# PENERAPAN ADABOOST UNTUK KLASIFIKASI SUPPORT VECTOR MACHINE GUNA MENINGKATKAN AKURASI PADA DIAGNOSA CHRONIC KIDNEY DISEASE

**Eka Listiana[1*], Much Aziz Muslim [1]**
[1]Program Studi Teknik Informatika, Fakultas MIPA, Universitas Negeri Semarang
[*]Email:ekalisti@students.unnes.ac.id

**Abstrak**

*Database masa kini berkembang dengan sangat pesat khususnya dalam bidang kesehatan. Data tersebut apabila tidak diolah dengan baik maka akan menjadi sebuah tumpukan data yang tidak bermanfaat, sehingga perlu adanya proses untuk mengolah data tersebut menjadi sebuah informasi yang bermanfaat. Proses tersebut biasa disebut dengan data mining yang merupakan suatu bidang ilmu penelitian yang mampu mengolah database menjadi pengetahuan yang dapat dimanfaatkan khusunya dalam penelitian ini akan digunakan untuk mendiagnosa penyakit, diantaranya chronic kidney disease. Salah satu metode data mining yang digunakan untuk memprediksi sebuah keputusan dalam suatu hal adalah klasifikasi, di mana dalam metode klasifikasi ada algoritma support vector machine yang bisa digunakan untuk mendiagnosa chronic kidney disease. Dalam penelitian ini untuk meningkatkan akurasi algoritma support vector machine dalam mendiagnosa chronic kidney disease menggunakan adaptive boosting (adaboost) sebagai ensemble learning dengan pemilihan kernel, nilai parameter C, dan iterasi yang sesuai. Dari hasil percobaan, menerapkan adaboost, dengan kernel linier dan pemilihan nilai parameter C pada algoritma support vector machine dalam mendiagnosa chronic kidney disease menunjukkan bahwa tingkat akurasi mempunyai peningkatan sebesar 37% dengan pemaparan hasil seperti berikut, 62,5% (SVM); 97,75% (SVM+linier kernel); 99,5% (SVM+linier kernel +adaboost).*

***Kata Kunci:*** *adaboost, data mining*, *SVM*, *Adaptive boosting*, *chronic kidney disease*

## 1. PENDAHULUAN

Perkembangan *database* pada masa kini tumbuh dengan sangat pesat, khususnya data dalam bidang kesehatan. Dari kumpulan banyak data tersebut apabila tidak digunakan maka hanya menjadi sebuah kumpulan data yang tidak bermanfaat. Oleh karena itu, dari tumpukan data yang tidak bermanfaat tersebut dapat dijadikan sebagai sumber data yang kemudian diolah supaya lebih bermanfaat yang biasa sering disebut dengan istilah *data mining*. *Data mining* menurut adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. Salah satu metode atau teknik yang digunakan untuk memprediksi sebuah keputusan, yaitu klasifikasi.

Menurut Han J, *et al*. (2012) Klasifikasi adalah suatu proses yang digunakan untuk menemukan model (atau fungsi) dengan menggambarkan dan membedakan kelas data atau konsep. Salah satu jenis algoritma pada klasifikasi yaitu *Supprot Vector Machine* (SVM). Menurut Dhayanand dan Vijayarani (2015) *Support vector machine* merupakan metode baru yang digunakan untuk klasifikasi data baik itu data linier maupun nonlinier, yang menciptakan *diskrit hyperplane* dalam ruang *descriptor* dari data pelatihan dan diklasifikaskan berdasarkan sisi *hyperplane* berada. Menurut Chezian dan Kumar (2015) *Support vector machine* memetakan data *input* nonlinier ke beberapa ruang dimensi yang lebih tinggi di mana data dapat dipisahkan secara linier, sehingga akan memberikan klasifikasi atau regresi data yang besar.

