkan secara linier (*non-linierly separable data)*, pencarian *hyperplane* yang optimal akan memperhatikan data-data yang tidak berada dalam kelasnya, yang dikembangkan dengan ξ. Berikut ini adalah gambar kasus yang dapat dipisahkan secara linier dan *non-linierly separable.*



(a) (b)

Gambar 1 Kasus *Linierly Separable* dan *Linierly Non-separable*

Pada kasus nyata, sangat jarang dijumpai kasus data yag dapat terpisah secara linier, olehkarena itu digunakan fungsi kernel untuk memetakan data ke dalam ruang vector yang berdimensi tinggi. Berikut ini adalah ilustrasi untuk data non-linier dan hasil transformasinya.

Beberapa fungsi kernel yang umum digunakan adalah sebagai berikut.

Tabel 1 Jenis-Jenis Kernel

Jenis Kernel	Fungsi Kernel
Kernel Linier	$K(\mathbf{x_i}, \mathbf{x}) = \mathbf{x_i^T x}$
Kernel Radial Basis Function	$K(\mathbf{x_i x}) = \exp\left(-\gamma \left\| \mathbf{x} - \mathbf{x_i} \right\|^2\right)$
Sigmoid Kernel	$K(\mathbf{x_i}, \mathbf{x}) = \tanh\left(\gamma \mathbf{x_i^T x} + r\right)$

III. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diambil dari *Uci Machine Learning.* Data yang digunakan yaitu ILDP (Indian Liver Patient Dataset). Terdapat 583 observasi dan 11 variabel yang digunakan. Variabel yang digunakan dalam penelitian ini antara lain.

Tabel 2 Variabel Penelitian

No	Variabel	Keterangan	Jenis Data
1	Age	Umur Pasien	Numerik
2	Gender	Jenis Kelamin	Kategori
3	TB	Total Bilirubin	Numerik
4	DB	Direct Bilirubin	Numerik
5	Alkaline	Alkphos Alkaline Phosphotase	Numerik
6	Alamine	Sgpt Alamine Aminotransferase	Numerik
7	Aspartate	Sgot Aspartate Aminotransferase	Numerik
8	TP	Total Protiens	Numerik
9	ALB	Albumin	Numerik
10	A/G	Rasio Albumin and Globulin	Numerik
11	Class	Menderita liver/ tidak menderita liver	Kategorik

Di dalam analisis ini akan dilakukan klasifikasi pada *Indian Liver Patient Dataset.*

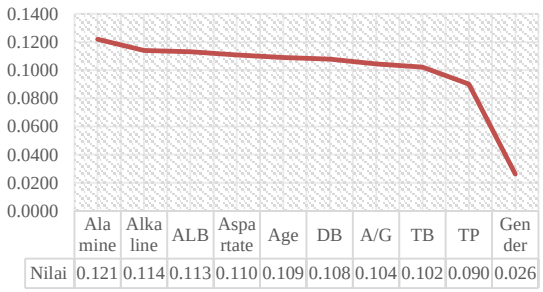
Adapun langkah-langkah analisis yang dilakukan adalah

1. Melakukan *preprocessing* pada data
2. Membagi data *training* dan *testing* (80% sebagai *training* dan 20% sebagai *testing*)
3. Melakukan klasifikasi dengan regresi logistic, *decision tree*, *naïve bayes*, *k-nearest neighbor dan support vecor machine*
4. Memilih metode tebaik

IV. HASIL DAN PEMBAHASAN

A. Preprossesing Dan Eksplorasi Data

Pada bagian ini seluruh variabel yang digunakan akan dilakukan *preprossesing*. Pertama dilakukan pengecekan apakah terdapat kasus *missing value*. Setelah dilakukan pengecekan didapatkan bahwa terdapat 4 observasi yang mengalami kasus *missing value* pada variabel *Ratio Albumin and Globuline* (A/G). Masalah ini selanjutnya akan diatasi dengan mengimputasikan nilai median ke dalam 4 observasi tersebut karena variabel ini mempunyai skala numerik. Pengimputasian dengan median dipilih juga karena nilai median lebih *robust* dibandingkan dengan nilai *mean*. Selanjutnya dilakukan *feature selection* dengan untuk mengetahui variabel apa saja yang mempengaruhi seseorang menderita penyakit liver. Nilai dari variabel *important* dapat dilihat pada Gambar 2.



