IMPLEMENTASI METODE K-NEAREST NEIGHBOR UNTUK KLASIFIKASI CITRA SEL PAP SMEAR MENGGUNAKAN ANALISIS TEKSTUR NUKLEUS

Toni.Arifin
Fakultas Teknik. Universitas BSI Bandung
Jalan Sekolah Internasional No. 1-6 Antapani Bandung 40282
toni.tfn@bsi.ac.id

Abstract *- Cervical cancer is the one of cause women death in the world. At least every 2 minutes 1 people death it cause of cervical cancer. One of prevention to early detection of cervical cancer is Pap Smear examination. Pap Smear test conducted to determine infection or abnormal cell that can turn into cancer cell. In this research used texture analysis data obtained from the result of image processing cell nucleus of normal and abnormal Pap Smear and 7 class Pap Smear cells is Normal Superficial (NS), Normal Intermediate (NI), Normal Columnar (NC), Mild (Light) Dysplasia (MLD), Severe Dysplasia (SD), Moderate Dysplasia (MD), Carcinoma In Situ (CIS). Image data derived from the data Harlev is totaling 280 images. The method of this research is used classification K-nearest neighbor method and for testing is used Confusion Matrix to see how much accuracy is generated by using K-nearest neighbor method. The result accuracy of normal and abnormal classification is 73,10% and for class classification is 33,33%.*

Keywords*: Texture Analysis, K-nearest neighbor , Classification, Pap Smear Cell, Cervical Cancer, Confusion Matrix.*

Abstrak - Kanker serviks merupakan salah satu penyebab kematian wanita di dunia. Setidaknya setiap 2 menit 1 orang di dunia meninggal karena kanker serviks. Salah satu cara pencegahan untuk mendeteksi secara dini kanker serviks adalah dengan melakukan Pemeriksaan Pap Smear. Tes Pap Smear dilakukan untuk melihat adanya infeksi atau sel-sel yang abnormal yang dapat berubah menjadi sel kanker. Pada penelitian ini menggunakan data analisis tekstur yang didapatkan dari hasil pengolahan citra inti sel Pap Smear normal dan abnormal dan 7 kelas sel Pap Smear yaitu Normal Superficial (NS), Normal Intermediate (NI), Normal Columnar (NC), Mild (Light) Dysplasia (MLD), Severe Dysplasia (SD), Moderate Dysplasia (MD), Carcinoma In Situ (CIS). Data citra berasal dari data Harlev yang berjumlah 280 citra. Metode yang digunakan dalam penelitian ini adalah metode klasifikasi K-nearest neighbor dan untuk pengujiannya menggunakan Confusion Matrix untuk melihat seberapa besar akurasi yang dihasilkan dengan menggunakan metode K-nearest neighbor . Akurasi yang dihasilkan dari klasifikasi normal dan abnormal adalah 73,10% dan untuk akurasi klasifikasi kelas adalah 33,33%.

Kata Kunci: Analisis Tekstur, K-nearest neighbor , Klasifikasi, Sel Pap Smear, Kanker Serviks, Confusion Matrix.

