Implementasi Algoritma *Modified K-Nearest Neighbor* (MKNN) untuk Klasifikasi Penyakit Kanker Payudara

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi S1 Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

1302140128

M Ikhsan Perdana Putra

Program Studi Sarjana Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

Bandung

2019

LEMBAR PENGESAHAN

Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Kanker Payudara

Implementation of K-Nearest Neighbor Modified Algorithm (MKNN) for Classification of Breast Cancer

NIM :1302140128
M Ikhsan Perdana Putra

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Ilmu Komputasi
Fakultas Informatika
Universitas Telkom

Bandung, 14 Januari 2019
Menyetujui

Pembimbing I,

Danang Triantoro Murdiansyah,S.Si.,M,T
NIP:14870045

Pembimbing II,

Annisa Aditsania,S.Si.,M.Si
NIP:15900046

Ketua Program Studi
Sarjana S1 Ilmu Komputasi,

Dr.Deni Saepudin,S.Si.,M.Si
NIP: 99750013

LEMBAR PERNYATAAN

Dengan ini saya, M Ikhsan Perdana Putra, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul "Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Kanker Payudara" beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 14 Januari 2019

Yang Menyatakan

M Ikhsan Perdana Putra

Implementasi Algoritma *Modified* K-*Nearest Neighbor* (MKNN) untuk Klasifikasi Penyakit Kanker Payudara

M Ikhsan Perdana Putra[1], Danang Triantoro Murdiansyah[2], Annisa Aditsania[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ikhsanperdana20@gmail.com, [2]danangtri@telkomuniversity.ac.id, [3]aditsania@telkomuniversity.ac.id

Abstrak

Kanker payudara adalah salah satu penyakit mematikan di dunia.Menurut data WHO tahun 2013,penderita kanker payudara di dunia meningkat dari 12,7 juta kasus pada tahun 2008 menjadi 14,1 kasus juta pada tahun 2012.Sedangkan jumlah kematian meningkat dari 7,6 juta orang tahun 2008 menjadi 8,2 juta pada tahun 2012[1]. Dikarenakan semakin tinggi penyakit kanker payudara penting untuk mengetahui dan mencegah penyakit tersebut. Penelitian ini menggunakan data dari "UCI – Machine Learning Repository Breast Cancer Winconsin". Data yang diklasifikasikan terbagi atas 2 kelas yaitu kanker payudara jinak dan kanker payudara ganas. Tujuan dari penelitian ini adalah mengelompokkan penyakit tersebut termasuk kategori jinak atau ganas berdasarkan data yang ada. Penelitian ini menggunakan dataset *breast cancer Wisconsin*. Metode yang digunakan dalam penelitian ini adalah algoritma *Modified K-Nearest Neighbor*(MKNN). Hasil pengujian menunjukkan bahwa nilai K sangat mempengaruhi akurasi. Rata-rata akurasi cenderung menurun jika nilai K dinaikkan dan akurasi akan meningkat jika data latihnya dinaikkan. Hasil akurasi tertinggi pada pengujian ini sebesar 97.61 % dengan K=1 dan data latih 90%.

Kata kunci : Kanker Payudara, Modified K-Nearest Neighbor(MKNN)

Abstract

Breast cancer is one of the deadliest diseases in the world. According to WHO data in 2013, breast cancer patients in the world increased from 12.7 million cases in 2008 to 14.1 million cases in 2012. While the number of deaths increased from 7.6 million people in 2008 became 8.2 million in 2012 [1]. Because the higher breast cancer is important to know and prevent the disease. This study uses data from "UCI - Machine Learning Repository Breast Cancer Wisconsin". Data classified are divided into 2 classes, namely benign breast cancer and malignant breast cancer. The purpose of this study is to classify the disease including benign or malignant categories based on existing data. This study uses the Wisconsin breast cancer dataset. The method used in this study is the Modified K-Nearest Neighbor (MKNN) algorithm. The test results show that the K value is very affect accuracy. Average accuracy tends to decrease if the value of K is increased and accuracy will increase if the training data is increased. The highest accuracy results in this test are 97.61% with K = 1 and training data 90%.