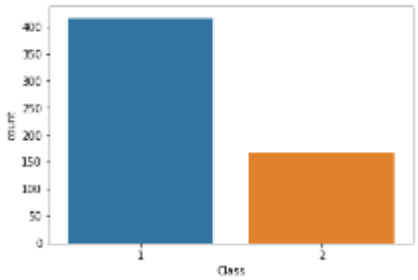
Gambar 2. *Important Variable*

Berdasarkan Gambar 2 dapat dilihat bahwa nilai variabel *important* untuk *Gender* turun tajam. Hal ini menunjukan bahwa variabel *Gender* tidak mempunyai pengaruh yang besar terhadap penentuan seseorang menderita penyakit liver atau tidak, sehingga variabel ini tidak diikutkan dalam analisis lebih lanjut. Setelah diketahui variabel apa saja yang akan digunakan untuk analisis maka selanjutnya dilakukan eksplorasi data. Langkah pertama dalam ekplorasi data pada penelitian ini adalah dengan melihat hasil statistika deskriptif dari variabel *independent*.

Tabel 3. Statistika Deskriptif Variabel Independen

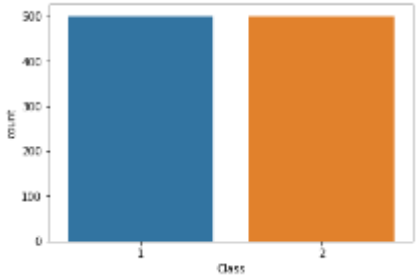
Variabel	Mean	Std	Min	Max
Age	44.746	16.190	4.00	90.00
TB	3.299	6.210	0.40	75.00
DB	1.486	2.808	0.10	19.70
Alkaline	290.576	242.938	63.00	2110.00
Alamine	80.714	182.620	10.00	2000.00
Aspartate	109.911	288.919	10.00	4929.00
TP	6.483	1.085	2.70	9.60
ALB	3.142	0.796	0.90	5.50
A/G	0.947	0.318	0.30	2.80

Selanjutnya dilakukan eksplorasi variabel *dependent* pada data yang hasilnya dapat dilihat pada Gambar 3.



Gambar 3. Jumlah Anggota pada Setiap Kategori

Gambar 3 menujukan bawa terdapat kasus *imbalance* pada kasus ini. Hal ini ditunjukan bahwa jumlah data kategori 1 (menderita penyakit liver) hampir tiga kali lipat jumlah data kategori 2 (tidak menderita penyakit liver). *Imbalance* data dapat menyebabkan kasus dimana hasil klasifikasi mempunyai tingkat akurasi yang tinggi namun hanya satu kategori yang terklasifikasi secara tepat sehingga ini mencerminkan bahwa model klasifikasi yang didapat tidakalah bagus. Karena itu, sebelum data dianalisis akan dilakukan resampel untuk menyeimbangkan data yang hasilnya dapat dilihat pada Gambar 4.

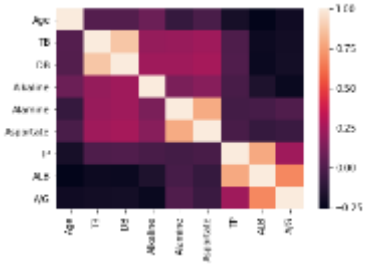


Gambar 4. Hasil *Resample*

Selanjutnya data hasil resampel dipisah menjadi dua yaitu data *training* dan *testing*. Pembagian ini dilakukan secara acak dengan ketentuan 80% data digunakan sebagai data *training* dan 20% data sebagai data *testing*. Sehingga didapatkan 801 observasi merupaka data *training* yang terdiri atas 400 observasi penderita liver dan 401 observasi bukan penderita liver dan 201 observasi sebagai data *testing*.

