PELITA TEKNOLOGI
Journal homepage: jurnal.pelitabangsa.ac.id, p-ISSN: 2301-475X, e-ISSN: 2656-7059

Perbandingan Dalam Memprediksi Penyakit Liver Menggunakan Algoritma *Naïve Bayes* Dan *K-Nearest Neighbor*

Muhammad Fatchan[1], Ir. Nanang Tedi K., MT[2] Alfiyan[3], Kurniawan[4], Edy Widodo[5]

Program Studi Teknik Informatika, Universitas Pelita Bangsa
Jl. Inspeksi Kalimalang Tegal Danas Arah DELTAMAS, Cikarang Pusat-Kab. Bekasi, Indonesia

fatchan@pelitabangsa.ac.id

Abstrak

Along with the rapid development of information technology, and also the increasing need for information in various fields including health sector. Based on data from the World Health Organization (WHO), chronic hepatitis B attacks 300 million people in the world including Southeast Asia and Africa which causes the death of more than 1 million people each year. So far, a lot of data in the hospital has not been used, even though this data can be used to predict liver disease if used. The purpose of this study was to determine the comparison of the accuracy value of the Naïve Bayes algorithm and K-Nearest Neighbor. One of the classifications is to use the Naïve Bayes and K-Nearest Neighbor algorithms and use the Rapid Miner tools in the tests used. The results of this study indicate that the Naïve Bayes algorithm has a higher accuracy rate of 84.00% in diagnosing liver disease compared to the K-Nearest Neighbor algorithm which only gets a value of 80.57%. From this research it can be concluded that the Naïve Bayes algorithm is 3.43% greater than K-Nearest Neighbor.

Informasi Artikel

Diterima: 20-02-2021
Direvisi: 04-03-2021
Dipublikasikan: 28-04-2021

Keywords

Data Mining, *Naïve Bayes*, *K-Nearest Neighbor*, Liver, RapidMiner

I. Pendahuluan

Seiring dengan perkembangan Tekonologi Informasi yang semakin pesat, dan juga kebutuhan informasi yang semakin meningkat pada berbagai bidang termasuk bidang kesehehatan, namun dalam suatu informasi sangat dibutuhkan tingkat keakuratannya dalam sebuah informasi tersebut. Untuk mendapatkan informasi yang tepat kita dapat melakukan pengolahan data dalam jumlah yang besar untuk mendapat sebuah pengetahuan yang baru atau yang biasa disebut dengan data mining. Data mining adalah salah satu teknik didalam ilmu komputer yang melibatkan beberapa proses seperti komputasi, teknik statistik, clustering, klasifikasi dan menemukan pola yang terdapat di dalam dataset tersebut. Tujuan utama dari data mining tersebut adalah mengekstrak dari dataset yang besar dan diubah menjadi format yang dapat dimengerti serta mudah dipahami [1].

Data Mining juga dapat digunakan untuk memprediksi suatu penyakit seperti penyakit Liver. Penyakit liver merupakan penyakit yang menyerang pada bagian peradangan hati yang disebabkan oleh infeksi virus, bakteri atau bahan yang beracun lainnya sehingga menyebabkan hati tidak dapat berfungsi dengan baik. Berdasarkan data World Health Organization (WHO), hepatitis B kronis menyerang 300 juta orang didunia termasuk Asia Tenggara dan Afrika yang menyebabkan kematian 1 juta orang lebih setiap tahunnya. Dari jumlah itu 15 sampai 25% yang terinfeksi kronis meninggal dunia baik dari penyakit komplikasi maupun penyakit liver [2].

Selama ini banyak data Institusi terkait medis yang datanya belum digunakan secara maksimal, padahal jika data itu dapat dimaksimalkan maka dapat digunakan untuk memprediksi suatu penyakit. Berkembangnya teknologi informasi, tuntutan untuk akan pengetahuan berbasis komputer sangat dibutuhkan untuk teknik analisis dalam mendiagnosa suatu penyakit menjadi semakin penting dalam menggantikan 2 analisis konvensional yang masih menggunakan cara manual [3].

