JURNAL ILMIAH INFORMATIKA GLOBAL VOLUME 11 No. 2 Desember 2020
ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes

Lastri Widya Astuti[1)], Imelda Saluza[2)], Faradilla[3)], M. Fadhiel Alie[4)]

[1)] Program Studi Teknik Informatika, Universitas IGM
[2)]Program Studi Manajemen Informatika, Universitas IGM
[3,4)]Program Studi Sistem Informasi, Universitas IGM
Jl. Jend.Sudirman No 629 Km 4 Palembang
Email : lastriwidya@uigm.ac.id[1)], Imeldasaluza@uigm.ac.id[2)], faradilla.hakim@uigm.ac.id[3)], fadhiel@uigm.ac.id[4)]

ABSTRACT

Breast cancer is a type of malignant tumor which is still the number one killer where the process of spread or metastasis takes a long time. The number of breast cancer sufferers increases every year so that if detected or caught early, prevention can be done early so as to reduce the number of breast cancer sufferers. To reduce the risk of increasing the number of cancer patients, it is necessary to do early detection, several methods can be used to assist the early detection process such as cancer screening, or computational methods. Several machine learning methods that have been chosen to solve cases of breast cancer prediction, especially the classification algorithm, including Naive Bayes have the advantage of being simple but having high accuracy even though they use little data. Weaknesses in Naive Bayes, namely the prediction of the probability result is not running optimally and the lack of selection of relevant features to the classification so that the accuracy is low. This research is intended to build a classification system for detecting breast cancer using the Naive Bayes method, by adding a forward selection method for feature selection from the many features that exist in breast cancer data, because not all features are features that can be used in the classification process. The result of combining the Naive Bayes method and the forward selection method in feature selection can increase the accuracy value of 96.49% detection of breast cancer patients.

Keywords : *Cancer, breast, naive bayes, forward selection, accuracy*

ABSTRAK

Kanker payudara merupakan jenis tumor ganas yang hingga kini masih menjadi pembunuh nomor satu dimana proses penyebaran atau metastasis memakan waktu yang lama. Jumlah penderita kanker payudara meningkat setiap tahunnya sehingga apabila dideteksi atau diketahui lebih awal bisa dilakukan pencegahan sejak dini sehingga dapat menekan angka penderita kanker payudara. Untuk mengurangi resiko peningkatan jumlah penderita kanker perlu dilakukan deteksi dini, beberapa metoda dapat digunakan untuk membantu proses pendeteksian diawal seperti *cancer screening*, atau dengan metode komputasi. Beberapa metode *machine learning* yang banyak dipilih untuk menyelesaikan kasus prediksi kanker payudara ini terutama algoritma klasifikasi diantaranya Naive bayes memiliki keunggulan sederhana tapi memiliki akurasi yang tinggi meskipun menggunakan data yang sedikit. Kelemahan dalam *Naive Bayes*, yaitu prediksi hasil probabilitas berjalan tidak optimal serta kurangnya pemilihan fitur yang relevan terhadap klasifikasi sehingga akurasi menjadi rendah. Penelitian ini dimaksudkan untuk membangun sistem klasifikasi untuk mendeteksi penyakit kanker payudara menggunakan metode *naive bayes*, dengan menambahkan metode *forward selection* untuk pemilihan fitur dari banyak fitur yang ada pada data kanker payudara, karena tidak semua fitur merupakan fitur yang dapat digunakan pada proses klasifikasi. Hasil pengabungan antara metode naive bayes dan metode forward selection dalam pemilihan fitur dapat meningkatkan nilai akurasi 96.49 % deteksi penderita kanker payudara.