B. Analisis Menggunakan Regresi Logistik

Metode klasifikasi yang pertama digunakan adalah metode Regresi Logistik Biner. Terdapat syarat yang harus dipenuhi jika menggunakan metode ini yaitu tidak ada kasus multikolinieritas pada data yang digunakan. Gambar 5 menunjukan variabel independen yang mempunyai korelasi tinggi dengan variabel lainnya.



Gambar 5. Nilai Korelasi Antar Variabel Independen

Gambar 5 menunjukan nilai korelasi dari setiap variabel independen yang digunakan setelah data dilakukan *feature selection*. Semakin terang warna pada gambar maka korelasi antar variabel tersebut semakin tinggi. Sehingga berdasarkan hasil pada Gambar 5 dapat disimpulkan bahwa terdapat kasus multikolinieritas pada data. Maka dari itu dilakukan analisis regresi logistik dengan metode *backward* hingga didapatkan data yang tidak terjadi kasus multikolinieritas dan seluruh variabel independennya signifikan berpengaruh terhadap penentuan seseorang menderita penyakit liver atau tidak. Setelah asumsi terpenuhi maka didapatkan model dengan nilai koefisien sebagai berikut

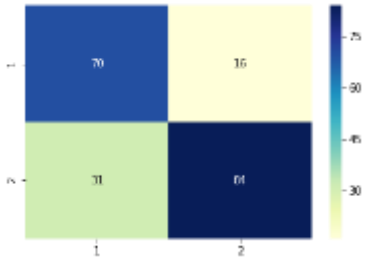
Tabel 4. Hasil Uji Parsial

Variabel	Koifisien	Statistik Hitung	P-*Value*
konstan	2.473	8.327	0.000
Age	-0.021	-4.126	0.000
DB	-0.600	-4.764	0.000
Alkaline	-0.002	-2.811	0.005
Alamine	-0.014	-5.126	0.000

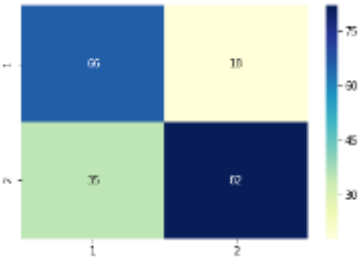
sehingga didapatkan model sebagai berikut

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = 8.327 - 4.126Age - 4.764DB - 2.811Alkaline - 5.126Alamine$$

Hasil *confusion matrix* dari seluruh data dan data yang telah direduksi dengan metode backward dapat dilihat pada Gambar 6 dan Gambar 7



Gambar 6. *Confusion* Matrik Data Sebelum Direduksi



Gambar 7. *Confusion* Matrik Data Setelah Direduksi

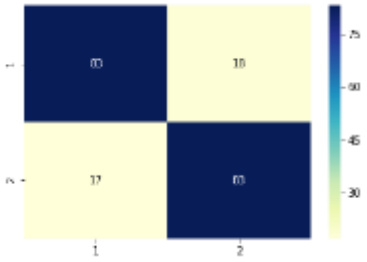
Selanjutnya akan dilihat peforma metode logistik biner terhadap dengan data *testing* yang hasilnya dapat dilihat pada pada Tabel 5.

