# *SMOOTH SUPPORT VECTOR MACHINE* DAN *MULTIVARIATE ADAPTIVE REGRESSION SPLINE* UNTUK MENDIAGNOSIS KANKER PAYUDARA

**[1]Shofi Andari, [2]Santi W. Purnami, [3]Bambang W. Otok**

*Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya*

Alamat e-mail : [1]shofi.andari11@mhs.statistika.its.ac.id

**ABSTRAK**

Kanker payudara merupakan kanker yang paling umum menyerang wanita dan menjadi kanker penyebab kematian utama bagi wanita di seluruh dunia. Penyebab dari kanker payudara masih belum dapat dipastikan sehingga metode preventif yang spesifik untuk penyakit ini juga belum dapat ditentukan, oleh karena itu diagnosis terhadap kanker payudara sedini mungkin menjadi sangat penting bagi para dokter dan tenaga medis untuk menyelamatkan pasien maupun orang-orang yang memiliki faktor risiko kanker payudara. Beberapa penelitian telah dikembangkan dengan ide dasar mengklasifikasikan kanker payudara berdasarkan rekaman gambar radiologi dan usia pasien terhadap hasil biopsi. Berdasarkan keunggulan *smooth* SVM (SSVM) serta potensi MARS dalam menyelesaikan permasalahan diagnosis kanker payudara, tulisan ini mengkaji dan memaparkan kedua metode tersebut digunakan untuk mengklasifikasikan kanker payudara ke dalam dua kelompok yaitu kelompok *malignant* dan kelompok *benign*. Secara umum baik SSVM maupun MARS mampu menghasilkan tingkat akurasi yang sama-sama tinggi. Tingkat akurasi kedua metode dalam mendiagnosis kanker payudara ke dalam kelompok *benign* dan *malignant* yang cukup tinggi dipercaya dapat mendukung prosedur pemeriksaan dan diagnosis kanker payudara.

**Kata Kunci :** kanker payudara, klasifikasi, *smooth* SVM, MARS

## PENDAHULUAN

Kanker payudara merupakan kanker yang paling umum menyerang wanita dan menjadi kanker penyebab kematian utama bagi wanita di seluruh dunia. Tercatat pada tahun 2008, satu perempat (23%) dari semua kanker yang diderita oleh kaum wanita adalah kanker payudara [1]. Berdasarkan data WHO tahun 2010, di Indonesia kanker payudara menduduki peringkat kedua kanker paling mematikan setelah kanker paru-paru [2]. Sampai saat ini, mengontrol kanker, khususnya kanker payudara, masih menjadi pekerjaan yang berat bagi pemerintah Indonesia.

Penyebab kanker payudara belum dapat dipastikan sehingga metode preventif yang spesifik untuk penyakit ini juga belum dapat ditentukan. Secara umum, pasien yang payudaranya didapati mengalami pengapuran (*calcification*) berdasarkan gambar mamografi akan dirujuk untuk melakukan biopsi agar mendapat kepastian mengenai diagnosis lebih lanjut dari pengapuran tersebut. Dengan mengusahakan diagnosis awal sejak tahap radiologi maka pasien-pasien yang dicurigai memiliki kanker tidak perlu melakukan biopsi (*unnecessary biopsy*). Diagnosis kanker payudara ini dilakukan dengan mengklasifikasikan kelainan sebagai *malignant* atau *benign.*

Beberapa penelitian telah dikembangkan dengan ide dasar mengklasifikasikan kanker payudara berdasarkan rekaman gambar radiologi. Hal ini seiring dengan berkembangnya metode dalam *data mining* dan *machine learning*, sehingga permasalahan mengenai pengenalan pola (*pattern recognition*) menjadi salah satu alat serta bahan penelitian yang populer dalam beberapa tahun terakhir. Penelitian tentang diagnosis kanker payudara telah dimulai sejak tahun 1990-an. Tahun 2002, [3] meneliti tentang diagnosis kanker payudara dengan *artificial neural network* dan *support vector machine*. Penelitian serupa dilakukan oleh [4] dengan mengintegrasikan *radial basis function* (RBF) dalam *neural network* kemudian membandingkannya dengan algoritma SVM. Tahun berikutnya, [5] meneliti tentang pendeteksian dan klasifikasi rekaman gambar ultrasonografi kanker payudara. Menggunakan data *benchmark* Wisconsin Breast Cancer Database (WBCD), [6] mengembangkan *feature selection* dan klasifikasi dengan *rough set-based* berdasarkan SVM. Model *hybrid* baru berdasarkan model SVM dikembangkan oleh [7] dengan mengintegrasi algoritma *fuzzy c-mean* dalam sistem klasifikasi SVM untuk diagnosis kanker payudara dengan data WBCD.

Metode pemulusan terhadap solusi SVM juga telah diaplikasikan untuk diagnosis kanker payudara oleh [8]-[10] dengan menggunakan data kanker payudara *benchmark* dan menyimpulkan bahwa metode *smooth* SVM (SSVM) menghasilkan akurasi yang lebih baik dibandingkan analisis diskriminan linier, *neural network*, *decision tree*, *genetic algorithm* dan *supervised fuzzy clustering*. Sementara itu penggunaan MARS untuk diagnosis kanker payudara belum banyak berkembang, pun demikian dalam tulisan ilmiahnya,[11] menyimpulkan bahwa MARS juga dapat mengatasi permasalahan diagnosis kanker payudara sama baik dengan analisis diskriminan maupun ANN. Penelitian-penelitian tersebut sebagian besaar dilakukan dengan memanfaatkan dataset yang disediakan oleh institusi penyedia database (*benchmark*).

