TESIS - KL142502

KLASIFIKASI MASSA PADA CITRA MAMMOGRAM MENGGUNAKAN KOMBINASI SELEKSI FITUR F-SCORE DAN LS-SVM

Muhammad Imron Rosadi
5113201024

PEMBIMBING I
Dr. Agus Zainal Arifin, S.Kom., M.Kom

PEMBIMBING 2
Anny Yuniarti, S.Kom., M.Comp. Sc

PROGRAM MAGISTER
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

THESIS - KL142502

CLASSIFICATION MASSES IN IMAGE MAMMOGRAM USING COMBINED FEATURE SELECTION F-SCORE AND LS-SVM

Muhammad Imron Rosadi
5113201024

SUPERVISOR I
Dr. Agus Zainal Arifin, S.Kom., M.Kom

SUPERVISOR 2
Anny Yuniarti, S.Kom., M.Comp. Sc

MASTER PROGRAM
DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M.Kom.)
di
Institut Teknologi Sepuluh Nopember Surabaya

oleh:
Muhammad Imron Rosadi
Nrp. 5113201024

Dengan judul :
Klasifikasi massa pada citra mammogram menggunakan kombinasi seleksi fitur F-Score dan LS-SVM

Tanggal Ujian : 22-6-2016
Periode Wisuda : 2015 Genap

Disetujui oleh:

Dr.Agus Zainal Arifin, S.Kom, M.Kom
NIP. 197208091995121001
(Pembimbing 1)

Anny Yuniarti, S.Kom., M.Comp.Sc
NIP. 198106222005012002
(Pembimbing 2)

Dr. Eng. Nanik Suciati, S.Kom, M.Kom
NIP. 197104281994122001
(Penguji 1)

Diana Purwitasari, S.Kom, M.Sc
NIP. 197804102003122001
(Penguji 2)

Arya Yudhi Wijaya, S.Kom, M.Kom
NIP. 198409042010121002
(Penguji 3)

Direktur Program Pasca Sarjana,

Prof.Ir.Djauhar Manfaat, M.Sc., Ph.D.
NIP. 196012021987011001

KLASIFIKASI MASSAPADA CITRA MAMMOGRAM MENGGUNAKAN KOMBINASI SELEKSI FITUR F-SCORE DAN LS-SVM

Nama mahasiswa	: Muhammad Imron Rosadi
NRP	: 5113201024
Pembimbing I	: Dr. Agus Zainal Arifin, S.Kom., M.Kom
Pembimbing II	: Anny Yuniarti, S.Kom., M. Comp. Sc

ABSTRAK

Kanker payudara adalah penyakit yang paling umum diderita oleh perempuan pada banyak negara. Pemeriksaan kanker payudara dapat dilakukan menggunakan citra mammogram. Sistem Computer-aided detection (CAD). Analisis CAD yang telah dikembangkan adalah Ekstraksi Ftur GLCM, reduksi/seleksi fitur dan SVM. Pada SVM (Support vector Machine) maupun LS-SVM (least Square Support vector Machine) terdapat tiga masalah yang muncul, yaitu; bagaimana memilih fungsi kernel, berapa jumlah fitur input yang optimal, dan bagaimana menentukan parameter kernel terbaik. Jumlah fitur dan nilai parameter kernel yang diperlukan saling mempengaruhi, sehingga seleksi fitur diperlukan dalam membangun sistem klasifikasi.

Pada penelitian ini bertujuan untuk mengklasifikasi massa *pada citra mammogram berdasarkan dua kelas yaitu kelas kanker jinak dan kelas kanker ganas. Ekstraksi fitur menggunakan gray level co-occurrence matrix (GLCM). Hasil proses ekstraksi fitur tersebut kemudian diseleksi mengunakan metode F-Score. F-Score diperoleh dengan menghitung nilai diskriminan data hasil ekstraksi fitur di antara data dua kelas pada data training.Nilai F-Score masing-masing fitur kemudian diurutkan secara descending. Hasil pengurutan tersebut digunakan untuk membuat kombinasi fitur. Kombinasi fitur tersebut digunakan sebagai input LS-SVM.*

Dari hasil ujicoba bahwa menggunakan kombinasi seleksi fitur sangat berpengaruh terhadap tingkat akurasi. Akurasi terbaik didapat menggunakan LS-SVM RBF dan SVM RBF dengan kombinasi seleksi fitur maupun tanpa kombinasi seleksi fitur dengan nilai akurasi yaitu 97,5%. Selain itu juga seleksi fitur mampu mengurasi waktu komputasi.

Kata kunci : Kanker payudara, GLCM, F-Score, LS-SVM

CLASSIFICATION MASSES IN IMAGE MAMMOGRAM USING COMBINED FEATURE SELECTION F-SCORE AND LS-SVM

Name	: Muhammad Imron Rosadi
Student Identity Number	: 5113201024
Supervisor I	: Dr. Agus Zainal Arifin, S.Kom., M.Kom
Supervisor 2	: Anny Yuniarti, S.Kom., M.Comp. Sc

ABSTRACT

Breast cancer is the most common disease suffered by women in many countries. Breast cancer screening can be done using a mammogram image. Computer-aided detection system (CAD). CAD analysis that has been developed is GLCM efficient feature extraction, reduction / feature selection and SVM. In SVM (Support Vector Machine) and LS-SVM (Support Vector Machine Square least) there are three problems that arise, namely; how to choose the kernel function, how many input features are optimal, and how to determine the best kernel parameters. The number of features and value required kernel parameters affect each other, so that the selection of the features needed to build a system of classification.

In this study aims to classify image of masses on digital mammography based on two classes benign cancer and malignant cancer. Feature extraction using gray level co-occurrence matrix (GLCM). The results of the feature extraction process then selected using the method F-Score. F-Score is obtained by calculating the value of the discriminant feature extraction results data between two classes of data in the data training. Value F-Score of each feature and then sorted in descending order. The sequencing results are used to make the combination of features. The combination of these features are used as input LS-SVM.

From the experiments that use a combination of feature selection affects the accuracy ting-kat. Best accuracy obtained using LS-SVM and SVM RBF RBF with combination or without the combination of feature selection with accuracy value is 97.5%. It also features a selection able to curate the computation time.

Keywords : *Breast Cancer, GLCM, F-Score, LS-SVM*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kanker payudara dianggap sebagai masalah kesehatan yang utama di negara-negara barat, dan merupakan kanker yang paling umum di kalangan perempuan di Uni Eropa (Eurostat, 2002). Di Amerika serikat sekitar 39.520 perempuan meninggal dunia disebabkan kanker tersebut. Kemajuan pengobatan, peningkatan kesadaran, dan deteksi sejak dini menghasilkan angka kematian menurun (Tai, Chen, dan Tsai, 2014).

Mammografi adalah alat *screening* yang paling efektif untuk mendeteksi kanker payudara (Zuckerman, 1987). Seorang ahli radiologi biasanya memeriksa mammogram untuk memeriksa tanda-tanda kanker. Secara mammografi, kanker payudara dikenali dengan keberadaan lesi massa atau biasa disebut massa, danmikrokalsifikasi (Pisano, Shtem, 1993). Deteksi massa lebih sulit daripada deteksi mikrokalsifikasi karena ukuran, bentuk, dan kepadatannya bervariasi dan menunjukkan kontras gambar yang buruk serta dikelilingi oleh *background* dengan karakteristik yang sama (Kom, Tiedeu, dan Kom: 2007).

Sistem *Computer-aided detection* (CAD) membantu ahli radiologi untuk mengevaluasi mammogram sebagai opini kedua untuk mengenali abnormalitas dan menghindari opsi yang tidak diperlukan. Oleh karena itu sistem CAD telah dikembangkan untuk membantu ahli radiologi dan meningkatkan akurasi diagnosis (Tai, Chen, dan Tsai, 2014).

Sebagian besar skema CAD untuk mendeksi massa melibatkan lima fase utama yaitu : praproses citra, segmentasi citra, ekstraksi fitur dan seleksi fitur, deteksi/klasifikasi, evaluasi performa (Ceng dkk, 2006).

Pada citra mammogram ada tiga jenis fitur utama untuk mendeteksi dan mensegmentasi massa yaitu fitur bentuk, fitur tekstur dan fitur tingkat keabuan.Fitur tekstur merupakan karakteristik intrinsik dari suatu citra yang terkait dengan tingkat kekasaran *(roughness),* granularitas *(granularity),* dan keteraturan *(regularity)* susunan

struktural piksel. Aspek tekstural dari sebuah citra digunakan untuk membedakan sifat-sifat fisik permukaan objek suatu citra (Haralic dkk., 1973). Analisa tekstur lazim dimanfaatkan sebagai proses untuk melakukan klasifikasi dan interpretasi citra. Suatu proses klasifikasi citra berbasis analisis tekstur pada umumnya membutuhkan metode ekstraksi fitur yaitu Statistikal, Geometri, *Model-Based* (Jain dkk., 1995):

Dalam analisis statistik tekstur, fitur tekstur dihitung dari kombinasi distribusi statistik dan intensitas pada posisi relatif tertentu terhadap satu sama lain dalam gambar. Menurut jumlah titik intensitas (pixel) di setiap kombinasi, statistik diklasifikasikan ke dalam orde pertama, orde kedua dan statistik tingkat tinggi (Albregtsen, 2008). Metode *Gray Level Co-ocurrence Matrix* (GLCM) adalah cara ekstraksi fitur tekstur statistik urutan kedua. Pendekatan ini telah digunakan dalam beberapa aplikasi (Albregtsen, 2008). Pengukuran nilai tekstur yang digunakan didasarkan pada persamaan Haralic dan Conner.

Sebagian besar klasifikasi yang ada menganggap seluruh ruang fitur yang ada pada citra mammogram sebagai masukan untuk klasifikasi. Namun, ruang fitur dengan jumlah yang besar dan berdimensi tinggi akan memberikan efek negatif terhadap proses analisis. Untuk menangani hal tersebut, mereduksi fitur menjadi hal yang sangat penting. Pengurangan fitur dapat menghindari *over-fitting*, mengurangi kompleksitas analisis dan meningkatkan kinerja analisis data. Fitur yang besar akan membuat tugas klasifikasi menjadi kompleks, karena *classifier*akan menghabiskan banyak waktu untuk mengklasifikasikan dataset. Efisiensi akan dicapai jika klasifikasi hanya menganalisis fitur penting atau fitur yang diperlukan saja, fitur yang tidak relevan akan membuat proses klasifikasi menjadi jauh lebih sulit. Salah satu teknik untuk mereduksi fitur adalah seleksi fitur dengan proses memilih *subset* dari fitur asli sehingga jumlah fitur berkurang secara optimal sesuai dengan kriteria yang ditentukan. (Yu, 2003).

Penelitian tentang pengaruh seleksi fitur terhadap peningkatan performa klasifikasi telah dilakukan. Hasil menunjukkan peningkatan akurasi yang signifikan dibandingkan klasifikasi tanpa penerapan seleksi fitur. Sahiner dkk,2001mengusulkan

kombinasi seleksi fitur *stepwise* dan LDA pada ekstraksi fitur morfologi menghasilkan kurva FROC 0,89 (Sahiner dkk, 2001). Chen & Lin, 2005 mengusulkan metode kombinasi seleksi fitur dengan SVM (Chen, 2005). Salah satu metode seleksi fitur yang diusulkan adalah F-Score. F-Score adalah sebuah teknik sederhana untuk menghitung diskriminan dari dua himpunan bilangan real. *F-score* yang memiliki tingkat subjektivitas tinggi dalam pemilihan fitur (Chen, 2005). Kombinasi metode SVM dan F-Score telah digunakan untuk mendiagnosis penyakit kanker payudara menggunakan dataset statistik dan menghasilkan tingkat akurasi sebesar 99,51% (Akay, 2009). Aarthi dkk (2011) mengusulkan metode K-Mean *Clustering* untuk pengelompokkan fitur sebagai fitur input SVM berdasarkan ekstraksi fitur tekstur dan fitur klinik. Menghasilkan akurasi 86,11% dengan *clustering* dan 80,0% tanpa *clustering*. *Clustering* juga mampu mengurangi waktu komputasi.