Kata Kunci : Kanker, payudara, naive bayes, forward selection, akurasi

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

1. Pendahuluan

Kanker payudara merupakan jenis tumor ganas yang hingga kini masih menjadi pembunuh nomor satu dimana proses penyebaran atau metastasis memakan waktu yang lama, sehingga apabila diketahui dan dilakukan pencegahan sejak dini maka dapat menekan angka penderita kanker payudara [1]. Kasus kanker payudara pada wanita di negara maju terjadi lebih sedikit daripada negara berkembang yakni sebanyak 794.000 kasus, sedangkan pada negara berkembang kasus kanker payudara sebanyak 833.000 kasus. Kanker payudara sendiri umumnya menyerang perempuan dan merupakan salah satu kanker terbanyak yang terjadi di Indonesia [2]. Jumlah penderita kanker payudara menunjukkan bahwa terdapat peningkatan setiap tahunnya, dimana diperkirakan akan terus meningkat hingga sebesar 4 (empat) kali lipat jumlahnya pada tahun 2020 [3]. Untuk mengurangi resiko peningkatan jumlah penderita kanker perlu dilakukan deteksi dini, beberapa metoda dapat digunakan untuk membantu proses pendeteksian diawal seperti *cancer screening*, atau dengan metode komputasi seperti machine learning [4].

Beberapa metode machine learning yang banyak dipilih untuk menyelesaikan kasus prediksi kanker payudara ini terutama algoritma klasifikasi diantaranya Support Vector Machine (SVM) [4], Artificial Neural Network (ANN) [5], Neural Network [6], dan naive bayes [7] ANN memiliki kelebihan untuk menemukan pola dari data yang terlalu rumit untuk diketahui oleh manusia atau dengan teknik komputasi lainnya, selain itu multi layer perceptron memiliki kekurangan yaitu sulit untuk menemukan pola bila data berdimensi tinggi, sementara SVM memiliki keunggulan mengatasi masalah klasifikasi dan regresi linier maupun nonlinier yang dapat menjadi satu kemampuan algoritma pembelajaran untuk klasifikasi serta regresi, selain itu SVM juga memiliki akurasi yang tinggi dan tingkat kesalahan yang relative kecil, dan kemampuan untuk mengatasi *overfitting*. Naive bayes memiliki keunggulan sederhana tapi memiliki akurasi yang tinggi meskipun menggunakan data yang sedikit.

Naïve Bayes adalah perhitungan probabilitas dengan metode pengklasifikasian. Model ini mudah untuk dibangun dan tidak *complicated*, sehingga dianggap tepat untuk database yang berukuran kecil sampai berukuran besar. Algoritma *Naive Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Kelemahan dalam *Naive Bayes*, yaitu prediksi hasil probabilitas berjalan tidak optimal serta kurangnya pemilihan fitur yang relevan terhadap klasifikasi sehingga akurasi menjadi rendah. Hal tersebut dapat diatasi dengan cara pemilihan fitur yang berguna untuk meningkatkan akurasi [8].

Feature selection merupakan suatu proses pemilihan bagian dari *variable* dari semua *variable* yang ada di *dataset*. Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi dataset akan menjadi besar hal ini membuat rendahnya nilai akurasi klasifikasi. Masalah dalam seleksi fitur adalah pengurangan dimensi, dimana awalanya semua atribut diperlukan untuk memperoleh akurasi yang maksimal. *Forward Selection* didasarkan pada model Regresi Linear. *Forward Selection* adalah salah satu teknik untuk mereduksi dimensi dataset dengan menghilangkan atribut-atribut yang tidak relevan atau redudan. Metode *Forward Selection* adalah pemodelan dimulai dari nol peubah (*empty* model), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi [9]

Penelitian ini dimaksudkan untuk membangun sistem klasifikasi untuk mendeteksi penyakit kanker payudara menggunakan metode naive bayes, dengan menambahkan metode forward selection untuk pemilihan fitur dari banyak fitur yang ada pada data kanker payudara, karena tidak semua fitur merupakan fitur yang dapat digunakan pada proses klasifikasi. Penggunaan metode seleksi fitur diharapkan dapat meningkatkan akurasi dalam memprediksi kanker payudara.