Penelitian mengenai implementasi metode klasifikasi untuk mendiagnosis kanker payudara menggunakan dataset lokal belum banyak diadakan. Data yang digunakan dalam peneltian ini sebelumnya telah digunakan dalam penelitian [12] untuk kepentingan yang sama menggunakan metode klasifikasi SVM (94,34%) dan regresi logistik (88,72%), sedangkan [13] pada penelitiannya mengimplementasikan metode CART (90,19%). Di samping itu, kedua penelitian tersebut tidak mengindahkan adanya data hilang (*missing value*) pada dataset dan menggantikannya dengan angka nol sehingga akurasi klasifikasi kurang representatif. Berdasarkan hal tersebut, sebelum mengimplementasikan metode klasifikasi pada penelitian ini dilakukan imputasi terhadap *missing value* menggunakan metode imputasi berganda untuk data kategorik. Mengingat pentingnya penyeleksian parameter dalam SVM dan metode-metode pengembangannya, maka dalam penelitian ini juga diulas teknik penyeleksian parameter dalam SSVM untuk fungsi kernel Gaussian dengan pendekatan *uniform design* dua tahap sebagaimana yang telah dilakukan dalam [14].

Diberikan permasalahan klasifikasi dari sebanyak $n$ objek dalam ruang dimensi $R^p$ sehingga susunan data berupa matriks $\mathbf{A}$ berukuran $n \times p$ dan keanggotaan tiap titik terhadap kelas {+1} atau {-1} yang didefinisikan pada diagonal matriks $\mathbf{D}$ berukuran $n \times n$, maka problem optimasi pada SSVM adalah:

$$\min_{\mathbf{w},b,\xi} \frac{C}{2}\xi'\xi + \frac{1}{2}(\mathbf{w}'\mathbf{w} + b^2)$$

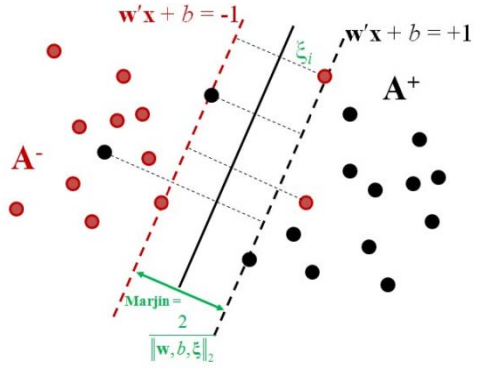
dengan kendala $\mathbf{D}(\mathbf{A}\mathbf{w} + \mathbf{e}b) + \xi \geq \mathbf{e}$ (1)

$\xi \geq \mathbf{0}$

Solusi problem 2.1 adalah

$$\xi = (\mathbf{e} - \mathbf{D}(\mathbf{A}\mathbf{w} + \mathbf{e}b))_+ \text{ a} \quad (2)$$

di mana $\xi$ merupakan variabel *slack* yang mengukur kesalahan klasifikasi. Permasalahan nonlinier ini dapat diilustrasikan seperti pada Gambar 1.




**Gambar 1.** Bidang pembatas $\mathbf{w}'\mathbf{x} + b = 0$ berada tepat di antara dua marjin $\mathbf{w}'\mathbf{x} + b = \pm 1$ dari solusi program nonlinier (2)

Melalui substitusi dan konversi, persamaan (2) dapat ditulis sebagai berikut:

$$\min_{\mathbf{w},b} \frac{C}{2} \| (\mathbf{e} - \mathbf{D}(\mathbf{A}\mathbf{w} - \mathbf{e}b))_+ \|_2^2 + \frac{1}{2}(\mathbf{w}'\mathbf{w} + b^2), \quad (3)$$

dengan fungsi plus didefinisikan sebagai $(x_+)_i = \text{maks}\{0, x_i\}$ untuk $i = 1, 2, \ldots, p$.

Fungsi objektif dalam persamaan (3) di atas tidak memiliki turunan kedua, teknik pemulusan yang diusulkan [18] dilakukan dengan menggantikan fungsi plus dengan $p(x,\alpha)$ yaitu integral dari fungsi sigmoid *neural network* $\left(1+\varepsilon^{-\alpha x}\right)^{-1}$ atau dapat dituliskan sebagai berikut:

$$p(x,\alpha) = x + \frac{1}{\alpha}\log(1+\varepsilon^{-\alpha x}), \ \alpha > 0 \quad (4)$$

di mana $\alpha$ adalah parameter penghalus. Dengan menggantikan fungsi plus dengan $p(x,\alpha)$ maka diperoleh model SSVM sebagai berikut:

$$\min_{(\mathbf{w},b)\in R^{p+1}} \Phi_\alpha(\mathbf{w},b) := \quad (2.6)$$

$$\min_{(\mathbf{w},\gamma)\in R^{p+1}} \frac{C}{2} \| p(\mathbf{e} - \mathbf{D}(\mathbf{A}\mathbf{w} - \mathbf{e}b),\alpha) \|_2^2 + \frac{1}{2}(\mathbf{w}'\mathbf{w} + b^2) \quad (5)$$