SVM (*Support Vector Machine*) merupakan suatu teknik yang relatif baru berbasis *machine learning* untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi, yang sangat populer belakangan ini. SVM memiliki dua fitur pembelajaran, pertama, data *training* pada penelitian dipetakan ke ruang fitur yang memiliki dimensi lebih tinggi (linear) melalui fungsi pemetaan fitur tidak linear. Kedua, metode optimisasi standar yang kemudian digunakan untuk menemukan solusi dalam memaksimalkan margin pemisah dari dua kelas yang berbeda dalam ruang fitur dengan meminimumkan *error* pada data training. Pada SVM, juga terdapat *quadratic programming* yang merupakan suatu kompleksitas komputasi dari algoritma SVM yang biasanya intensif untuk digunakan, karena dengan *quadratic programming* dapat diperoleh solusi optimal dalam menentukan variabel lagrange yang nantinya digunakan dalam perhitungan nilai beta dan bias. Tetapi *quadratic programming* tidak efisien apabila diterapkan pada dimensi ruang yang lebih tinggi, oleh karena itu , Suykens dkk,. (2002) melakukan modifikasi terhadap rumusan asli dari SVM, dan rumusan baru tersebut diperkenalkan sebagai *Least Squares Support Vector Machines* (LS-SVM). Kinerja LS-SVM lebih baik dibandingkan SVM dalam hal proses perhitungan, konvergensi cepat dan presisi yang tinggi. Saat ini, LS-SVM banyak dilakukan pada

klasifikasi dan estimasi fungsi. Jika SVM dikarakteristikkan dengan permasalahan *quadratic programming* dengan fungsi constrain berupa pertidaksamaan, LS-SVM sebaliknya, diformulasikan dengan menggunakan fungsi *constrain* yang hanya berupa persamaan. Sehingga solusi LS-SVM dihasilkan dengan menyelesaikan persamaan linier (Suykens dkk,. 2002).

Berdasarkan uraian kelebihan metode yang diusulkan sebelumnya, peneliti mengusulkan kombinasi seleksi fitur F-Score dan LS-SVM untuk klasifikasi massa pada citra mammogram. Dengan sistem ini diharapkan mampu meningkatkan hasil akurasi, mengurangi waktu komputasi pada *classifier*, serta mendapatkan seleksi fitur dengan akurasi terbaik di antara seleksi fitur yang ada.

1.2 Perumusan Masalah

Permasalahan dalam penelitian ini adalah sebagai berikut :

1. Bagaimana cara menentukan seleksi fitur dengan F-Score?
2. Bagaimana pengaruh penggunaan seleksi fitur terhadap tingkat akurasi dan waktu komputasi pada LS-SVM?

1.3 Batasan Masalah

Sistem menggunakan dataset 118 massa (68 kanker jinak, 50 kanker ganas) pada mammogram tampilan *medio lateral oblique* (MLO) dari database *Mammographic Image Analysis Society* (MIAS) untuk data *training* dan *testing*.

1.4 Tujuan dan Manfaat Penelitian

Tujuan diadakannya penelitian ini adalah mengimplementasikan seleksi fitur sebagai solusi peningkatan keakuratan klasifikasi massa pada citra mammogram serta dapat mengurangi waktu pengujian klasifikasi. Dalam rangka mencapai tujuan tersebut, ada beberapa tujuan yang harus dicapai terlebih dahulu antara lain sebagai berikut.

1. Metode F-score sebagai seleksi fitur untuk meningkatkan performa klasifikasi massa pada citra mammogram.
2. Mengevaluasi performa klasifikasi LS-SVM terhadap subset fitur hasil seleksi metode F-score,serta mendapatkan seleksi fitur dengan akurasi terbaik diantara seleksi fitur yang ada.

Manfaat dilakukannya penelitian ini adalah untuk meningkatkan keakuratan performa diagnosis massa pada citra mammogram dengan menerapkan metodeseleksi fitur F-Score dan klasifikasi LS-SVM. Selain itu Penelitian inimengembangkan sebuah diagnosis otomatis berbasis komputer yang membantumemudahkan para ahli medis untuk meningkatkan keakuratan dan kecepatananalisis data medis.

1.5 Kontribusi Penelitian

Kontribusi pada penelitian ini adalah memberikan solusi untuk klasifikasi massa pada citra mammogram menggunakan kombinasi seleksi fitur F-score dan LS-SVM.

[Halaman ini sengaja dikosongkan]

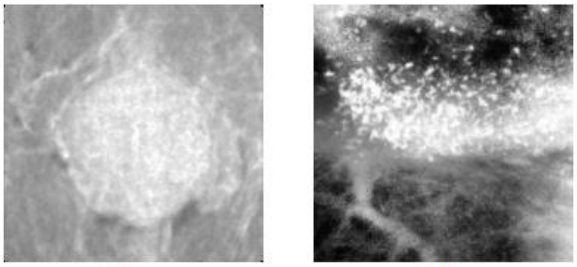
BAB II
KAJIAN PUSTAKA DAN DASAR TEORI

Pada bab ini dibahas dasar teori yang menjadi acuan penelitian ini. Tinjauan pustaka yang dijelaskan meliputi kanker payudara, dan mamografi yang menjadi dasar ilmu dalam pengerjaan penelitian ini. Selain itu, juga dibahas metode-metode yang digunakan dalam setiap tahap yaitu praproses, ekstraksi fitur, seleksi fitur dan LS-SVM.

2.1 Kanker Payudara

Kanker payudara merupakan jenis kanker yang paling umum diderita oleh wanita saat ini. Kanker payudara merupakan jenis kanker dengan angka kematian tertinggi pada wanita. Menurut Timp (2006) kisaran 22% dari semua jenis kanker yang terjadi pada wanita adalah kanker payudara Penyakit ini terjadi dimana sel-sel tidak normal (kanker) terbentuk pada jaringan payudara. Secara mammografi, kanker payudara dikenali dengan keberadaan lesi massa atau biasa disebut massa, atau keberadaan mikrokalsifikasi.

1. **Massa** adalah area terdapatnya pola tekstur dengan bentuk serta batas area tertentu pada proyeksi foto mammografi. Biasanya massa tampak dari dua proyeksi foto mammografi yang berbeda. Pada sebuah proyeksi mammografi saja, massa sering kali sulit dibedakan dari jaringan padat (*fibroglandular*) jika bentuk dan batas areanya tidak tampak jelas.
2. **Mikrokalsifikasi.** Fitur lainnya dari kanker adalah keberadaan mikrokalsifikasi. Mikrokalsifikasi berbentuk seperti noda berukuran kecil dan terkadang berupa titik-titik, terdapat di dalam *lobula* atau *ductal*. Bentuknya terkadang lingkaran maupun titik-titik yang seragam. Baik massa maupun mikrokalsifikasi, tidaklah mudah dikenali dalam jaringan payudara. Hal ini disebabkan baik karena jaringan payudara Baik massa maupun mikrokalsifikasi, tidaklah mudah dikenali dalam jaringan payudara.



Gambar 2.1 (a) Potongan citra massa. (b) Potongan citra mikrokalsifikasi

2.2 Mammografi

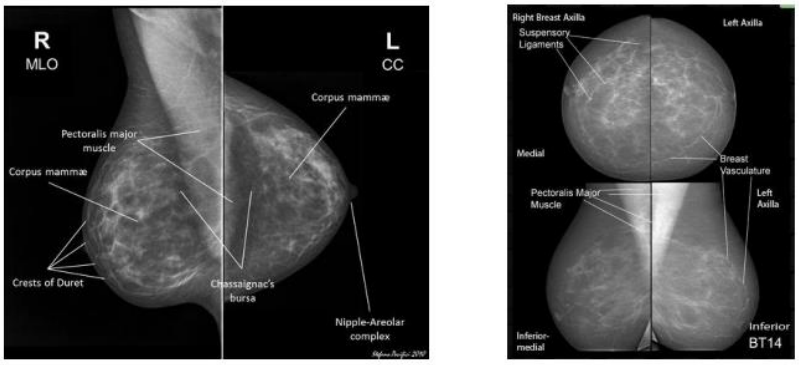
Mammografi merupakan pemeriksaan radiologi untuk pencitraan payudara dengan menggunakan sinar-x dosis rendah (rentang dosis 0,07-0,89 mSv, dosis rata-rata 0,48 mSv). Unit mammografi seperti pada Gambar 2.2. Tujuan dari mammografi adalah untuk deteksi dini kanker payudara, biasanya melalui deteksi karakteristik *lesion* dan atau bentuk kalsifikasi (holmes, 2014).

Gambar 2.2 Unit mammografi

Mammografi memegang peranan penting dalam deteksi dini kanker payudara, hal ini karena mammografi mampu mendeteksi hampir 75% kankerpayudara kurang lebih satu tahun sebelum pasien merasakan gejala. Terdapatdua tipe pemeriksaan mammografi, yaitu skrining dan diagnostik. skrining Mammografi dilakukan pada wanita yang tidak memiliki gejala pada payudara, sedangkan mammografi diagnostik

dilakukan pada wanita dengan gejala pada payudara, yaitu ketika ditemukan benjolan payudara atau *nipple discharge* selama pemeriksaan payudara sendiri atau abnormalitas payudara ditemukan ketika dilakukan pemeriksaan *screening* mammografi. Pemeriksaan Mammografi digunakan untuk menentukan ukuran yang tepat dan lokasi dari abnormalitaspayudara serta untuk menggambarkan jaringan sekitar dan limfonodi (Disha, dkk., 2009).

Selama prosedur pemeriksaan mammografi, payudara dikompresi menggunakan pelat paralel pada alat mammografi. Kompresi pelat paralel akan meratakan ketebalan jaringan payudara yang bertujuan untuk meningkatkan kualitas gambar, dengan cara mengurangi ketebalan jaringan yang akan ditembus oleh sinar-x, mengurangi jumlah radiasi hambur (karena radiasi hambur dapat menurunkan kualitas gambar), mengurangi dosis radiasi yang diperlukan, dan menahan payudara untuk mencegah *motion blur*. Pencitraan mammografi diambil dalam dua *view*, yaitu *craniocaudal* (CC) dan *medio lateral oblique* (MLO) seperti pada Gambar 2.3 Pada keadaan yang membutuhkan gambar yanglebih fokus dan jelas maka dilakukan magnifikasi dan atau *spot* kompresi padaarea tertentu yang menjadi perhatian. Deodoran, bedak atau *lotion* mungkin muncul pada gambar mammografi sebagai bintik-bintik kalsium, dan pasien disarankan untuk tidak memakai deodoran, bedak atau *lotion* pada haripemeriksaan untuk menghindari timbulnya artefak tersebut (**Anonymous, 2014**).



Gambar 2.3. (a) Mammografi normal MLO dan CC view (b) Mammografi normal MLO dan CC view pada fatty breast.

Mammografi diketahui memiliki angka negatif palsu. Berdasarkan data dari *Breast Cancer Detection Demonstration Project*, angka negatif palsu pada mammografi sekitar 8-10%. Kurang lebih 1-3% wanita yang secara klinis memiliki abnormalitas payudara yang mencurigakan, dengan hasil mammografi dan hasil ultrasonografi yang negatif, masih mungkin menderita kanker payudara. Kemungkinan yang menjadi penyebab hal tersebut adalahparenkim payudara yang padat menutupi gambaran lesi, posisi atau teknikmammografi yang kurang baik, kesalahan persepsi, interpretasi yang salah dari temuan yang dicurigai suatu abnormalitas, gambaran lesi keganasan yangsamar, dan lambatnya pertumbuhan lesi (Disha, dkk., 2009).