.

Keywords: breast cancer, Modified K-Nearest Neighbor(MKNN)

1. Pendahuluan

Latar Belakang

Kanker payudara adalah keganasan yang berasal dari sel kelenjar, saluran kelenjar, dan jaringan penunjang payudara,tidak termasuk kulit payudara dan merupakan salah satu penyebab utama kematian diakibatkan oleh kanker pada perempuan di seluruh dunia.Setiap 2 dari 10.000 perempuan di dunia diperkirakan mengalami kanker payudara setiap tahunnya[1]. Di Indonesia, estimasi insiden kanker payudara sebesar 40 per 100.000 perempuan[2].

Estimasi angka kejadian kanker payudara yang cukup tinggi tersebut disebabkan oleh kurangnya kesadaran perempuan untuk segera memeriksakan diri jika terjadi sesuatu pada payudara[2].Salah satu kelainan pada payudara adalah tumor.Menurut *National breast cancer foundation*, tumor pada payudara diklasifikasikan menjadi tumor payudara jinak dan tumor payudara ganas.

Karena banyaknya jenis penyakit kanker payudara tersebut sangat penting untuk mengetahui gejala untuk mencegah kesalahan dalam melakukan diagnosa.Dengan mempelajari pola dari data hasil pemeriksaan pasien kanker payudara sehingga gejala-gejala tersebut dapat diklasifikasikan berdasarkan kedekatan antara data lama dengan data baru. Proses klasifikasi menggunakan komputer dapat diterapkan dengan menggunakan algoritma, antara lain *naïve bayes* , *fuzzy tsukamoto*, dan SVM.

Dalam penelitian sebelumnya menggunakan algoritma *K-Nearest Neighbor* (KNN) tentang klasifikasi penyakit kulit diperoleh akurasi sebesar 65% [7] dan juga penelitian yang dilakukan oleh Made Bela Pramesthi Putri, Edy Santoso, dan Marji tentang penyakit kulit menggunakan algoritma *Modified K-Nearest Neighbor(MKNN)* diperoleh akurasi 86.67% [8].Dalam penelitian ini algoritma yang digunakan adalah algoritma *Modified K-Nearest Neighbor(MKNN)*.Implementasi algoritma MKNN untuk klasifikasi kanker payudara bertujuan untuk memudahkan ahli medis dalam dalam memperoleh diagnosa sementara dengan mengklasifikasi gejala-gejala yang dialami oleh pasien dengan K data tetangga terdekat untuk mendapatkan nilai akurasi yang tinggi.

MKNN merupakan algoritma yang dikembangkan dari algoritma KNN, algoritma MKNN menambahkan proses baru untuk melakukan klasifikasi yaitu perhitungan nilai validitas untuk mempertimbangkan validitas antar data latih dan perhitungan *weighted voting* untuk menghitung bobot dari masing-masing terdekat.Penambahan 2 proses baru dalam MKNN diharapkan dapat memperbaiki setiap kesalahan pada proses KNN.

Topik dan Batasannya

Berdasarkan latar belakang diatas, permasalahan yang dibahas dalam penelitian ini adalah bagaimana implementasi algoritma *Modified K-Nearest Neighbor* dalam pengklasifikasian penyakit kanker payudara dan bagaimana performansi algoritma *Modified K-Nearest Neighbor* dalam mengklasifikasikan penyakit kanker payudara.

Adapun Batasan masalah dalam penelitian ini adalah data yang digunakan adalah data set yang digunakan bersumber dari "UCI – Machine Learning Repository Breast Cancer Winconsin".Data yang diklasifikasikan terbagi atas 2 kelas yaitu kanker payudara jinak dan kanker payudara ganas.

Tujuan

Adapun tujuan dari penelitian ini adalah mengimplementasikan algoritma *Modified K-Nearest Neighbor* dalam mengklasifikasikan penyakit kanker payudara dan menganalisis algoritma *Modified K-Nearest Neighbor* dalam mengklasifikasikan penyakit kanker payudara.

.