II. Metodologi

2.1 Data Mining

Data mining adalah proses yang meggunakan tehnik statistika, matematika, machine learning dan kecerdasan buatan untuk mengekstrak dan mengidentifikasi informasi sehingga mendapatkan pengetahuan yang berkaitan dengan berbagai database [4]. Data mining merupakan cara untuk mendapatkan pengetahun yang terbaru dengan cara menggunakan jumlah data yang besar. Data mining merupakan cara untuk menggali nilai tambah dari suatu kumpulan data yang berupa pengetahuan yang selama ini tidak diketahui secara manual. Salah satu Teknik yang digunakan dalam data mining adalah classification [5]. Beberapa Teknik ini telah ditingkatkan dan implementasikan untuk mengekstrak pengetahun yang berguna untuk mengambil keputusan. Teknik pengekstrakan tersebut diantaranya adalah untuk pengenalan pola, clustering, memperlambat proses pencarian data yang dibutuhkan. Maka penulis akan menghitung dan asosiasi, prediksi dan klasifikasi.

2.2 Naïve Bayes

Naïve Bayes adalah suatu pengklasifikasikan *probabilistic* sederhana yang menghitung dengan menggunakan cara jumlah frekuensi dan kombinasi nilai dari dataset yang telah diberikan [6]. Naïve Bayes merupakan algortima yang dapat digunakan untuk memprediksi keanggotaan dari suatu class berdasarkan teorema bayes yang bekerja seperti decision tree dan neural network. Naïve Bayes adalah pengklasifikasian dengan cara menggunakan metode probablititas dan statistik yang ditemukan oleh ilmuwan Inggris yang bernama Thomas Bayes, yaitu dengan cara memprediksi peluang yang akan datanag berdasarkan pengamalan pada masa senelumnya [7].

Naïve Bayes didasarkan pada asumsi sederhana melalui nilai atribut secara

kondisional yang saling berkaitan berdasarkan nilai output, probabilitas mengamati secara bersama merupakan produk dari probabilitas individu. Keuntungan menggunakan metode Naïve Bayes adalah membutuhkan data training yang kecil dalam menentukan estimasi yang diperlukan dalam prosespengklasifikasian. Naïve Bayes dapat bekerja jauh lebih baik di dalam dunia nyata dan bekerja lebih kompleks [8].

Algoritma naïve bayes adalah salah satu algoritma yang tergolong kedalam *statistical classifier* algoritma tersebut pertama kali diperkenalkan oleh Thomas Bayes dimana algoritma ini merupakan adaptasi dari theorem bayes. Algoritma ini mengandalkan sebuah peluang kemungkinan suatu objek [9].

Persamaan dari theorem bayes adalah:

$$P\ (H \mid X) = \frac{P(X|H).\,P(H)}{P(H)} \quad ..(1)$$

Dimana :

X : Data dengan class yang belum diketahui
H : Hipotesis data menggunakan suatu class spesifik
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X) : Probabilitas H
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (parteriori probabilitas)
P(X|H): Probabilitas X berdasarkan kondisi pada hipotesis H

2.3 K-Nearest Neighbor

k-Nearest Network(k-NN) adalah algoritma klasifikasi yang bekerja berdasarkan k instancesterdekat dengan query instancesyang diberikan, kemudian melakukan pemilihan antara k tetangga terdekat untuk memperoleh keluaran label dari *query instances* [10]. k-NN menyimpan semua instances pada tempat yang sama, dimana n merupakan fitur instances yang telah didefinisikan sebelumnya. Matrik distancesyang dipakai untuk mengukur jarak antara instances [11].

Algoritma K-Nearest Neighbor merupakan metode klasifikasi yang menggunakan data kelompok baru berdasarkan estimasi jarak data baru ke beberapa tetangga terdekat [12]. Algoritma K-Nearest Neighbord (K-NN) merupakan metode dengan cara melaksanakan klasifikasi terhadap obyek yang berdasarkan data pembelajaran yang jarak obyek keduannya saling berdekatan [13]. Berdasarkan 2 penjelasan diatas, dapat ditarik kesimpulan bahwa Algoritma K-Nearest Neighbor merupakan metode klasifikasi yang mengelompokan obyek melalui data yang jaraknya paling dekat dengan obyek tersebut.