Agoritma tersebut dapat dimanfaatkan dan membantu ahli medis dalam mendiagnosa suatu penyakit, salah satunya adalah *Chronic Kidney Disease* (CKD). Menurut Levey dan Coresh (2012) CKD adalah gangguan heterogen yang mempengaruhi struktur dan fungsi ginjal secara progresif dan sukar untuk pulih kembali, dimana tubuh tidak mampu memelihara metabolisme dan gagal memelihara keseimbangan cairan dan elektrolit yang berakibat pada peningkatan ureum. Sedangkan menurut Muslim, dkk., (2015) CKD juga merupakan proses fisiologi patogen dengan berbagai variasi yang akan menyebabkan turunnya fungsi renal (kelenjar pada ginjal) secara signifikan dimana yang pada akhirnya akan terjadi gagal ginjal. CKD pada umumnya merupakan suatu penyakit yang dapat dideteksi melalui tes urin dan tes darah dari uji laboratorium secara rutin

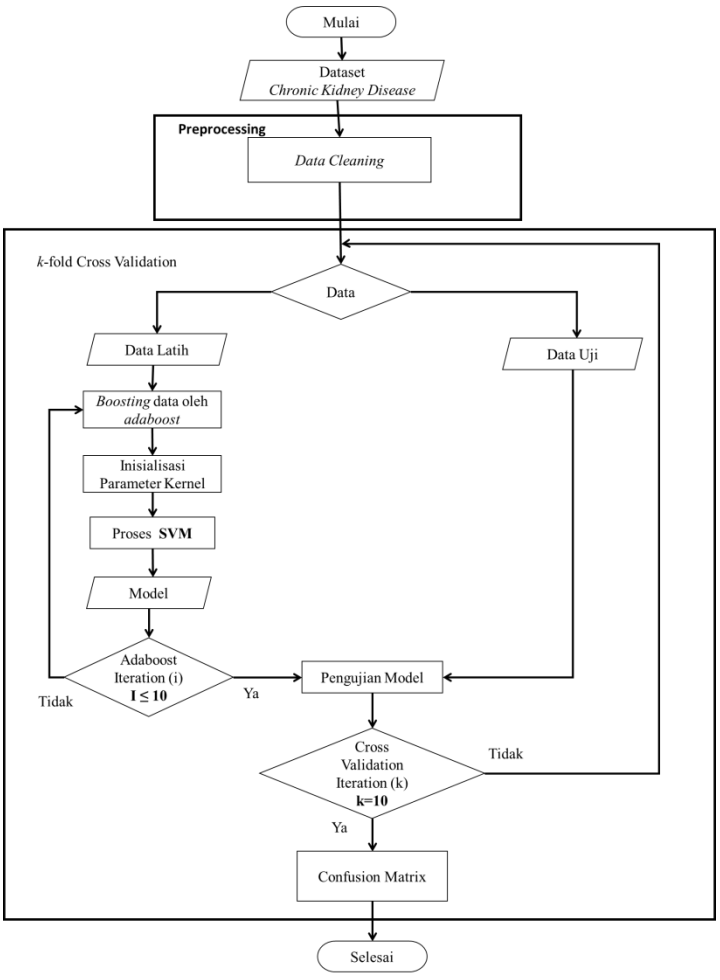
dan beberapa perawatan untuk mencegah pengembangan dan memperlambat perkembangan penyakit.

Penelitian diagnosa ini menggunakan *dataset chronic kidney disease*. *Dataset* yang digunakan dalam penelitian ini diperoleh dari UCI *repository of machine learning datasets*. *Datasets* merupakan kumpulan dari objek dan sifat atau karakteristik dari suatu objek itu sendiri (atribut). Penelitian ini menggunakan *ensemble learning adaboost* sebagai teknik untuk memperbaiki tingkat klasifikasi yang diterapkan untuk mengklasifikasikan teks pada komentar dari review suatu restoran untuk meningkatkan akurasi dalam mendiagnosa *chronic kidney disease*.

Tujuan dari penelitian ini yaitu meningkatkan akurasi dari algoritma support vector machine dengan menerapkan adaboost dan pemilihan kernel dan nilai parameter C dalam memprediksi *chronic kidney disease* dengan membandingkan hasil sesudah dan sebelum diterapkan *information gain* dan *adaboost*. Validasi dalam penelitian ini menggunakan 10 *fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confussion matrix*.

## 2. METODOLOGI

Penelitian ini bertujuan untuk mengoptimalisasi atribut dari *dataset* yang digunakan guna meningkatkan akurasi dengan menerapkan *adaboost* serta pemilihan parameter C dan kernel linier. Penelitian ini diterapkan pada algoritma *support vector machine* untuk mendiagnosa *chronic kidney disease*, agar dapat membuktikan bahwa *adaboost* serta dengan pemilihan parameter C dan kernel linier dapat meningkatkan akurasi, maka penelitian ini akan membandingkan performa akurasi dari algoritma *suppport vector machine* yang tidak diterapkan adaboost dengan yang menerapkannya. *Flowchart* penerapan *adaboost* pada algoritma *support vector machine* dapat dilihat pada Gambar 1.