Organisasi Tulisan

Penulisan tugas akhir ini tersusun dalam beberapa bagian, yaitu sebagai berikut: bagian pertama berisi latar belakang, batasan masalah, hingga tujuan mengenai penelitian ini. Bagian kedua berisi studi terkait, yang menjelaskan hal-hal yang berkaitan dengan penelitian ini. Kemudian, bagian ketiga berisi sistem yang dibangun, akan menjelaskan rancangan sistem yang dibangun. Pada bagian keempat berisi evaluasi, mengenai hasil pengujian dan evaluasi sistem, dan bagian kelima berisi kesimpulan dari penelitian.

2. Studi Terkait

2.1 Kanker Payudara

Kanker Payudara adalah keganasan yang berasal dari sel kelenjar, saluran kelenjar dan jaringan penunjang payudara, tidak ternasuk kulit payudara. Kanker payudara banyak menyerang wanita, namun tidak menutup kemungkinan pria juga dapat terjangkit penyakit ini. Penyakit ini oleh *World Health*

Organization(WHO) dimasukkan kedalam *International Classificaion of Disseas*(ICD) dengan kode nomor 174 untuk wanita dan 175 untuk pria[1].

2.2 Data Mining

Data mining didefinisikan sebagai proses menemukan pola dalam data.proses itu harus otomatis atau semi otomatis.Pola yang ditemukan harus berguna sehingga dapat memberikan keuntungan[3].Data mining adalah proses menemukan hubungan yang bermakna ,pola dan tren dengan memilah-memilah sejumlah besar data dengan menggunakan teknologi pengenalan pola serta statistik dan teknik matematika[3].

2.3 Algoritma Modified K-Nearest Neighbor

Algoritma modified k-nearest neighbor (MKNN) merupakan pengembangan dari metode KNN dengan penambahan beberapa proses yaitu, perhitungan nilai validitas dan perhitungan bobot. Algoritma k-nearest neighbor (KNN) merupakan algoritma clustering yang sangat sederhana dengan cara mengelompokkan data baru dengan K tetangga terdekat [11]. Berikut langkah proses klasifikasi Algoritma modified k-nearest neighbor:

1. Perhitungan Jarak *Eucledian*

   Untuk mendefinisikan jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y) maka digunakan rumus Euclidean, seperti yang ditunjukkan pada persamaan 2

   d (xi,yi) = $\sqrt{\sum_{i=0}^{n}(xi - yi)^2}$ (1)

   dengan d adalah jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi, dimana x=x1,x2,…,xi dan y=y1,y2,…,yi dan I merepresentasikan nilai atribut serta n merupakan dimensi atribut.

2. Perhitungan Nilai Validitas

   Dalam algoritma MKNN, setiap data pada data training harus divalidasi terlebih dahulu pada awalnya. Validitas setiap data tergantung pada setiap tetangganya. Proses validasi dilakukan untuk semua data pada data training. Setelah dihitung validitas tiap data maka nilai validitas tersebut digunakan sebagai informasi lebih mengenai data tersebut. Persamaan yang digunakan untuk menghitung nilai validitas pada setiap data training adalah seperti persamaan dibawah ini Persamaan 3 [11].

   Validity(x)= $\frac{1}{H}\sum_{i=0}^{n} S(lbl(x), lbl(Ni(x)))$ (2)

   Dimana: H : jumlah titik terdekat
   lbl(x) : kelas x
   lbl(Ni(x)) : label kelas titik terdekat x
   Fungsi S digunakan untuk menghitung kesamaan antara titik x dan data ke i dari tetangga terdekat. Yang dituliskan dalam persamaan di bawah ini mendefinisikan fungsi S pada persamaan 4.

   $S(a, b)$ = { 1 $a = b$,0 $a \neq b$ } (3)

   Keterangan:
   a = kelas a pada data training.
   b = kelas lain selain a pada data training.

3. Perhitungan Weighted Voting

   Dalam metode MKNN, pertama weight masing-masing tetangga dihitung dengan menggunakan 1 / (de + 0.5). Kemudian, Validitas dari tiap data pada data training dikalikan dengan weighted berdasarkan pada jarak Euclidian. Dalam metode MKNN, weight voting tiap tetangga Persamaan 5 [11].

