Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan *Support Vector Machine* (SVM) Berdasarkan Hasil Mamografi

Fourina Ayu Novianti dan Santi Wulan Purnami
Jurusan Statistika, Fakultas MIPA, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111
E-mail: santi_wp@statistika.its.ac.id

Abstrak*—Kanker payudara merupakan jenis kanker yang sering ditemukan oleh kebanyakan wanita. Di Indonesia Kanker payudara menempati urutan pertama pada pasien rawat inap di seluruh rumah sakit. Diagnosis dini pada payudara merupakan salah satu upaya untuk meminimumkan kanker *malignant* (ganas) yaitu dengan cara melakukan pemeriksaan mamografi. Pada penelitian ini akan dilakukan pengklasifikasian diagnosis keadaan pasien kanker payudara *benign* (jinak) dan *malignant* (ganas) berdasarkan hasil mamografi dan melakukan analisis faktor-faktor apa saja yang mempengaruhi kanker payudara menggunakan metode regresi logistik dan *support vector machine* (SVM). Pengklasifikasian menggunakan regresi logistik biner menghasilkan ketepatan klasifikasi sebesar 88,72% dimana terdapat dua faktor yang berpengaruh terhadap kanker payudara *malignant* yaitu *intermediate findings* dan BIRADS. Sedangkan dengan menggunakan seleksi variabel L_1*-Norm* SVM, semua variabel prediktor yang digunakan berpengaruh terhadap kanker payudara *malignant* dengan kontribusi terbesar adalah *intermediate findings*, kemudian BIRADS, *suspicious for malignancy*, letak abnormal, dan usia dengan ketepatan klasifikasi sebesar 94,34%.

***Kata Kunci*—Klasifikasi, Regresi Logistik, SVM, Kanker Payudara, Mamografi**

I. PENDAHULUAN

KANKER payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel-sel jaringan payudara. Kanker payudara merupakan jenis kanker yang sering ditemukan oleh kebanyakan wanita. Menurut WHO pada tahun 2005 dilaporkan sebanyak 506.000 wanita meninggal disebabkan oleh kanker payudara [1]. Sedangkan di Indonesia menurut profil kesehatan Departemen Kesehatan Republik Indonesia Tahun 2007 kanker tertinggi yang diderita wanita Indonesia adalah kanker payudara dengan angka kejadian 26 per 100.000 perempuan [2].

Deteksi dini kanker payudara melalui mamografi dapat meningkatkan kesempatan untuk bertahan hidup [3]. Mamografi dapat mengidentifikasi kanker untuk beberapa tahun dan merupakan metode pemeriksaan kanker payudara yang paling efektif saat ini.

Penelitian tentang kanker payudara berdasarkan faktor resiko dengan menggunakan regresi logistik pernah dilakukan oleh Purwantaka [4]. Penelitian tersebut yang diklasifikasikan adalah penderita dan non penderita kanker payudara. Ketepatan klasifikasi yang didapatkan dari model regresi logistik pada kasus ini hanya sebesar 37%.

Selain itu telah dilakukan beberapa penelitian tentang diagnosis kanker payudara berbasis *Support Vector Machine* [5]-[7]. Penelitian-penelitian tersebut menunjukkan *Support Vector Machine* memberikan ketepatan klasifikasi di atas 95 %. Hal ini menunjukkan keunggulan metode *Support Vector Machine* yang menghasilkan akurasi yang tinggi. Makadari itu pada penelitian ini akan dilakukan analisis perbandingan antara metode regresi logistik dan SVM dengan data mamografi pada pasien kanker payudara di rumah sakit 'X' Surabaya pada tahun 2011 dimana dilakukan perbandingan ketepatan klasifikasi dari kedua metode dan memperoleh faktor-faktor yang menggambarkan kanker payudara *benign* (jinak) dan *malignant* (ganas) pada kanker payudara. Sehingga nantinya diharapkan dapat dijadikan sebagai bahan pertimbangan dokter untuk pemeriksaan lebih lanjut.

II. LANDASAN TEORI

A. Kanker Payudara

Kanker payudara adalah pertumbuhan sel yang abnormal pada jaringan payudara seseorang. Payudara wanita terdiri dari lobulus (kelenjar susu), duktus (saluran susu), lemak dan jaringan ikat, pembuluh darah dan *limfe*. Sebagian besar kanker payudara bermula pada sel-sel yang melapisi duktus (kanker duktal), beberapa bermula di lobulus (kanker lobular), serta sebagian kecil bermula di jaringan lain [8].