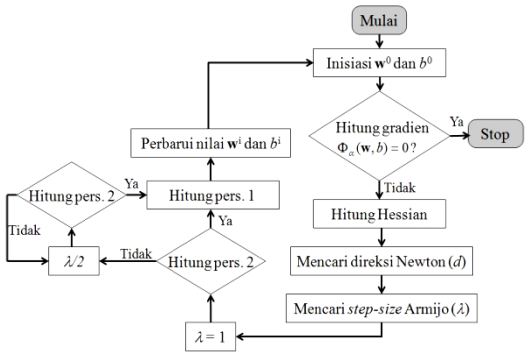
Secara umum, problem optimasi SSVM dapat ditulis sebagai berikut: (2.7)

$$\min_{(\mathbf{u},b)\in R^{p+1}} \Phi_\alpha(\mathbf{w},b) :=$$

$$\min_{(\mathbf{u},b)\in R^{p+1}}$$

$$\frac{C}{2} \| p(\mathbf{e} - \mathbf{D}(K(\mathbf{A},\mathbf{A}')\mathbf{D}\mathbf{u} - \mathbf{e}b),\alpha) \|_2^2 + \frac{1}{2}(\mathbf{u}'\mathbf{u} + b^2) \quad (6)$$

yang diselesaikan dengan iterasi Newton-Armijo (Gambar 2) dan $K(\mathbf{A},\mathbf{A}')$ merupakan fungsi kernel yang dalam penelitian ini digunakan kernel Gaussian, atau bisa dirumuskan berikut $K(A_i, A_j) = \exp\left(-\gamma \left\| A_i, A_j \right\|_2^2\right)$ dengan parameter kernel $\gamma$.




**Gambar 2.** Diagram alir algoritma Newton-Armijo

Pers.1:

$$\Phi_\alpha(\mathbf{w}^i, b^i) - \Phi_\alpha((\mathbf{w}^i, b^i) + \lambda_i d^i) \geq -\delta\lambda_i \nabla\Phi_\alpha(\mathbf{w}^i, b^i) d^i$$

Pers. 2: $(\mathbf{w}^{i+1}, b^{i+1}) = (\mathbf{w}^i, b^i) + \lambda_i d^i$

Saat iterasi pada algoritma Newton - Armijo berhenti, diperoleh nilai $\mathbf{w}$ dan $b$ yang konvergen. Dengan demikian fungsi pemisah yang diperoleh untuk kasus klasifikasi linier adalah

$$f(\mathbf{x}) = sign(\mathbf{w}'\mathbf{x} + b), \quad (7)$$

sedangkan fungsi pemisah untuk kasus klasifikasi nonlinier adalah sebagai berikut

$$f(\mathbf{x}) = sign(\mathbf{w}'\mathbf{x} + b) = sign(\mathbf{u}'\mathbf{D}'K(\mathbf{A},\mathbf{A}') + b) \quad (8)$$

Perumusan program linier SVM 1-norm telah ditunjukkan dalam [19] sebagai salah satu cara untuk memilih atribut (*feature selection*) di antara varian-varian norm SVM, problem linier tersebut adalah sebagai berikut

$$\min_{(\mathbf{w},b,\mathbf{s},\xi)\in R^{(2p)+1+n}} C\mathbf{e}'\xi + \mathbf{e}'\mathbf{s}$$

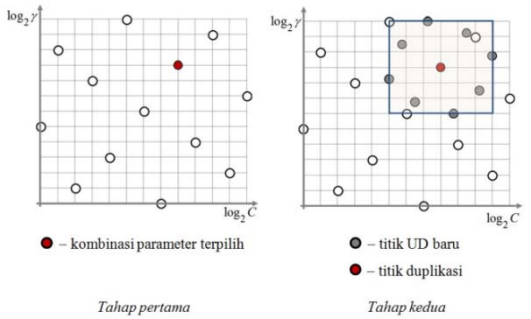
dengan kendala $\mathbf{D}(\mathbf{A}\mathbf{w}+\mathbf{e}b)+\xi \geq \mathbf{e}$ (9)

$$-\mathbf{s} \leq \mathbf{w} \leq \mathbf{s}$$

$$\xi \geq \mathbf{0}.$$

Solusi dari **w** mampu menghasilkan model yang parsimoni dan bersifat *sparsity*. Jika nilai dari elemen vektor $\mathbf{w}_p = 0$, maka variabel *p* tidak berkontribusi dalam penentuan kelas. Kontribusi atribut atau variabel prediktor dapat dinilai dari besarnya nilai $w_l$ untuk masing-masing atribut, dengan $l = 1, 2, ..., p$.

Penerapan *uniform design* (UD) dalam penentuan parameter SVM dijabarkan dalam [14]. Pada dasarnya tahap pertama digunakan untuk mencobakan kombinasi-kombinasi parameter *C* dan $\gamma$ kemudian mekanisme tersarang yang digunakan pada tahap kedua berfungsi untuk mempersempit ruang penyeleksian. Dengan kata lain, tahap pertama merupakan tahap untuk menentukan kombinasi parameter terpilih secara kasar dengan wilayah pencarian yang lebih luas dan kemudian pada tahap kedua penyeleksian dibatasi pada titik-titik di sekitar kombinasi parameter terpilih pada tahap pertama.