2.3 Praproses

Data yang digunakan dalam penelitian adalah dataset yang diambil dari dari hasil *screening mammography*. Proses pra-pengolahan atau lebih dikenal dengan *preprocessing* adalah langkah memperbaiki citra untuk menonjolkan citra yang ingin di ekstraksi.

2.4 Ekstraksi Fitur Statistik

Tekstur merupakan karakteristik dari suatu citra yang terkati dengan tingkat kekasaran, granularitas, dan keteraturan susunan structural piksel. Tekstur difiturkan sebagai distribusi spasial dari derajat keabuan di dalam sekumpulan piksel-piksel yang bertetangga. Analisis tekstur penting dan berguna dalam bidang *computer vision*. Dari elemen tekstur, sebuah citra akan dapat dimanfaatkan dalam proses segmentasi, klasifikasi, maupun interpretasi citra (Jain dkk, 1995).

Analisa tekstur lazim dimanfaatkan sebagai proses untuk melakukan klasifikasi dan interpretasi citra. Suatu proses klasifikasi citra berbasis analisis tekstur pada umumnya membutuhkan metode ekstraksi fitur yaitu Statistikal, Geometri, *Model-Based* (Jain dkk., 1995):

2.5.1 *Gray level co-occurrence Matric (GLCM)*

Dalam analisis statistik tekstur, fitur tekstur dihitung dari kombinasi distribusi statistik dan intensitas pada posisi relatif tertentu terhadap satu sama lain dalam gambar. Menurut jumlah titik intensitas (pixel) disetiap kombinasi, statistik diklasifikasikan ke dalam orde pertama, orde kedua dan statistik tingkat tinggi (Albregtsen, 2008).

Metode *Gray Level Coocurrence Matrix* (GLCM) adalah cara ekstraksi fitur tekstur statistik urutan kedua. Pendekatan ini telah digunakan dalam beberapa aplikasi (Albregtsen, 2008).

GLCM adalah matriks di mana jumlah baris dan kolom sama dengan jumlah tingkat abu-abu *(G)* dalam gambar. Elemen matriks P(i, j|Δx, Δy) adalah frekuensi yang relatif dengan dua piksel, dipisahkan oleh jarak pixel (Δx, Δy), terjadi dalam lingkungan tertentu, satu dengan intensitas *i* dan lainnya dengan intensitas *j*. Satu juga dapat mengatakan bahwa elemen matriks P(i, j|d, θ) berisi urutan kedua nilai probabilitas statistik untuk perubahan antara tingkat abu-abu *I* dan *j* pada khususnya jarak perpindahan *(d)* dan pada sudut tertentu (θ) (Albregtsen, 2008).

Mengingat area M×N dari suatu gambar masukan yang mengandung tingkat abu-abu (G) dari 0 sampai G-1, gunakan *f(m, n)* sebagai intensitas pada contoh *m*, garis *n* pada area sekitar.

Kemudian

$$P(i,j|\Delta x,\Delta y) = WQ(i,j|\Delta x,\Delta y) \tag{2.1}$$

Dimana

$$W = \frac{1}{(M-\Delta x)(N-\Delta y)} \tag{2.2}$$

$$Q(i,j|\Delta x,\Delta y) = \sum_{n=1}^{N-\Delta y} \sum_{m=1}^{M-\Delta x} A \tag{2.3}$$

Dan

$$A = \begin{cases} 1 \; iff(m,n) = 1 \; and f(m + \Delta x, n + \Delta y = j \\ \\ 0 \; elsewhere \end{cases} \quad (2.4)$$

Ukuran kecil (5 ×5) bagian gambar dengan 4 tingkat abu-abu dan *gray level co-ocrurrence matrix* P(i, j|Δx=1, Δy=0) diilustrasikan di bawah ini.

0	1	1	2	3
0	0	2	3	3
0	1	2	2	3
1	2	3	2	2
2	2	3	3	2

(a)

	j=0	1	2	3
i=0	1	2	1	0
1	0	1	3	0
2	0	0	3	5
3	0	0	2	2

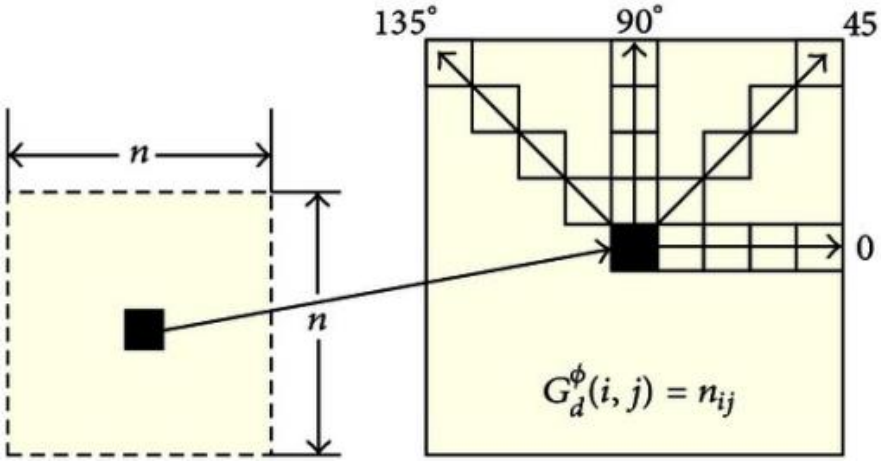
(b)

Gambar2.4: a) Matrik asal, Matrik A, b) Matrik co-occurance dari matrik A

Menggunakan sejumlah besar tingkat intensitas *G* menyiratkan menyimpan banyak data sementara, yaitu matriks G×G untuk setiap kombinasi jarak piksel (Δx, Δy) atau (d, θ). Satu kadang-kadang memiliki situasi paradoks bahwa matriks dari manafitur tekstur yang diekstrak lebih produktif dari pada gambar asli dari mana mereka berasal. Hal ini juga jelas bahwa karena dimensi yang besar, GLCM sangat sensitif terhadap ukuran sampel tekstur yang mereka perkirakan. Dengan demikian, jumlah tingkat abu-abu sering berkurang. Bahkan secara visual, kuantisasi menjadi 16 tingkat abu-abu sering kali cukup untuk diskriminasi atau segmentasi tekstur. Menggunakan beberapa tingkat setara dengan melihat gambar pada skala kasar, sedangkan tingkat lebih memberikan gambar dengan lebih detail. Namun, kinerja dari fitur berbasis GLCM, serta peringkat fitur, mungkin tergantung pada jumlah tingkat abu-abu yang digunakan.

Karena matriks G×G harus diakumulasikan untuk setiap jendela bagian gambar dan untuk setiap set parameter pemisahan (d, θ), biasanya komputasi diperlukan untuk

membatasi (d, θ) nilai yang akan diuji untuk sejumlah nilai. Gambar 2.9 di bawah ini menggambarkan hubungan geometris pengukuran GLCM dibuat untuk empat jarak *d*(d =max{| Δx|, |Δy|}) dan sudut θ=0, π/4, π/2 dan 3π/4 radian dengan asumsi simetri sudut.



Gambar 2.5. Geometri untuk pengukuran *gray level co-occurance matrix* (GLCM) untuk 4 jarak d dan 4 sudutθ.

Untuk mendapatkan perkiraan statistik yang dapat diandalkan dari distribusi probabilitas gabungan, matriks harus berisi tingkat hunian rata-rata cukup besar. Hal ini dapat dicapai baik dengan membatasi jumlah tingkat nilai kuantisasi abu-abu atau dengan menggunakan jendela yang relatif besar. Sebelumnya hasil pendekatan dalam kehilangan akurasi deskripsi tekstur dalam analisis tekstur amplitudo rendah, sedangkan yang kedua penyebab ketidakpastian dan kesalahan jika perubahan tekstur atas jendela besar. Sebuah kompromi yang khas adalah dengan menggunakan 16 tingkat abu-abu dan jendela sekitar 30 sampai 50 piksel di setiap sisi.

Hubungan sederhana ada di antara pasangan tertentu dari perkiraan distribusi probabilitas P(d, θ). Biarkan P^t(d, θ) menyatakan transpose dari matriks P(d, θ). Yaitu P(d, 0^0) = P^t(d, 180^0), P(d, 45^0) = P^t(d, 225^0), P(d, 90^0) = P^t(d, 270^0), P(d, 135^0) = P^t(d, 315^0).Dengan demikian, pengetahuan tentang P(d, 180^0), P(d, 225^0), P(d, 270^0), dan P(d, 315^0) tidak ada penambahan spesifikasi tekstur.

Pengukuran nilai tekstur yang digunakan didasarkan pada persamaan (Haralic et al, 1973 dan Conner et al. 1984). Menggunakan notasi berikut: G adalah jumlah tingkat abu-abu yang digunakan,μ adalah nilai rata-rata dari P,μ_x, μ_y, $\sigma_x$dan σ_y adalah *means* dan *standard deviations* Px dan Py. i dan j adalah masukan dalam matriks tepi probabilitas yang diperoleh dengan menjumlahkan baris dan kolom P(i, j).

Berikut ini fitur yang digunakan :

1. Energi (Energi)

   Menunjukkan ukuran dari *local homogeneity* dan merupakan kebalikan dari *entropy*. Persamaan :

$$Energy = \sum_{i,j} \mathrm{P}(i,j)^2 \tag{2.5}$$

2. Kontras (Contrast)

$$Contrast = \sum_{i=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G1} \sum_{j=1}^{G1} P(i,j) \right\}, |i-j| = n \tag{2.6}$$

3. Homogenitas (*Homogeneity*), *Angular Second Moment* (ASM)

   ASM adalah ukuran homogenitas dari suatu gambar. Didefinisikan :

$$ASM = \sum_{i=0}^{G-1} \quad \sum_{j=0}^{G1} \{p \quad (i,j)\}^2 \tag{2.7}$$

4. Korelasi (Correlation)

Korelasi menunjukkan ketergantungan linear derajat keabuan dari piksel-piksel yang saling bertetangga dalam suatu citra abu-abu. Persamaan :

$$Correlation = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\{ixj\}xP(i,j) - \{\mu_x x \mu_y\}}{\sigma_x x \sigma_y} \tag{2.8}$$

dimana :

μ_x= nilai rata-rata elemen kolom pada matriks Pθ(i,j)
μ_y= nilai rata-rata elemen baris pada matriks Pθ(i,j)
σ_x= nilai standar deviasi elemen kolom pada matriks Pθ(i,j)
σ_y= nilai standar deviasi elemen kolom pada matriks Pθ(i,j)

5. *Autocorrelation*

$$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right] \tag{2.9}$$

6. Jumlah Rata-rata (*Sum Average*)

$$AVER = \sum_{I=0}^{2G-2} Ip_{X+Y}(i) \tag{2.10}$$

7. Jumlah Entropi (*Sum Entropy*)

$$SEN = \sum_{i=0}^{2G-2} p_{x+y}(i)\log(p_{x+y}(i)) \tag{2.11}$$

8. Sum Varians (*Sum Variance*)

$$VARIANCE = \sum_{i=0}^{G-1} \quad \sum_{i=0}^{G1} (k-\mu)^2 p \quad (i,j) \tag{2.12}$$

