 pISSN: 2087-8893
eISSN: 2527-3671

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

Muhammad I. Rosadi[1)], Agus Z. Arifin[2)], dan Anny Yuniarti[3)]
[1)]Teknik Informatika, Fakultas Teknik, Universitas Yudharta, Pasuruan
Kompleks Ponpes Darul 'Ulum Peterongan Jombang 61481
[2,3)]Magister Teknik Informatika, Fakultas Teknologi Informasi, ITS, Surabaya
Jl. Yudharta No. 7, Sengon Agung, Purwosari, Pasuruan, Jawa Timur 67162

e-mail: imron.rosadi13@mhs.if.its.ac.id[1)], agusza@cs.its.ac.id[2)], anny@if.its.ac.id[3)]

ABSTRAK

Kanker payudara adalah penyakit yang paling umum diderita oleh wanita pada banyak negara. Pemeriksaan kanker payudara dapat dilakukan menggunakan citra *Mammogram*. Analisis CAD (*Computer-Aded Detection system*) yang telah dikembangkan adalah ekstraksi fitur GLCM *(Gray Level Co-occurrence Matrix)*, reduksi/seleksi fitur dan SVM (*Support vector Machine*). Pada SVM 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 yang terbaik. Jumlah fitur dan nilai parameter kernel yang diperlukan saling mempengaruhi, sehingga seleksi fitur diperlukan dalam membangun sistem klasifikasi. Penelitian ini bertujuan untuk mengklasifikasi massa pada citra *Mammogram*berdasarkan dua kelas yaitu kelas kanker jinak dan kelas kanker ganas. Ekstraksi fitur menggunakan 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*. 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 uji coba didapatkan kesimpulan bahwa penggunaan kombinasi seleksi fitur sangat berpengaruh terhadap tingkat akurasi. Akurasi terbaik didapat dengan menggunakan LS-SVM dengan *kernel* RBF *(Radial Basis Function)* dan SVM RBF baik dengan kombinasi seleksi fitur, maupun tanpa kombinasi seleksi fitur dengan nilai akurasi yaitu 97,5%. Selain itu, seleksi fitur mampu mengurasi waktu komputasi.
Kata kunci: F-Score, GLCM, kanker payudara, LS-SVM.

ABSTRACT

Breast cancer is the most common disease suffered by women in many countries. Breast cancer screening can be done using a mammogram image. Analysis CAD (Computer-Aded Detection system) that has been developed is the feature extraction by GLCM (Gray Level Co-occurrence Matrix), reduction / feature selection and SVM (Support Vector Machine). In SVM and LS-SVM (Least Square Support Vector Machine), 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. This study aims to classify the mass on the mammogram image based on two categories of classes. They are benign and malignant cancer. Feature extraction is done by using GLCM. The results of the feature extraction process is then selected based on F-Score. F-Score is obtained by calculating the value of the discriminant feature extraction results data between the data of two classes in the training data. Then, the F-Score of each feature is 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 test results it was concluded that the use of a combination of feature selection affects the level of accuracy. Best accuracy is obtained by using LS-SVM with kernel RBF (Radial Basis Function) and SVM RBF well with a combination of feature selection, or without the combination of feature selection with accuracy score is 97.5%. In addition, the feature selection is able to reduce the computation time.
Keywords: *Breast Cancer, F-Score, GLCM, LS-SVM.*

I. PENDAHULUAN

KANKER payudara dianggap sebagai masalah kesehatan yang utama di negara-negara barat, dan merupakan kanker yang paling umum di kalangan perempuan di Uni Eropa [1]. Di Amerika Serikat, sekitar 39.520 perempuan meninggal dunia disebabkan kanker tersebut. Kemajuan pengobatan, peningkatan kesadaran, dan deteksi sejak dini menghasilkan angka kematian menurun. *Mammografi* adalah alat *screening* yang paling efektif untuk mendeteksi kanker payudara [2]. Seorang ahli radiologi biasanya memeriksa *Mammogram* untuk memeriksa tanda-tanda kanker. Secara *Mammografi*, kanker payudara dikenali dengan keberadaan lesi massa atau biasa disebut massa, dan mikrokalsifikasi [3]. 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 [4].