PENDAHULUAN

Kanker serviks atau disebut juga kanker mulut rahim merupakan salah satu penyakit kanker yang paling banyak ditakuti kaum wanita. Berdasarkan data dari WHO tahun 2013 dari sekian banyak penderita kanker di indonesia, penderita kanker serviks mencapai sepertiganya, dan setiap tahun ribuan wanita meninggal karena penyakit kanker serviks ini yang merupakan jenis kanker yang menempati peringkat teratas sebagai penyebab kematian wanita dunia. Human Papilloma Virus (HPV) merupakan penyebab dari kanker seviks. Sedangkan penyebab banyak kematian pada kaum wanita adalah virus HPV tipe 16 dan 18. Virus ini sangat mudah berpindah dan menyebar, tidak hanya melalui cairan, tapi juga berpindah melalui sentuhan kulit. Selain itu penggunaan wc umum yang sudah terkena virus HPV dapat menjangkit seseorang yang menggunakan jika tidak membersihkan dengan baik. Selain itu, kebiasaan hidup yang kurang baik juga bisa menyebabkan terjangkitnya kanker serviks ini, seperti kebiasaan merokok, kurangnya asupan vitamin terutama vitamin c dan vitamin e serta kurangnya asupan asam folat. Kebiasaan buruk lainnya yang dapat menyebabkan kanker seviks adalah seringnya melakukan hubungan intim dengan pria yang sering beganti pasangan dan melakukan hubungan intim pada usia dini (melakukan hubungan intim pada usia <16 tahun bahkan dapat meningkatan resiko 2x terkena kanker serviks). Salah satu pencegahan kanker serviks adalah melakukan tes Pap Smear. Pap Smear merupakan sebuah langkah pengujian medis untuk mendeteksi ada atau tidaknya ganguan pada leher rahim. Tes Pap Smear memberikan fakta medis keberadaan virus papiloma yang merupakan virus penyebab kanker seviks. Pengamatan visual citra Pap Smear secara manual memiliki banyak keterbatasan, membutuhkan waktu yang lama, dan rawan kesalahan saat melakukan pengamatan dibawah mikroskop. Oleh karena itu analisis secara otomatis akan memudahkan dalam proses pengamatan sel Pap Smear. Pada penelitian sebelumnya pendeteksian dilakukan melalui pengolahan citra. Pada proses pengolahan citra kanker serviks dilihat dari warna dan tekstur sel Pap Smear. Pada analisis tekstur sel Pap Smear menggunakan metode GLCM (*Gray level Co-occurrence Matrix*). Tujuan dari penelitian ini adalah mengembangkan penelitian sebelumnya menggunakan data mining klasifikasi *k-nearest neighbor* berdasarkan data hasil dari analisis tekstur nukleus dan menemukan metode yang paling akurat dalam klasifikasi Sel Nukleus.

METODE PENELITIAN

Dalam konteks penelitian, metode yang dilakukan mengacu kepada pemecahan masalah yang meliputi mengumpulkan data, merumuskan hipotesis atau proposisi, pengujian hipotesis, menafsirkan hasil, dan kesimpulan Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

Pengumpulan data

Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.

Pengolahan data awal

Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model

Metode yang diusulkan

Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan

model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (training data) dan data uji (testing data) juga diperlukan untuk pembuatan model.

Eksperimen dan pengujian metode

Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa rule yang akan dimanfaatkan dalam pengambilan keputusa.

Evaluasi dan validasi

Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

Jenis penelitian yang dilakukan pada penelitian ini adalah penelitian eksperimen. Pengertian penelitian eksperimen yakni sebuah penelitian yang dimana terdapat unsur perubahan terhadap data (objek) yang didapat, disertai dengan control yang ketat terhadap faktor-faktor luar serta dengan melibatkan objek lain sebagai pembanding.

Dibawah ini adalah sampel data yang digunakan.

Tabel 1
Sampel Data

No	Nama Kelas	Jumlah Data	Kelas
1	Normal Superficial (NS)	40	1
2	Normal Intermediate (NI)	40	2
3	Normal Columnar (NC)	40	3
4	Mild (Light) Dysplasia (MLD)	40	4
5	Severe Dysplasia (SD)	40	5
6	Moderate Dysplasia (MD)	40	6
7	Carcinoma In Situ (CIS)	40	7
	Total Data	280	

Sampel data yang digunakan dalam penelitian ini adalah hasil analisis tekstur dari penelitian sebelumnya yang menggunakan pengolahan citra pada proses awal dan menghasilkan data set analisis tekstur, data yang digunakan berasal dari 280 citra Harlev yang terdiri dari citra normal dan abnormal yang terbagi ke dalam 7 kelas masing-masing kelas terdiri dari 40 citra. Pada penelitian ini data di kelompokkan ke dalam 2 bagian yaitu data Traning dan data Testing jumlah data yang digunakan adalah 70% data training dan 30% data testing. Berikut ini adalah sampel data analisis tekstur yang digunakan dalam proses penelitian.