9. Selisih Entropi (*Difference Entropy*)

$$DENT = -\sum_{i=0}^{G-1} P_{x+y}(i)\, log(p_{x+y}(i)) \tag{2.13}$$

10. *Sum of Squares*

$$VARIANCE = \sum_{i=0}^{G-1} \quad \sum_{i=0}^{G1} (k-\mu)^2 p \quad (i,j) \tag{2.14}$$

11. *Cluster Shade*

$$\text{SHADE} = \sum_{\text{i}=0}^{\text{G}-1} \sum_{\text{j}=0}^{\text{G}-1} \left\{\text{i} + \text{j} - \mu_{\text{x}}\mu_{\text{y}}\right\}^3 \text{x P(i,j)} \tag{2.15}$$

12. *Cluster prominence*

$$\text{PROM} = \sum_{\text{i}} \sum_{\text{j}} \left\{\text{i} + \text{j} - \mu_{\text{x}}\mu_{\text{y}}\right\}^4 \text{x P(i,j)} \tag{2.16}$$

2.6 Seleksi fitur

Seleksi fitur adalah salah teknik terpenting dan sering digunakan dalam *pre-processing* aplikasi *machine learning*. Seleksi fitur adalah proses memilih *subset* dari fitur asli sehingga jumlah fitur berkurang secara optimal sesuai dengan kriteria yang ditentukan. Teknik ini terbukti efektif mengurangi fitur-fitur yang tidak relevan dan berlebihan, meningkatkan efisiensi dalam proses *learning*, dan meningkatkan kinerja *learning* seperti akurasi prediksi. Data dimensi tinggi dapat berisi banyak sekali informasi yang tidak relevan dan berlebihan yang sangat mungkin menurunkan kinerja dari algoritma *learning*. Oleh karena itu, seleksi fitur menjadi sangat diperlukan oleh aplikasi *machine learning* ketika menghadapi data dengan dimensi yang tinggi. (Yu, 2003). Dengan jumlah fitur yang sedikit, penjelasan tentang keputusan klasifikasi yang rasional lebih mudah diperoleh. Pada diagnosis medis, jumlah fitur yang kecil berarti mengurangi biaya tes dan biaya diagnostik (Akay, 2009).

Beberapa metode seleksi fitur yang digunakan adalah:

1. *Principal component analysis* (PCA). PCA memproyeksikan fitur untuk mendapatkan jumlah fitur yang lebih sedikit. PCA melakukan tranformasi linier ortogonal data ke sistem koordinat baru.
2. Metode genetika dan evolusi. Ini merupakan metode *unsupervised* yang menggunakan pendekatan evolusioner untuk memangkas jumlah fitur yang ada.
3. *Hill climbing*. Dengan asumsi jumlah *p* fitur, metode ini dimulai dengan memilih satu fitur dan membangun *classifier* berdasarkan fitur tersebut. Fitur dengan akurasi tertinggi dipertahankan dan seterusnya sehingga tersisa *p-1* fitur yang dipilih dan dikombinasikan dengan fitur sebelumnya. Hal tersebut diulang sampai semua fitur telah digabungkan. Jika didapatkan himpunan/kombinasi dengan akurasi tertinggi, maka kombinasi fitur tersebut adalah kombinasi yang optimal.
4. *Hill descent*. Metode ini adalah kebalikan dari metode *hill climbing*, yaitu langkah pertama adalah semua fitur *p* digunakan, kemudian satu fitur dihilangkan dan sisanya digunakan untuk melatih *classifier*.
5. *Receiver operating characteristics area*. Metode sederhana *thresholding* yang

dapat digunakan untuk menghitung daerah *receiver operating characteristics* (ROC) fitur tunggal. Daerah dengan nilai yang cenderung menyatu/mengumpul, menunjukkan keterpisahan fitur yang lebih tinggi dan cenderung berisi informasi yang lebih diskriminatif (Begg, 2008).

2.5.1 F-Score

Menurut Chen, (2005) F-score *(Fisher score)* adalah teknik sederhana yang mengukur diskriminan dua himpunan bilangan real. Pada vektor *training* x_k, dengan k = 1, 2, . . . , m, jika jumlah kasus positif dan negatif adalah n_+ dan n_-, maka F-score masing-masing fitur i didefinisikan sebagai:

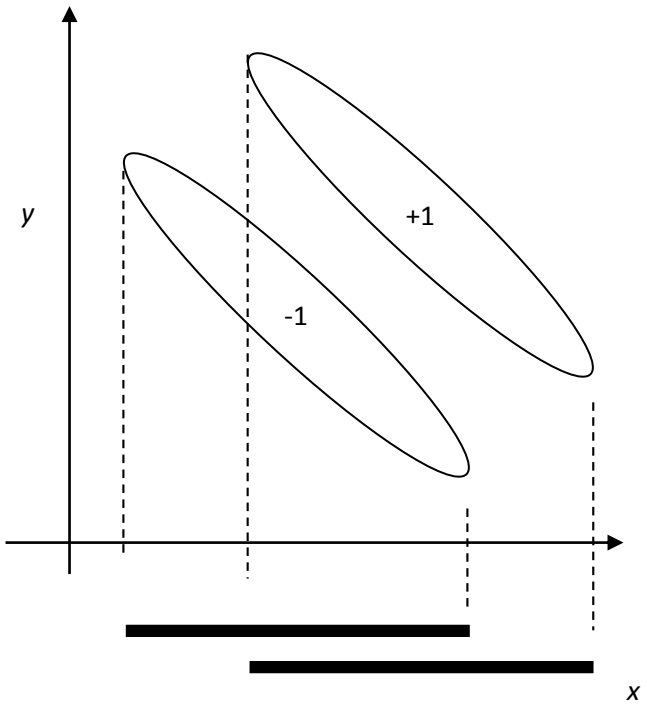
$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_+ - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}, \qquad (2.17)$$

di mana $\bar{x}_i$, $\bar{x}_i^{(+)}$,$\bar{x}_i^{(-)}$ adalah rata-rata dari fitur ke-i keseluruhan, dataset positif, dan negatif, $x_{k,i}^{(+)}$ adalah fitur ke-i dari kasus positif ke-k, dan $x_{k,i}^{(-)}$ adalah fitur ke-i dari kasus negatif ke-k. Pembilang menunjukkan diskriminasi antara himpunan positif dan negatif, dan penyebut menunjukkan fitur-fitur dalam dua himpunan. Semakin besar F-score, kemungkinan fitur lebih diskriminatif semakin besar pula.

Kekurangan F-Score adalah tidak mengungkapkan informasi timbal balik antar fitur. Ilustrasi sederhana dapat dilihat pada Gambar 2.6. Gambar 2.6 menunjukkan bahwa kedua fitur tersebut mempunyai nilai F-Score yang rendah, karena sesuai dengan rumus 2.17, penyebut yaitu jumlah varian dari set positif dan negatif mempunyai nilai yang jauh lebih besar daripada pembilang. Meskipun terdapat kekurangan, F-Score adalah metode yang sederhana dan cukup efektif (Chen, 2005).

Metode seleksi fitur dengan F-Score dilakukan dengan menghitung nilai F-Score semua fitur. Nilai F-Score masing-masing fitur tersebut kemudian diurutkan secara *descending*. Sesuai dengan metode *hill climbing*, dipilih satu fitur dengan nilai F-Score tertinggi, kemudian dimasukkan ke *classifier*. Fitur tersebut kemudian dikombinasikan dengan satu fitur dari fitur sisa. Hal tersebut diulang sampai semua

fitur telah digabungkan. Jika didapatkan kombinasi dengan akurasi tertinggi, maka kombinasi fitur tersebut adalah kombinasi yang optimal.



Gambar 2.6. Ilustrasi *Dataset* dengan Nilai *F-Score* Rendah

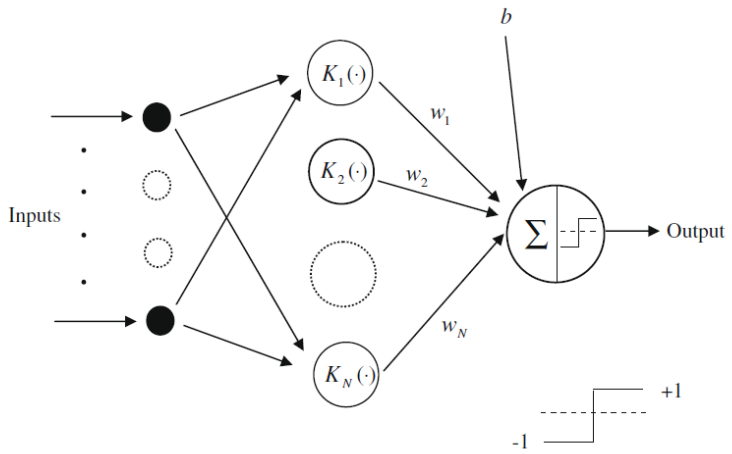
2.6 *Support Vector Machines* (SVM)

SVM yang diusulkan oleh Vapnik (1995) telah dipelajari secara ekstensif untuk klasifikasi, regresi dan estimasi kepadatan. Gambar 2.7. adalah arsitektur SVM. SVM memetakan pola input ke ruang fitur dimensi yang lebih tinggi melalui pemetaan *non linear* berdasar teori yang dipilih. Bidang pemisah linear ini kemudian dibangun dalam ruang fitur dimensi tinggi. Dengan demikian, SVM adalah *linear classifier* di ruang parameter, tapi itu menjadi *non linear classifier* sebagai akibat dari pemetaan *non linear* dari ruang pola input ke ruang fitur dimensi tinggi. Bila data pelatihan berdimensi *m* adalah x_i (i = 1, ..., M) dan masing-masing kelas labelnya adalah y_i, di mana y_i = 1 dan y_i = -1 untuk kelas 1 dan 2. Jika data input terpisah secara linear di ruang fitur, maka fungsi keputusan dapat ditentukan:

$$D(x) = w^t g(x) + b \qquad (2.18)$$

di mana *g(x)* adalah fungsi pemetaan yang memetakan *x* ke dalam ruang dimensi 1, *w* adalah vektor dimensi dan 1, dan *b* adalah skalar. Untuk memisahkan data secara linier, fungsi keputusan memenuhi kondisi berikut:

$$y_i(w^t g(x_i) + b) \geq 1 \text{ untuk } i = 1, ..., M \quad (2.19)$$



Gambar 2.7. Arsitektur SVM

Jika masalah terpisah secara linier dalam ruang fitur, maka fungsi keputusan jumlahnya tak terbatas. Di antara fungsi-fungsi tersebut, diperlukan *hyperplane* dengan margin terbesar antara dua kelas. Margin adalah jarak minimum yang memisahkan *hyperplane* terhadap data input dan ini dihasilkan dari $|\boldsymbol{D(x)}|/||\boldsymbol{w}||$. Sehingga didapatkan *hyperplane* pemisah dengan margin maksimal yang optimal memisahkan *hyperplane*.