Rumus Algoritma K-Nearest Neighbor (K-NN):

$$d_i = \sqrt{\sum_{i=1}^{p}(X_{2i} - X_{1i})^2 \ldots\ldots (1)} \quad \ldots(2)$$

Keterangan :

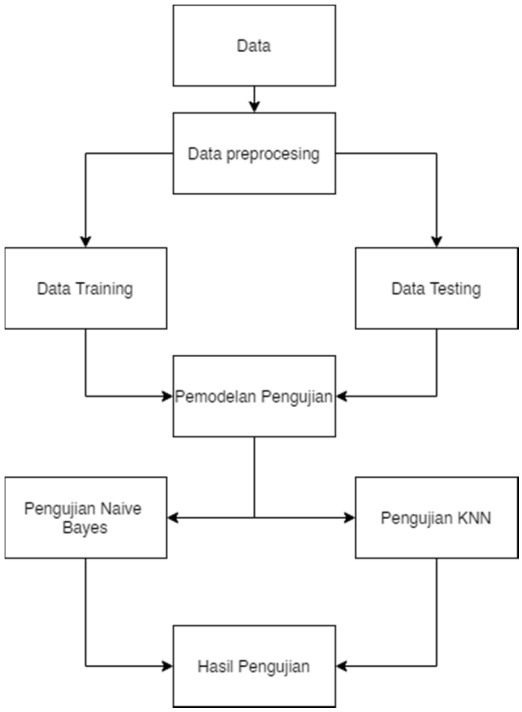
X_1 : Sampel Data
X_2 : Data Uji/Testing
i : Variabel Data
d : Jarak
p : Dimensi Data

Pada penelitian ini akan akan dilakukan tahapan-tahapan penelitian sebagai berikut:

a. Data yang digunakan merupakan dataset pasein penyakit liver yang berasal dari data publik yaitu dari : Kaggle.com dengan jumlah data sebanyak 583 data.
b. Pemilihan data Pada tahap ini dilakukan pemilihan variabel data yang akan dianalisis, dari total 583 data dengan 10 Variabel dan 1 Class.
c. *Preprocessing* Data dilakukan pada pada data ILPD (Indian Liver Patient) yang berasal dari UCI dataset. Atribut *class* sendiri mempunyai 2 nilai yaitu pasein dan non pasein yang direpresentasikan dengan angka 1 sebagai pasein dan angka 0 sebagai non

pasein. Namun dalam dataset tersebut juga masih terkandung beberapa data dengan nilai yang inkonsisten dan missing value, sehingga perlu dilakukannya tahap data *preprocessing.*

d. Split Data digunakan untuk membagi dataset menjadi dua, yaitu untuk data training dan data testing. Pembagian data menggunakan tools split data yang ada pada aplikasi Rapid Miner.
e. Metode Usulan, dalam penelitian ini akan dilakukan analisa menggunakan metode algoritma algoritma *Naïve Bayes* dan K-Nearest Neighbor. Data dihitung dengan menggunakan algoritma Naïve Bayes kemudian dibandingkan metode K-Nearest Neighbor dengan melihat perbandingan tertinggi. Dalam tahapan ini akan dilakukan beberapa langkah pengujian data yaitu seperti berikut :



Gambar 1 Metode usulan

2.4 Confusion Matrix

Confusion Matrix adalah alat *(tools) visualisasi* yang biasa digunakan pada *supervised learning*. Tiap kolom pada matrix adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya. Matriks ini menginformasikan hasil prediksi secara keseluruhan dari nilai akurasi dan untuk melihat kinerja pengklasifikasi, yaitu seberapa sering kasus class X yang benar diklasifikasikan sebagai class X atau kesalahan klasifikasi class yang lainnya [14]. Confusion matrix melakukan pengujian untuk memperkirakan obyek yang benar dan salah. Urutan pengujian ditabulasikan dalam confusion matrix dimana kelas yang diprediksi ditampilkan di bagian atas matriks dan kelas yang diamati di bagian kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi[15].