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 [5]. 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*. Metode *Gray Level Co-ocurrence Matrix* (GLCM) adalah cara ekstraksi fitur tekstur statistik urutan kedua. Pendekatan ini telah digunakan dalam beberapa aplikasi [6].

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 [7].

Penelitian tentang pengaruh seleksi fitur terhadap peningkatan performa klasifikasi telah dilakukan. Hasil menunjukkan peningkatan akurasi yang signifikan dibandingkan klasifikasi tanpa penerapan seleksi fitur. Chen dan Lin [8] mengusulkan metode kombinasi seleksi fitur dengan SVM. 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 [8]. Kombinasi metode SVM dan F-Score telah digunakan untuk mendiagnosis penyakit kanker payudara menggunakan dataset statistik dan menghasilkan tingkat akurasi yang lebih baik dari LS-SVM [9]. Aarthi 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 [10].

Dari penelitian sebelumnya diketahui bahwa jumlah fitur yang optimal adalah salah satu dari tiga masalah yang muncul pada SVM. Tiga masalah pada SVM termasuk juga LS-SVM adalah: 1) Bagaimana memilih fungsi kernel; 2) Menentukan berapa jumlah fitur *input* yang optimal; 3) Bagaimana menentukan parameter kernel terbaik. Masalah-masalah tersebut penting karena jumlah fitur dan nilai parameter kernel yang diperlukan saling mempengaruhi. Dengan demikian, seleksi fitur diperlukan dalam membangun sistem klasifikasi, karena dengan pembatasan/pengurangan jumlah fitur *input* dalam *classifier*, maka akan mengurangi kompleksitas komputasi. 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. Untuk masalah klasifikasi *dua spiral* yang kompleks dapat ditemukan dengan LS-SVM RBF dengan kinerja yang sangat baik dan komputasi rendah [11].

Berdasarkan uraian dari penelitian sebelumnya, dalam penelitian ini diusulkan 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.

II. TINJAUAN PUSTAKA

A. 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. kisaran 22% dari semua jenis kanker yang terjadi pada wanita adalah kanker payudara. Penyakit ini terjadi di mana 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 [12].

B. *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). Tujuan dari *Mammografi* adalah untuk deteksi dini kanker payudara, biasanya melalui deteksi karakteristik lesion dan atau bentuk kalsifikasi [13]. *Mammografi* memegang peranan penting dalam deteksi dini kanker payudara, hal ini karena *Mammografi* mampu mendeteksi hampir 75% kanker payudara kurang lebih satu tahun sebelum pasien merasakan gejala. Terdapat dua tipe pemeriksaan *Mammografi*, yaitu skrining dan diagnostik.

 pISSN: 2087-8893
eISSN: 2527-3671

C. *Gray Level Co-occurance Matric* (GLCM)

Metode *Gray Level Coocurrence Matrix* (GLCM) adalah cara ekstraksi fitur tekstur statistik urutan kedua. Pendekatan ini telah digunakan dalam beberapa aplikasi [6]. GLCM adalah matriks di mana jumlah baris dan kolom sama dengan jumlah tingkat abu-abu (G) dalam gambar. Elemen matriks $P(i, j|\Delta x, \Delta y)$ 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 $P(i, j|d, \theta)$ 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).

Pengukuran nilai tekstur yang digunakan didasarkan dengan notasi berikut: G adalah jumlah tingkat abu-abu yang digunakan, μ adalah nilai rata-rata dari *P*, μ*x*, μ*y*, σ*x*, 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 (*Energy*)
   Menunjukkan ukuran dari *local homogeneity* dan merupakan kebalikan dari entropi. Untuk mendapatkan energi digunakan formula pada Persamaan 1.

$$Energy = \sum_{i,j} \mathrm{P}(i,j)^2 \tag{1}$$

2. Kontras (*Contrast*)
   Formula *Contrast* dapat dilihat pada Persamaan 2.

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

3. Homogenitas (*Homogeneity*), *Angular Second Moment* (ASM)
   ASM adalah ukuran homogenitas dari suatu gambar. Lebih detail dapat dilihat pada Persamaan 3.