**Gambar 3.** UD dua tahap: 13-titik UD pada tahap pertama dan 9-titik UD pada tahap kedua

MARS diperkenalkan oleh [20] untuk pendekatan model nonparametrik antara variabel respon dan beberapa variabel prediktor pada regresi *piecewise*. Regresi *piecewise* merupakan regresi yang memiliki sifat tersegmen atau terpotong-potong. Prosedur pembentukan modelnya didasari oleh ide dari *recursive partition regression* atau RPR [21] dan *generalized additive modeling* [22]. RPR merupakan metode yang men-janjikan, tetapi masih memiliki beberapa kelemahan antara lain himpunan bagian yang saling lepas menyebabkan model RPR tidak kontinyu pada batas-batas setiap himpunan bagian, RPR juga tidak mampu mengidentifikasi fungsi $f(\mathbf{x})$ linier atau aditif, dan RPR cenderung sulit diinterpretasikan apabila variabel prediktor terlalu banyak [20].

Hasil modifikasi model *recursive partitioning regression* dengan kombinasi spline adalah model *multivariate adaptive regression splines* atau MARS yang berbentuk:

$$\hat{f}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[ s_{km} \left( x_{p(k,m)} - t_{km} \right) \right]_+ \quad (10)$$

dengan $a_0$ adalah koefisien dari basis fungsi $BF_0$ sedangkan penjumlahan basis-basis fungsi yang diper-oleh dari algoritma *forward* dan berhasil bertahan dari strategi penghapusan pada algoritma *backward* dan $s_{km} = \pm 1$.

Persamaan (10) dapat pula ditulis sebagai berikut:

$$\hat{f}(\mathbf{x}) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + ... \quad (11)$$

Penjumlahan pertama adalah untuk semua basis fungsi yang mengandung satu variabel prediktor. Penjumlahan kedua untuk semua basis fungsi yang mengandung tepat dua variabel prediktor, menunjukkan (jika ada) interaksi dua-variabel. Sama halnya dengan penjumlahan ketiga yang menunjukkan

(jika ada) kontribusi dari interaksi tiga-variabel dan seterusnya. Persamaan MARS dapat disederhanakan sebagai berikut:

$$\hat{f}(\mathbf{x}) = a_0 + a_1 BF_1 + a_2 BF_2 + \ldots + a_m BF_m \quad (12)$$

dengan $\hat{f}(\mathbf{x})$ merupakan variabel respon, $a_0$ adalah konstanta, $a_m$ adalah koefisien untuk basis fungsi ke-*m*, di mana $\{a_m\}_{m=0}^{M}$ merupakan penaksir untuk $\{\alpha_m\}_{m=0}^{M}$ yang diperoleh dengan pendekatan kuadrat terkecil (OLS) sedangkan $BF_m$ adalah basis fungsi ke-*m.*

Model MARS untuk nilai variabel respon biner merupakan pendekatan regresi logistik linier, yaitu

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i = \hat{f}(\mathbf{x}), \quad (13)$$

Dengan $\pi$ adalah probabilitas untuk respon bernilai paling besar (yaitu 1 apabila respon bernilai 0 dan 1). Koefisien-koefisien $\beta_i$ dengan $i = 1,2,\ldots,n$ ditaksir secara numerik dengan memaksimumkan *likelihood* dari data dan $\hat{f}(\mathbf{x})$ didekati dengaan MARS [20].

Berkaitan dengan evaluasi performasi klasifikasi, s*ensitivity* dan *specificity* merupakan statistik yang mengukur performansi klasifikasi biner. *Sensitivity* mengukur proporsi dari kondisi yang benar-benar positif, yaitu yang teridentifikasi sakit dengan benar dan *specificity* mengukur proporsi negatif, yaitu yang teridentifikasi sehat dengan benar [23] [24]. Hasil klasifikasi dapat diringkas dalam tabulasi silang yang disebut juga *confusion matrix* seperti pada Tabel 1 dengan ***tp*** untuk *true positive* (sebenarnya positif dan diklasifikasikan positif), ***fp*** adalah *false positive* (sebenarnya negatif tetapi diklasifikasikan positif), ***tn*** adalah *true negative* (sebenarnya negatif dan diklasifikasikan negatif) dan ***fn*** yaitu *false negative* (sebenarnya positif tetapi diklasifikasikan negatif).

**Tabel 1.**Tabulasi silang (*confusion matrix*) untuk hasil klasifikasi biner

| **Kelas sebenarnya** | **Kelas prediksi** | |
|---|---|---|
| | Positif | Negatif |
| Positif | *tp* | *fn* |
| Negatif | *fp* | *tn* |

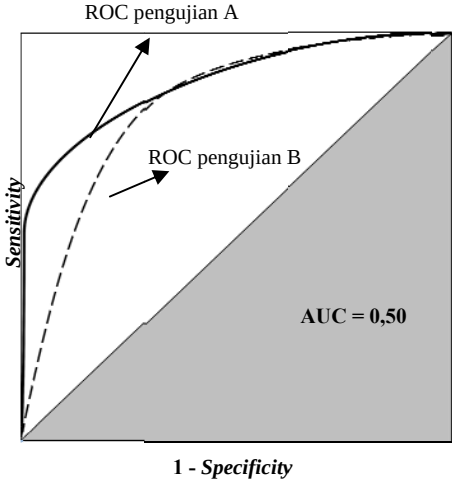
Kedua ukuran *sensitivity-specificity* menjelaskan akurasi diagnosis dengan lebih bermakna daripada indeks prosentasi akurasi tunggal.

$$\text{Akurasi klasifikasi (\%)} = \frac{tp+tn}{tp+fp+tn+fn} \quad (14)$$

$$\textit{Sensitivity}\ (\%) = \frac{tp}{tp+fn} \quad (15)$$

$$\textit{Specificity}\ (\%) = \frac{tn}{fp+tn} \quad (16)$$

Membuat plot ROC merupakan cara yang umum digunakan untuk mengambarkan akurasi diskriminasi dari suatu pengujian diagnosis untuk menentukan apakah seseoran gmenderita suatu penyakit tertentu atau tidak. Secara teori, kurva ROC meru-pakan plot dari *sensitivity* terhadap 1-*specificity* untuk bebe-rapa nilai *threshold* [25].