Tabel 2
Sampel Data Analisis Tekstur

Kerne_A	Cyto_A	K/C	Kerne_Ycol	Cyto_Ycol	CytoShort	CytoLong	CytoElong
803,5	27804,13	0,028087	85,86608	192,5246	0,843403	181,5749	242,0434
610,125	18067,88	0,032665	81,53135	153,4398	0,818583	171,1088	197,5702
990,375	79029,88	0,012377	77,84365	118,0012	0,858397	290,2502	355,8033
554,5	98941	0,005573	70,05455	139,3598	0,793271	327,9831	425,6195
636,375	99663,25	0,006345	76,95095	137,7483	0,829332	418,4945	344,1933
689,5	62074,25	0,010986	76,46793	138,342	0,791613	242,7274	347,0807
722,25	127313,8	0,005641	73,70891	146,4226	0,770828	414,8917	428,7528
562,375	69395,88	0,008039	67,45617	148,9644	0,708949	323,0749	319,412
520,75	92213,63	0,005616	70,11219	144,9375	0,95815	307,6737	446,9508
646,875	67751,75	0,009457	71,93146	135,5057	0,728873	276,5613	353,6467
704	54104,38	0,012845	73,71286	142,6313	0,79324	268,3177	336,8991

Data Mining

Data mining adalah proses menemukan korelasi baru yang bermakna, pola dengan memilah-milah sejumlah besar data yang tersimpan dalam repositori, menggunakan teknologi penalaran pola serta teknik-teknik statistik dan matematika (Larose, 2005).

Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menggali pengetahuan dari data atau informasi yang kita miliki. Data mining, sering disebut juga sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan pola atau hubungan dalam set data berukuran besar (Prasetyo, 2010). Data mining dibagi menjadi beberapa kelompok berdasarkan tugas/pekerjaan yang dapat dilakukan yaitu :

Deskripsi

Para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola dan trend yang tersembunyi dalam data.

Estimasi

Estimasi mirip dengan klasifikasi kecuali variabel tujuan yang lebih kearah numerik dari pada kategori.

Prediksi

Prediksi memiliki kemiripan dengan estimasi dan klasifikasi, hanya saja prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di masa depan).

Klasifikasi

Dalam klasifikasi, terdapat terget variabel kategori. sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

Klaster

Pengklasteran merupakan pengelompokan *record*, pengamatan atau memeperhatikan dan membentuk kelas obyek-obyek yang memiliki kemiripan. klaster adalah kumpulan *record* yang memiliki kemiripan satu dengan lainnya dan memiliki ketidak miripan *record* dalam klaster yang lain

Asosiasi

Tugas Asosiasi dalam data mining adalah untuk menemukan atribut yang muncul dalam satu waktu. salah satu implementasi dari asosiasi adalah market basket analysis atau analisis keranjang belanja (Larose, 2005).

Pada penelitian ini data mining yang digunakan adalah data mining klasifikasi adapun metode yang digunakan adalah metode klasifikasi *k-nearest neighbor* .

Metode *K-Nearest Neighbor*

K-nearest neighbor termasuk kelompok *instance-based learning*. Algortima ini juga merupakan salah satu teknik *Lazy learning*. *k-nearest neighbor* bekerja dengan mencari kelompok K objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing (Wu, 2008).