Dengan asumsi bahwa margin adalah ρ, kondisi berikut harus memenuhi:

$$\frac{y_i D(x_i)}{\| w \|} \geq \rho \text{ untuk } i = 1, ..., M \quad (2.20)$$

Hasil perkalian produk dari ρ dan $\|w\|$ adalah tetap:

$$\rho \, \|w\| = 1 \quad (2.21)$$

Untuk mendapatkan *hyperplane* pemisah yang optimal dengan margin maksimal, *w* dengan $\|w\|$ yang memenuhi persamaan (2.22) harus ditemukan. Persamaan (2.23)

mengarahkan ke pemecahan masalah optimasi berikutnya. Dengan meminimalkan

$$\frac{1}{2} w^t w \tag{2.22}$$

dan mengikuti batasan:

$$y_i(w^t g(x_i) + b) \geq 1 \text{ untuk } i = 1, ..., M \tag{2.23}$$

Bila data pelatihan tidak linier dipisahkan, digunakan *slack variable* ξ_i ke persamaan (2.24):

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i \xi_i \geq 0 \text{ untuk } i = 1, ..., M \tag{2.24}$$

Hyperplane pemisah yang optimal telah ditentukan sehingga maksimalisasi dari margin dan meminimalisasi dari kesalahan *training* didapatkan. Dengan meminimalkan

$$\frac{1}{2} w^t w + \frac{C}{2} \sum_{i=1}^{n} \xi_i^{\rho} \tag{2.25}$$

mengikuti batasan:

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i \xi_i \geq 0 \text{ untuk } i = 1, ..., M \tag{2.26}$$

di mana *C* adalah parameter yang menentukan *tradeoff* antara margin maksimum dan kesalahan klasifikasi minimum dan ρ adalah 1 atau 2. Jika $\rho = 1$, SVM disebut SVM dengan *soft margin* L1 (L1-SVM), dan jika $\rho = 2$, SVM dengan *soft margin* L2 (L2-SVM). Pada SVM konvensional, *hyperplane* pemisah yang optimal diperoleh dengan memecahkan masalah pemrograman kuadratik.

Fungsi kernel memungkinkan operasi yang akan dilakukan di ruang input bukan di ruang fitur dimensi tinggi. Beberapa contoh fungsi kernel adalah $K(u, v) = v^T u$ (SVM linier); $K(u, v) = (v^T u + 1)^n$ (SVM polinomial derajat n); $K(u, v) = \exp(-\|u - v\|^2 / 2\sigma^2)$ (SVM fungsi *radial bases* – SVM RBF); $K(u, v) = \tanh(Kv^T y + o)$ (*neural* SVM dua *layer*) di mana σ, κ, *o* adalah konstanta [Vapnik, 1995; Cortes, 1995]. Namun, fungsi kernel yang tepat untuk suatu masalah tertentu tergantung pada data, dan sampai

saat ini belum ada metode yang baik tentang cara memilih fungsi kernel.

2.7 *Least Squares Support Vectors Machine* (LS-SVM)

Least Squares Support Vectors Machine (LS-SVM) adalah salah satu modifikasi dari SVM (Suykens, 1999). Jika SVM dikarakteristik oleh permasalahan konveks *quadratic programming* dengan pembatas berupa pertidaksamaan, LS-SVM sebaliknya, diformulasikan dengan menggunakan pembatas yang hanya berupa persamaan. Sehingga solusi LS-SVM dihasilkan dengan menyelesaikan persamaan linier. Hal ini tentulah berbeda dengan SVM yang mana solusinya dihasilkan melalui penyelesaian *quadratic programming*. Saat ini, LS-SVM banyak dilakukan pada klasifikasi dan estimasi fungsi (Suykens, 1999).

LS-SVM di-*training* dengan meminimalkan

$$\frac{1}{2}w^t w+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2\,, \tag{2.27}$$

dan mengikuti batasan persamaan:

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i \xi_i \geq 0 \text{ untuk } i = 1, ..., M. \tag{2.28}$$

Pada LS-SVM, batasan persamaan digunakan sebagai pengganti pertidaksamaan yang digunakan pada SVM konvensional. Karena itu, solusi yang optimal dapat diperoleh dengan menyelesaikan sekumpulan persamaan linier bukan dengan penyelesaian *quadratic programming*. Untuk menurunkan dua masalah persamaan (2.22) dan (2.23) digunakan *Lagrangemultiplier*, yaitu :

$$Q(w,b,\alpha,\xi)=\frac{1}{2}w^t w+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2-\sum_{i=1}^{n}\alpha_i\{y_i(w^t g(x_i)+b)-1+\xi_i\}, \tag{2.29}$$

di mana $\alpha = (\alpha_1, ..., \alpha_M)^t$ adalah *Lagrange multiplier* yang bisa bernilai positif atau negatif pada rumus LS-SVM. Kondisi yang optimum diperoleh dengan mendiferensialkan persamaan di atas pada persamaan (2.30). nilai w, ξ_i, *b*, dan $\alpha_i$sebagian besarnilai-nilaiyangsama dengan nol (Suykens, 1999).

$$\begin{cases} \dfrac{\partial \mathcal{L}_3}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \\ \dfrac{\partial \mathcal{L}_3}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}_3}{\partial \xi_i} = 0 \rightarrow \alpha_i = \gamma \xi_i i = 1, \ldots, N \\ \frac{\partial \mathcal{L}_3}{\partial \alpha_i} = 0 \rightarrow y_i [w^T \varphi(x_i) + b] - 1 + \xi_i = 0, \ i = 1, \ldots, N \end{cases} \quad (2.30)$$

bisa ditulis dengan solusi persamaan linear (2.31)

$$\left[\begin{array}{ccc|c} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & \gamma I & -I \\ \hline Z & Y & I & 0 \end{array}\right] \left[\begin{matrix} w \\ b \\ \xi \\ a \end{matrix}\right] = \left[\begin{matrix} 0 \\ 0 \\ 0 \\ \vec{1} \end{matrix}\right] \quad (2.31)$$

Dimana$Z = \lceil \varphi(x_1)^T y_{1\,;\ldots;} \varphi(x_N)^T y_N \rceil, Y = \lceil y_1; \ldots; y_N \rceil, \vec{1} = [1; \ldots; 1], \xi = \lceil \xi_1; \ldots; \xi_N \rceil, a = \lceil a_1; \ldots; a_N \rceil$.Solusi ini juga bisa ditulis dengan

$$\left[\begin{array}{c|c} 0 & -Y^T \\ \hline Y & ZZ^T + \gamma^{-1} I \end{array}\right] \left[\begin{matrix} b \\ a \end{matrix}\right] = \left[\begin{matrix} b \\ \vec{1} \end{matrix}\right] \quad (2.32)$$

Kondisi Mercer dapat diterapkan lagi pada matrikΩadalah definitif positif,$\Omega = ZZ^T$,dimana

$$\Omega_{il} = y_i y_l \varphi(x_i)^T \varphi(x_l) \quad (2.33)$$
$$= y_i y_l \Psi(x_i, x_l).$$

Seperti pada SVM konvensional, fungsi kernel memungkinkan operasi yang akan dilakukan di ruang input bukan di ruang fitur dimensi tinggi. Beberapa penelitian menggunakan LS-SVM dan fungsi kernel RBF (LS-SVM RBF) secara empiris menghasilkan hasil yang optimal (Suykens, 1999). Untukmasalahklasifikasidua-spiral yang kompleks dapat ditemukan dengan LS-SVM RBF dengankinerjayang sangat baikdan komputasirendah (Suykens, 1999).

2.8 Fungsi Kernel

Salah satu karakteristikk dari SVM adalah menggunakan teknik yang disebut kernel (Suykens, 1999). Didefinisikan pada persamaan (2.34)

$$K(x, x') = \mathrm{g}(x^t)\,\mathrm{g}(x'), \tag{2.34}$$

dimana $K(x, x')$ adalah fungsi kenel, sehingga dapat menghindari memberlakuan variabel dalam ruang fitur. Ada beberapa fungsi kernel dalam SVM, antara lain :

- Kernel dot product:$K(x, x') = x^t x'$
- Kernel polynomial :$K(x, x') = (x^t x')^d$, dimana *d* adalah bilangan bulat positif
- Kernel RBF :$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$, dimana γ adalah parameter positif.

Jika memiliki masalah yang sangat besar pada variable input, nilai fungsi kernel menjadi sangat kecil atau besar. Bahwa *training* SVM menjadi sulit. Untuk kernel polynomial dengan tingkat *d*, nilai maksimum adalah $(m+1)^d$ jika *range vaiable* input adalah [0,1]. Dengan demikian, saat nilai *m* sangat besar, maka kernel polynomial dinormalisasi dengan persamaan (2.35).

$$K(x, x') = \frac{(\mathrm{x^t x'})^d}{(m+1)^d} \tag{2.35}$$

demikian juga untuk kernel RBF, nilai maximum $\|x - x'\|^2$ adalah *m* dan kemudian dinormalisasi dengan persamaan (2.36).

$$K(x, x') = \exp\left(-\frac{\gamma}{m}\|x - x'\|^2\right) \tag{2.36}$$

BAB III
METODE PENELITIAN

Dalam bab ini akan diuraikan tentang rancangan penelitian, rancangan sistem, dan rancangan uji coba.

3.1 Rancangan Penelitian

Secara umum, penelitian ini dilakukan dalam beberapa tahap yaitu diawali dari studi literatur, perumusan masalah, perancangan metode dan implementasi, serta uji coba dan evaluasi. Sedangkan penulisan laporan penelitian dimulai dari awal sampai akhir penelitian ini.

1. Studi literatur

   Mempelajari berbagai literatur tentang sistem klasifikasi massa mulai dari metode praproses, metode ekstraksi fitur, metode seleksi fitur, dan metode klasifikasi.
2. Perumusan masalah

   Merumuskan permasalahan yang akan diteliti dan mencari solusinya.
3. Perancangan metode dan implementasi

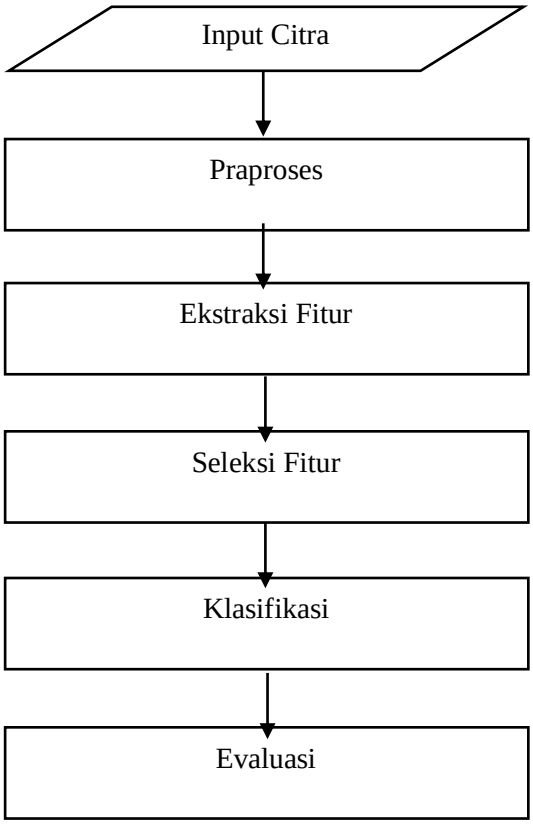
   Merancang dan mengimplementasikan metode penyelesaian dari permasalahan yang telah dirumuskan berdasarkan pengetahuan yang diperoleh dari studi literatur. Rancangan metode penyelesain yang diusulkan akan dijelaskan pada sub bab 3.2.
4. Uji coba dan evaluasi

   Melakukan pengujian dan evaluasi terhadap metode yang telah dirancang dengan menerapkan beberapa skenario. Uji coba dan evaluasi akan dijelaskan pada sub bab 3.3.
5. Penyusunan laporan

   Penyusunan laporan dilakukan mulai dari awal sampai akhir penelitian ini. Penyusunan laporan ditulis dalam bentuk laporan tesis berdasarkan ketentuan yang berlaku.

3.2 Rancangan Sistem

Pada rancangan CAD untuk mendeteksi massa mempunyai 4 tahap : *preprocessing*, ekstraksi fitur, seleksi fitur dan klasifikasi. Sesuai dengan gambar 3.1.