Accuracy merupakan persentase antara nilai prediksi dengan nilai sebenarnya yang ada. Recall merupakan persentase nilai kinerja keberhasilan algoritma yang dipakai. Precision merupakan nilai akurasi dengan class yang telah diprediksi [13]. Berikut adalah persamaan model confusion matrix:

Confusion Matrix		Nilai Prediksi	
		Positif	Negatif
Nilai Sebenarnya	Positif	(a) TP	(b) FP
	Negatif	(c) FN	(d) TN

Keterangan :

a : jika nilai prediks positif dan kelas sebenarnya positif (True Positif)
b : jika nilai prediks negatif dan kelas sebenarnya positif (False Positif)
c : jika nilai prediks positif dan kelas sebenarnya negative (False Negatif)
d : jika nilai prediks negatif dan kelas sebenarnya negative (True Negatif)

III.Hasil dan Pembahasan

3.1 Pengujian Naïve Bayes

Berdasarkan hasil pengujian Naïve Bayes menggunakan tools RapidMiner 5.3, maka diperoleh *accuracy* sebesar 84%, precision sebesar 88,37%, dan recall sebesar 80,85%. Atau bisa dilihat pada gambar dibawah ini:

accuracy: 84.00%

	true 0	true 1	class precision
pred. 0	71	18	79.78%
pred. 1	10	76	88.37%
class recall	87.65%	80.85%	

Gambar 2 Hasil *Accuracy* Naïve Bayes

Dari **Gambar 2** dapat dilihat bahwa nilai *accuracy* yang didapat dari pengujian algoritma *Naïve Bayes* yaitu sebesar 84.00%.

precision: 88.37% (positive class: 1)

	true 0	true 1	class precision
pred. 0	71	18	79.78%
pred. 1	10	76	88.37%
class recall	87.65%	80.85%	

Gambar 3 Hasil *Precision* Naïve Bayes

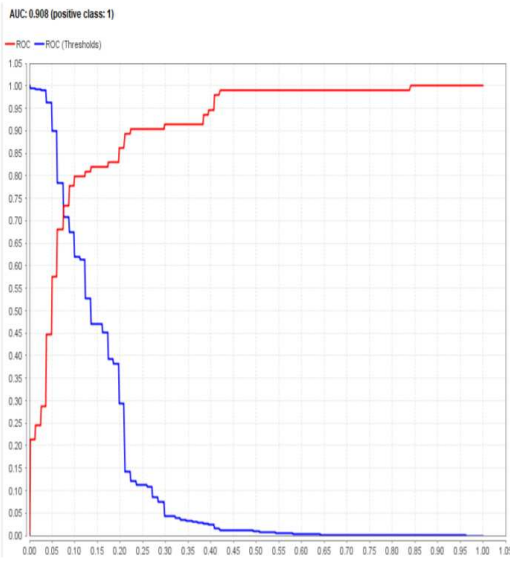
Dari **Gambar 3** dapat dilihat bahwa nilai *precision* yang didapat dari pengujian algoritma *Naïve Bayes* yaitu sebesar 88,37 %.

recall: 80.85% (positive class: 1)

	true 0	true 1	class precision
pred. 0	71	18	79.78%
pred. 1	10	76	88.37%
class recall	87.65%	80.85%	

Gambar 4 Hasil Recall Naïve Bayes

Dari **Gambar 4** dapat dilihat bahwa nilai recall yang didapat dari pengujian algoritma Naïve Bayes yaitu sebesar 80,85 %.



Gambar 5. Hasil AUC Naïve Bayes

Dari **Gambar 5** maka dapat dilihat bahwa nilai AUC dari pengujian algoritma Naïve Bayes yaitu sebesar 0,908 dan termasuk dalam kategori excellent classification karena berada pada range 0,900-1,000 yang artinya menunjukan hasil yang sangat baik dalam akurasi.

3.2 Pengujian K-Nearest Neighbor

Berdasarkan hasil pengujian Naïve Bayes K-Nearest Neighbor menggunakan tools RapidMiner 5.3, maka diperoleh accuracy sebesar 80,57%, precision sebesar 83,33%, dan recall sebesar 79,79%. Atau bisa dilihat pada gambar dibawah ini:

accuracy: 80.57%

	true 0	true 1	class precision
pred. 0	66	19	77.65%
pred. 1	15	75	83.33%
class recall	81.48%	79.79%	