Algoritma *k-nearest neighbor* merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Dekat atau jauhnya lokasi (jarak) biasanya dihitung berdasarkan jarak *Euclidean* atau jarak terdekat dengan rumus sebagai berikut (Han dan Kamber, 2006).

$$d\ (x1,x2) = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

x_1 = Data Sampel
x_2 = Data uji
I = Variabel Data
d = Jarak
p = Dimensi Data

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max normalization* atau *Z-score standardization* (Larose, 2005). Jika data training terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* (Larose, 2005).

$$x^* = \frac{x - \min\ (x)}{max\ (x) - \min\ (x)} \quad (2)$$

*X** = Standarisasi min-max
min (x) = Nilai minimum data sampel
max (x)= Nilai maximum data sampel

Pengujian

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain agar diketahui di kelas mana objek data tersebut berdasarkan model yang telah disimpan dalam memori. Sistem dalam klasifikasi diharapkan mampu melakukan klasifikasi semua set data dengan benar, namun tidak dapat dipungkiri bahwa kesalahan akan terjadi dalam proses pengklasifikasian tersebut sehingga perlunya dilakukan pengukuran kinerja dari sistem klasifikasi tersebut. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*). *confusion*

matrix merupakan tabel pencatat hasil kerja klasifikasi (Prasetyo, 2010). Contoh dari *confusion matrix* untuk dua kelas (biner) dapat dilihat pada Tabel berikut.

Tabel 3
Confusion matrix

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 2
Kelas asli (i)	Kelas = 1	f_{11}	f_{12}
	Kelas = 2	f_{21}	f_{22}

Setiap sel fij dalam matriks menyatakan jumlah rekord atau data dari kelas i yang hasil prediksinya masuk ke kelas j. Dari matriks konfusi dapat diketahui jumlah data pemetaan yang diprediksi benar dengan cara menjumlahkan nilai f11 dan f22 (f11 + f22) dan jumlah data pemetaan yang diprediksi salah dengan menjumlahkan nilai f21 dan f12 (f21 + f12). Akurasi hasil prediksi dapat dihitung ketika jumlah data yang diklasifikasi secara benar maupun salah telah diketahui. Untuk menghitung akurasi digunakan formula:

$$\text{Akurasi} = \frac{\text{jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} \quad (3)$$

PEMBAHASAN

Perhitungan Manual Metode *k-nearest neighbor*

Dibawah ini dijelaskan cara perhitungan manual menggunakan metode *k-nearest neighbor* untuk mendapatkan nilai terdekat dari data uji dengan data sampel mengunakan data analisis tekstur sel Pap Smear.

Setelah menghitung jarak terdekat maka proses perhitungan selanjutnya adalah menghitung nilai *min-max standarization* yang bertujuan untuk mendapatkan hasil klasifikasi sesuai jarak terdekat.

Hasil dan Pengujian

Dibawah ini adalah hasil pengujian klasifikasi *k-nearest neighbor* (k=1) menggunakan *confusion matrix*.

Tabel 5
confusion matrix k-nearest neighbor jika K=1
(Normal dan Abnormal)

accuracy: 75.86%

	true normal	true abnormal	class precision
pred. normal	22	6	78.57%
pred. abnormal	8	22	73.33%
class recall	73.33%	78.57%	

Tabel 6
Confusion Matrix k-nearest neighbor jika K=1
(Klasifikasi Kelas)

accuracy: 34.92%

	true 1.0	true 2.0	true 3.0	true 4.0	true 5.0	true 6.0	true 7.0	class precision
pred. 1.0	5	2	0	0	0	1	4	41.67%
pred. 2.0	2	10	1	3	0	1	2	52.63%
pred. 3.0	5	1	0	0	0	0	0	0.00%
pred. 4.0	1	1	1	3	0	2	0	37.50%
pred. 5.0	3	0	0	0	2	3	0	25.00%
pred. 6.0	0	1	1	1	0	0	1	0.00%
pred. 7.0	1	0	0	0	1	2	2	33.33%
class recall	29.41%	66.67%	0.00%	42.86%	66.67%	0.00%	22.22%	

Kedua tabel diatas menunjukan hasil dari klasifikasi untuk 2 kelas yaitu normal dan abnormal dan klasifikasi untuk 7 kelas sel Pap Smear, jika k=1. terlihat hasil akurasi untuk k=1, bahwa klasifikasi 2 kelas (normal dan abnormal) sebesar 75,86% dan untuk klasifikasi 7 kelas akurasinya sebesar 34,92%. dibawah ini adalah tabel yang menerangkan hasil akurasi jika k= 1, 3, 5, 7, dan 9.