Gambar 3.1 Rancangan Sistem Klasifikasi Massa pada Citra Mammogram

3.2.1 Dataset Kanker Payudara

Dataset yang digunakan pada penelitian ini adalah diambil dari database mini-MIAS (*MAMMOGRAPHIC IMAGE ANALYSIS SOCIETY*) digitalkan pada 50 mikron piksel tepi yang telah direduksi menjadi 200 mikron piksel tepi dan setiap gambar dipotong menjadi 1024x1024 piksel. Hanya tampilan MLO yang dianalisis pada penelitian ini. Gambar di rubah ke format *.png. system ini dievaluasi menggunakan 118 massa (68 kanker jinak dan 50 kanker ganas). Untuk pelatihan, menggunakan 88 massa (48 kanker jinak, 40 kanker ganas), Untuk pengujian, meggunakan 40 massa (30 kanker jinak, 10 kanker ganas).

3.2.2 Praproses *(Preprocessing)*

Praproses pada penelitian ini dilakukan pemotongan secara manual untuk mendeteksi massa (ROI) secara proporsional seperti pada Gambar 3.1. Tujuan proposes ini adalah untuk mengurangi kesalahan dalam proses klasifikasi.

3.1a Citra Asli

3.1b Hasil Pemotongan

3.2.3 Ekstraksi Fitur

Setelah ROI diseleksi kemudian beberapa fitur diekstraksi untuk mengetahui karakteristik wilayah massa. Ekstraksi fitur berdasarkan fitur tekstur yang digunakan pada penelitian ini adalah metode GLCM. GLCM terdiri dari dua belas nilai fitur tekstur yaitu: *Energy, Correlation, Contrast, Autocorrelation, Cluster_Prominence, Cluster_Shade, Sum_variance, Difference_entropy, Homogeneity, Sum_average, Sum_of_squares,* dan *Sum_entropy*.

3.2.4 Seleksi Fitur

Seleksi fitur merupakan isu penting dalam membangun sistem klasifikasi. Keuntungan dengan membatasi jumlah fitur yang digunakan dalam *classifier* adalah untuk meningkatkan akurasi dan mengurangi komputasi. Seleksi fitur adalah tahap keempat dari metode penelitian ini.

Proses seleksi fitur dilakukan dengan menghitung nilai *F-Score* menggunakan persamaan (2.46) dari data *training*. Perhitungan nilai *F-Score* dari data *training* tersebut berbeda dengan metode yang kombinasi seleksi fitur yang diusulkan oleh Chen

& Lin (Chen, 2005). Jika pada metode Chen & Lin, perhitungan *F-Score* dilakukan untuk seluruh data, baik data *training* maupun *testing*. Sehingga seleksi fitur yang dihasilkan dari beberapa uji coba yang dilakukan adalah sama.

Perhitungan nilai *F-Score* berdasarkan dari jumlah fitur yang dipakai dalam penelitian ini adalah 12, maka jumlah hasil perhitungan nilai *F-Score* adalah 12. Nilai masing-masing *F-Score* yang telah dihasilkan diurutkan secara menurun (*descending*). Hasil pengurutan tersebut digunakan untuk menentukan seleksi fitur yang akan digunakan baik untuk *training* maupun *testing*.

Seleksi fitur pertama dibuat dari fitur dengan nilai *F-Score* terbesar. Seleksi fitur kedua dibuat dari fitur dengan nilai *F-Score* terbesar kedua, dan seterusnya sehingga didapatkan dua belasseleksi*F-Score*. Sebagai contoh, misal hasil pengurutan secara *descending* untuk *F-Score* dari data *training* adalah Fitur 4 (F_4), Fitur 1 (F_1), Fitur 3 (F_3), Fitur 7 (F_7), Fitur 5 (F_5), Fitur 10 (F_{10}), Fitur 8 (F_8), Fitur 2 (F_2), Fitur 11 (F_{11}), Fitur 6 (F_6), dan Fitur 9 (F_9) sampai fitur ke-12. Urutan tersebut dapat ditulis (F_4, F_1, F_3, F_7, F_5, F_{10}, F_8, F_2, F_{11}, F_6, $F_{9, \ldots.,}$ F_{12}). Berdasarkan hasil pengurutan tersebut dapat dibuat 12 kombinasi fitur yaitu F_4, F_4F_1, $F_4F_1F_3$, $F_4F_1F_3F_7$, $F_4F_1F_3F_7F_5F_{10}F_8F_2F_{11}F_6F_{9, \ldots..,}$ $F_4F_1F_3F_7F_5F_{10}F_8F_2F_{11}F_6F_{9\ldots}F_{12}$. Dua belas seleksi fitur tersebut secara lengkap dapat dilihat pada Tabel 3.1.

Dua belas macam seleksi tersebut kemudian digunakan sebagai input pada LS-SVM. Pertama, seleksi fitur model #1 digunakan sebagai input pada LS-SVM RBF baik untuk proses *training* mapun *testing*. Proses *training* mapun *testing* tersebut kemudian diulang lagi untuk seleksi fitur model #2, #3, #4, dan seterusnya sampai dengan model #12. Jika diperhatikan pada tabel 3.1, seleksi fitur model #12, yaitu $F_4F_1F_3F_7F_5F_{10}F_8F_2F_{11}F_6F_{9\ldots}F_{12}$ merupakan kombinasi input LS-SVM pada penelitian ini. Bentuk *pseudo code* perhitungan *F-Score* adalah:

rata2_xi=mean(xtrain)

rata2_xp=mean(xtrain[groupp])

rata2_xn=mean(xtrain[groupn])

varian_xp=var(xtrain[groupp])

varian_xn=var(xtrain[groupn])

fscore=((rata2_xp-rata2_xi)^2+(rata2_xn-rata2_xi)^2)/

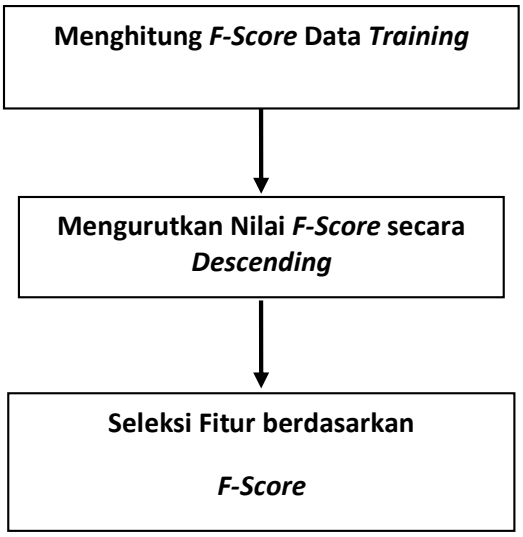
(varian_xp+varian_xn);

xi=fitur ke-i, xtrain=fitur pada data training, xp=fitur pada kelas positif, xn=fitur pada kelas negatif, groupp=golongan pada kelas positif, groupn=golongan pada kelas negatif, varian=vukuran variasi fitur.

Tahapan proses seleksi fitur dalam bentuk diagram seperti yang dijelaskan sebelumnya secara lengkap dapat dilihat pada Gambar 3.1.

Tabel 3.1. Kombinasi Fitur untuk *F-Score*

No.	Urutan Nilai *F-Score*	Kombinasi Fitur
#1	F_4	F_4
#2	F_1	F_4F_1
#3	F_3	$F_4F_1F_3$
#4	F_7	$F_4F_1F_3F_7F_5$
#5	F_{10}	$F_4F_1F_3F_7F_5F_{10}$
….	……..	……..
#12	F_{12}	$F_4F_1F_3F_7F_5F_{10}F_8F_2F_{11}F_6F_{9\ldots.}F_{12}$

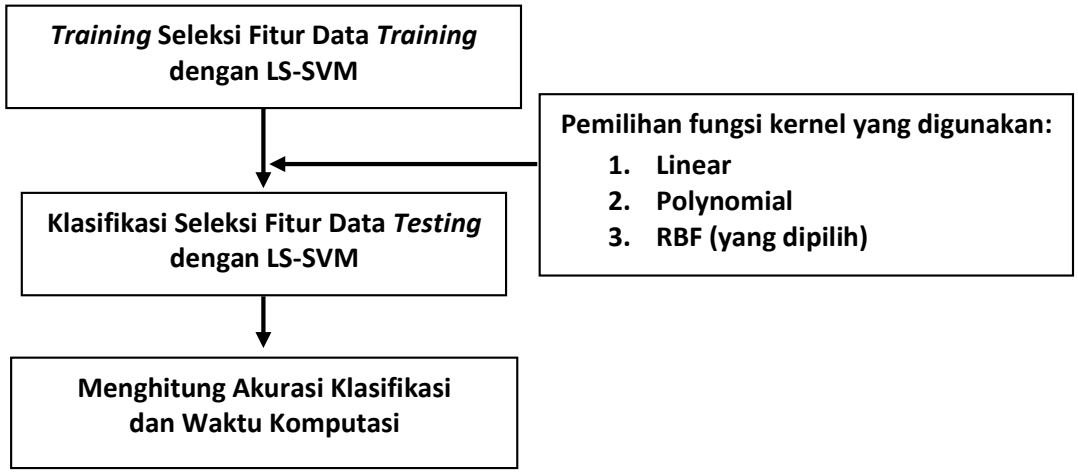


Gambar 3.2. Tahap Seleksi Fitur

3.2.5 Klasifikasi Seleksi Fitur dengan LS-SVM

Tahapan setelah seleksi fitur pada metode penelitian adalah melakukan klasifikasi seleksi fitur dengan LS-SVM dengan pemilihan kernel. Data *training* untuk masing-masing seleksi fitur yang dihasilkan selanjutnya di-*training* dengan LS-SVM. Proses *training* dilakukan dengan nilai parameter LS-SVM (γ dan σ^2) pada kernel RBF. γ adalah adalah parameter regulerisasi, yang menentukan *trade-off* antara margin maksimum dan kesalahan klasifikasi minimum. Pada beberapa penelitian sebelumnya nilai γ disebut sebagai *C penalty* (Akay, 2009).

Hasil proses dari masing-masing seleksi fitur *training* pada *classifer* LS-SVM digunakan untuk menguji seleksi fitur *data testing* dengan LS-SVM. Hasil klasifikasi berupa *class label* tersebut dibandingkan dengan *class label* sebenarnya. Penelitian ini disebut sebagai *supervised learning* karena *class label* telah diketahui sebelumnya. Pengujian tersebut menggunakan nilai parameter γ dan σ^2 yang sama dengan saat *training.* Hasil pengujian tersebut digunakan untuk proses evaluasi dari sistem yang telah dikembangkan. Proses *training* dilakukan menggunakan fungsi **trainlssvm** dan proses *testing* menggunakan fungsi **latentlssvm** yang telah disediakan oleh *toolbox* Matlab **LS-SVMlab1.5** (Pelckmans 2002, 2003). Tahapan seluruh proses klasifikasi seleksi fitur dengan LS-SVM seperti yang telah dijelaskan tersebut dapat dilihat pada Gambar 3.3.



Gambar 3.3. Tahap Klasifikasi Seleksi Fitur

3.3 Rancangan Uji Coba

Uji coba akan dilakukan menggunakan parameter percobaan dan evaluasi dari klasifikasi dari seleksi fitur.

3.3.1 Parameter Percobaan

Penentuan parameter untuk LS-SVM RBF dilakukan secara *trial and error*. Nilai parameter γ ditentukan dengan nilai 1. Nilai σ^2 ditentukan dengan 0,1. Nilai γ dan σ^2 tersebut merujuk pada *guide* dari *toolbox* Matlab **LS-SVMlab1.5** (Pelckmans 2002, 2003).

Proses penentuan parameter ini menggunakan seluruh fitur hasil ekstraksi yaitu 12 fitur. Tingkat akurasi adalah perbandingan jumlah *class label* yang benar hasil prediksi dibandingkan dengan jumlah *class label* sesungguhnya. Sedangkan waktu komputasi adalah waktu yang diperlukan untuk proses *training* dan *testing*.