Tabel 5. Kebaikan Model Regresi Logistik

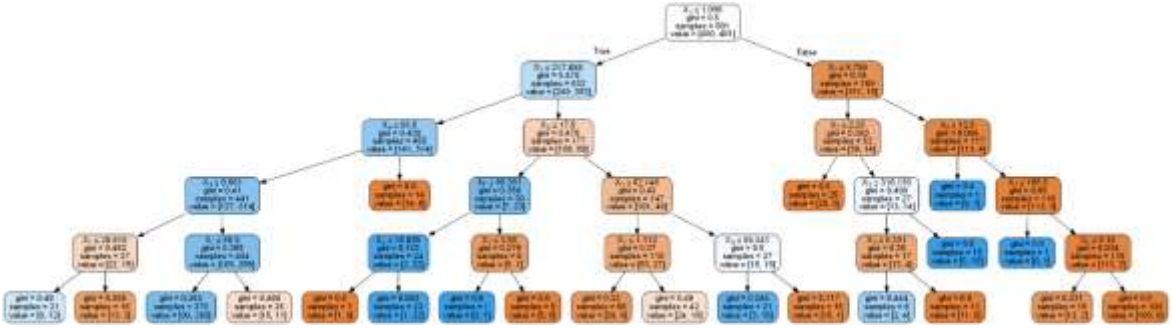
	Semua Variabel	Age, DB, Alkaline, Alamine
Akurasi	76.620%	73.630%
Presisi	63.310%	65.350%
Recall	81.390%	78.570%

C. Analisis Menggunakan *Decision Tree*

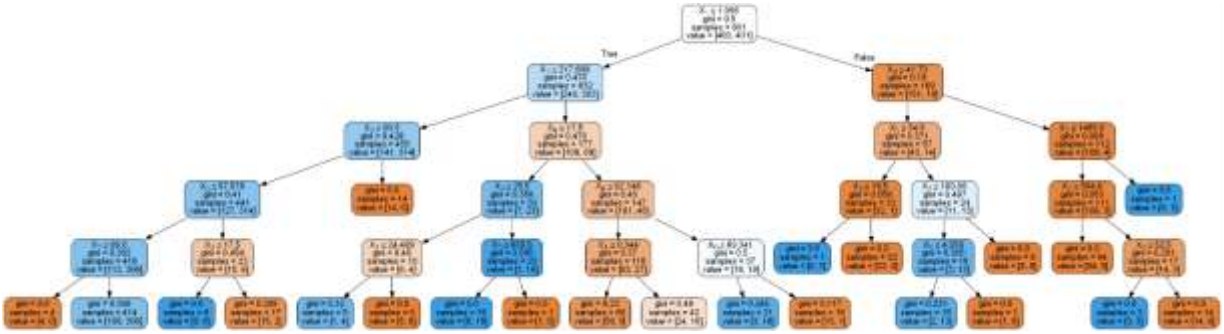
Metode klasifikasi kedua yang digunakan adalah *Decicion Tree*. Berbeda dengan metode Regresi Logistik, metode ini tidak membutuhkan asumsi yang harus dipenuhi sehinga data hasil dari *preprossesing* dapat langsung dimodelkan dengan metode *Decision Tree*. Namun karena penelitian ini bertujuann untuk membandingakan metode klasifikasi maka akan dianalisis pula metode *Decicion Tree* yang hanya menggunkan variabel *Age*, DB, Alkaline dan Alamine. Hasil pohon klasifikasi untuk semua variabel dapat dilihat pada Gambar 8. Berdasarkan Gambar 8 diketahui bahwa *output Decicion Tree* disimbolkan dengan X_0-X_8. Ini berpasangan dengan urutan variabel yan dimasukan yaitu X_0 (*Age*), X_1 (TB), X_2 (DB), X_3 (Alkaline), X_4 (Alamine), X_5 (Aspartate), X_6 (TP), X_7 (ALB) dan X_8 (A/G). Dari 801 observasi yang dimodelkan jika nilai DB kurang dari atau sama dengan 1.096 maka pengklasifikasian harus melihat variabel Alkaline dan jika sebaliknya maka harus melihat varibel TP. Untuk kasus pertama jika nilai Alkaline kurang dari sama dengan 217.688 maka harus dilihat variabel Alamine dan jika sebaliknya maka harus dilihat variabel Age. Ketika pada kasus Alkaline kurang dari sama dengan 217.688 dan dilihat Alamine bernilai kurang dari sama dengan 93.5 maka harus dilihat variabel umur dan jika sebaliknya nilai Alamine lebih dari 93.5 maka observasi tersebut akan terklasifikasi sebagai penderita liver, begitu seterusnya hingga seluruh observasi terklasifikasi ke dalam dua kategori yang ada. Model ini kemudian diterapkan padaa data *testing*. Dari hasil klasifikasi data *testing* didapatkan *confusion matrix* yang dapat dilihat pada Gambar 9.