$$ASM = \sum_{i=0}^{G-1} \sum_{j=0}^{G1} \{p \quad (i,j)\}^2 \tag{3}$$

4. Korelasi (*Correlation*)
   Korelasi menunjukkan ketergantungan linear derajat keabuan dari piksel-piksel yang saling bertetangga dalam suatu citra abu-abu. Lebih detail dapat dilihat pada Persamaan 4.

$$Correlation = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\{ixj\}xP(i,j) - \{\mu_x x \mu_y\}}{\sigma_x x \sigma_y} \tag{4}$$

   di mana:
   μ_x= nilai rata-rata elemen kolom pada matriks *P*θ (*i*,*j*)
   μ_y= nilai rata-rata elemen baris pada matriks *P*θ (*i*,*j*)
   σ_x= nilai standar deviasi elemen kolom pada matriks *P*θ (*i*,*j*)
   σ_y= nilai standar deviasi elemen kolom pada matriks *P*θ (*i*,*j*)

5. *Autocorrelation*
   Formula *Autocorrelation* dapat dilihat pada Persamaan 5.

$$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right] \tag{5}$$

6. Rata-rata jumlah (*Sum Average*)
   Formula *Sum Average* dapat dilihat pada Persamaan 6.

$$AVER = \sum_{I=0}^{2G-2} Ip_{X+Y}(i) \tag{6}$$

7. Entropi Jumlah (*Sum Entropy*)
   Formula *Sum Entropy* dapat dilihat pada Persamaan 7.

$$SEN = \sum_{i=0}^{2G-2} p_{x+y}(i) \log(p_{x+y}(i)) \tag{7}$$

8. Sum Varians (Sum *Variance*)
   Formula *Sum Variance* dapat dilihat pada Persamaan 8.

$$VARIANCE = \sum_{i=0}^{G-1} \sum_{i=0}^{G1} (k-\mu)^2 p \quad (i,j) \tag{8}$$

9. Entropi Selisih (*Difference Entropy*)
   Formula *Difference Entropy* dapat dilihat pada Persamaan 9.

$$DENT = -\sum_{i=0}^{G-1} P_{x+y}(i)\, log(p_{x+y}(i)) \tag{9}$$

10. *Sum of Squares*
    Formula *Sum of Squares* dapat dilihat pada Persamaan 10.

$$VARIANCE = \sum_{i=0}^{G-1} \sum_{i=0}^{G1} (k-\mu)^2 p \quad (i,j) \tag{10}$$

11. *Cluster Shade*
    Formula *Cluster Shade* dapat dilihat pada Persamaan 11.

$$\mathrm{SHADE} = \sum_{\mathrm{i}=0}^{\mathrm{G}-1} \sum_{\mathrm{j}=0}^{\mathrm{G}-1} \{\mathrm{i} + \mathrm{j} - \mu_{\mathrm{x}}\mu_{\mathrm{y}}\}^3 \mathrm{x}\, \mathrm{P(i,j)} \tag{11}$$

12. *Cluster prominence*
    Formula *Cluster prominence* dapat dilihat pada Persamaan 12.

$$\mathrm{PROM} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \{\mathrm{i} + \mathrm{j} - \mu_{\mathrm{x}}\mu_{\mathrm{y}}\}^4 \mathrm{x}\, \mathrm{P(i,j)} \tag{12}$$

D. F-Score

F-score adalah teknik sederhana yang mengukur diskriminan dua himpunan bilangan real. Pada vektor *training* x_k, dengan $k = 1, 2, \ldots, m$, jika jumlah kasus positif dan negatif adalah n_+ dan n_-, maka F-score masing-masing fitur i didefinisikan sebagai Persamaan 13.

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - x_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_+ - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{13}$$

di mana x_i, $x_i^{(+)}$, $x_i^{(-)}$ 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 [8].