3.3.2 Ujicoba

Ujicoba dilakukan dengan perbandingan klasifikasi LS-SVM dengan SVM serta pemilihan kernel (linear, polynomial, dan RBF) baik menggunakan seleksi fitur maupun tanpa menggunakan seleksi fitur

Seleksi fitur dilakukan dengan *F-Score*. Setelah proses *training* dan *testing* dari seleksi fitur *F-Score*. Data yang dihasilkan selama proses uji coba adalah tingkat akurasi, sensitivitas, spesifitas, waktu komputasi, dan kombinasi fitur.

3.3.3 Evaluasi

Evaluasi dilakukan dengan tujuan untuk mengevaluasi efektivitas metode dan sistem yang telah dibuat. Evaluasi dilakukan terhadap tingkat akurasi klasifikasi dan tingkat kesalahan klasifikasi. Ukuran atau parameter yang digunakan untuk evaluasi antara lain akurasi klasifikasi, sensitivitas, spesifisitas, dan matriks konfusi (*confusion matrix*). Matriks konfusi berisi informasi tentang klasifikasi yang sebenarnya dan yang diperkirakan dari hasil sistem klasifikasi. Tabel 3.2 menunjukkan matriks konfusi

untuk dua kelas klasifikasi. Akurasi klasifikasi, sensitivitas, spesifisitas, nilai prediksi positif dan nilai prediksi negatif dapat didefinisikan menggunakan elemen-elemen matriks konfusi sebagai berikut:

- Klasifikasi akurasi (%) = $\frac{TP+TN}{TP+FP+FN+TN}$, (3.1)
- Sensitivitas (%) = $\frac{TP}{TP+FN}$, (3.2)
- Spesifisitas (%) = $\frac{TN}{FP+TN}$, (3.3)
- Nilai prediksi positif = $\frac{TP}{TP+FP}\times 100$, (3.4)
- Nilai prediksi negatif = $\frac{TN}{FN+TN}\times 100$. (3.5)

Selain itu juga analisis data hasil uji coba dilakukan terhadap waktu komputasi dari kombinasi fitur yang ada. Evaluasi dilakukan dengan melihat perubahan waktu komputasi yang dibutuhkan oleh tiap kombinasi fitur.

Tabel 3.2. Matriks Konfusi

Aktual	Prediksi	
	Positif	Negatif
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (TP)	*True Negative* (TN)

[Halaman ini sengaja dikosongkan]

BAB IV
HASIL DAN PEMBAHASAN

Bab ini menjelaskan lingkungan uji coba, uji coba, dan evaluasi. Uji coba dibagi menjadi tiga sub bab, yaitu proses ekstraksi fitur, perangkingan dan kombinasi fitur, dan Klasifikasi.

4.1 Lingkungan Uji Coba

Spesifikasi perangkat keras dan lunak yang digunakan dalam implementasi adalah komputer dengan prosesor Intel$^{(R)}$Core i3 M360 @2.53 GHz, memori 2 GB, harddisk 500 GB, sistem operasi Windows 7 Ultimate 32bit dan Matlab (R2013a) dilengkapi dengan *toolbox* LS-SVMlab 1.5 (Pelckmans 2002, 2003).

4.2 Uji Coba

Uji coba dilakukan terhadap sistem yang telah dikembangkan. Uji coba dilakukan dalam empat tahap, yaitu :

1. Uji coba terhadap proses ekstraksi fitur dengan GLCM untuk mengetahui hasil proses ekstraksi fitur.
2. Uji coba terhadap proses perangkingan dan kombinasi fitur menggunakan *F-Score*
3. Uji coba terhadap klasifikasi LS-SVM serta kombinasi seleksi fitur untuk mengetahui kombinasi dari fitur-fitur yang menghasilkan akurasi yang terbaik serta dilakukan dengan perbandingan.

4.2.1 Ekstraksi Fitur

Proses ekstraksi fitur dilakukan terhadap 88 data *training* dan 40 data *testing* yang mana setiap data menghasilkan 12 fitur menggunakan metode GLCM. Dari hasil ekstraksi 12 fitur tersebut yang nantinya dijadikan untuk seleksi fitur klasifikasi. Tabel 4.1 adalah salah satu ektraksi fitur.

Tabel 4.1 Hasil ekstraksi fitur dari salah satu citra dataset *training*

No	Fitur ciri	Nilai
1	*Energy*	0.995740
2	*Correlation*	0.057935
3	*Contrast*	0.517210
4	*Autocorrelation*	14.028000
5	*Cluster_Prominence*	1461.300000
6	*Cluster_Shade*	132.650000
7	*Sum_variance*	46.201000
8	*Difference_entropy*	0.093267
9	*Homogeneity*	0.990750
10	*Sum_average*	5.389000
11	*Sum_of_squares*	13.972800
12	*Sum_entropy*	1.0216800

4.2.2 Seleksi Fitur

Proses seleksi fitur dilakukan dengan menghitung nilai *F-Score* dari data *training*. Perhitungan nilai *F-Score*. Dari perhitungan nilai *F-Score* diperoleh dua belas fitur. Nilai masing-masing *F-Score* yang telah dihasilkan diurutkan secara menurun (*descending*) dengan fungsi **sort** yang telah disediakan oleh **Matlab**. Hasil pengurutan tersebut digunakan untuk menentukan kombinasi fitur yang akan digunakan baik untuk proses *training* maupun *testing*.

Contoh hasil perhitungan nilai *F-Score* Tabel 4.2. Berdasarkan tabel *F-Score* yang sudah diurutkan tersebut dibuat kombinasi fitur seperti terlihat pada Tabel 4.3 untuk *F-Score*. . Dari Tabel 4.3 dapat dilihat bahwa kombinasi fitur model #1, dibuat dari fitur 1 (F_2), karena F_1 mempunyai nilai *F-Score* terbesar. Sedangkan kombinasi fitur model #2, dibuat dari F_1 dan F_{11}, karena F_2 dan F_{11} mempunyai nilai *F-Score* terbesar pertama dan kedua. Demikian seterusnya sehingga didapatkan 12 macam kombinasi fitur untuk *F-Score*.

Tabel 4.2. Nilai *F-Score* untuk masing-masing Fitur

No. Fitur	Fitur	*F-Score*
1	F2	0.021877
2	F11	0.015198
3	F8	0.010540
4	F1	0.004878
5	F5	0.004833
6	F3	0.004129
7	F6	0.002604
8	F9	0.001306
9	F10	0.000626
10	F7	0.000183
11	F12	0.000028
12	F4	0.000010

Tabel 4.3. Kombinasi Fitur untuk *F-Score*

Model	Jumlah Fitur	*F-Score*	Kombinasi Fitur
#1	1	0.021877	F2
#2	2	0.015198	F2F11
#3	3	0.010540	F2F11F8
#4	4	0.004878	F2F11F8F1
#5	5	0.004833	F2F11F8F1F5
#6	6	0.004129	F2F11F8F1F5F3
#7	7	0.002604	F2F11F8F1F5F3F6
#8	8	0.001306	F2F11F8F1F5F3F6F9
#9	9	0.000626	F2F11F8F1F5F3F6F9F10
#10	10	0.000183	F2F11F8F1F5F3F6F9F10F7
#11	11	0.000028	F2F11F8F1F5F3F6F9F10F7F12
#12	12	0.000010	F2F11F8F1F5F3F6F9F10F7F12 F4

4.2.3 Uji Coba Klasifikasi

Uji coba menggunakan SVM maupun LS-SVM dengan penentuan kernel linear, Polynimial, dan RBF dengan parameter γ sebesar 1 dan nilai $\sigma 2$ sebesar 0,1 digunakan untuk membandingkan akurasi, sensifitas dan spesifitas dan waktu komputasi yang terbaik menggunakan seleksi fitur maupun tanpa menggunakan seleksi fitur. Hasil yang didapat dapat dilihat pada tabel 4.5

Tabel 4.4 Hasil Klasifikasi terbaik tanpa menggunakan seleksi fitur

Klasifikasi	Akurasi (%)	Spesifitas (%)	Sensivitas (%)	Waktu (detik)
SVM-linear	35	13	100	0.037
SVM-Polynimial	70	70	70	0.628
SVM-RBF	**97.5**	**100**	**90**	**0.043**
LS-SVM linear	57.5	66.6	30	0.234
LS-SVM Polynomial	75	100	0	0.054
LS-SVM RBF	**97.5**	**100**	**90**	**0.047**

Tabel 4.5 Hasil Klasifikasi terbaik menggunakan seleksi fitur

Klasifikasi	Model Fitur	Akurasi (%)	Spesifitas (%)	Sensivitas (%)	Waktu (detik)
SVM-linear	7	40	20	100	0.016
SVM-Polynimial	11	72.5	73.3	70	0.512
SVM-RBF	**8**	**97.5**	**100**	**90**	**0.026**
LS-SVM linear	1	75	100.0	0	0.014
LS-SVM Polynomial	1	75	100	0	0.015
LS-SVM RBF	**10**	**97.5**	**100**	**90**	**0.023**

4.3 Evaluasi

Sesuai dengan tujuan penelitian ini yaitu menghasilkan sistem klasifikasi massa pada citra mammografi menggunakan kombinasi seleksi fitur, dan LS-SVM, maka evaluasi dilakukan terhadap tingkat akurasi, waktu komputasi dan model kombinasi yang dihasilkan. Evaluasi dilakukan dengan membandingkan hasil uji coba dengan SVM.

4.3.1 Tingkat Akurasi Klasifikasi

Berdasarkan hasil ujicoba yang dilakukan akurasi terbaik terletak pada klasifikasi SVM dan LS-SVM menggunakan kernel RBF dengan tingkat akurasi 97,5%. Pada tabel 4.4 dan 4.5 dapat disimpulkan bahwa sensitivitas hasil klasifikasi lebih kecil dari tingkat spesifitas. Perbedaan tingkat sensitifitas dan spesifitas ini terjadi karena persamaan nilai varian antara fitur hasil ekstraksi citra kanker ganas dengan citra kanker jinak. Hasil konfusi klasifikasi terbaik bisa dilihat pada tabel 4.6. dari hasil klasifikasi ada satu data testing jenis kanker ganas tidak bisa diklasifikasi karena nilai varian dataset mirip dengan nilai varian pada dataset kanker jinak.

Dari hasil perbandingan diatas bahwa menggunakan seleksi fitur mampu meningkatkan akurasi klasifikasi dikarenakan tidak semua fitur digunakan. Namun, untuk kombinasi seleksi fitur pada LS-SVM dengan kernel RBF tingkat akurasi terbaik nilainya stabil mulai dari kombinasi fitur model #10 #11 #12. Selain itu juga pada klasikasi SVM dengan penggunaan kernel RBF tingkat akurasi terbaik didapat pada kombinasi seleksi fitur model #8 #9 #10 #11 #12.

Tabel 4.6. Matriks Konfusi untuk Hasil Klasifikasi Terbaik

Aktual	Prediksi	
	Ganas	Jinak
Ganas	9	1
Jinak	0	30

4.3.2 Waktu Komputasi

Bentuk tabulasi data waktu yang dibutuhkan untuk proses klasifikasi (proses *training* dan *testing*) terhadap model kombinasi dari uji coba untuk *F-Score* dan tanpa seleksi fitur masing-masing dapat diketahui yaitu rata-rata waktu komputasi F-Score dengan LS-SVM yaitu 0,023 detik dan untuk LS-

SVM tanpa seleksi fitur diketahui yaitu 0,047 detik. Rata-rata waktu komputasi F-Score dengan SVM membutuhkan waktu 0,026 detik dan SVM tanpa seleksi fitur membutuhkan waktu rata-rata 0,046. Hal tersebut dibuktikan bahwa seleksi fitur sangat bepengaruh terhadap waktu komputasi.