B. Mamografi

Mamografi adalah foto payudara dengan sinar X dosis rendah. Pada mammografi dapat dilihat gambaran payudara secara keseluruhan. Mamografi merupakan alat yang terbaik untuk deteksi dini kanker payudara, karena sinar X pada mamografi mempunyai kemampuan menembus jaringan payudara yang mengalami kelainan berupa tumor dan menunjukkan kelainan dalam payudara tersebut secara memuaskan. Faktor-faktor yang dilihat pada saat pemeriksaan mamografi antara lain.

1. *Intermediate Findings*
   Variabel yang menjelaskan keadaan sel atau jaringan yang terdapat dalam payudara, dimana variabel ini terdiri dari lima indikator yaitu *well defined, developing, architectural, skin thickening,* dan *asymetry*. Seorang wanita yang melakukan pemeriksaan mamografi memungkinkan untuk memiliki lebih dari satu indikator atau tidak sama sekali pada variabel ini.

2. *Suspicious for Malignancy*
   Variabel yang menjelaskan bentuk tumor yang terdapat dalam payudara atau tanda-tanda keganasan yang terlihat pada payudara, dimana variabel ini terdiri dari tiga indikator yaitu *mass*, *calcification*, dan *speculated sign*.
3. BIRADS *Category*
   Breast Imaging Reporting and Data System (BIRADS) digunakan untuk memprediksi tingkat keganasan pasien kanker payudara dalam skrining mamografi.
4. Letak abnormal
   Akan dilihat letak dimana ada perubahan yang tidak wajar pada payudara kanan atau payudara kiri.

Prediksi malignansi dapat dipermudah dengan menerapkan kategori BIRADS (*Breast Imaging Reporting and Data System*). Adapun kategori BIRADS adalah sebagai berikut [9].

C-0 : perlu pemeriksaan lanjut
C-1 : normal
C-2 : kelainan jinak
C-3 : kelainan yang mungkin jinak, disarankan untuk evaluasi ketat
C-4 : kelainan yang mungkin mengarah keganasan
C-5 : sangat mungkin ganas

C. Regresi Logistik Biner

Regresi logistik merupakan suatu metode analisis data yang mendeskripsikan antara variabel respon dengan satu atau lebih variabel prediktor. Regresi logistik biner variabel responnya yang bersifat dikotomus yang terdiri dari dua kategori yaitu 0 dan 1, sehingga variabel respon akan mengikuti distribusi Bernoulli dengan fungsi probabilitas sebagai berikut [10].

$f(y_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$ dengan $y_i = 0,1$

Berdasarkan [10] model regresi logistik adalah sebagai berikut.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \quad (1)$$

Persamaan (1) tersebut kemudian ditransformasi yang dikenal dengan tranformasi logit $\pi(x)$ untuk memperoleh fungsi g(x) yang linear dalam parameternya, sehingga mempermudah pendugaan parameter regresi yang dirumuskan sebagai berikut

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (2)$$

Metode *Maximum Likelihood Estimator* (MLE) adalah metode yang digunakan untuk menduga parameter-parameter yang terdapat dalam model regresi logistik. Metode ini menduga β dengan meterbesarkan fungsi *likelihood*. Fungsi *likelihood* yang diterbesarkan adalah

$$L(\beta) = \ln(l(\beta)) = \sum_{j=0}^{p}\left[\sum_{i=1}^{n} y_i x_{ij}\right]\beta_j - \sum_{i=1}^{n}\ln\left[1+\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)\right] \quad (3)$$

Persamaan (2) dideferensialkan terhadap β, setelah dideferensialkan terhadap β kemudian disamakan dengan nol, namun cara ini sering kali diperoleh hasil yang implisit sehingga dilakukan metode iterasi Newton Rhapson untuk meterbesarkan fungsi *likelihood* [11].

Pengujian parameter dalam regresi logistik biner dilakukan baik secara serentak maupun individu. Statistik uji yang digunakan dalam uji serentak adalah statistik uji G atau *likelihood ratio test*. Sedangkan statistik uji yang digunakan dalam uji parsial adalah statistik uji Wald [10].

Salah satu ukuran yang digunakan untuk menginterpretasi koefisien variabel prediktor adalah *Odds ratio*. *Odds ratio* menunjukkan perbandingan peluang munculnya suatu kejadian dengan peluang tidak munculnya kejadian tersebut. Jika nilai *odds ratio* < 1, maka antara variabel prediktor dan variabel respon terdapat hubungan negatif setiap kali perubahan nilai variabel prediktor (X) dan jika nilai *odds ratio*>1, maka antara variabel prediktor dan variabel respon terdapat hubungan positif setiap kali perubahan nilai variabel prediktor (X).