Gambar 9. *Confusion* Matrik Metode *Decision Tree* Semua Variabel



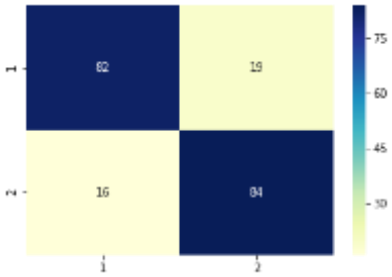
Gambar 8. Pohon Klasifikasi Semua Variabel



Gambar 10. Pohon Klasifikasi Variabel Age, DB, Alkaline dan Alamine

Sedangkan hasil pohon klasifikasi hanya dengan menggunkan variabel *Age*, DB, Alkaline dan Alamine dapat dilihat pada Gambar 10.

Cara membaca pohon klasifikasi pada Gambar 10 sama dengan pada Gambar 9. Yang berbeda adalah kode Variabelnya pada Gambar 10 penyimbolan output adalah X_0 (*Age*), X_1 (DB), X_2 (Alkaline) dan X_3 (Alamine). Hasil *confusion matrix* data *testing* untuk model *decicion tree* yang kedua dapat dilihat pada Gambar 11.



Gambar 11. *Confusion* Matrik Metode *Decision Tree* dengan 4 Variabel

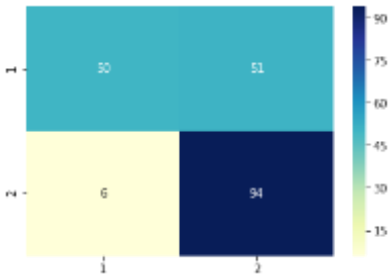
Berdasarkan hasil *confusion matrix* maka akan didapatkan nilai akurasi, presisi dan *recall* yang hasilnya dapat dilihat pada Tabel 6. Sebelumnya bila dilihat bahwa *confusion matrix* yang dihasilkan *balance* yang tercermin pada nilai presisi dan *recall* yang didapat.

Tabel 6. Kebaikan Model *Decision Tree*

	Semua Variabel	Age, DB, Alkaline, Alamine
Akurasi	82.587%	82.587%
Presisi	83.000%	83.673%
Recall	82.178%	81.188%

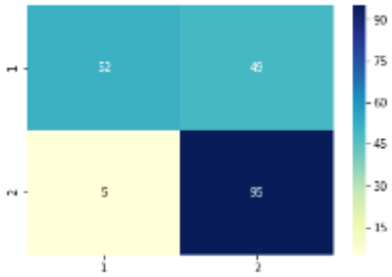
D. Analisis Menggunakan *Naive Bayes*

Metode selanjutnya adalah *Naive Bayes*. Seperti dengan metode *Decision Tree*, metode ini juga tidak memerlukan asumsi khusus. Hasil *confusion matrix* pada data *testing* menggunakan seluruh variabel dapat dilihat pada Gambar 12.



Gambar 12. *Confusion* Matrik *Naïve Bayes* Semua Variabel

Sedangkan hasil *cofusion matrix* yang hanya menggunakan variabel *Age*, DB, Alkaline dan Alamine dapat dilihat pada Gambar 13.



Gambar 13. *Confusion* Matrik *Naïve Bayes* Variabel Age, DB, Alkaline, dan Alamine

Berdasarkan hasil *confusion matrix* diperoleh nilai akurasi, presisi dan *recall* yang hasilnya dapat dilihat pada Tabel 7.