**Gambar 1. *Flowchart* penerapan *adaboost* pada algoritma *support vector machine***

***Support Vector Machine* (SVM)**

SVM dikembangkan pada tahun 1992 oleh Vladimir Vapnik dan rekannya, Bernhard Boser dan Isabelle Guyon yang dikembangkan dari teori *structural risk minimaization*. Menurut Li, *et al.* (2008) Dengan menggunakan trik kernel untuk memetakan sampel pelatihan dari ruang input ke ruang fitur dimensi tinggi. Xuchun, *et al.* (2008) menjelaskan dasar untuk SVM telah ada sejak 1960-an metode ini menjadikan SVM sebagai metode baru yang menjanjikan untuk mengklasifikasi data, baik data *linear* maupun *nonlinear*. SVM adalah sebuah algoritma yang bekerja menggunakan pemetaan *nonlinear* untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi, dalam dimensi yang baru, kemudian akan mencari *linear* optimal pemisah *hyperplane* (yaitu, "*decision boundary*" yang memisahkan tupel dari satu kelas dengan kelas lainnya). Dengan pemetaan *nonlinear* yang tepat untuk dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan dengan *hyperplane*. SVM menemukan *hyperplane* ini menggunakan *support vector* ("*essential*" *training tupels*) dan *margins* (didefinisikan oleh *support vectors*).

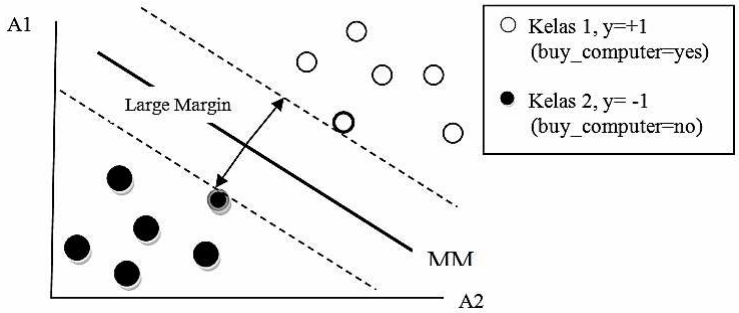
Langkah awal suatu algoritma SVM adalah pendefinisian persamaan suatu *hyperplane* pemisah yang dituliskan dengan Persaman 1.

$$w.X + b = 0 \quad (1)$$

Di mana w merupakan suatu bobot vektor, yaitu $w = \{w_1, w_2, \ldots, w_n\}$; n adalah jumlah atribut dan b merupakan suatu skalar yang disebut dengan bias. Jika berdasarkan pada atribut A1, A2 dengan permisalan tupel pelatihan X = $(x_1, x_2)$, x1 dan x2 merupakan nilai dari atribut A1 dan A2, dan jika b dianggap sebagai suatu bobot tambahan $w_0$, maka persamaan suatu *hyperplane* pemisah dapat ditulis ulang seperti pada Persamaan 2.

$$w_0 + w_1 x_1 + w_2 x_2 = 0 \quad (2)$$

Setelah persamaan dapat didefinisikan, nilai $x_1$ dan $x_2$ dapat dimasukkan ke dalam persamaan untuk mencari bobot $w_1, w_2$, dan $w_0$ atau b. Grafik pemisahan dua kelas data dengan margin maksimum dapat dilihat pada Gambar 2.