Statistik uji yang dipakai untuk uji kesesuaian model adalah statistik *Hosmer-Lemeshow Test*($\hat{C}$)

D. Seleksi Variabel SVM

SVM dapat digunakan untuk melakukan pemilihan variabel sekaligus melakukan tugas klasifikasi. SVM yang digunakan adalah L_1*-norm*. Misalkan data berdimensi *p*, maka kelas dari suatu titik baru **x** ditentukan dengan memasukkan **x** ke dalam *hyperplane* atau fungsi $z=\mathbf{w}\mathbf{x}+b$ yang didapatkan selama *training*. *Hyperplane z* didefinisikan sebagai berikut [12].

$$z = \mathbf{w}\mathbf{x} + b = \sum_{i=1}^{p}\mathbf{w}_p\mathbf{x}_p + b = \sum_{wp\neq 0}\mathbf{w}_p\mathbf{x}_p + b = 0 \quad (4)$$

Jika nilai dari elemen vektor bobot $\mathbf{w}_p = 0$, maka variabel ke-*p* dalam vektor input tidak menentukan kelas dari **x** dalam penentuan kelas **x**. Jadi hanya variabel-variabel dimana $\mathbf{w}_p \neq 0$ yang mempunyai kontribusi dalam penentuan kelas suatu data. Dalam kasus dimana masalah klasifikasinya adalah *infeasible* atau beberapa data tidak bisa diklasifikasikan secara tepat, maka perlu menambah variabel *slack* t_i.

$$\min_{w,b} \|\mathbf{w}\|_1 + C\sum_{i=1}^{\lambda} t_i \quad (5)$$

dengan batasan : $y_i(\mathbf{w}\mathbf{x}_i + b) + t_i \geq 1$

$t_i \geq 0,1,\ldots,\lambda$

Formulasi persamaan (4) dapat diubah ke dalam bentuk L_1*-norm* dengan mendefinisikan variabel baru v_p, u_p dimana $\mathbf{w}_p = u_p - v_{p'}$ sehingga $|\mathbf{w}_p| = u_p - v_{p'}$. Jadi L_1*-norm* dari $\|\mathbf{w}\|_1 = \sum_{i=1}^{p} + u_p + v_p = u + v$. Formulasi problem optimasi dari SVM dalam persamaan (5) menjadi sebagai berikut.

$$\min\ u + v + C\sum_{i=1}^{\lambda} t_i \quad (6)$$

$$y_i((u-v)\mathbf{x}_i + b) + t_i \geq 1$$

dengan batasan: $t_i \geq 0,\ i = 1,\ldots,\lambda$

$u_p, v_p \forall_p = \{1,\ldots,p\}$

Dimana nilai C ditentukan oleh peneliti. Pada seleksi variabel ini bekerja dalam *primal space* dan tidak memerlukan *kernel-map* seperti dalam SVM regular [12].

E. Support Vector Machine (SVM)

Support vector machine (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1992 pada saat dipresentasikan di *Annual Workshop on Computational Learning Theory* [13].

Prinsip dasar SVM adalah linier *classifier*, yaitu kasus klasifikasi yang secara linier dapat dipisahkan. Misalkan diberikan himpunan $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, dinyatakan sebagai kelas positif jika $f(\mathbf{x}) \geq 0$ dan yang lainnya termasuk ke dalam kelas negatif. SVM melakukan klasifikasi himpunan vektor *training* berupa set data berpasangan dari dua kelas, [14]

$$(\mathbf{x}_i, y_i),\ \mathbf{x}_i \in R^n,\ y_i \in \{1,-1\},\ i = 1,\ldots,n, \tag{7}$$

Pemisahan *hyperplane* dengan bentuk *canonical* mengikuti *constraint* atau batasan berikut,

$$y_i\left[\left(\mathbf{w}^T\mathbf{x}_i\right) + b\right] \geq 1, \quad i = 1,2,\ldots,n. \tag{8}$$

Hyperplane yang optimal diperoleh dengan meterbesarkan $\frac{2}{\|\mathbf{w}\|}$ atau meminimumkan $\varphi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ Kemudian permasalahan optimasi ini dapat diselesaikan dengan menggunakan Fungsi Lagrange berikut.

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right] \tag{9}$$

dimana α_i adalah pengganda fungsi Lagrange. Persamaan (9) merupakan *primal space* sehingga perlu ditransformasi menjadi *dual space* agar lebih mudah dan efisien untuk diselesaikan. Sehingga dual problemnya menjadi seperti berikut.