**Gambar 4.** Kurva ROC dari dua pengujian diagnosis (A dan B), masing-masing uji paling sedikit memiliki AUC seluas 0,50

Kanker payudara (*Carcinoma mammae*) adalah suatu penyakit neoplasma yang berasal dari *parenchyma*. Kanker payudara ditandai dengan adanya pertumbuhan sel yang abnormal pada jaringan payudara. Pada

stadium lanjut, tidak jarang payudara harus diangkat demi menyelamatkan nyawa pasien. Bagi kaum wanita, kanker ini menjadi salah satu penyakit yang paling menakutkan karena mengenai organ yang dapat dilihat dan menjadi simbol kewanitaan.

Mamografi merupakan metode yang umum digunakan dalam diagnosis awal kanker payudara [26]. Pemeriksaan mamografi merupakan salah satu pemeriksaan sensitif untuk mendeteksi lesi yang tidak teraba (*nonpalpable*). Pengambilan gambar dengan mamografi telah meningkatkan jumlah kanker payudara yang terdeteksi *nonpalpable* dan bahkan sering pula *noninvasive* [3]. Laporan *radiologist* setelah melengkapi prosedur mamografi umumnya disertai dengan BI-RADS (*Breast Imaging Reporting and Data Systems*) yang terdiri atas 6 kategori. Selain itu terdapat beberapa hal yang dapat dilihat saat pemeriksaan dengan mamografi seperti berikut:

1. *Intermediate findings* menjelaskan keadaan jaringan payudara dan sel-sel di dalamnya berdasarkan lima indikator yaitu *well defined*, *developing*, *architectural distortion*, *skin thickening*, dan *symmetry*. *Well defined* menunjukkan adanya sel yang memiliki potensi untuk menjadi sel kanker namun tidak menginfiltrasi sel lainnya. Keadaan *developing* menunjukkan kondisi *well defined* di atas ambang batas tertentu. *Architectural distortion* merupakan keadaaan di mana sel-sel dalam jaringan payudara tidak membentuk jaringan sebagaimana mestinya. *Skin thickening* merupakan indikasi adanya penebalan kulit payudara. *Asymmetry* adalah keadaan payudara tidak simetris antara payudara kiri dan kanan.
2. *Suspicion of malignancy* atau indikasi kecurigaan malignansi menjelaskan bentuk kelainan yang terdapat dalam payudara atau melihat tanda-tanda keganasan (malignansi) yang tampak pada payudara. Indikator dalam pemeriksaan ini antara lain *mass*, *calcification*, dan *speculated sign*. *Mass* menunjukkan adanya penggumpalan (*lump*) dalam payudara. *Calcification* berarti telah terjadi proses pengapuran pengapuran berupa titik-titik pada jaringan payudara. *Speculated sign* merupakan penanda batas tumor di mana batas tumor tidak beraturan.
3. Letak kelainan dicatat oleh radiologist dengan menan-dai pada bagian payudara sebelah mana yang didapati kelainan.

## METODE PENELITIAN

### Sumber Data dan Variabel Penelitian

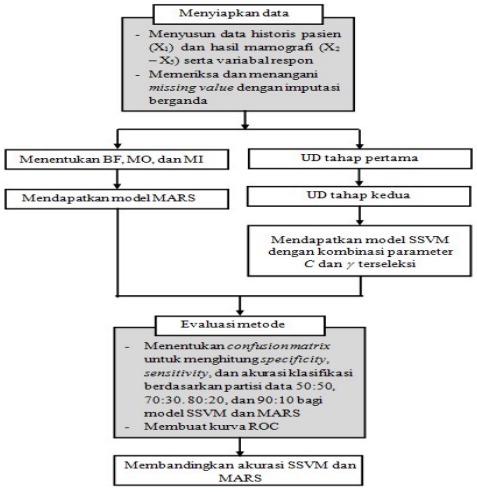
Seluruh data merupakan data sekunder yang dikumpulkan dari rekaman gambar mamografi pasien yang menjalani pemeriksaan payudara dan biopsi di salah satu rumah sakit kanker di Surabaya, Indonesia. Variabel respon ($Y$) merupakan variabel yang berisi kelas yang terdiri atas dua kategori yaitu kategori (-1) untuk *malignant* dan kategori (+1) untuk *benign*. Pengukuran variabel respon diperoleh dari hasil biopsi, sedangkan variabel-variabel prediktor merupakan data historis pasien catatan hasil mamografi. Variabel-variabel prediktor dijelaskan dalam Tabel 2.

### Metode Analisis

Langkah-langkah penelitian secara umum digambarkan dalam diagram alir Gambar 5.