Tabel 7. Kebaikan Model *Naïve Bayes*

Metode	Semua Variabel	Age, DB, Alkaline, Alamine
Akurasi	71.642%	73.134%
Presisi	89.286%	91.228%
Recall	49.505%	51.485%

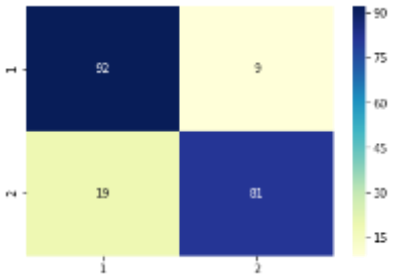
E. Analisis Menggunakan *K-Nearest Neighbor*

Metode keempat pada analisis ini adalah *K-Nearest Neighbor*. Konsep dari metode ini adalah mengklasifikasikan observasi berdasarkan jarak terdekat. Pada penelitian ini ukuran jarak yang digunakan adalah *Minkowski*. Sedangkan untuk banyak *neighbors* yang akan digunakan karena tidak diketahui berapa jumlah yang paling optimal maka akan dianalisis peforma metode ini dengan jarak 1 s/d 5. Hasil peforma metode ini pada data *testing* baik menggunkan semua variabel maupun variabel *Age*, DB, Alkaline dan Alamine saja dapat dilihat pada Tabel 8.

Tabel 8. Kebaikan Model *K-Nearest Neighbor*

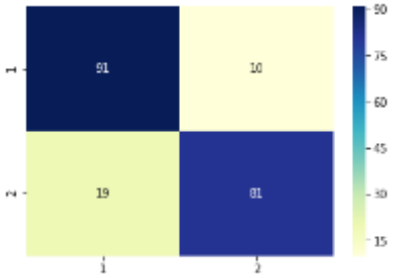
Metode	Neighbor				
	1	2	3	4	5
Semua Variabel					
Akurasi	**86.070%**	78.109%	81.592%	78.109%	79.602%
Presisi	**82.883%**	70.803%	78.070%	72.441%	77.778%
Recall	91.089%	**96.040%**	88.119%	91.089%	83.168%
Age, DB, Alkaline, Alamine					
Akurasi	**85.572%**	77.612%	78.109%	75.124%	78.109%
Presisi	**82.727%**	70.896%	77.143%	71.429%	78.218%
Recall	90.099%	**94.059%**	80.198%	84.158%	78.218%

Berdasarkan Tabel 8 akurasi dan presisi terbaik didapatkan ketika jumlah *neighbor*-nya satu sedangkan recall terbaik didapatkan ketika *neigbor*-nya dua. Namun ketika *neighbor*-nya dua nilai akursi dan presisinya terkecil dibanding dengan nilai akurasi dan presisi jika menggunakan jumah *neighbor* lainnya. Sehingga neigbor terbaik pada metode ini adalah 1. Selanjutnya akan dilihat hasil *confusion matrix* dari data *testing* menggunakan metode KNN untuk semua variabel yang dapat dilihat pada Gambar 14.



Gambar 14. *Confusion* Matrik KNN Semua Variabel

Sedangkan hasil confusion matrix untuk variabel *Age*, DB, Alkaline dan Alamine dapat dilihat pada Gambar 15.



Gambar 15. *Confusion* Matrik KNN Age, DB, Alkaline dan Alamine

Berdasarkan hasil *confusion matrik* metode K-*Nearest Neighbor* diperoleh bahwa data dapat terklasifikasi ke tiap kategori. Sehingga dapat disimpulkan bahwa hasil klasifikasi ini baik digunakan.

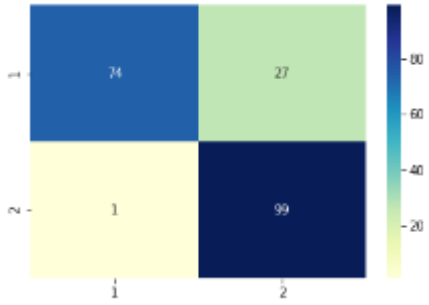
F. Analisis Menggunakan *Support Vector Machine (SVM)*

Metode terkahir yang digunakan dalam peneilitian ini adalah *Support Vector Machine* (SVM). *Support Vector Machine* (SVM) merupakan salah satu metode klasifikasi non-linier. Pada penelitian ini akan digunakan tiga fungsi kernel yang berbeda yaitu *linear*, *sigmoid* dan *radial basis function* (RBF) dengan parameter yang digunakan adalah *cost* bernilai 1, *degree* bernilai 3 dan *gamma* merupakan *default software* yang digunakan. Hasil akurasi, presisi dan *recall* model SVM dengan ketiga jenis kernel pada data *testing* dapat dilihat pada Tabel 9.