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\mathbf{x}_i^T\mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i \tag{10}$$

dengan batasan,

$$\alpha_i \geq 0, \quad i = 1,\ldots,n \ \text{ dan } \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{11}$$

Pada kasus non-separabel beberapa data mungkin tidak bisa dikelompokkan secara benar atau terjadi *misclassification*. Sehingga fungsi obyektif maupun kendala dimodifikasi dengan mengikutsertakan variabel *slack* $\xi > 0$. Formulasinya menjadi sebagai berikut [14].

$$\Phi(\mathbf{w}, \xi)\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\lambda}\xi_i \tag{12}$$

dengan kendala

$$y_i\left[\left(\mathbf{w}^T\mathbf{x}_i\right) + b\right] + \xi_i \geq 1, \quad i = 1,2,\ldots,n \tag{13}$$

Pada kasus separabel dan kasus non-separabel perbedaan keduanya hanya terletak dengan adanya penambahan kendala $0 \leq \alpha_i \leq C$ pada masalah non-separabel**.**

Pada kasus non-linier optimasi persamaan (10) menjadi sebagai berikut [15].

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i \tag{14}$$

dengan batasan: $0 \leq \alpha_i \leq C, \quad i = 1,\ldots,n$ dan $\sum_{i=1}^{n}\alpha_i y_i = 0$

$K\left(\mathbf{x}_i, \mathbf{x}_j\right)$ adalah fungsi kernel yang digunakan untuk mengatasi data non-linier. Berdasarkan langkah langkah yang telah dijelaskan dalam kasus linier, diperoleh fungsi sebagai berikut

$$\begin{aligned} f(x) &= \text{sign}\left(\sum_{i=1}^{n} y_i\hat{\alpha}_i\left(\Phi\left(\mathbf{x}_i\right), \Phi\left(\mathbf{x}_j\right)\right) + \hat{b}\right) \\ &= \text{sign}\left(\sum_{i=1}^{n} y_i\hat{\alpha}_i\left(K\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) + \hat{b}\right) \end{aligned} \tag{15}$$

dengan fungsi sign semua nilai $f(x) < 0$ diberi label -1 dan nilai $f(x) > 0$ diberi label +1.

Fungsi kernel yang biasanya digunakan dalam literatur SVM [12].

1. Kernel Linier : $\left(\mathbf{x}^T\mathbf{x}\right)$
2. Kernel Polinomial : $\left(\mathbf{x}^T\mathbf{x} + 1\right)^p$
3. Kernel RBF : $K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$

F. Evaluasi Performansi Model

Evaluasi performansi model yang digunakan pada penelitian ini adalah *classification accuracy*, *sensitivity*, dan *specificity* [5]. *Classification accuracy* merupakan ketepatan klasifikasi yang diperoleh. *Sensitivity* merupakan ukuran ketepatan dari suatu kejadian yang diinginkan. *Specificity* merupakan suatu ukuran yang menyatakan persentase kejadian-kejadian yang tidak diinginkan. *Classification accuracy*, *sensitivity*, dan *specificity* dapat ditentukan menggunakan nilai yang terdapat dalam *confusion matrix*. *Confusion matrix* adalah klasifikasi tentang aktual dan prediksi yang dilakukan dengan sistem klasifikasi. *Confusion matrix* ditunjukkan pada Tabel 1.

Tabel 1.
Confusion Matrix

Actual	*Predicted*	
	Positive = class 0	*Negative= class* 1
Positive = class 0	*True Positive* (TP)	*False Negative* (FN)
Negative = class 1	*False Positive* (FP)	*True Negative* (TN)

$$\textit{Classsification accuracy}\ (\%) = \frac{TP + TN}{TP + FP + FN + TN} \tag{16}$$

$$\textit{Sensitivity}\ (\%) = \frac{TP}{TP + FN} \tag{17}$$

$$\textit{Specificity}\ (\%) = \frac{TN}{FP + TN} \tag{18}$$

III. METODOLOGI

Data yang digunakan dalam penelitian ini adalah data sekunder pasien kanker payudara yang diperoleh dari Rumah Sakit 'X' Surabaya tahun 2011 sebanyak 267 data dengan jumlah pasien kanker payudara *benign* sebanyak 100 data dan jumlah pasien kanker payudara *malignant* sebanyak 167 data.

Data tersebut adalah data pasien yang melakukan pemeriksaan mamografi dengan kategori BIRADS C-2 sampai dengan C-5.

Variabel respon (Y) dalam penelitian terdiri dari dua kategori yaitu kategori 1 adalah pasien kanker payudara didiagnosis kanker (*benign*) dan kategori 2 adalah kanker payudara (*malignant*) . Sedangkan variabel prediktor (X) pada penelitian ini ditunjukkan pada Tabel 2.