E. *Support Vector Machines* (SVM)

SVM yang diusulkan oleh Vapnik [14] [15]telah dipelajari secara ekstensif untuk klasifikasi, regresi dan estimasi kepadatan. Gambar 1 adalah arsitektur SVM. SVM memetakan pola *input* ke ruang fitur dimensi yang lebih tinggi melalui pemetaan nonlinier berdasar teori yang dipilih. Bidang pemisah linier ini kemudian dibangun dalam ruang fitur dimensi tinggi. Dengan demikian, SVM adalah *linear classifier* di ruang parameter, tapi itu menjadi *nonlinear classifier* sebagai akibat dari pemetaan *nonlinear* dari ruang pola *input* ke ruang fitur dimensi tinggi. Bila data *training* 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 linier di ruang fitur, maka fungsi keputusan dapat ditentukan seperti pada Persamaan 14.

$$D(x) = w^t g(x) + b, \tag{14}$$

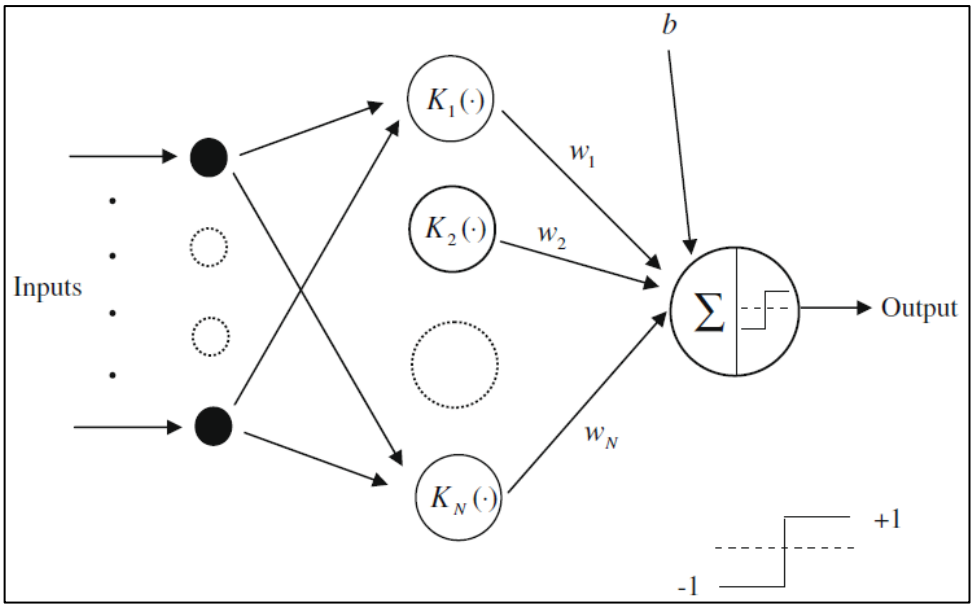
di mana $g(x)$ adalah fungsi pemetaan yang memetakan x ke dalam ruang dimensi 1, w adalah vektor dimensi dan 1, dan b adalah skalar. Untuk memisahkan data secara linier, fungsi keputusan memenuhi kondisi Persamaan 15.

$$y_i(w^t g(x_i) + b) \geq 1 \text{ untuk } i = 1, ..., \mathrm{M}. \tag{15}$$

Jika masalah terpisah secara linier dalam ruang fitur, maka fungsi keputusan yang memenuhi Persamaan 15 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

 pISSN: 2087-8893
eISSN: 2527-3671

dihasilkan dari $|\boldsymbol{D(x)}|/\|\boldsymbol{w}\|$. Sehingga didapatkan *hyperplane* pemisah dengan margin maksimal yang optimal memisahkan *hyperplane*.



Gambar 1. Portofolio aplikasi saat ini

Dengan asumsi bahwa margin adalah ρ, kondisi berikut harus memenuhi Persamaan 16.

$$\frac{y_i D(x_i)}{\|w\|} \geq \rho \text{ untuk } i = 1, ..., M. \quad (16)$$

Hasil perkalian produk dari ρ dan $\|w\|$ adalah tetap dengan Persamaan 17.

$$\rho \|w\| = 1. \quad (17)$$

Untuk mendapatkan *hyperplane* pemisah yang optimal dengan margin maksimal, w dengan $\|w\|$ yang memenuhi Persamaan 17 harus ditemukan. Persamaan 17 mengarahkan ke pemecahan masalah optimasi berikutnya. Dengan meminimalkan Persamaan 18.

$$\frac{1}{2} w^t w, \quad (18)$$

dan mengikuti batasan untuk Persamaan 19.

$$y_i(w^t g(x_i) + b) \geq 1 \text{ untuk } i = 1, ..., M. \quad (19)$$

Bila data latih tidak linier dipisahkan, digunakan *slack variable* ξ_i ke Persamaan 20.