**Tabel 2.** Definisi operasional variabel prediktor diagnosis kanker payudara

| Nama variabel | Kategori | Skala variabel |
|---|---|---|
| Usia ($X_1$) | - | Rasio |
| *Intermediate findings* ($X_2$) | 1. Tidak ada kelainan<br>2. Tepat didapati satu indikasi kelainan<br>3. Terdapat lebih dari satu indikasi kelainan | Nominal |
| Kecurigaan malignansi ($X_3$) | 1. Tidak ada tanda kanker<br>2. *Mass*<br>3. *Calcification*<br>4. *Speculated sign*<br>5. *Mass*, *Calcification*<br>6. Mass, Speculated sign<br>7. *Calcification*, *Speculated sign*<br>8. *Mass*, *Calcification*, *Speculated sign* | Nominal |
| BIRADS ($X_4$) | 1. C1, C2<br>2. C3<br>3. C4<br>4. C5 | Ordinal |
| Letak kelainan ($X_5$) | 1. Sebelah kanan<br>2. Sebelah kiri<br>3. Kedua payudara | Nominal |




**Gambar 5.** Kerangka penelitian diagnosis kanker payudara dengan SSVM dan MARS

## HASIL PENELITIAN

### Imputasi Berganda untuk *Missing Value*

Kelima variabel prediktor yang digunakan dalam penelitian ini memiliki data lengkap ($n$ = 267) kecuali variabel prediktor $X_3$ yaitu variabel yang menggam-barkan kecurigaan terhadap malignansi yang mengandung 11% data hilang. Ketiga metode klasifikasi dicobakan untuk data tidak lengkap yang nilai *missing value*-nya digantikan dengan nilai nol dan data lengkap yakni data yang telah diterapkan imputasi berganda terhadap *missing value*-nya, sesuai dengan prosedur yang dilakukan [12] dan [13] (Tabel 3). Imputasi berganda telah meningkatkan performansi metode klasifikasi dalam mendiagnosis malignansi kanker payudara. Pembahasan selanjutnya, metode klasifikasi baik menggunakan SSVM maupun MARS merujuk pada penggunaan data lengkap yang telah diterapkan imputasi berganda.

**Tabel 3.** Imputasi berganda pada $X_3$ meningkatkan akurasi klasifikasi (%)

| | Reg. Logistik Biner | SVM | CART |
|---|---|---|---|
| Tanpa imputasi berganda | 89,52 | 88,36 | 95,14 |
| Dengan imputasi berganda | 94,02 | 93,99 | 95,47 |

### Diagnosis Kanker Payudara dengan SSVM

Sesuai dengan parameter ($C$; $\gamma$) yang dipilih dalam *uniform design* dua tahap, yakni parameter SSVM yang menghasilkan akurasi tertinggi, Tabel 4 merupakan ringkasan perolehan akurasi tertinggi SSVM dalam mengklasifikasikan kanker payudara.

**Tabel 4.** Parameter SSVM yang menghasilkan akurasi tertinggi berdasarkan *uniform design* dua tahap

| Data | Akurasi (%) | $C$ | $\gamma$ |
|---|---|---|---|
| 5-*fold cv* | 99,63 | 464,16 | 0,1998 |
| 50:50 | 94,78 | 2,15* | 0,004588 |
| | | 121,15 | 0,000695 |
| 70:30 | 96,25 | 0,56* | 0,1065 |
| 80:20 | 94,34 | 0,56* | 0.1065 |
| | | 31,62 | 0,1707 |
| | | 464,16 | 0,1998 |
| 90:10 | 96,15 | 0,56* | 0.1065 |

* titik duplikasi

Penentuan variabel yang berpengaruh terhadap klasifikasi dilakukan dengan SVM 1-norm dengan menghitung **w**. Tabel 4.3 menunjukkan bahwa pada data dengan ukuran *training* 70, 80 dan 90% variabel prediktor yang menjelaskan letak kelainan pada payudara ($X_5$) tidak berpengaruh dalam penentuan kelas malignansi.

**Tabel 5.**Hasil perhitungan nilai *w* dengan SVM 1-*norm*

| Data | 50:50 | 0:30 | 80:20 | 90:10 |
|---|---|---|---|---|
| *C* | 2,15 | 0,56 | 0,56 | 0,56 |
| $w_1$ | 0,0396 | 0,0245 | 0,0177 | 0,0216 |
| $w_2$ | 0,6980 | 0,4657 | 0,4645 | 0,4838 |
| $w_3$ | 0,3218 | 0,4559 | 0,4433 | 0,4536 |
| $w_4$ | 0,8020 | 0,6128 | 0,6631 | 0,6436 |
| $w_5$ | 0,0495 | 0 | 0 | 0 |

## Diagnosis Kanker Payudara dengan MARS

Setiap data training memiliki model MARS yang berbeda. Data *training* dengan ukuran 50% dan 70%, sesuai dengan perolehan GCV minimum, memiliki derajat interaksi maksimum dua, sedangkan data training berukuran 80% dan 90% memiliki derajat maksimum interaksi sebanyak tiga.