Tabel 6
Confusion Matrix k-nearest neighbor
(Klasifikasi Normal dan Abnormal)

KELAS	*K-NEAREST NEIGHBOR* 2 KELAS (Normal dan Abnormal)				
	K=1	**K=3**	**K=5**	**K=7**	**K=9**
Hasil Akurasi	75.86%	70.69%	74.14%	72.41%	72.41%

Tabel 7
Confusion Matrix K-nearest neighbor
(Klasifikasi Kelas)

KELAS	*K-NEAREST NEIGHBOR* 7 KELAS				
	K=1	**K=3**	**K=5**	**K=7**	**K=9**
Hasil Akurasi	34.92%	36.51%	33.33%	34.92%	26.98%

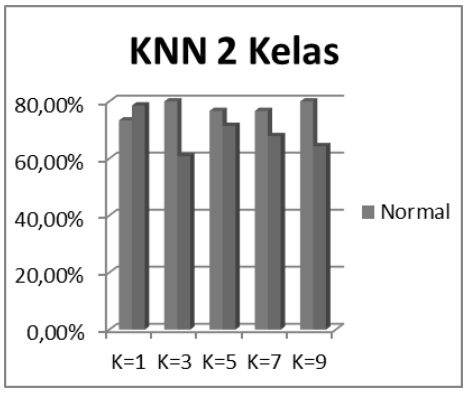
Tabel diatas menunjukkan hasil keseluruhan untuk akurasi Normal dan Abnormal dan akurasi untuk klasifikasi 7 kelas. Berikut ini dijelaskan tabel detail akurasi untuk masing-masing kelas.

Tabel 8
Detail klasifikasi untuk 2 kelas (Normal dan Abnormal)

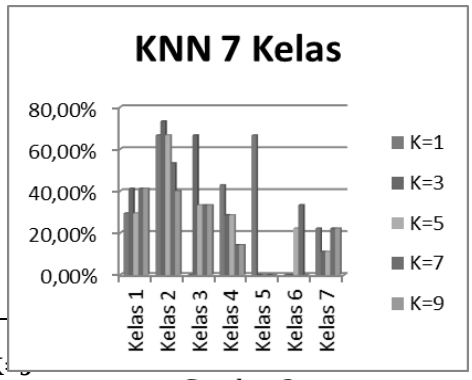
KELAS	*K-NEAREST NEIGHBOR* 2 KELAS				
	K=1	**K=3**	**K=5**	**K=7**	**K=9**
Normal	73.33%	80%	76.67%	76.67%	80%
Abnormal	78.57%	60.75%	71.43%	67.86%	64.29%

Tabel 9
Detail klasifikasi untuk 2 kelas
(7 kelas)

KELAS	*K-NEAREST NEIGHBOR* 7 KELAS				
	K=1	**K=3**	**K=5**	**K=7**	**K=9**
Kelas 1	29.41%	41.18%	29.41	41.18	41.18
Kelas 2	66.67%	73.33%	66.67%	53.33%	40%
Kelas 3	0%	66.67%	33.33%	33.33%	33.33%
Kelas 4	42.86%	28.57%	28.55%	14.29%	14.29%
Kelas 5	66.67%	0%	0%	0%	0%
Kelas 6	0%	0%	22.22%	33.39%	0%
Kelas 7	22.22%	11.11%	11.11%	22.22%	22.22%