Pembahasan mengenai teori-teori yang berhubungan dengan sistem secara umum sangat penting diuraikan terlebih dahulu, untuk pemahaaman sebelum memasuki inti masalah yang dibahas. Klasifikasi adalah teknik yang memetakan data kedalam kelompok-kelompok yang telah ditetapkan atau kelas. Klasifikasi adalah proses penemuan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelas nya tidak diketahui [10]. Klasifikasi (taksonomi) adalah proses menempatkan objek tertentu (konsep) dalam satu set kategori, berdasarkan masing-masing objek (konsep) *property*. Proses klasifikasi didasarkan pada empat komponen mendasar yaitu kelas, prediktor, *training set*, dan pengujian *dataset* [11].

Naïve Bayes adalah suatu metode yang digunakan untuk dapat memperkirakan atau memprediksi suatu class dari suatu objek yang kelasnya tidak diketahui dari masing-masing kelompok atribut yang ada, dan menentukan class mana yang paling optimal berdasarkan pengaruh yang didapat dari hasil pengamatan. Klasifikasi naive b*ayes merupakan pengembangan dari klasifikasi bayes. Klasifikasi bayes* adalah klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class* [12]. Prediksi Bayes didasarkan pada formula teorema Bayes dengan formula umum sebagai berikut:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad \text{..... (1)}$$

Keterangan:
B = Data dengan *class* yang belum diketahui
A = Hipotesis data *B* merupakan suatu *class* spesifik
P(A|B) = Probabilitas *A* berdasarkan kondisi *B*
P(B|A) = Probabilitas *B* berdasarkan kondisi *A*
P(A) = Probabilitas dari *A*
P(B) = Probabilitas dari *B*

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Tahapan penggunaan Algoritma *Naïve Bayes*: (1) Baca data *training;* (2) Hitung jumlah *class*; (3) Hitung jumlah kasus yang sama dengan *class* yang sama (Probabilitas); (4) Kalikan semua nilai hasil sesuai dengan data baru yang dicari *class*nya. Bandingkan hasil *class* pilih dengan nilai terbesar.

Metode seleksi fitur yang dipilih adalah metode *forward selection* dimana metode *forward selection* merupakan salah satu metode dari kategori metode pembungkus (*wrap method*) dalam seleksi fitur dimana dalam seleksi fitur terdapat tiga kategori yaitu penyaring/ *filter*, pembungkus/ *wrapper*, dan tertanam/ *embedded* [13]. Algoritma *forward selection* didasarkan pada model regresi linear, prosedur *forward* selection dapat dirumuskan sebagai berikut:

a. Menentukan model awal

$\hat{y}=b_0$ .....(2)

Memasukkan variable respon dengan setiap variable berprediktor, misalnya $X1,X2,\ldots.Xn$ yang terkait dengan $\hat{y}$. Misalkan $X1$ sehingga membentuk model:

$\hat{y} = b_0 + b_1X_1$ ..... (3)

b. Uji F terhadap peubah pertama yang terpilih;
Jika $Fhitung < Ftabel$ maka peubah terpilih dibuang dan proses dihentikan;
Apabila $Fhitung > Ftabel$ maka peubah terpilih memiliki pengaruh nyata terhadap peubah terkait y;
sehingga layak untuk diperhitungkan di dalam model;

c. Masukan peubah bebas terpilih (yang paling signifikan) ke dalam model. Misalkan $X2$, sehingga membentuk suatu model:

$\hat{y} = b_0 + b_1X_1 + b_2X_2$ ..... (4)

d. Uji F, jika $Fhitung < Ftabel$ maka proses dihentikan dan model terbaik adalah model sebelumnya;
Namun jika $Fhitung \geq Ftabel$, variable peubah bebas layak untuk dimasukan ke dalam model dan kembali ke langkah C;
proses akan berakhir jika tidak ada lagi peubah yang tersisa yang bias dimasukkan ke dalam model.