**Gambar 2. Pemisahan dua kelas data dengan margin maksimum**

Pada Gambar 2, SVM menemukan *hyperlane* pemisah maksimum, yaitu *hyperlane* yang mempunyai jarak maksimum antara tupel pelatihan terdekat. *Support vector* ditunjukkan dengan batasan tebal pada titik tupel. Dengan demikian, setiap titik yang terletak di atas *hyperplane* pemisah memenuhi Persamaan 3.

$$w_0 + w_1 x_1 + w_2 x_2 > 0 \quad (3)$$

Sedangkan, titik yang terletak di bawah *hyperlane* pemisah memenuhi rumus seperti pada Persamaan 4.

$$w_0 + w_1 x_1 + w_2 x_2 < 0 \quad (4)$$

Melihat dua kondisi di atas, maka didapatkan dua persamaan *hyperplane*, seperti pada Persamaan 5 dan 6.

$$H_1: w_0 + w_1 x_1 + w_2 x_2 \geq 0 \quad (5)$$

$$H_2: w_0 + w_1 x_1 + w_2 x_2 \leq 0 \quad (6)$$

Artinya, setiap *tuple* yang berada di atas $H_1$ memikiki kelas +1, dan setiap *tuple* yang berada di bawah $H_2$ memiliki kelas -1. Perumusan model SVM menggunakan trik matematika yaitu *lagrangian formulation.* Perumusan model SVM menggunakan trik matematika yaitu *lagrangian formulation*. Berdasarkan *lagrangian formulation*, *Maksimum Margin Hyperplane* (MMH) dapat ditulis ulang sebagai suatu batas keputusan (*decision boundary*) dituliskan dengan Persamaan 7.

$$d(X^T) = \sum_{i=1}^{l} y_i a_i \; X_i X^T + b_0 \quad (7)$$

$y_i$ adalah label kelas dari *support vector* $X_i$, $X^T$ merupakan suatu tupel *test*. $a_i$ dan $b_o$ adalah parameter numerik yang ditentukan secara otomatis oleh optimalisasi algoritma SVM dan $l$ adalah jumlah *vector support*.

***Adaptive Boosting* (*Adaboost*)**

*Adaptive* boosting (*adaboost*) merupakan salah satu dari beberapa varian pada algoritma *boosting* [1]. *Adaboost* merupakan *ensemble learning* yang sering digunakan pada algoritma *boosting*. *Boosting* bisa dikombinasikan dengan *classifier* algoritma yang lain untuk meningkatkan performa klasifikasi. Tentunya secara intuitif, penggabungan beberapa model akan membantu jika model tersebut berbeda satu sama lain [1].

*Adaboost* dan variannya telah sukses diterapkan pada beberapa bidang (*domain*) karena dasar teorinya yang kuat, presdiksi yang akurat, dan kesederhanaan yang besar. Langkah-langkah pada algoritma *adaboost* adalah sebagai berikut.

a. *Input*: Suatu kumpulan *sample* penelitian dengan label $\{(x_i, y_i), \ldots, (x_N, y_N)\}$, suatu *component learn* algoritma, jumlah perputaran T.
b. *Initialize*: Bobot suatu sampel pelatihan $w_i^1 = 1/N$, untuk semua i=1,...,N
c. Do for t= 1, ...,T
d. Gunakan c*omponent learn* algoritma untuk melatih suatu komponen klasifikasi, $h_t$, pada *sample* bobot pelatihan.
e. Hitung kesalahan pelatihannya pada $h_t$: $\varepsilon_t = \sum_{i=1}^{N} w_i^t , y_i \neq h_t(x_i)$
f. Tetapkan bobot untuk *component classifier* $h_t = \alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
g. *Update* bobot *sample* pelatihan $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{C_t}, i = 1, \ldots, N$ $C_t$ adalah suatu konstanta normalisasi.
h. *Output*: $f(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