Tabel 2 .
Variabel Prediktor

Variabel	Definisi	Kategori	Skala
X_1	*Intermediate Findings*	1. tidak ada tanda 2. terdapat 1 tanda 3. terdapat >1 tanda	Ordinal
X_2	BIRADS *category*	1. C-2 2. C-3 3. C-4 4. C-5	Ordinal
X_3	*Suspicious for Malignancy*	1. tidak ada ciri keganasan 2. *Mass* 3. *Calcification* 4. *Speculated Sign* 5. *Mass,Calcification* 6. *Mass,Speculated Sign* 7. *Calcification, Speculated Sign* 8. *Mass,Calcification, Speculated Sign*	Nominal
X_4	Usia	-	Rasio
X_5	Letak abnormal	1. Kanan 2. Kiri	Nominal

Langkah-langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut.

1. Melakukan pengumpulan data sekunder, yaitu data pasien kanker yang melakukan mamografi di Rumah Sakit 'X' Surabaya tahun 2011
2. Melakukan pengkodingan terhadap data sekunder Melakukan analisis statistika deskriptif untuk mengetahui karakteristik pasien kanker payudara
3. Membagi data menjadi data *training* dan *testing* dengan beberapa persentase partisi yaitu 50:50, 70:30, dan 80:20.
4. Memodelkan menggunakan analisis regresi logistik untuk mengetahui faktor-faktor yang mempengaruhi pasien kanker payudara dalam pengklasifikasian kanker jinak atau ganas dengan langkah sebagai berikut.
   a) Melakukan analisis regresi logistik secara serentak terhadapa data *training*
   b) Melakukan analisis regrsei logistik secara parsial terhadap data *training*
   c) Membentuk model regresi logistik menggunakan metode *Enter*
   d) Menginterpretasi *odds ratio* untuk mengetahui besarnya pengaruh masing-masing variabel yang signifikan berpengaruh dari data *training*
   e) Melakukan uji kesesuain model yang diperoleh dari data *training*
   f) Menghitung ketepatan klasifikasi dari data *testing*
5. Melakukan seleksi variabel dari data *training* dengan menggunakan L_1*-norm*
6. Melakukan pengklasifikasian pasien kanker payudara dengan menggunakan metode SVM. Berikut adalah algoritma dari metode SVM.
   a) Melakukan transformasi data sesuai dengan format *software* SVM yang akan digunakan
   b) Menentukan fungsi kernel untuk permodelan
   c) Menentukan nilai-nilai parameter kernel dan parameter *cost* untuk optimasi
   d) Memilih nilai parameter terbaik untuk optimasi data *training* untuk klasifikasi data *testing*
   e) Menghitung ketepatan klasifikasi
7. Membandingkan ketepatan klasifikasi yang diperoleh dari metode regresi logistik dengan SVM
8. Membuat kesimpulan dan saran

IV. ANALISIS DAN PEMBAHASAN

A. Analisis Deskriptif

Analisis tabulasi silang digunakan untuk menyajikan data kualitatif dalam bentuk tabulasi yang mempunyai hubungan secara deskriptif sebagai berikut. Dari analisis tabulasi silang yang telah dilakukan menunjukkan bahwa pada variabel *intermediate findings* pada kategori 1 dari 53,2% wanita yang melakukan mamografi, wanita yang tidak terdeteksi memiliki tanda sel didiagnosis kanker payudara *malignant* sebesar 46,8% dan sebesar 6,4% hasil diagnosisnya *benign*. Wanita dengan hasil diagnosis *malignant* mayoritas terdekteksi memiliki kategori BIRADS C-5 yaitu sebesar 42,7%. Wanita yang memiliki ciri keganasan kategori 8 (*mass*, *calcification*, dan *speculated sign*) didiagnosis *malignant* sebesar 19,5%. Dari 47,6% wanita yang memiliki letak abnormal payudara sebelah kiri, 31,8% didiagnosis *malignant*.

Usia wanita yang melakukan pemeriksaan mamografi pada tahun 2011 di rumah sakit 'X' rata-rata berumur 48 tahun dengan usia paling muda adalah 19 tahun dan usia paling tua adalah 87 tahun.

B. Analisis Diagnosis Kanker Payudara dengan Regresi Logistik Biner

Analisis regresi logistik biner data dibagi menjadi *training* dan *testing* dengan beberapa persentase partisi yaitu 50:50, 70:30, dan 80:20. Berikut merupakan analisis regresi logistik biner pada tiap-tiap partisi yang memberikan ketepatan klasifikasi terbesar.

Dari ketiga data partisi yang telah dilakukan uji serentak diketahui bahwa *P-value*=0,000. Sehingga tolak H_0 karena *P-value*$<\alpha$ yang berarti secara serentak terdapat satu atau lebih faktor pasien kanker payudara yang berpengaruh signifikan terhadap diagnosis kanker payudara.