   $W(i)$ = $Validity(i)$ x $\frac{1}{de(i)+0.5}$ (4)

   dimana:
   $W(i)$: Perhitungan Weight Voting
   $Validity(i)$: Nilai Validitas
   $de(i)$: Jarak Euclidean

Teknik *weighted voting* ini mempunyai pengaruh yang lebih penting terhadap data yang mempunyai nilai validitas lebih tinggi dan paling dekat dengan data. Selain itu, dengan mengalikan validitas dengan jarak dapat mengatasi kelemahan dari setiap data yang mempunyai jarak dengan *weight* yang memiliki banyak masalah dalam *outlier*. Jadi, algoritma MKNN diusulkan secara signifikan lebih kuat daripada metode KNN tradisional yang didasarkan hanya pada jarak. (Parvin, 2008).

2.4 Confusion matrix

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek *testing* mana yang diprediksi benar dan tidak benar. Perhitungan ini ditabulasikan kedalam table yang disebut *confusion matrix* (gorunescu, 2011). *Confusion matrix* merupakan dataset yang memiliki dua kelas, kelas yang satu positif dan kelas lain sebagai negatif. Terdiri dari empat sel yaitu *true positif, false positif, true negative, false negative* (Max Bramer, 2007).

Tabel 1. Confusion matrix untuk 2 model kelas (Gorunescu, 2011)

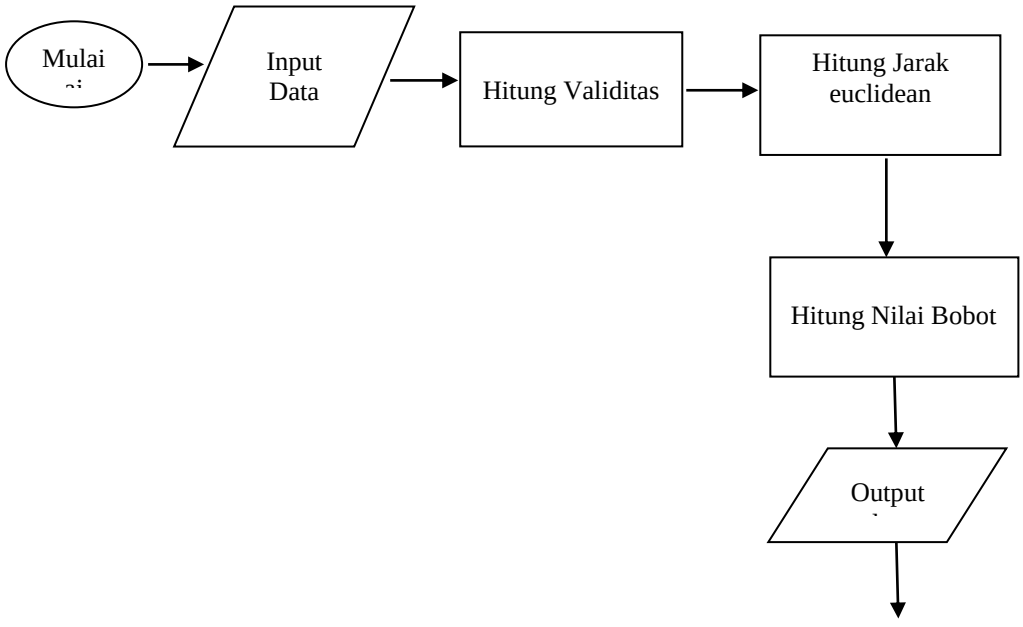
KELAS OBSERVASI	KELAS PREDIKSI	
KELAS AKTUAL	Kelas = YA	Kelas = TIDAK
Kelas = YA	True positif (TP)	False Negatif (FN)
Kelas = TIDAK	False positif (FP)	True Negatif (TN)

Untuk menghitung akurasi menggunakan rumus (Gorunescu, 2011):

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

Dimana:
TP : jumlah data positif yang terklasifikasi benar oleh sistem.
TN : jumlah data negatif yang terklasifikasi benar oleh sistem.
FN : jumlah data data negatif namun terklasifikasi salah oleh sistem.
FP : jumlah data positif namun terklasifikasi salah oleh sistem.