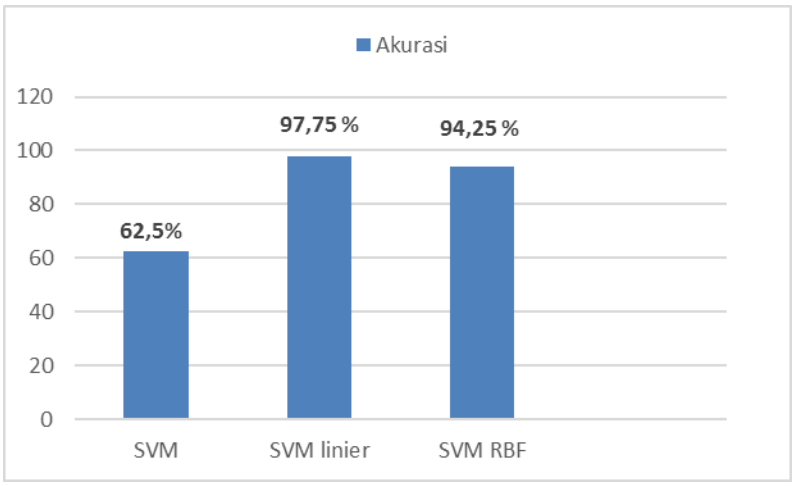
***Dataset***

*Dataset* merupakan kumpulan dari objek dan atributnya. Atribut merupakan sifat atau karakteristik dari suatu objek. Contohnya: warna mata seseorang, suhu, dsb. Atribut juga dikenal sebagai variabel, *field*, karakteristik atau fitur. Kumpulan dari atribut menggambarkan sebuah objek bisa disebut *record*, titik, kasus, sample, entitas atau *instance*[1].

Selain itu, *dataset* merupakan suatu kumpulan data atau satu data statistik di mana setiap atribut data dapat mewakili suatu variabel dan setiap instance memiliki deskripsi sendiri (Tang, *et al*.(2014) dan Ian, *et al*. (2011)).

## 3. HASIL DAN PEMBAHASAN

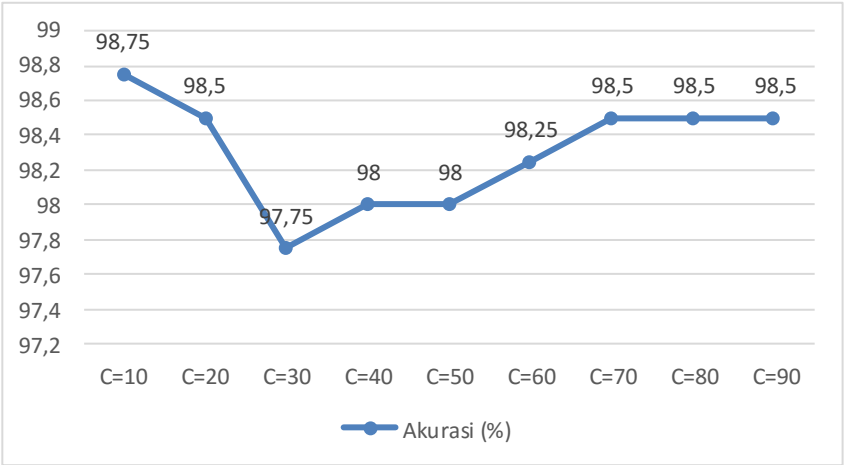
Proses eksperimen ini menggunakan aplikasi weka 3.6. Untuk pengujian model dilakukan menggunakan *dataset* chronic *kidney disease* yang diambil dari UCI *repository of machine learning*. Spesifikasi yang digunakan untuk penelitian ini, yaitu *processor intel(R) Core(TM) i3-3227U CPU @1.90GHz (4 CPUs), ~1,9GHz*; memori 2,00 GB; dan sistem operasi *Windows 8.1*. Langkah awal dari penelitian ini, yaitu pemilihan kernel yang dilakukan pada pengklasifikasian SVM. Perbandingan antara SVM dan SVM linier kernel dapat dilihat pada Gambar 2.




**Gambar 3. Perbnadingan akurasi SVM, SVM Linier, dan SVM RBF**

Hasil perbandingan akurasi diatas SVM dengan kernel linier lebih baik daripada SVM RBF dengan data chronic kidney disease. Maka, SVM dengan kernel linier akan dicari lebih lanjut nilai parameter C untuk memperoleh akurasi yang lebih baik dengan menerapkan *adaboost* untuk dapat memperoleh akurasi yang lebih baik.

Penilaian nilai parameter C akan dipilih dengan dengan melihat implikasi terhadap hasil akurasi yang paling tinggi terhadap nilai C yang diberikan. Nilai C ditentukan sebelumnya dengan mencari tahu nilai C dari 10 sampai dengan 90. Berhenti pada 90 disebabkan pada nilai C dari 70 sampai dengan 90 mengalami nilai yang *stagnan* (tidak aa perubahan). Hasil akurasi perbandingan nilai parameter C pada SVM linier dapat dilihat pada Gambar 4.




**Gambar 4. Hasil akurasi perbandingan nilai parameter C pada SVM linier**

Hasil percobaan nilai C dengan akurasi yang paling tinggi didapat oleh C dengan nilai 10 dengan akurasi 98,75%. Maka langkah selanjutnya SVM kernel linier dan nilai parameter C dengan nilai 10 akan digunakan sebagai *base classifier* dari *ensemble learning adaboost*.