Analisis regresi logistik parsial dengan menggunakan data partisi 50:50 , menunjukkan bahwa parameter dari kelima variabel prediktor yang digunakan yaitu *intermediate findings* (X_1), kategori BIRADS (X_2), *suspicious for malignancy* (X_3), usia (X_4) dan letak abnormal (X_5) signifikan terhadap model secara parsial karena *P-value* $< \alpha$ Sedangkan analisis regresi logistik parsial untuk data partisi 70:30 dan 80:20 hanya parameter variabel letak abnormal (X_5) yang tidak signifikan terhadap model secara parsial.

Metode yang digunakan dalam pembentukan model adalah metode *Enter* dengan memasukkan semua variabel prediktor. Dengan menggunakan partisi data *training* dan *testing* 50:50 diagnosis *malignant* pada kanker payudara dipengaruhi oleh

intermediate findings dan BIRADS. Model logit adalah sebagai berikut.

$$g(x) = 1,297 + 2,948\ X_{1(1)} - 4,059\ X_{2(1)} - 2,793\ X_{2(2)}$$

Sedangkan dengan menggunakan partisi data 70:30, diagnosis malignant pada kanker payudara dipengaruhi oleh *intermediate findings*, BIRADS, dan usia. Model logit yang diperoleh sebagai berikut.

$$g(x) = -2,537 + 2,625\ X_{1(1)} - 4,157\ X_{2(1)} - 5,402\ X_{2(2)} + 0,096\ X_4$$

dengan menggunakan partisi 80:20 terdapat tiga variabel yang berpengaruh yaitu faktor *intermediate findings*, BIRADS, *suspicious for malignancy*, dan usia dengan ketepatan klasifikasi sebesar 84,9%. Sehingga model logitnya adalah

$$g(x) = -2,49 + 4,428\ X_{1(1)} + 2,624\ X_{1(2)} - 5,098\ X_{2(1)} - 3,043\ X_{2(2)} - 2,721\ X_{2(3)} + 3,043\ X_{3(2)} + -3,043\ X_{3(6)} + -3,043\ X_4$$

Berikut merupakan interpretasi koefisien parameter berdasarkan nilai *odds ratio* dengan menggunakan partisi data *training* dan *testing* 50:50

a) *Intermediate findings*
   Pasien kanker payudara dengan *intermediate findings* yang tidak terdeteksi tanda apapun cenderung memiliki diagnosis *malignant* 19,065 kali dibandingkan dengan yang memiliki lebih dari 1 tanda pada sel payudaranya.
b) Kategori BIRADS
   Pasien kanker payudara yang terdeteksi C-2 dalam pemeriksaan mamografi cenderung akan memiliki diagnosis malignant 0,017 kali dibandingkan dengan pasien yang terdeteksi C-5. Sedangkan pasien kanker payudara yang terdeteksi C-3 cenderung memiliki diagnosis malignant 0,061 kali dibandingkan dengan pasien yang terdeteksi C-5.

Interpretasi yang sama juga dilakukan pada partisi data *training testing* 70:30 dan 80:20. Tabel 3 merupakan nilai odds ratio yang diperoleh dari *training testing* 70:30 dan 80:20.

Tabel 3.
Nilai *Odds Ratio*

Variabel	**Persentase Partisi**	
	70:30	**80:20**
	Exp(B)	
Intermediate findings (X_1)		
$X_1(1)$	13,804	19,065
Kategori BIRADS (X_2)		
$X_2(1)$	0,016	0,006
$X_2(2)$	0,005	0,048
		0,066
Suspicious for Malignancy (X_3)		
$X_3(2)$	*	39,882
$X_3(6)$	*	19,586
Usia (X_4)	1,101	1,060

*)Ket : tidak berpengaruh signifikan

Pada uji kesesuaian model diketahui bahwa artinya dari ketiga data partisi tersebut model yang diperoleh sesuai atau tidak terdapat perbedaan nyata antara observasi dengan prediksi model. Hal ini ditunjukkan karena nilai *P-value* dari ketiga data partisi > α (5%).

Setelah dilakukan uji kesesuaian model, maka dilakukan pengukuran ketepatan kalsifikasi model dengan menggunakan Tabel *confusion matrix*, sehingga diperoleh classification accuracy.

Tabel 4.
Confusion Matrix

	Partisi (%)		
	50:50	70:30	80:20
***Classification accuracy* (%)**	88,72	86,4	84,90
Sensitivity (%)	73,07	81,25	85,71
Specificity (%)	98,76	89,79	84,37

Berdasarkan Tabel 4 dapat diketahui bahwa *classification accuracy* terbesar diberikan oleh partisi data *training* dan *testing* 50:50 yaitu sebesar 88,72%, kemudian diikuti partisi 70:30, 80:20 yaitu masing-masing sebesar 86,4 dan 84,90.