**Tabel 6.** Akurasi dan variabel prediktor dalam model MARS

| Data terpartisi | Akurasi (%) | Variabel prediktor yang ada dalam model |
|---|---|---|
| 5- *fold cv* | 95,88 | - |
| 50 | 97,76 | $X_1, X_2,$dan $X_3$ |
| 70 | 96,25 | $X_1, X_2, X_3,$ dan $X_4$ |
| 80 | 94,33 | $X_1, X_2, X_3,$ dan $X_4$ |
| 90 | 96,15 | $X_1, X_2, X_3,$ dan $X_5$ |

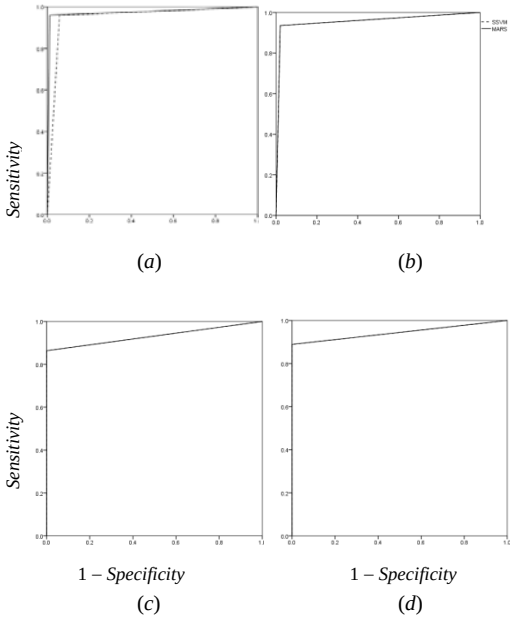
Secara umum, variabel prediktor yang menjelaskan usia pasien ($X_1$), tanda-tanda kelainan pada pemeriksaan awal atau *intermediate findings* ($X_2$) dan tanda-tanda kecurigaan malignansi ($X_3$) selalu muncul dalam model untuk keempat kelompok data *training*. Variabel prediktor yang menjelaskan letak atau posisi kelainan ($X_5$) hanya muncul dalam model untuk data *training* 90% (Tabel 6). Akurasi yang dihasilkan oleh model MARS untuk data *training* berukuran 70, 80 dan 90% sama dengan akurasi yang dihasilkan oleh SSVM pada data *training* berukuran sama secara berurutan.

## Evaluasi Performansi Diagnosis Kanker Payudara dengan SSVM dan MARS

Evaluasi klasifikasi dapat dilihat dari nilai *sensitivity* dan *specificity* yang ditunjukkan pada Tabel 7. Tingkat akurasi yang dihasilkan dari data 5-*fold cross validation* adalah 99,63% untuk SSVM dan 95,88% untuk MARS.

**Tabel 7.** *Sensitivity* dan *specificity* dari hasil klasifikasi SSVM dan MARS

| Data | Akurasi (%) | | *Sensitivity* (%) | | *Specificity* (%) | |
|---|---|---|---|---|---|---|
| | SSVM | MARS | SSVM | MARS | SSVM | MARS |
| 5-*fold* | 99,63 | 95,88 | - | - | - | - |
| 50:50 | 94,78 | 97,76 | 94,19 | 98,84 | 95,83 | 95,83 |
| 70:30 | 96,25 | 96,25 | 97,96 | 97,96 | 93,55 | 93,55 |
| 80:20 | 94,34 | 94,33 | 100,00 | 100,00 | 86,36 | 86,36 |
| 90:10 | 96,15 | 96,15 | 100,00 | 100,00 | 88,89 | 88,89 |




**Gambar 6.** Kurva ROC untuk data terpartisi (a) 50:50, (b) 70:30, (c) 80:20 dan (d) 90:10

Kurva ROC (*receiver operating characteristics*) juga disajikan pada Gambar 6. Area di bawah kurva (AUC) dihitung, semakin luas area menunjukkan performansi klasifikasi yang semakin baik. Secara umum berdasarkan kurva ROC pada Gambar 6 (*a* sampai dengan *d*), klasifikasi menggunakan metode SSVM lebih baik performansinya sebab wilayah di bawah kurva lebih luas dibandingkaan MARS pada dua kurva ROC terakhir.

Luas wilayah di bawah kurva secara akurat diringkas dalam Tabel 8. Berdasarkan luasan-luasan tersebut dapat ditarik kesimpulan bahwa pada ukuran data *training* yang lebih besar, SSVM memiliki performansi lebih baik daripada MARS untuk mengklasifikasikan diagnosis kanker payudara pada penelitian ini.

**Tabel 8** Luas area di bawah kurva ROC hasil klasifikasi SSVM dan MARS

| Data | AUC SSVM | AUC MARS |
|---|---|---|
| 50:50 | 0,950 | 0,973 |
| 70:30 | 0,958 | 0,958 |
| 80:20 | 0,932 | 0,932 |
| 90:10 | 0,944 | 0,944 |

## KESIMPULAN

Pencegahan terhadap tingginya angka penderita kanker payudara di Indonesia dapat dilakukan dengan mengupayakan prosedur identifikasi dan diagnosis kelainan pada payudara secara efisien dan akurat. Implementasi SSVM dan MARS pada penelitian ini menunjukkan bahwa metode *machine learning* dapat mengklasifikasikan diagnosis kanker payudara dengan tingkat akurasi yang cukup tinggi (lebih dari 90%).