3. Sistem yang Dibangun



Gambar 1. *Flowchart* sistem

1. Input Data
   Data yang diambil adalah data dari UCI – Machine Learning Repository Breast Cancer Winconsin yang terdiri dari 31 gejala penyakit dan jumlah data 568. Terdiri dari dua kelas yaitu kanker jinak dan kanker ganas.

Table 2. Atribut WDBC dataset

Nama atribut	Domain
Radius	Data hasil FNA
Texture	Data hasil FNA
Perimeter	Data hasil FNA
Area	Data hasil FNA
Smoothness	Data hasil FNA
Compactness	Data hasil FNA
Concavity	Data hasil FNA
Concave points	Data hasil FNA
symetry	Data hasil FNA
Fractal dimension	Data hasil FNA

Keterangan:
FNA : merupakan landasan pengambilan keputusan bedah pada kanker.

2. Hitung Validitas
   Dalam algoritma MKNN, setiap data pada data training harus divalidasi terlebih dahulu pada awalnya. Validitas setiap data tergantung pada setiap tetangganya. Proses validasi dilakukan untuk semua data pada data training..
3. Hitung Jarak Euclidean
   Pada proses ini akan dihitung jarak Euclidean data training dengan data latih.
4. Hitung Nilai Bobot
   Pada proses ini akan dihitung nilai bobot berdasarkan nilai validitas dan jarak Euclidean yang dihitung tadi.
5. Output Data Klasifikasi Kanker Setelah proses klasifikasi MKNN di dapat hasil klasifikasi penderita kanker payudara.

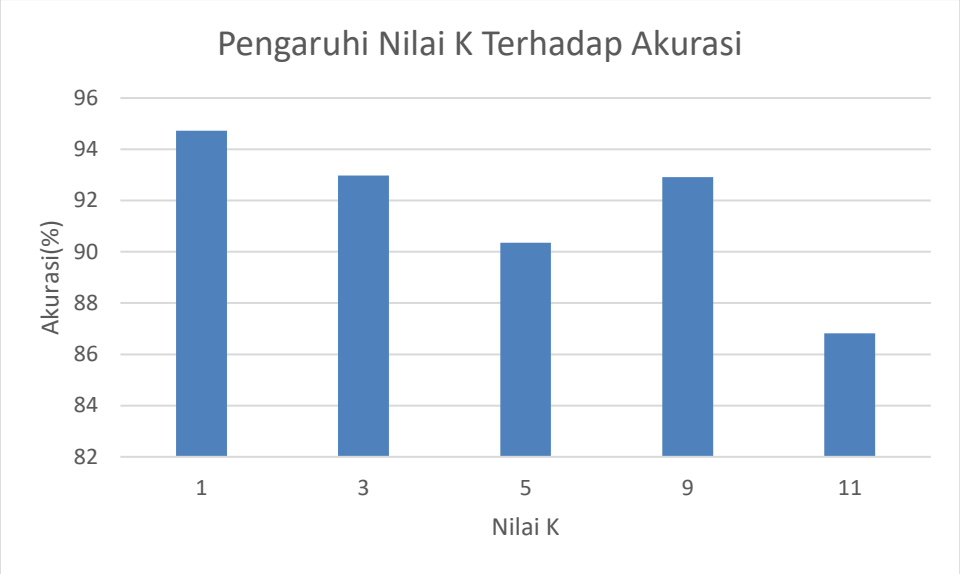
4. PENGUJIAN DAN ANALISIS

4.1 Skenario Pengujian

Penelitian ini menggunakan 2 skenario pengujian, yang pertama yaitu pengaruh nilai K terhadap akurasi dan kedua yaitu pengaruh data latih terhadap akurasi. Pada pengujian pengaruh nilai K terhadap akurasi dilakukan percobaan sebanyak 5 kali yaitu dengan K=1, K=3, K=5, K=9, dan K=11. Pada pengujian pengaruh data latih terhadap akurasi juga dilakukan 5 kali percobaan untuk data latih 60%, 70%, 80%, dan 90%.