Seleksi Variabel Menggunakan SVM L_1-norm

Hasil seleksi variabel menunjukkan bahwa SVM memilih semua variabel prediktor untuk masuk ke dalam proses klasifikasi yang ditunjukkan pada Tabel 5.

Tabel 5.
Nilai **w** dan b untuk masing-masing partisi

Partisi (%)	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	*b*
50:50	0,8678	0,7831	0,3409	0,0248	0,3616	-3,5868
70:20	0,8632	0,7088	0,3158	0,0351	0,4140	-3,7930
80:20	0,8678	0,7831	0,3409	0,0248	0,3161	-3,5868

Berdasarkan Tabel 5 dapat diketahui bahwa dengan menggunakan partisi data training testing 50:50, 70:30, dan 80:20 kelima variabel berpengaruh karena nilai $\mathbf{w} \neq 0$, dimana nilai **w** meupakan vektor bobot dan nilai *b* merupakan bias. $\mathbf{w}_1$ merupakan vektor bobot yang dihasilkan oleh variabel *intermediate findings*, begitu juga untuk $\mathbf{w}_1$,$\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$, $\mathbf{w}_5$ adalah BIRADS, *suspicious for malignancy*, usia, dan letak abnormal. Variabel prediktor yang memberikan pengaruh paling kuat adalah variabel yang menghasilkan vektor bobot $\mathbf{w}_i$ paling besar yaitu *intermediate findings*, kemudian diikuti kategori BIRADS, *suspicious for malignancy*, letak abnormal dan usia.

Perbandingan seleksi variabel antara SVM dan regresi logistik diketahui bahwa variabel yang selalu ada pada tiap partisi adalah variabel *intermediate findings* dan BIRADS.

C. Klasifikasi Menggunakan SVM

Klasifikasi SVM pada penelitian ini menggunakan fungsi kernel linear, polynomial, dan *Radial Basis Function* (RBF) yang ditunjukkan pada Tabel 6. Data *training* dan *testing* dipartisi menjadi beberapa bagian yaitu 50:50, 70:30, dan 80:20, nilai parameter kernel dan nilai C berdasarkan *trial and error*. Ketepatan klasifikasi terbesar yang dihasilkan oleh metode SVM dari partisi data *training* dan *testing* 80:20 yaitu sebesar 94,34% dengan menggunakan fungsi kernel RBF dimana nilai C=100 dan σ =35. Untuk partisi data traing dan testing 70:30 ketepatan klasifikasi terbesar yang diperoleh sebesar 88,89% dengan fungsi kernel linier dan nilai C=10 atau C=100. Sedangkan untuk partisi data training dan testing 50:50 ketepatan klasifikasi yang terbesar sebesar 92,48 dengan menggunakan fungsi kernel RBF dimana nilai C=100 dan σ =35.

Tabel 6.
Tingkat Akurasi Klasifikasi SVM

Kernel	Parameter		Persentase Partisi		
		C	50:50	70:30	80:20
Linier	0	1	90,23	87,65	88,68
		10	90,23	**88,89**	90,57
		100	86,56	**88,89**	90,57
Polynomial	p=1	1	89,47	85,19	88,68
		10	89,47	85,19	88,68
		100	89,47	85,19	88,68
	p=2	1	89,47	85,19	88,68
		10	89,47	85,19	88,68
		100	89,47	85,19	88,68
	p=3	1	89,47	85,19	88,68
		10	89,47	85,19	88,68
		100	89,47	85,19	88,68
RBF	σ =5	1	87,97	81,48	83,33
		10	89,47	80,25	83,02
		100	89,47	77,78	83,33
	σ =10	1	87,22	80,25	81,13
		10	89,47	81,48	90,57
		100	90,98	80,25	90,57
	σ =20	1	87,97	77,78	81,13
		10	90,23	83,95	90,57
		100	90,98	82,72	92,45
	σ =35	1	75,94	77,78	81,13
		10	87,97	83,95	92,45
		100	**92,48**	85,19	**94,34***

*) Ketepatan klasifikasi terbesar

Tabel 7 menujukkan perbandingan akurasi klasifikasi yang diperoleh dari regresi logistik biner dan SVM.