Berdasarkan tingkat akurasi yang dihasilkan melalui validasi silang (*cross validation*) dengan 5-*fold*, SSVM menghasilkan akurasi sebesar 99,63% sedangkan MARS menghasilkan 95,88%. Secara umum untuk partisi data 50:50, 70:30, 80:20 maupun 90:10, SSVM tidak lebih baik dibandingkan MARS. Hal ini menunjukkan bahwa kedua metode sama baiknya dalam menentukan kelas malignansi kanker payudara. Tingkat akurasi kedua metode dalam mendiagnosis kanker payudara ke dalam kelompok *benign* dan *malignant* yang cukup tinggi dipercaya dapat mendukung prosedur pemeriksaan dan diagnosis kanker payudara.

## DAFTAR PUSTAKA

[1]Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM, (2008), GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 10 [Internet]. Lyon, France: International Agency for Research on Cancer 2010. Tersedia pada http://globocan.iarc.fr. Diakses terakhir Juni 2012.

[2]World Health Ranking, http://www.worldlifeexpectancy.com/country-health-profile/indonesia, Diakses terakhir pada Juli 2012.

[3]Gajdos C, Tartter PI, Bleiweiss IJ, Hermann G, de Csepel J, Estabrook A, Rademaker AW, (2002), Mammography appearance of nonpalpable breast cancer reflects pathologi characteristics, *Annals of Surgery*, Vol. 235, No. 2, hal. 246 – 251.

[4]Subashini TS, Ramalingam V, Palanivel S, (2009), Breast mass classification based on cytological patterns using RBFNN and SVM, *Expert Systems and Applications*, 36, hal. 5284 – 5290.

[5]Shi X, Cheng HD, Hu L, Ju W, Tian J, (2010), Detection and Classification of masses in breast ultrasound images, *Digital Signal Processing*, 20, hal. 824 – 836.

[6]Chen HL, Yang B, Liu J, Liu DY, (2011), A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, *Expert Systems with Applications*, 38, hal. 9014 – 9022.

[7]Addeh A dan Ebrahimzadeh A, (2012), Breast Cancer Recognition Using a Novel Hybrid Intelligent Method, *Journal of Medical Signal and Sensors*, Vol. 2, No. 2, hal. 22 – 30.

[8]Purnami SW dan Embong A, (2008), Feature selection and classification of breast cancer diagnosis based on SVM, *The 3rd International Symposium of Information Technology (ITSIM08) KLCC, Kuala Lumpur Malaysia*.

[9]_____, (2008) Smooth Support vector machine for breast cancer classification, *The 4th IMT-GT 2008 Conference of Mathematics, Statistics and Its Application (ICMSA 2008), Banda Aceh, Indonesia*.

[10] Purnami SW, Embong A, Zain JM, (2009) Application of data mining technique using best polynomial smoot support vector machine in breast cancer diagnosis, *International Conference in Robotics, Vision, Signal Symposisum and Power Application (Rovisp 2009) Langkawi Kedah, Malaysia*.

[11] Chou SM, Lee TS, Shao YE, Chen IF, (2004), Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines, *Journal of Expert System with Application*, 20, hal. 133 – 142.

[12] Novianti FA, (2012), *Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi, Studi Kasus: RS 'X' Surabaya*, Skripsi ITS.

[13] Pertiwi YD, (2012), *Klasifikasi Diagnosa Kanker Payudara (Patologi Anatomi) Pasien Kanker Payudara di RS 'X' Surabaya dengan Pendekatan CART*, Skripsi ITS.

[14] Huang CM, Lee YJ, Lin DKJ, Huang SY, (2007), Model selection for support vector machie via uniform design, *Computational Statistics and Data Analysis*, Vol. 52, hal. 335 – 346.

[15] Hair Jr JF, Black WC, Babin BJ, Anderson RE, (2010), *Multivariate Data Analysis* 7th Ed, Prentice Hall.

[16] Yuan Y, (2011), Multiple Imputation Using SAS Software, *Journal of Statistics Software*, Vol. 45, No. 6.

[17] Rubin DB, (1987), *Multiple Imputation for Nonresponse Surveys*, John Wiley and Sons.

[18] Lee YJ dan Mangasarian OL, (2001), A Smooth Support Vector Machine, *Journal of Computational Optimization and Applications*, 20, hal. 5 – 22.

[19] Zhu J, Rosset S, Hastie T, Tibshirani R,(2003), 1-norm support vector machines, *Neural Information Proceeding Systems* 16.

[20] Friedman JH, (1991), Multivariate adaptive regression splines, *Annals of Statistics*, 19, hal. 1 – 67.

[21] Breiman L, Friedman JH, Olshen RA, Stone CJ, (1984), *Classification and Regression Trees*, Wadsworth, Pacific Grove, CA.

[22] Hastie T dan Tibshirani R, (1990), *Generalized Additive Models*, Chapman & Hall, London.

[23] Agresti A, (2006), *An Introduction to Categorical Data*

*Analysis* 2nd Ed, John Wiley & Sons: New Jersey.

[24] Menendez LA, de Cos Juez FJ, Lasheras FS, Riesgo JAA, (2010), Artificial neural networks applied to cancer detection in a breast screening programme, *Journal of Mathematical and Computer Modelling*, Vo. 52, hal. 983 – 991.

[25] Faraggi D dan Reiser B, (2002), Estimation of the area under the ROC curve, *Journal of Statistics in Medicine*, Vol. 21, hal. 3093 – 3106.

[26] Sun Z, Liang HW, Xu HM, (2005), Classification of breast cancer microcal-cification, *Chinese Medical Journal*, Vol. 118, No. 17, hal. 1429 – 1435.