4.2 Pengujian Pengaruh Nilai K (ketetanggan) Terhadap Akurasi

Berdasarkan hasil pegujian yang telah dilakukan dapat dilihat bahwa semakin besar nilai K rata-rata nilai akurasinya semakin menurun. Hal ini terjadi sebab semakin besar nilai K maka semakin banyak tetangga yang digunakan untuk melakukan proses klasifikasi sehingga kemungkinan *noise* semakin tinggi.



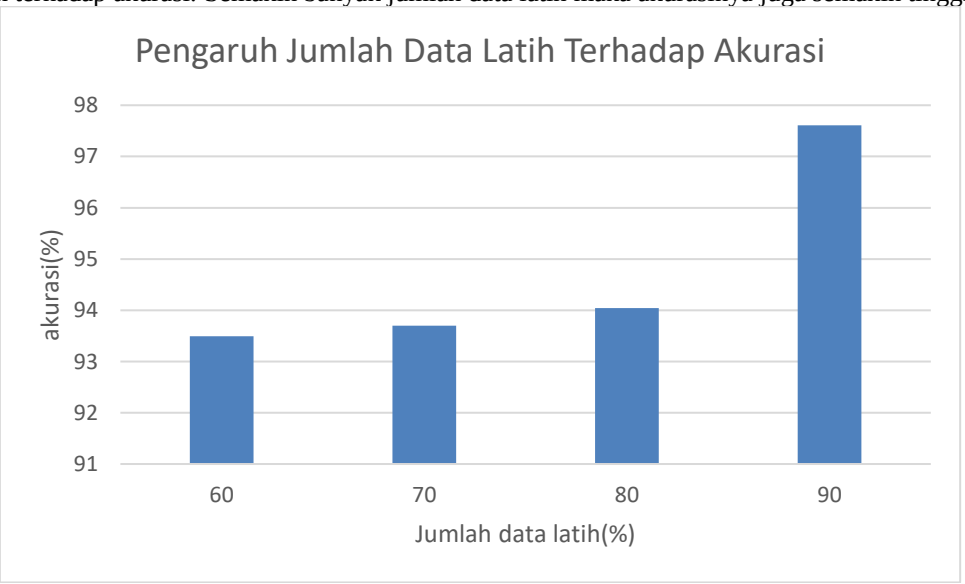
Gambar 2.Grafik Pengaruh Nilai K Terhadap Akurasi

Tabel 3. Confusion matrix nilai k=1

Jumlah data testing =114	Aktual jinak	Aktual ganas
Prediksi jinak	73	2
Prediksi ganas	4	35

4.3 Pengaruh Jumlah Data Latih Terhadap Akurasi

Berdasarkan hasil percobaan yang telah dilakukan dapat dilihat bahwa junlah data latih sangat berpengaruh terhadap akurasi. Semakin banyak jumlah data latih maka akurasinya juga semakin tinggi.



Gambar 3.Grafik Pengaruh Jumlah Data Latih Terhadap Akurasi

Tabel 4. Confusion matrix jumlah data latih 90(%)

Jumlah data testing = 42	Aktual jinak	Aktual ganas
Prediksi jinak	23	0
Prediksi ganas	1	18

4.4 Analisis Hasil Pengujian

Berdasarkan dari hasil pengujian dapat dilihat bahwa tingkat akurasi sistem dipengaruhi dengan nilai K dan jumlah data latih. Semakin tinggi nilai K maka akurasi sistem cenderung menurun karena semakin besar nilai K maka semakin banyak tetangga yang digunakan untuk melakukan proses klasifikasi sehingga kemungkinan *noise* semakin tinggi. Karena dalam perhitungan validitas setiap data latih itu tergantung pada setiap tetangganya. Karena jika tetangga banyak digunakan membuat data validitas semakin kecil dan itu sangat berpengaruh nanti dalam perhitungan pembobotan.