Gambar 2
Grafik Dari Detail Hasil Klasifikasi Normal dan Abnormal



Gambar 3
Grafik Dari Detail Hasil Klasifikasi 7 Kelas

Tabel dan grafik menujukkan hasil akurasi yang dihasilkan dengan menggunakan metode K - Nearest Neighbor dengan nilai K =1, 3, 5, 7, dan 9 untuk akurasi normal dan abnormal akurasi terbesar didapat dari nilai k= 1 dan untuk klasifikasi kelas menunjukan akurasi yang baik jika nilai k=3. Hasil keseluruhan yang didapat diambil dari perhitungan dengan menggunakan metode K- Neighbor untuk klasifikasi normal dan abnormal dikatakan baik karena akurasi rata-rata mencapai

73,10% dan untuk klasifikasi kelas masih perlu perbaikan lagi karena nilai rata-rata akurasi yang dihasilkan adalah sebesar 33,33%.

KESIMPULAN

Berdasarkan analisis dan pembahasan yang telah dipaparkan sebelumnya, maka dapat diambil kesimpulan sebagai berikut:

1. Penelitian ini menggunakan data analisis tekstur nukleus yang diambil dari sel Pap Smear dari penelitian sebelumnya.
2. Data sampel dibagi 2 data training dan testing, pembagiannya meliputi 70% data training dan 30% data testing.
3. Pada metode K- Nearest Neighbor hasil akurasi untuk 2 kelas yaitu normal dan abnormal, akurasi terbesar dihasilkan dari nilai k=1 yaitu sebesar 75,86% dan yang paling kecil dihasilkan dari nilai k=3 yaitu 70,69%.
4. Nilai rata-rata akurasi untuk klasifikasi kelas normal dan abnormal adalah 73,10%
5. Hasil akurasi untuk 7 kelas, akurasi terbesar dihasilkan dari nilai k=3 yaitu sebesar 36,51% dan akurai terkecil dihasilkan dari nilai k=9 yaitu sebesar 26,98%
6. Nilai rata-rata akurasi untuk klasifikasi 7 kelas adalah 33,33%
7. Untuk hasil seluruh klasifikasi, klasifikasi kelas normal dan abnormal dikatakan baik dan untuk klasifikasi kelas masih kurang baik dan membutuhkan perbaikan. Perbaikan bisa dilakukan dengan menggunakan metode yang berbeda atau menambahkan data training dan data tesnya.

REFERENSI

Arifin, T. (2014). Klasifikasi Inti Sel Pap Smear Berdasarkan Analisis Tekstur Menggunakan Correlation-Based Feature Selection Berbasis Algoritma C4.5. *Jurnal Informatika*, 1, 123-129.

Arifin, T., Riana, D., & Hapsari, G. I. (2014). Klasifikasi Statistikal Tekstur Sel Pap Smear Dengan Decesion Tree. *Jurnal Informatika, 1*, 38-43.

Bidanku. (n.d.). *Kanker Serviks: Ciri-ciri, Penyebab, dan Pencegahan Kanker Serviks.* Retrieved February 1, 2015, from http://bidanku.com/kanker-serviks-ciri-ciri-penyebab-dan-pencegahan-kanker-serviks.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2003). Digital Image Processing Using Matlab. 11-12.

Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques.* San Fransisco: Morgan Kauffman.

Jantzen, J., Norup, G. J., Dounias, & Bjerregaard, B. (2005). Pap Smear Benchmark Data For Cervical Cell Types in Pap Smear Digital Images . 1-7.

Larose, D. T. (2005). *Discovering Knowledge in Data.* New Jersey: John Willey & Sons, Inc.

Martin, E. (n.d.). *Pap Smear Classification From Technical University of Denmark.* Retrieved January 25, 2014, from http://labs.fme.aegean.gr/decision/downloads/.

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta: ANDI Yogyakarta.

WHO. (n.d.). *WHO Guidance Note.* Retrieved January 28, 2015, from http://www.who.int/reproductivehe

alth/publications/9789241505147/en/index.html.

Wu, x. (2008). *Top 10 algorithms in data mining. Knowledge Inference System*.