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, \text{ untuk } i = 1, ..., M. \quad (20)$$

Hyperplane pemisah yang optimal telah ditentukan sehingga maksimalisasi dari margin dan meminimalisasi dari kesalahan *training* didapatkan. Dengan meminimalkan Persamaan 21.

$$\frac{1}{2} w^t w + \frac{C}{2} \sum_{i=1}^{n} \xi_i^{\rho} \quad (21)$$

mengikuti batasan Persamaan 22.

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \text{ untuk } i = 1, ..., M \quad (22)$$

di mana C adalah parameter yang menentukan *trade-off* 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 [14] [16]. Namun, fungsi kernel yang tepat untuk

suatu masalah tertentu tergantung pada data, dan sampai saat ini belum ada metode yang dianggap terbaik tentang cara memilih fungsi kernel.

F. *Least Squares Support Vectors Machine* (LS-SVM)

Least Squares Support Vectors Machine (LS-SVM) adalah salah satu modifikasi dari SVM [17]. Jika SVM dikarakteristik oleh permasalahan konveks *quadratic programming* dengan pembatas berupa pertidaksamaan, LS-SVM sebaliknya, diformulasikan 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 [11].

LS-SVM dilatih dengan meminimalkan dengan Persamaan 23.

$$\frac{1}{2}w^t w + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \quad (23)$$

dan mengikuti batasan Persamaan 24.

$$y_i(w^t g(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \text{ untuk } i = 1, ..., M. \quad (24)$$

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 23 dan Persamaan 24 digunakan *Lagrange multiplier*, yaitu dengan Persamaan 25.

$$Q(w, b, \alpha, \xi) = \quad (25)$$

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

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 mendifferensialkan persamaan di atas terhadap w, ξ_i, b, α_i, dan persamaan dihasilkan sama dengan nol [11].

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. Untuk masalah klasifikasi *dua spiral* yang kompleks dapat ditemukan dengan LS-SVM RBF dengan kinerja yang sangat baik dan komputasi rendah [11].

III. METODE

Pada penelitian ini untuk perancangan klasifikasi massa pada citra *Mammogram* menggunakan 4 tahap yaitu, praproses, ekstraksi fitur, seleksi fitur, dan klasifikasi.

A. Dataset

Dataset yang digunakan pada penelitian ini adalah diambil dari *database* mini-MIAS (*Mammographic Image Analysis Society*) yang didigitalkan pada 50 mikron piksel tepi yang telah direduksi menjadi 200 mikron piksel tepi dan setiap gambar dipotong menjadi 1024×1024 piksel. Hanya tampilan MLO yang dianalisis pada penelitian ini. Gambar dirubah ke format *.png. Sstem ini dievaluasi menggunakan 118 massa (68 kanker jinak dan 50 kanker ganas). Untuk pelatihan, menggunakan 88 massa (48 kanker jinak dan 40 kanker ganas), untuk pengujian, meggunakan 40 massa (30 kanker jinak dan 10 kanker ganas).

B. Praproses

Praproses pada penelitian ini dilakukan pemotongan secara manual untuk mendeteksi massa (ROI, *Region of Interest*) secara proporsional. Tujuan proposes ini adalah untuk mengurangi kesalahan dalam proses klasifikasi. Hasil pemotongan citra dari citra asli dapat dilihat pada Gambar 2.