Gambar 6. Hasil Accuracy K-Nearest Neighbor

Dari Gambar 6 dapat dilihat bahwa nilai accuracy yang didapat dari pengujian algoritma K-Nearest Neighbor yaitu sebesar 80.57 %

precision: 83.33% (positive class: 1)

	true 0	true 1	class precision
pred. 0	66	19	77.65%
pred. 1	15	75	83.33%
class recall	81.48%	79.79%	

Gambar 7 Hasil Precision K-Nearest Neighbor

Dari Gambar 7 dapat dilihat bahwa nilai precision yang didapat dari pengujian algoritma K-Nearest Neighbor yaitu sebesar 83.33 %

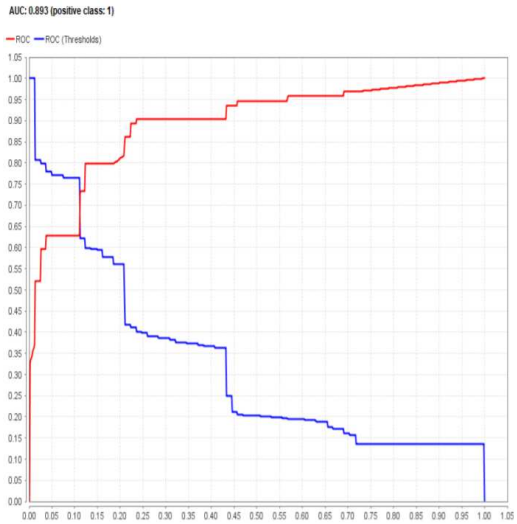
recall: 79.79% (positive class: 1)

	true 0	true 1	class precision
pred. 0	66	19	77.65%
pred. 1	15	75	83.33%
class recall	81.48%	79.79%	

Gambar 8. Hasil Accuracy K-Nearest Neighbor

Dari Gambar 8 dapat dilihat bahwa nilai recall yang didapat dari pengujian

algoritma K-Nearest Neighbor yaitu sebesar 79,79 %.

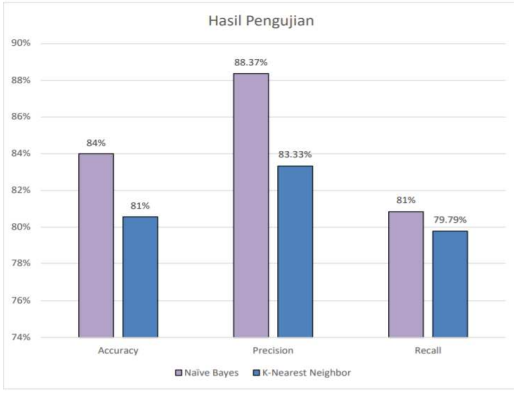


Gambar 9 Hasil AUC K-Nearest Neighbor

Dari Gambar 9 maka dapat dilihat bahwa nilai AUC dari pengujian algoritma K-Nearest Neighbor yaitu sebesar 0,893 dan termasuk dalam kategori *good classification* karena berada pada range 0,800-0,900 yang artinya menunjukan hasil yang baik dalam akurasi.

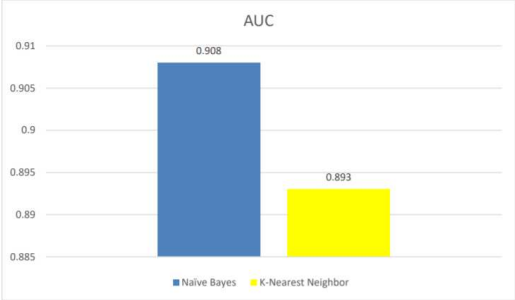
3.3 Pembahasan Hasil pengujian

Penelitian yang dilakukan dengan 2 pengujian, pengujian pertama menggunakan algoritma klasifikasi Naïve Bayes dan pengujian kedua menggunakan algoritma klasifiksi K-Nearest Neighbor menghasilkan nilai accuracy, recall, precision. Berikut tabel hasil dari 2 penelitian yang telah dilakukan dengan menggunakan dataset Indian Liver Patient.