Berdasarkan tingkat kesalahan base classifier dapat dikurangi dengan memperbesar iterasi yang tepat pada adaboost. Maka percobaan selanjutnya, pemilihan iterasi K digunakan untuk memperoleh hasil yang lebih baik. Perbandingan hasil akurasi dari setiap iterasi dapat dilihat pada Tabel 1. Di mana pada tabel tersebut menunjukkan proses peningkatan akurasi algoritma support vctor machine dengan iterasi adaboost yang dimulai dari angka 1 sampai dengan default putarannya yaitu 10. Sehingga, setelah diiterasi ke-3 akurasi baru meningkat, dalam hal ini berarti adaboost telah dapat memperbaiki kesalahan pada SVM.

**Tabel 1.** Perbandingan hasil akurasi dari setiap iterasi

| **Iterasi** | **K=1** | **K=2** | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** | **K=10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Akurasi (%)** | 98,75 | 98,75 | 99,5 | 99,5 | 99,5 | 99,25 | 99,25 | 99,25 | 99,25 | 99,25 |

Berdasarkan percobaan yang telah dilakukan pada pengklasifikasi dasar SVM, yaitu SVM linier, SVM dioptimasi dengan pemilihan nilai parameter C=10 dan adaboost-SVM. Maka hasil perbandingan akurasi dapat dilihat pada Tabel 2.

**Tabel 2.** Hasil perbandingan akurasi model

| **Perbandingan Model** | **SVM** | **SVM linier** | **SVM Optimasi (C=10)** | **Adaboost-SVM** |
|---|---|---|---|---|
| Akurasi (%) | 62,5 | 97,75 | 98,75 | 99,5 |

Dari pengolahan data yang sudah dilakukan dengan metode *boosting* yaitu *adaboost*, terbukti dapat meningkatkan akurasi algoritma support vector machine pada diagnosa *chronic kidney disease*. Data yang digunakan dapat diklasifikasikan dengan baik ke dalam bentuk positif dan negatif.

## 4. KESIMPULAN

Pengolahan *dataset chronic kidney disease* yang diambil dari UCI *repository of machine learning* peneliti menggunakan algoritma support vector machine. Dan ternyata, jika hanya *support vector machine* saja yang digunakan hanya mencapai 62,5%, jika pemilihan kernel menggunakan kernel linier mencapai 97,75%, setelah diberi nilai parameter C mencapai 98,75%, dan setelah ditambah dengan *adaboost* mencapai 99,5%. Dari penelitian yang dilakukan dengan menggunakan *boosting* dan pemeilihan kernel serta nilai parameter C memiliki peningkatan akurasi sebesar 37%.

## DAFTAR PUSTAKA

Chezian, D. R. M., & Kumar, K. S. 2014. Support Vector Machine and K-Nearest Neighbor Based Analysis for the Prediction of Hypothyroid. *International Journal of Pharma and Bio Sciences*. 5(4): (B) 447- 453.

Dhayanand, M, S., & Vijayarani, S. 2015. Data Mining Classification Algorithms for Kidney Disease Prediction. *International Journal on Cybernetics & Informatics (IJCI)*. Vol. 4 (4): 13-25.

Han, J., Micheline, K. & Jian, P. 2012. *Data mining: Concepts and Techniques* (3th ed.). Waltham, MA: Elsevier/Morgan Kaufmann.

Ian H Witten, Eibe Frank, and Mark A Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. USA: Morgan Kaufmann Publishers, 2011.

Levey, A. S., & Coresh, J. 2012. Chronic kidney disease. *The Lancet*. 379(9811): 165-180.

Li, X., Wang, L., & Sung, E. 2008. AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*. Vol. 21 (5): 785-795.

Muslim, M. A., Kurniawati, I., & Sugiharti, E. 2015. Expert System Diagnosa Chronic Kidney Disease Based On Mamdani Fuzzy Inference System. *Journal of Theoretical and applied Information Technology*, 78(1), 70.

Xuchun Li, Lei Wang, & Eric Sung. 2008. AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*. vol. 21, pp. 785–795.

Tang, J., Alelyani, S., & Liu, H. 2014. Feature selection for classification: A review. *Data Classification: Algorithms and Applications.* (37).