C. 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 *Gray level co-occurrence matric* (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.* Dari kedua belas ciri fitur ini yang nantinya seleksi menggunakan *F-Score*.

pISSN: 2087-8893
eISSN: 2527-3671

(a) (b)

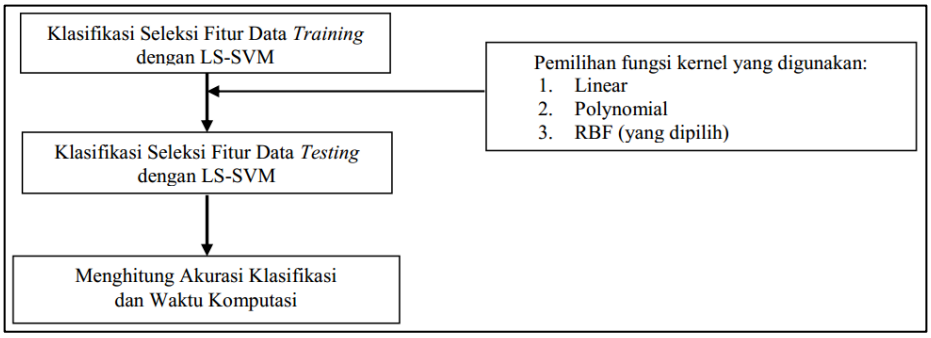
Gambar 2. a) Citra asli, b) Hasil pemotongan

D. Seleksi Fitur

Perhitungan F-Score dilakukan baik data *training* maupun *testing*, sehingga kombinasi seleksi fitur yang dihasilkan dari beberapa uji coba adalah sama. Perhitungan nilai F-Score menggunakan Persamaan 2. Nilai F-Score tersebut diurutkan secara *descending* untuk membuat kombinasi fitur yang akan digunakan baik untuk *training* maupun *testing*. Kombinasi fitur pertama dibuat dari fitur dengan nilai F-Score terbesar. Kombinasi fitur kedua dibuat dari fitur dengan nilai F-Score terbesar pertama dan terbesar kedua, dan seterusnya sehingga didapatkan dua belas macam kombinasi fitur. 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), Fitur 12 (F_{12}), dan Fitur 9 (F_9). 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_{12}, F_9). Berdasarkan hasil pengurutan tersebut dapat dibuat 11 kombinasi fitur yaitu F_4, F_4F_1, $F_4F_1F_3$, $F_4F_1F_3F_7$, ..., $F_4F_1F_3F_7F_5F_{10}F_8F_2F_{11}F_6F_9$.

Dua belas macam kombinasi tersebut menjadi *input* pada LS-SVM. Kombinasi fitur model #1 digunakan sebagai *input* pada LS-SVM baik untuk proses *training* maupun *testing*. Proses *training* maupun *testing* tersebut kemudian diulang lagi untuk kombinasi fitur model #2, #3, #4, dan seterusnya sampai dengan model #12.

E. Klasifikasi Kombinasi Fitur dengan LS-SVM



Gambar 3. Tahap Klasifikasi Seleksi Fitur

Setelah fitur dihitung nilai F-Score dan diurutkan secara *descending* langkah berikutnya adalah klasifikasi kombinasi fitur dengan LS-SVM menggunakan kernel yang pada data *training* untuk masing-masing kombinasi fitur yang dihasilkan dilatih dengan LS-SVM (Gambar 3). Proses *training* dilakukan dengan nilai parameter LS-SVM RBF (γ dan σ^2). γ adalah adalah parameter regulerisasi, yang menentukan *trade-off* antara margin maksimum dan kesalahan klasifikasi minimum. Pada beberapa penelitian lain nilai γ disebut sebagai *C penalty*. Sedangkan σ^2 adalah *bandwidth* untuk fungsi kernel RBF. Nilai parameter γ dan σ^2 yang dipilih untuk proses *training* tiap kombinasi fitur adalah yang menghasilkan akurasi tertinggi dan waktu komputasi terendah. Hasil *training* dari masing-masing kombinasi fitur pada *classifier* LS-SVM RBF digunakan untuk menguji kombinasi fitur *data testing*. Hasil prediksi *class label* tersebut dibandingkan dengan *class label* sebenarnya, sehingga penelitian ini termasuk *supervised learning*. Pengujian dilakukan dengan nilai parameter γ dan σ^2 yang sama dengan saat *training* yaitu γ=1 dan σ^2=0,1.

F. Uji Coba

Ujicoba dilakukan dengan membandingkan klasifikasi LS-SVM dengan SVM baik menggunakan seleksi fitur maupun tanpa seleksi fitur. Proses uji coba adalah tingkat akurasi, sensitivitas, spesifitas, waktu komputasi, dan kombinasi fitur.