Pada pengujian pengaruh data latih terhadap akurasi sistem dapat dilihat bahwa junlah data latih sangat berpengaruh terhadap akurasi. Semakin banyak jumlah data latih maka akurasinya juga semakin tinggi, karena semakin banyak data latih yang digunakan maka makin banyak data yang akan dibandingkan dan itu sangat berpengaruh terhadap hasil akurasi yang dihasilkan.

5 Kesimpulan

Berdasarkan pengujian diatas penulis dapat menarik beberapa kesimpulan.

1. Metode algoritma Modified K-Nearest Neighbor dapat diimplementasikan dalam mengklasifikasikan kanker payudara karena akurasinya rata-rata diatas 90%.
2. Berdasarkan hasil implementasi algoritma Modified K-Nearest Neighbor didapatkan hasil sebagai berikut:
    a. Berdasarkan pengujian pengaruh nilai K terhadap akurasi, semakin tinggi nilai K maka akurasi sistem cenderung menurun. Akurasi tertinggi terdapat pada K=1 dengan akurasi 95.61%.
    b. Pada pengujian pengaruh data latih terhadap akurasi dapat dilihat semakin banyak data latih semakin tinggi akurasi.hal ini terjadi semakin banyak data latih, semakin banyak data yang mendekati kelas prediksi. Akurasi tertinggi terdapat pada K=1 dengan jumlah datih 90% dengan akurasi 97.61%

Daftar Pustaka

[1] Depkes RI (2009). Buku Saku Pencegahan Kanker Leher Rahim dan Kanker Payudara.Diunduh dari http://www.pppl.depkes.go.id/

[2] Globocan/IARC (2012). Breast Cancer Estimated Incidence, Mortality and Prevelance Worldwide in 2012.Diakses dari http://globocan.iarc.fr/old/FactSheets/cancers/breastnew.asp.

[3] Zainuddin, S., Hidayat, N & Soebroto, A., 2014. Penerapan Algoritma Modified KNearest Neighbor (M-KNN) pada Pengklasifikasian Tanaman Kedelai. S1. Universitas Brawijaya.

[4] Gevorkian, D, Egiazarian, K & Astola, J., 2000. Modified K-nearest neighbor filters for simple implementation. 2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No.00CH36353), Geneva. pp. 568- 565 vol.4.

[5] Fakihatin Wafiyah , Nurul Hidayat & Rizal Setya Perdana.2017. Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.

[6] Made Bela Pramesthi Putri, Edy Santoso, & Marji.2017. Diagnosis Penyakit Kulit Pada Kucing Menggunakan Algoritma Modified K-Nearest Neighbor (MKNN). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. [9] Panjaitan, A., Hidayat, B & Sujatmoko, K., 2013. Analisis Diskriminan Linear dalam Klasifikasi Data pada Teori Informasi dengan Metode Cross Validation. S1. Universitas Telkom.

[7] Sebastian Rori Listyanto Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Mengenali Pola Citra Dalam Mendeteksi Penyakit Kulit.

[8] Made Bela Pramesthi Putri, Edy Santoso, & Marji.2017. Diagnosis Penyakit Kulit Pada Kucing Menggunakan Algoritma Modified K-Nearest Neighbor (MKNN). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.

[9] Annisa, C.D., Putri, R.R & Marji., 2016. Sistem Pakar Diagnosa Awal Penyakit DBD, Malaria dan Tifoid menggunakan Metode Fuzzy K-Nearest 14 Neighbor (FKNN). S1. Universitas Brawijaya.

[10] Ao Li, Lirong Wang, Yunzhou Shi, Minghui Wang, Zhaohui Jiang and Huanqing Feng. 2005. Phosphorylation Site Prediction with A Modified k-Nearest Neighbor Algorithm and BLOSUM62 Matrix. IEEE Engineering in Medicine and Biology 27th Annual Conference. Shanghai.

[11] Parvin, H., Alizadeh, H & Bidgoli, B., 2008. MKNN: Modified K-Nearest Neighbor. Proceedings of the World Congress on Engineering and Computer Science 2008. San Fransisco. USA.