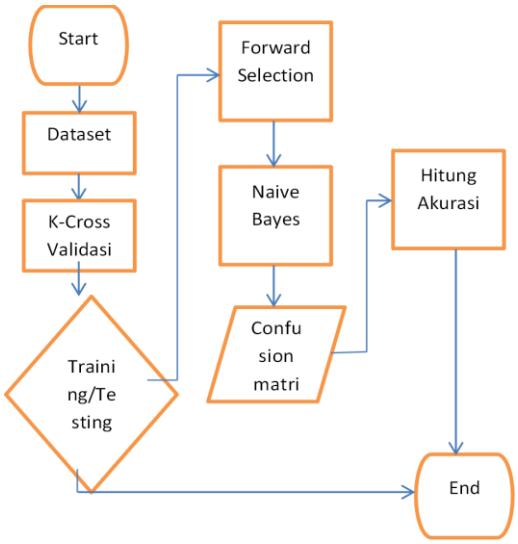
Pada metode tertanam/*Embedded* proses pencarian fitur tertanam kedalam algoritma klasifikasi, dan proses pembelajaran dengan proses pemilihan fitur tidak dapat dipisahkan [13]. Mirip seperti metode pembungkus (*wrapper*), metode tertanam mencakup interaksi dengan algoritma pengklasifikasi, sementara pada saat yang sama, metode tertanam dapat menghemat biaya komputasi/ *computational cost*.

2. Pembahasan

Perancangan format data masukan dan format keluaran dari sistem merupakan hal yang signifikan karena hal ini berkaitan dengan bagaimana implementasi dari algoritma gabungan pada proses klasifikasi yang dilakukan. Dalam penelitian ini data yang menjadi masukan ke dalam sistem dapat dibagi menjadi dua yaitu: data training dan data testing. Pada penelitian ini, data yang digunakan adalah dataset kanker payudara yang didapat dari situs UCI Machine Learning Repository yang dapat diunduh pada laman

https://archive.ics.uci.edu/ml/machine-learning-databases/00451/.

Pada dataset kanker payudara terdapat 30 atribut label attribut terdiri dari 1 = normal dan 2 = kanker. Validasi silang dilakukan pada dataset kanker payudara. Validasi silang (*cross validation*) adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua bagian, yaitu: untuk pelatihan dan yang lain digunakan untuk validasi model [14].



Gambar 1. Diagram Optimalisasi Kanker Payudara

Umumnya untuk validasi silang, dataset *training* dan validasi harus diputar berurutan sehingga setiap jalur data memiliki kesempatan tervalidasi kembali. Metode dasar dari validasi *cross-over* adalah *k - fold cross validation.*

K-fold cross validation membagi data menjadi k subset yang ukurannya hampir sama satu sama lain. Himpunan bagian yang dihasilkan, yaitu: S_1, S_2, ……, S_k yang digunakan sebagai data pelatihan dan data pengujian. Dalam metode ini dilakukan perulangan sebanyak *k* kali. Setiap kali perulangan, salah satu subset dijadikan data uji dan k-l subset lainnya dijadikan sebagai data latih. Pada iterasi ke-i, himpunan bagian S_i digunakan sebagai data pengujian dan himpunan bagian lainnya digunakan data pelatihan, dan seterusnya.

Pada tahap pertama pada penelitian ini, dataset kanker payudara dirubah record labelnya menjadi nominal. 1 diganti menjadi normal, dan 2 diganti menjadi kanker. Pada tahapan kedua sebelum dataset dilatih (training) dan diuji (testing), dataset akan dipecah terlebih dahulu dengan menerapkan 10-fold cross validation untuk membagi data menjadi dua bagian yaitu 80% training dan 20% testing. Kemudian percobaan dilakukan dengan tiga tahapan yaitu yang pertama memasukkan training dataset

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

ke dalam algoritma naive bayes dan mengukur performa percobaan. Percobaan kedua dilakukan dengan menerapkan teknik seleksi fitur dengan metode forward selection dan naive bayes

Tahapan ketiga pada penelitian ini adalah membandingkan hasil akurasi , antara akurasi naive bayes dengan akurasi naive bayes yang ditambahkan *forward selection*.