G. Evaluasi

Evaluasi dilakukan dengan tujuan untuk mengevaluasi efektivitas metode dan sistem yang telah dibuat. Ukuran atau parameter yang digunakan untuk evaluasi antara lain akurasi klasifikasi, sensitivitas, spesifisitas, nilai prediksi positif, dan nilai prediksi negatif. Matriks konfusi berisi informasi tentang klasifikasi yang sebenarnya dan yang diperkirakan dari hasil sistem klasifikasi. Tabel 1 adalah matriks konfusi untuk dua kelas klasifikasi. Akurasi klasifikasi (Persamaan 26), sensitivitas (Persamaan 27), spesifisitas (Persamaan 28), nilai prediksi positif (Persamaan 29), dan nilai prediksi negatif (Persamaan 30) dapat didefinisikan menggunakan elemen-elemen matriks konfusi sebagai berikut:

TABEL I
MATRIKS KONFUSI

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

- Akurasi (%) = $\frac{TP+TN}{TP+FP+FN+TN}$, (26)
- Sensitivitas (%) = $\frac{TP}{TP+FN}$, (27)
- Spesifisitas (%) = $\frac{TN}{FP+TN}$, (28)
- Nilai prediksi positif = $\frac{TP}{TP+FP}\times 100$, (29)
- Nilai prediksi negatif = $\frac{TN}{FN+TN}\times 100$. (30)

IV. HASIL DAN PEMBAHASAN

A. Ektraksi Fitur

Proses ekstraksi fitur dilakukan terhadap seluruh dataset. Proses ektraksi fitur menggunakan metode GLCM dengan menghasilkan dua belas ciri fitur. Contoh hasil ekstraksi fitur salah satu citra dari dataset pada Tabel 2.

TABEL II
HASIL EKSTRAKSI FITUR GLCM

No	Fitur ciri	Nilai
1	*Energy*	0.99574
2	*Correlation*	0.057935
3	*Contrast*	0.51721
4	*Autocorrelation*	14.028
5	*Cluster_Prominence*	1461.3
6	*Cluster_Shade*	132.65
7	*Sum_variance*	46.201
8	*Difference_entropy*	0.093267
9	*Homogeneity*	0.99075
10	*Sum_average*	5.389
11	*Sum_of_squares*	0.80097
12	*Sum_entropy*	7.5991

B. Seleksi Fitur

Seleksi fitur dilakukan dengan menghitung nilai F-Score dari data *training*. Contoh hasil perhitungan nilai F-Score dapat dilihat pada Tabel 2. Berdasarkan tabel F-Score yang sudah diurutkan tersebut dibuat kombinasi fitur seperti terlihat pada Tabel 3.

pISSN: 2087-8893
eISSN: 2527-3671

C. Klasifikasi

Hasil ujicoba klasifikasi LS-SVM dan SVM dengan pemilihan kernel linear, *Polynomial*, dan RBF serta menggunakan kombinasi seleksi fitur dapat dilihat pada Tabel 3 dan tanpa menggunakan kombinasi seleksi fitur dapat dilihat pada Tabel 4. Dari hasil ujicoba bahwa menggunakan kombinasi seleksi fitur sangat berpengaruh terhadap tingkat akurasi dan waktu komputasi. Selain itu juga pemilihan kernel berpengaruh terhadap tingkat akurasi dan waktu komputasi. Untuk hasil akurasi terbaik diperoleh dengan kernel RBF yang menggunakan nilai parameter γ sebesar 1 dan nilai σ^2 sebesar 0,1. Akurasi terbaik didapat menggunakan LS-SVM RBF dan SVM RBF dengan menggunakan kombinasi seleksi fitur maupun tanpa menggunakan kombinasi seleksi fitur dengan nilai akurasi sama yaitu 97,5%, namun untuk penggunaan kernel linear dan *Polynomial* hasil akurasi yang didapat meningkat. Waktu yang dibutuhkan untuk proses klasifikasi (proses *training* dan *testing*) terhadap model kombinasi dari uji coba atau tanpa seleksi fitur masing-masing dapat diketahui bahwa kombinasi seleksi fitur sangat berpengaruh terhadap waktu komputasi. Matrik konvolusi terbaik dari hasil klasifikasi bisa dilihat pada Tabel 5.