Gambar 10 Grafik Hasil Pengujian

Dari Gambar 10 dapat dilihat bahwa nilai accuracy yang didapat dari pengujian naïve bayes sebesar 84 % sedangkan pada pengujian k-nearest neighbor mendapatkan nilai sebesar 80,57 %. Pada nilai precision hasil yang didapat pada naïve bayes adalah sebesar 88,37 % sedangkan pada pengujian k-nearest neighbor mendapatkan nilai sebesar 83,33 %. Pada nilai recall hasil yang didapat pada naïve bayes adalah sebesar 80,85 % sedangkan pada pengujian k-nearest neighbor mendapatkan nilai sebesar 79,79 %.



Gambar 11 Grafik Hasil AUC

IV. Kesimpulan

Berdasarkan penelitian yang telah dilakukan algoritma naïve bayes memiliki tingkat akurasi yang lebih tinggi yaitu sebesar 84,00% dalam memprediksi penyakit liver dibandingkan dengan algoritma k-nearest neighbor yang hanya mendapatkan nilai sebesar 80,57%. Maka dari dapat disimpulkan bahwa algoritma Naïve Bayes lebih besar 3,43% dari K-Nearest Neighbor dan lebih baik digunakan untuk memprediksi penyakit liver.

Daftar Pustaka

[1]Dimas Anggara, “Algoritma Decision Tree C4.5 Dan K-Nearest Neighbor (K-Nn) Dalam Mendiagnosa Penyakit Liver,” 2018.

[2]M. Neshat, M. Sargolzaei, A. Nadjaran Toosi, And A. Masoumi, “Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning And Particle Swarm Optimization,” Isrn Artif. Intell., Vol. 2012, Pp. 1–6, 2012.

[3] M. Yuli, “Data Mining : Klasifikasi Menggunakan Algoritma C4.5,” J. Edik Inform., Vol. 2, No. 2, Pp. 213–219, 2017.

[4] H. Sujaini, “Analisis Asosiasi Pada Transaksi Obat Menggunakan Data Mining Dengan Algoritma A Priori,” Justin, Vol. 4, No. 2, P. 6, 2016.

[5] Agus Nur Khormarudin, “Teknik Data Mining : Algoritma K-Means Clustering,” Pp. 1–12, 2016.

[6] A. Saleh, “Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga,” Vol. 2, No. 3, Pp. 207–217, 2015.

[7] C. V Sumber And U. Telekomunikasi, “Penerapan Data Mining Dengan Algoritma *Naive Bayes* Clasifier Untuk Mengetahui Minat Beli Pelanggan Terhadap Kartu Internet Xl Studi Kasus Di,” Pp. 81–92, 2016.

[8] D. Hastuti et al., “Algoritma Naïve Baiyes Untuk Prediksi Profesi Berdasarkan Skill Job Seeker,” No. April, 2017

[9] M. Ayu, D. Widyadara, And R. H. Irawan, “Implementasi Metode Naïve Bayes Dalam Penentuan Tingkat Kesejahteraan Keluarga,” Vol. 2, No. 1, Pp. 19–24, 2019.

[10] A. Rane, N. Naik and J. Laxminarayana, “Performance Enhancement of K Nearest Neighbor Classification Algorithm Using 8-Bin Hashing and Feature Weighting,” ACM 978-1-4503-2908-8/14/08,2014.http://wwww.cs.umass.edu/lfw/,Wild, Labeled Faces in the, 2016.

[11] A. Rohman, “Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa,” 2012.

[12] W. Yustanti, “Algoritma K-Nearest Neighbour Untuk Memprediksi Harga Jual Tanah,” Vol. 9, No. 1, Pp. 57–68, 2012.

[13] N. Musyaffa And B. Rifai, “Model Support Vector Machine Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Liver,” Jitk (Jurnal Ilmu Pengetah. Dan Teknol. Komputer), Vol. 3, No. 2, Pp. 189–194, 2018

[14] A. H. Intan Setiawati, Adityo Permana Wibowo, “Implementasi Decision Tree Untuk Mendiagnosis Penyakit Liver,” Vol. 1, No. 1, Pp. 13–17, 2019

[15] Z. Niswati, A. Paramita, And F. A. Mustika, “Aplikasi Fuzzy Logic Dalam Diagnosa Penyakit Liver,” J. Nas. Teknol. Dan Sist. Inf., Vol. 2, No. 3, Pp. 21–30, 2016.