Metode evaluasi kinerja, dimaksudkan untuk mengevaluasi pendekatan yang diusulkan. Analisa akan menampilkan hasil eksperimen dan memeriksa kinerja pengklasifikasi yang diusulkan untuk Dataset kanker payudara. Evaluasi kinerja dari metode yang diusulkan adalah *confusion matrix*. *Confusion matrix* berisi informasi detail tentang hasil pengenalan atau klasifikasi (prediksi) oleh sistem terhadap data *testing* yang telah diketahui kelasnya (aktual), dan biasanya disusun membentuk matrik. Elemen pada diagonal utama (↘) *confusion matrix* menunjukkan jumlah data *testing* yang dikenali dengan benar (sesuai kelasnya) oleh sistem, sedangkan yang di luar diagonal utama adalah yang salah dikenali oleh sistem.

Tabel 1 *Confusion Matrix* Dua Kelas

		Prediksi	
		Positif	Negatif
Actual	Positif	TP	FN
	Negatif	FP	TN

dengan:

TP (*True Positive*) = jumlah prediksi benar untuk data *testing* positif
FN (*False Negative*) = junlah prediksi salah (sebagai negatif) untuk data *testing* positif
FP (*False Positive*) = jumlah prediksi salah (sebagai positif) untuk data *testing* negatif
TN (*True Negative*) = jumlah prediksi benar untuk data *testing* negatif

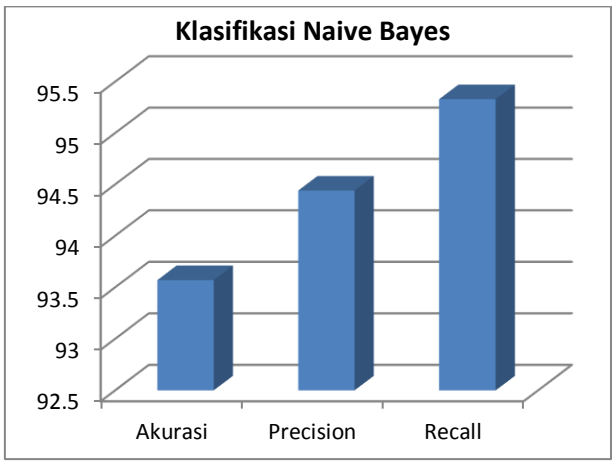
Evaluasi kinerja akan dilakukan terhadap tiga hal: *precision, recall* dan akurasi. Ketiga istilah ini didefinisikan sebagai berikut [15] :

a. *Precision:*
$$\frac{TN}{TP+TN} \quad \text{.....} \quad (5)$$

b. Recall
$$\frac{TN}{TN+FN} \quad \text{.....} \quad (6)$$

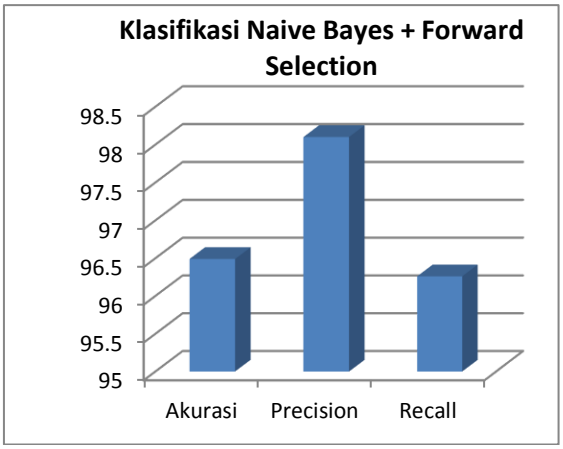
c. Akurasi:
$$\frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....} \quad (7)$$

Hasil pemilihan fitur mengunakan *forward selection*, menghasilkan 3 (tiga) fitur yang digunakan pada proses klasifikasi menggunakan *naive bayes*. Fitur atribut terpilih: *concave*, *texture*, dan *symetry*.



Gambar 2. Hasil Klasifikasi Naive Bayes

Uji coba 1 pada dataset kanker payudara menggunakan algoritma *naive bayes* menghasilkan nilai akurasi 93,57 %, *Precision* 94,44% dan *recall* sebesar 95,33%.



Gambar 3. Hasil Klsifikasi Naive Bayes dan Forward selection

Uji coba 2 pada dataset kanker payudara menggunakan algoritma *naive bayes* dengan mengunakan *forward selection* untuk memilih fitur atribut menghasilkan nilai akurasi 96,49 %, *Precision* 98,10% dan *recall* sebesar 96,26%.