Tabel 9. Kebaikan Model SVM Semua Variabel

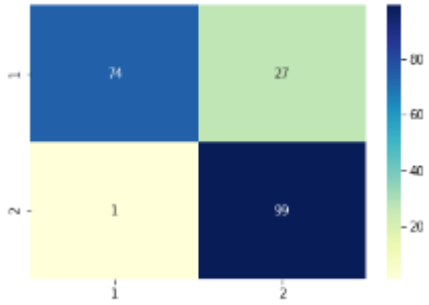
Metode	Kernel		
	Linear	*Sigmoid*	*Radial Basis Function*
Semua Variabel			
Akurasi	76.617%	49.751%	**86.070%**
Presisi	88.571%	0.000%	**98.667%**
Recall	61.386%	0.000%	**73.267%**
Age, DB, Alkaline, Alamine			
Akurasi	74.627%	49.751%	**86.070%**
Presisi	84.722%	0.000%	**98.667%**
Recall	60.396%	0.000%	**73.267%**

Berdasarkan Tabel 9, SVM dengan fungsi kernel *radial basis function* menghasilkan nilai akurasi, presisi dann *recall* terbesar. Terlihat bahwa akurasi, presisi dan *recall* yang didapatkan dengan menggunakan semua variabel dan hannya variabel *Age*, DB, Alkaline dan Alamine adalah sama. Dapat dilihat pula bahwa terdapat keanehan pada hasil presisi dan *recall* jika menggunakan kernel *sigmoid*. Hal ini disebabkan karena fungsi kernel *sigmoid* hanya dapat mengklasifikasikan observasi pada data *testing* ke satu kategori. Pada kasus ini observasi pada data *testing* hanya terklasifikasi pada kategori 2 yaitu bukan penderita liver. Berbeda dengan menggunakan fungsi kernel *sigmoid*, jika menggunakan fungsi kernel *radial basis* maka observasi pada data *testing* dapat diklasifikasikan baik ke kategori 1 (menderita penyakit liver) dan kategori 2 (bukan penderita liver). Sehingga didapatkan *confusion matrix* yang dapat dilihat pada Gambar 16 untuk semua variabel.



Gambar 16. *Confusion* Matrik SVM Semua Variabel

Sedangkan hasil *confusion matrix* untuk klasifikasi menggunakan SVM hanya menggunakan variabel *Age*, DB, Alkaline dan Alamine dapat dilihat pada Gambar 17.



Gambar 17. *Confusion* Matrik SVM Age, DB, Alkaline dan Alamine

Berdasarkan Gambar 16 dan Gambar 17 metode SVM dengan fungsi kernel *radial basis* dengan menggunkan semua variabel dan hanya variabel *Age*, DB, Alkaline dan Alamine menghasilkan *confusion matrix* yang sama dalam yang mengklasifikasikan observasi ke tiap kategori. Sehingga hasil klasifikasi ini juga baik digunakan..

G. Metode Terbaik

Setelah dilakukan klasifikasi menggunakan metode Regresi Logistik, *Decision Tree*, *Naive Bayes*, K-NN dan SVM baik menggunakan semua variabel hasil *preprossesing* maupun hanya menggunakan variabel *Age*, DB, Alkaline dan Alamine dan didapatkan performa setiap model yang digunakan, maka langkah selanjutnya adalah membandingkan performa yang telah didapatkan. Perbandingan performa hasil analisis dapat dilihat pada Tabel 10.