TABEL III
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 IV
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

TABEL V
MATRIKS KONFUSI UNTUK HASIL KLASIFIKASI TERBAIK

Aktual	Prediksi	
	Ganas	Jinak
Ganas	9	1
Jinak	0	30

V. KESIMPULAN DAN SARAN

Penambahan metode kombinasi seleksi fitur, pemilihan kernel, serta penggunaan parameter terbukti berpengaruh pada tingkat akurasi dan penurunan waktu komputasi. 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 *Polynomial*.

Penelitian tentang penggunaan ektraksi fitur GLCM menggunakan 12 fitur masih belum bisa memperoleh fitur terbaik sebagai *input* klasifikasi. Penelitian tentang pengaruh parameter γ dan σ^2 terhadap tingkat akurasi dan waktu komputasi dapat diperluas dengan menambah rentang nilai γ dan σ^2 yang digunakan. Serta dibutuhkan perluasan dengan penggunaan K-fold *validation* untuk mengetahui pengaruhnya terhadap tingkat akurasi dan waktu komputasi.

TABEL VI
HASIL KLASIFIKASI TERBAIK MENGGUNAKAN KOMBINASI SELEKSI FITUR

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

TABEL VII
HASIL KLASIFIKASI TERBAIK TANPA KOMBINASI SELEKSI FITUR

Klasifikasi	Akurasi (%)	Spesifitas (%)	Sensivitas (%)	Waktu (detik)
SVM-linear	35	13	100	0.037
SVM-*Polynomial*	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**

VI. DAFTAR PUSTAKA

[1] Eurostat, "Healt statistic: atlas on mortaly in the European Union," Eurostat, Luxembourg, 2002.

[2] H. C. Zuckerman, The role of mammography in the diagnosis of breast cancer. Breast cancer, diagnosis and treatment, New York: McGraw-Hill, 1987.

[3] E. D. PISANO and F. SHTERN, "Image Processing And Computer Aided Diagnosis In Digital Mammography: A Clinical Perspective," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 6, pp. 1493-1503, 1993.

[4] S. Tai, Z. Chen and W. Tsai, "An Automatic Mass Detection System in Mammograms based on Complex Texture Features," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 618-627, 2014.

[5] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," *Journal Pattern Recognition*, vol. 39, no. 4, pp. 646-668, 2006.

[6] F. Albregtsen, "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," University of Oslo, Oslo, 2008.

[7] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

[8] Y.-W. Chen and C.-J. Lin, "Combining SVMs with Various Feature Selection Strategies," in *Feature Extraction*, vol. 207, Berlin Heidelberg, Springer, 2006, pp. 315-324.

[9] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications* , vol. 36, no. 2009, p. 3240–3247, 2009.

[10] R. Aarthi, K. Divya, N. Komala and S. Kavitha, "Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine," in *Third International Conference on Advanced Computing*, Chennai, 2011.

[11] J. A. K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters* , vol. 9, no. 3, pp. 293-300, 1999.

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

[13] E. B. Holmes, G. L. White and D. K. Gaffney, "Ionizing Radiation Exposure, Medical Imaging," Medscape, 2010.

[14] V. Vapnik, The nature of statistical learning theory, New York: Springer Science & Business Media, 2013.

[15] L. Hakim, S. Mutrofin dan E. K. Ratnasari, "Segmentasi Citra menggunakan Support Vector Machine (SVM)dan Ellipsoid Region Search Strategy (ERSS) Arimoto Entropy berdasarkan Ciri Warna dan Tekstur," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 2, no. 1, pp. 11-16, 2016.

[16] K. Pelckmans, J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, L. Lukas, B. Hamers, B. D. Moor and J.Vandewalle, "LS-SVMlab: a MATLAB/C toolbox for Least Squares Support Vector Machines," Leuven, Belgium, 2002.

[17] K. Pelckmans, J. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle, "LS-SVMlab toolbox user's guide," *Pattern recognition letters*, vol. 24, no. 2003, pp. 659-675, 2003.