Tabel 7.
Perbandingan Akurasi Klasifikasi

	Akurasi (%)			Rata-rata
	50:50	**70:30**	**80:20**	
Regresi Logistik	88,72	86,42	84,90	86,67
SVM	**92,48**	**88,89**	**94,34**	**91,9**

Berdasarkan Tabel 7 dapat diketahui performansi akurasi klasifikasi terbaik dimiliki oleh SVM yaitu untuk partisi data *training* dan *testing* 50:50 sebesar 92,48%, partisi data *training* dan *testing* 70:30 sebesar 88,89%, dan untuk partisi data *training* dan *testing* 80:20 sebesar 94,34% dengan rata-rata ketepatan klasifikasi sebesar 91,9%. Hal ini menunjukkan akurasi klasifikasi dengan menggunakan SVM lebih baik daripada regresi logistik.

V. KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa performansi akurasi klasifikasi terbaik dimiliki oleh SVM yaitu sebesar 94,34% sedangkan regresi logistik sebesar 84,90% dengan menggunakan partisi 80:80. Pada regresi logistik biner, kanker payudara *malignant* dipengaruhi oleh faktor *intermediate findings* BIRADS, *Suspicious for malignancy*, dan usia. Sedangkan pada metode SVM, kanker payudara *malignant* dipengaruhi oleh semua variabel prediktor.

Data pada *intermediate findings* dan *suspicious for malignancy* terdapat beberapa data yang *missing value*, oleh karena itu disaranakan kepada pihak rumah sakit 'X' memperhatikan data-data *missing value* sehingga diharapkan nantinya akan diperoleh analisis yang lebih tepat. Selain itu untuk metode *Support Vector Machine* dalam penentuan parameter SVM sebaiknya tidak menggunakan *trial and error* agar efisien dan menghasilkan akurasi yang optimum. Namun apabila data *missing value* tersebut tidak dapat dihindarkan maka untuk penelitian selanjutnya dapat dilakukan pengembangan metode SVM untuk data *missing value* dan penentuan parameter SVM tanpa *trial and error* yang diharapkan nantinya akan memberikan akurasi yang lebih tinggi.

DAFTAR PUSTAKA

[1] WHO. (2005). Data penderita kanker payudara di dunia. Dikases pada tanggal 3 Februari 2012 dari [http://www.who.int/cancer/detection/braestcancer/en/index1.html].

[2] Dinas Kesehatan Nasional.(2007). Data penderita kanker payudara di Indonesia. Diakses pada tanggal 31 januari 2011 dari [http://www.depkes.go.id/index.php/berita/press-release/1060-jika-tidak-dikendalikan-26-juta-orang-di-dunia-menderita-kanker-.html]

[3] Keles, A., Keles, A., dan Yavuz, U. (2011). Expert System Based On Neuro-Fuzzy Rules For Diagnosis Breast Cancer. *Expert Systems with Applications. 38.* 5719–5726.

[4] Purwantaka, R. I. (2010). [Tugas Akhir] *Faktor-Faktor Yang Mempengaruhi Resiko Penyebab Penderita Kanker Payudara Dengan Menggunakan Pendekatan Regresi Logistik*. Surabaya: Institut Teknologi Sepuluh Nopember Surabaya.

[5] Purnami, S. W., dan Embong, A. (2008). Smooth Support Vector Machine For Breast Cancer Classification. *The 4th IMT-GT 2008 Conference on Mathematics, Statistics, and Their Applications (ICMSA08)*, Banda Aceh, Indonesia.

[6] Wang, D., Shi, L., dan Heng, P. A. (2009). Automatic Detectiom of Breast Cancer in Mammogrmas using Support Vector Machines. *Neurocomputing 72.* 3296-3302.

[7] Huang, C-L., Liao, H-C., dan Chen, M-C. (2008). Prediction Model Building and Feature Selection With Support Vector Machine*. Expert System with Application 34.* 578-587.

[8] Ellis, E.O., Schnitt, S.J., S.-Garau, X., Bussolati, G., Tavassaoli, F.A., Eusebi, V. Pathology and Genetic of Tumours of The Breast and Female Genital Organs / WHO Classification of Tumours. Washington: IARC Press; 2003. P.10, 34-6.

[9] Kardinah (2002). *Penatalaksanaan Kanker Payudara Terkini oleh Penanggulangan & Pelayanan Kanker Payudara Terpadu Paripurna R.S. Kanker Dharmais*. Jakarta: Pustaka Populer Obor.

[10] Hosmer, D. W., dan Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons, Inc.

[11] Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. John Willey & Sons, New York.

[12] Santosa, B. (2006). *Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

[13] Nugroho, A.S., Witarto, A.B., Handoko, D., 2003. Support Vector Machine –Teori dan Aplikasi dalam Bioinformatika. Diakses pada tanggal 9 Maret 2012 dari http://www.Ilmukomputer.com

[14] Gunn, Steve. (1998). *Support Vector Machine for Clasification and Regression*. Southampton: University of Southaton.