Tabel 10. Perbandingan Semua Metode

Variabel pada Model	Metode	Akurasi	Presisi	*Recall*
Semua Variabel	Regresi Logistik	76.620%	69.310%	81.390%
	Decision Tree	82.587%	83.000%	82.178%
	Naive Bayes	71.642%	89.286%	49.505%
	K-NN	**86.070%**	82.883%	**91.089%**
	SVM	**86.070%**	**98.667%**	73.267%
Age, DB, Alkaline, Alamine	Regresi Logistik	73.630%	65.350%	78.570%
	Decision Tree	82.587%	83.673%	81.188%
	Naive Bayes	73.134%	91.228%	51.485%
	K-NN	**85.572%**	82.727%	**90.099%**
	SVM	**86.070%**	**98.667%**	73.267%

Berdasarkan Tabel 10 dapat dilihat bahwa hasil klasifikasi menggunakan semua variabel atau hanya menggunakan variabel Age, DB, Alkaline dan Alamine memilki nilai akurasi, presisi dan recall yang tidak berbeda jauh, sehingga kelima variabel yang tidak digunakan tidak memberikan pengaruh yang signifikan dalam penentuan klasifikasi apakah seseorang menderita penyakit liver atau tidak. Jika hanya digunakan variabel Age, DB, Alkaline dan Alamine, maka metode yang memberikan nilai akurasi dan presisi terbaik adalah metode SVM namun jika menggunakan metode ini maka akan terjadi ketimpangan yang sangat besar antara nilai presisi dan recall yang menunjukan bahwa metode ini lebih cenderung mengklasifikan hasil klasifikasi ke salah satu kategori. Jika menggunakan recall, maka metode K-Nearest Neighbor memberikan hasil terbaik.

V. KESIMPULAN DAN SARAN

Pada bab 4 telah dijelaskan tentang hasil dari analisis yang telah dilakukan. Berdasarkan pemaparan pada bab 4 tersebut, dapat ditarik kesimpulan bahwa berdasarkan nilai akurasi dan presisi, maka metode SVM memberikan hasil yang terbaik, tapi berdasarkan recall maka metode K-Nearest Neighbor memberikan hasil terbaik. Walaupun SVM memberikan hasil nilai akurasi dan presisi tertinggi tetapi terdapat ketimpangan yang besar antara nilai presisi dan recall yang dihasilkan, jika dibandingkan selisih nilai akurasi dan recall dari metode K-Nearest Neighbor.

DAFTAR PUSTAKA

[1] R.-H. Lin, "An Intelligent Model for Liver Disease Diagnosis," *Artificial Intelligence in Medicine*, 2009.

[2] I. Rish, "An Empirical Study of The Naïve Bayes Classifier.," in *International Joint Confeence on Artificial Inteliggence*, 2006.

[3] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd ed, USA: Morgan Kaufmann, 2012.

[4] B. V. Ramana, " A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *International Journal of Database Management System*, vol. 3, 2011.

[5] S. Kafelegn and P. Kamat, "Prediction and Analysis of Liver Disorder Disease by using Data Mining Technique: Survey," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 9, pp. 765-770, 2018.

[6] Hosmer and Lemeshow, Applied Logistic Regression, USA: John Wiley & Sons, 2000.

[7] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd ed., USA: Morgan Kaufman, 2012. .

[8] X. Wu and V. Kumar, The Top Ten Algoritms in Data Mining., New York: CRC Press, 2009.

[9] F. Gorunescu, Data Mining Concepts, Models and Technique., Jerman: Springer, 2011.

[10] E. Prasetyo, Data Mining Konsep dan Aplikasi Menggunakan MATLAB., Yogyakarta: ANDI Yogyakarta., 2012.

[11] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machine., Cambridge: Cambridge University Press, 2000.

[12] S. Ramkishore, P. Madhumitha and P. Palanichamy, "Comparison of Logistic Regression and Support Vector Machine for The Classification of Microstructure and Interfacial Defects in Zircaloy-1," in *International Conference on Soft Computing and Machine Intelligence*, 2014.