3. Kesimpulan

Simpulan hasil penelitian ada peningkatan nilai akurasi setelah menggunakan metode *forward selection* yang digunakan untuk memilih fitur/mereduksi dimensi. Selisih akurasi meningkat sebesar 2,92%, precision meningkat sebesar 3,66% dan recall sebesar 0,93 % setelah menggunakan metode *forward selection* pada

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

metode *naive bayes* untuk mengklasifikasi kanker payudara.

Daftar Pustaka

[1] Sarina, et al,.(2020). Faktor Yang Berhubungan Dengan Perilaku Sadari Sebagai Deteksi Dini Kanker Payudara Pada Mahasiswi FKM UNHAS, Hasanuddin Journal of Public Health Volume 1 Issue 1 | Februari 2020 | Hal 61-70 DOI: http://dx.doi.org/10.30597/hjph.v1i1.9513

[2] Kementerian Kesehatan Republik Indonesia.(2015), Pedoman Teknis Pengendalian Kanker Payudara Dan Kanker Leher Rahim. Jakarta: Departemen Kesehatan

[3] American Cancer Society. 2016. Breast Cancer Fact and Figures 2016. [Online] Available at http://www.cancer.org/research/cancerfactsfigure.

[4] Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2018). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*. https://doi.org/10.1016/j.cegh.2018.10.003

[5] Jafari-Marandi, R., Davarzani, S., Soltanpour Gharibdousti, M., & Smith, B. K. (2018). An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. *Applied Soft Computing Journal*, *72*, 108–120. https://doi.org/10.1016/j.asoc.2018.07.060

[6] Ellmann, S., Seyler, L., Evers, J., Heinen, H., Bozec, A., Prante, O., … Bäuerle, T. (2019). Prediction of early metastatic disease in experimental breast cancer bone metastasis by combining PET/CT and MRI parameters to a Model-Averaged Neural Network. *Bone*, *120*, 254–261. https://doi.org/10.1016/j.bone.2018.11.008

[7] Agustin, Riska, et al, (2019), *Metode Naive Bayes Dalam Mendeteksi Sel Kanker Payudara*, Jurnal Statistika dan Aplikasinya (JSA) Vol. 3 No.1, Juni 2019

[8] Fanani, Rudi .(2020). *Algoritma Naïve Bayes Berbasis Forward Selection Untuk Prediksi Bimbingan Konseling Siswa*, Jurnal Disprotek Volume 11 Nomor 1, Januari 2020, ISSN. 2088-6500 e-ISSN. 2548-4168

[9] Nugroho, MF, et al.(2017).*Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes*, Jurnal Informatika Upgris Vol. 3, No. 1, (2017) P/E-ISSN: 2460-4801/2447-6645

[10] Rafiska, R., dkk. (2018).*Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4.5*, 391-396.

[11] Septiani, W. D. (2017).*Kompirasi Metode Klasifikasi Data Mining Algoritma C4.5 dan Naive Bayes untuk Memprediksi Penyakit Hepatitis*, 76-84.

[12] Lieng, J., Kencana, I., & Oka, T. 2014. Analisis Sentimen Menggunakan Metode Naive Bayes Classifier dengan Seleksi Fitur Chi Square. *Jurnal Matematika Vol. 3*, 92-99.

[13] Zhu, M., & Song, J. (2013). An embedded backward feature selection method for MCLP classification algorithm. *Procedia Computer Science*, *17*, 1047–1054. https://doi.org/10.1016/j.procs.2013.05.133

[14] R. Payam, Tang. L (2008), "Cross Validation", *Arizona State University*, File path://ppdys1108/womat3/production/PRODEN/00 0000005/0000008302/0000000016/0000875816.3D

[15] Kemal Polat, Bayram Akdemir, Salih Güne. (2008), "Computer aided diagnosis of ECG data on the least square support vector machine", *Digital Signal Process*. 18, hal 25–32.