4.3.3 Model Kombinasi

Evaluasi model kombinasi ini bertujuan untuk menguji apakah model kombinasi dengan tingkat akurasi tertinggi tersebut merupakan kombinasi fitur yang tetap. Model kombinasi untuk klasifikasi SVM RBF #8 yaitu F2F11F8F1F5F3F6F9F10F7 dan untuk klasifikasi LS-SVM RBF #10 yaitu F2F11F8F1F5 F3F6F9. Hasil lebih lengkap bisa dilihat dilampiran

4.3.4 Hubungan Kernel dengan Tingkat Akurasi Klasifikasi

Hubungan kernel dengan tingat akurasi klasifikasi sangat berpengaruh terhadap tingkat akurasi. Terbukti bahwa penggunaan kernel RBF mampu menghasilkan akurasi terbaik daripada penggunaan kernel linear dan Polynomial. Karena pemilihan kernel akan menentukan *feature space* dimana fungsi klasifier akan dicari. Selagi fungsi kernelnya lagimate, SVM maupun LS-SVM akan beroperasi secara benar meskipun tidak tahu map apa yang digunakan untuk satu per satu data.

4.3.5 Hubungan Kernel dengan Waktu Komputasi

Hubungan kernel dengan waktu komputasi sangat berpengaruh. Itu terbukti bahwa waktu yang dihasilkan untuk klasifikasi masing-masing kernel mempunyai nilai waktu yang berbeda. Bisa dilihat pada tabel 4.5 dan 4.6 terbukti bahwa penggunaan kernel RBF waktu yang dibutuhkan lebih baik daripada kernel Linear dan Polynomial pada klasifikasi SVM dan LS-SVM dengan seleksi fitur maupun tanpa seleksi fitur.

4.3.6 Hubungan Jumlah Fitur dengan Tingkat Akurasi Klasifikasi

Hubungan jumlah fitur dengan tingkat akurasi pada klasifikasi LS-SVM RBF dapat dilihat pada Gambar 4.5, 4.6. Bahwa jumlah fitur berpengaruh terhadap tingkat akurasi yang dihasilkan. Semakin banyak fitur yang digunakan semakin tinggi tingkat akurasi yang dihasilkan, tetapi setelah mencapai model #10, tingkat akurasi yang dihasilkan cenderung tetap sampai model #12 begitu juga untuk klasifikasi SVM RBF akurasi terbaik pada model #8, tingkat akurasi yang dihasilkan cenderung tetap sampai model #12.

4.3.7 Hubungan Jumlah Fitur dengan Waktu Komputasi

Hubungan jumlah fitur dengan waktu komputasi pada klasifikasi SVM RBF diperoleh dari hasil uji coba klasifikasi dengan kombinasi seleksi fitur. Menunjukkan bahwa jumlah fitur berpengaruh terhadap waktu komputasi.

4.3.8 Hubungan Parameter γ dan σ^2 dengan Tingkat Akurasi

Hubungan penggunaan nilai parameter terbukti mempengaruhi tingkat akurasi. Hasilnya dapat dilihat pada tabel 4.4 dan tabel 4.5. Hasil ujcoba diketahui bahwa nilai hasil terbaik didapat pada kernel RBF dengan penggunaan nilai gamma 1 dan sigma 0,1

4.3.9 Hubungan Parameter γ dan σ^2 dengan Waktu Komputasi

Hubungan penggunaan nilai parameter terbukti mempengaruhi waktu komputasi. Hasilnya dapat dilihat pada tabel 4.4 dan tabel 4.5. Hasil ujcoba diketahui bahwa nilai waktu komputasi terbaik dengan penggunaan nilai gamma 1 dan sigma 0,1 pada kernel RBF.

[Halaman ini sengaja dikosongkan]

BAB 5
KESIMPULAN DAN SARAN

Bab ini menguraikan kesimpulan yang dapat diambil dari penelitian ini dan saran-saran yang dapat digunakan untuk pengembangan selanjutnya.

5.1 Kesimpulan

1. Penggunaan fitur ciri dari ektraksi fitur GLCM untuk input klasifikasi masih belum mencapai akurasi maksimal.
2. Penambahan metode kombinasi seleksi fitur, pemilihan kernel, serta penggunaan parameter terbukti berpengaruh pada tingkat akurasi dan penurunan waktu komputasi.
3. Klasifikasi LS-SVM dengan seleksi fitur maupun tanpa seleksi fitur yaitu sama, begitu juga klasifikasi SVM dengan penggunaan kernel RBF yaitu nilai akurasi tertinggi 97,5% daripada dengan kernel Linear maupun Polynimial.

5.2 Saran

1. Diperlukan penambahan atau penggunaan fitur ektraksi fitur GLCM lainnya.
2. Pengaruh parameter γ dan σ^2 terhadap tingkat akurasi dan waktu komputasi dapat diperluas dengan menambah rentang nilai γ dan σ^2 yang digunakan.
3. Dibutuhkan perluasan dengan penggunaan *K-fold validation* untuk mengetahui pengaruhnya terhadap tingkat akurasi dan waktu komputasi.

Daftar Pustaka

Aarthi, R., Divya, K., Komala, N., & Kavitha, S. (2011). "Application of Feature Extraction and Clustering in Mammogram Classification using Support Vector Machine", *Advanced Computing (ICoAC)Third International Conference onIEEE*,hal. 62–67.

Akay, M. F. (2009),"Support vector machines combined with feature selection for breast cancer diagnosis",*Expert Systems With Applications*, vol. 36no. 2, hal. 3240–3247.

Albregtsen, F. (2008). :"Statistical Texture Measures Computed from Gray Level Coocurrence Matrices". *Image Processing Laboratory Department of Informatics*. University of Oslo,hal 1-14.

Anonymous. Mammography. Tersedia di www.wikipedia.org (diaksespada 3 Maret 2015)

B. Sahiner, N. Petrick, H.P. Chan (2001) "Computer-aided characterization of mammographic massa: accuracy of mass segmentation and its effects on characterization", *IEEE Trans. Med. Imaging*, vol. 20, no. 12, hal. 1275–1284.

Begg, R., Lai, D.T.H. & Palaniswami, M. (2008). Computational intelligence in biomedical engineering. First Edition. CRC Press.

Chen, Y. W., & Lin, C. J. (2005). Combining SVMs with various feature selection strategies. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/ features.pdf.

Cortes, C., & Vapnik, V. (1995). "Support vector networks. Machine Learning", vol. 20,no.3, hal. 273–297.

Disha ED, Kërliu SM, Ymeri H, Kutllovci A. (2009). "Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density". *Bosnian Journal of Basic Medical Sciences*, vol. 9, no. 2, hal. 131-36.

E.d. Pisano, F. Shtem, (1993). "Image processing and computer aided diagnosis in digital mammography",*a clinical perspective, Int. J. Pattern Recog. Artific. Intell.*Vol. 7,no. 6, hal. 1493–1503.

Eurostat (2002). Healt statistic atlas on mortaly in the European Union, Official J Eur Union.

H.C. Zuckerman (1987). "The role of mammography in the diagnosis of breast cancer", *in: I.M. Ariel, J.B. Clearly (Eds.), Breast Cancer: Diagnosis and Treatment*, McGraw-Hill, New York, , hal. 152–172.

H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai, H.N. Du (2006) "Approaches for automated detection and classification of massa in mammograms", *Pattern Recognition*, vol. 39, hal. 646-668.

Holmes EB. Ionizing radiation exposure with medical imaging. Available at Medscape Radiology, www.Medscape.org (diakses pada15 Maret 2015)

Holmes EB. Ionizing radiation exposure with medical imaging. Available at Medscape Radiology, www.Medscape.org (diakses pada 15 maret 2015)

Islam M.J, Ahmadi M, Sid-Ahmed A.M (2010), *"*An Efficient Automatic Mass Classification Method in Digitized Mammograms Using Artificial Neural Network", *International Journal of Artificial Intelligence & Applications (IJAIA)*, vol.1, no.3, hal. 1–13.

Jain, R., Kasturi, R., & Schunck, B. G.(1995). "Machine vision". *McGraw-Hill, Inc. Chapter 7 Texture.* (n.d.), hal 234–248.

Kom, G., Tiedeu, A., & Kom, M. (2007). "Automated detection of massa in mammograms by local adaptive thresholding", *Computers in Biology and Medicine*, vol.37, hal. 37–48.

Liu, X., Tang, J (2014). "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method". *Systems Journal, IEEE*,vo. 8, no. 3, hal. 910 – 920.

Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., & Denton, E. R. E. (2010). "A review of automatic mass detection and segmentation in mammographic images". *Medical Image Analysis*, vol. 14, no. 2, hal. 87–110.

P. Undrill, R. Gupta, S. Henry, M. Downing. (1996). "Texture analysis and boundary refinement to outline mammography massa", in: *Proceedings of theIEEE Colloquium on Digital Mammography*, vol.5, hal. 1-6.

Pelckmans K., Suykens J.A.K., Van Gestel T., De Brabanter J., Lukas L., Hamers B., De Moor B. & Vandewalle J. (2002). LS-SVMlab : a Matlab/C toolbox for

Least Squares Support Vector Machines. Internal Report 02-44, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), (presented at NIPS2002 Vancouver in the demo track), 2002.

Pelckmans, K., Suykens, J.A.K., Van Gestel, T., De Brabanter, J., Lukas, L., Hamers B., De Moor, B. & Vandewalle, J. (2003). LS-SVMlab Toolbox User's Guide version 1.5. Katholieke Universiteit Leuven Department of Electrical Engineering, ESAT-SCD-SISTA Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, http://www.esat.kuleuven.ac.be/sista/lssvmlab/ ESAT-SCD-SISTA Technical Report 02-145.

S. Timp and N. Karssemeijer. (2006). "Interval change analysis to improve computer aided detection in mammography," *Medical Image Analysis*, vol. 10, no. 1, hal. 82 – 95.

Sameti, M., Member, S., Ward, R. K., & Morgan-parkes, J. (2009). Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer, *signal processing: IEEE*, vol. 3, no. 1,hal. 46–52.

Suykens, J. A. K., & Vandewalle, J (1999). "Least squares support vector machine *classifier*s". *Neural Processing Letters*, vol. 9, no.3, hal. 293–300.

Tai, S., Chen, Z., & Tsai, W. (2014). "An Automatic Mass Detection System in Mammograms based on Complex Texture Features", *Biomedical and Health Informatics, IEEE,*vol. 18, no. 2, hal. 618 – 627.

Vapnik, V. (1995). The nature of statistical learning theory. New York: Springer-Verlag.

Yu, L. & Liu, H. (2003). "Feature selection for high-dimensional data: a fast correlation-based filter solution". *Proceedings of the Twentieth International Conference on Machine Learning,ICML*, Washington DC.

.

BIODATA

Muhammad Imron Rosadi, Anak ke-3 dari Pasangan Bpk. M.Khozin dan Ibu Kholifah pendidikan TK- SD Tunggulwulung Pandaan kemudian lulus SD berangkat mondok ke Ponpes Ngalah sengonagung Purwosari pasuruan dibawah asuhan KH. Sholeh Bahrudin di pondok tersebut saya menempuh Pendidikan Formal dan Nonformal mulai MTs Darut Taqwa lulus 2004, Jurusan TKJ SMK Darut Taqwa lulus 2007, S1 Jurusan Teknik Informatika Univ Yudharta Lulus 2011 melanjutkan pendidikan Pasca Sarjana di Jurusan Teknik Informatika FTIf - ITS Surabaya mengambil Bidang Minat Komputasi Cerdas dan Visi.