JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 5
No. 3
Desember
2019

Prediksi Bedah Toraks Menggunakan Seleksi Fitur *Forward Selection* dan *K-Nearest Neighbor*

Rangga Sanjaya[#1], Fitriyani[#2]

[#]*Program Studi Sistem Informasi, Universitas BSI*
Ters.Jl.Jakarta, Jl.Sekolah Internasional No.1-6 Antapani – Bandung, Jawa Barat
[1]fitriyani.fyn@gmail.com
[2]rangga.rsj@gmail.com

Abstrak*— Kanker paru merupakan penyakit yang memerlukan tindakan penanganan yang cepat dan terarah, dimana penyebab paling tinggi dari kanker paru adalah merokok. Bedah toraks merupakan operasi yang paling umum untuk kanker paru. Bedah toraks dapat mengobati kanker paru, akan tetapi usia hidup pasien pasca operasi yang menjadi masalah, sehingga sebelum melakukan operasi dokter harus dapat memilih pasien dengan tepat berdasarkan resiko dan manfaat. Penelitian ini menggunakan dataset *thoracic surgery* dengan menggunakan algoritma *K-Nearest Neighbor*. Pada dataset *thoracic surgery* terdapat kelas atau fitur yang tidak relevan sehingga dilakukan seleksi fitur menggunakan *Forward Selection*. Eksperimen dan pengolahan data yang dilakukan, dibantu oleh *software Rapidminer*. Pada penelitian ini akan dilakukan perbandingan performa antara algoritma *K-Nearest Neighbor* tanpa seleksi fitur dengan *K-Nearest Neighbor* dengan seleksi fitur *Forward Selection*. Berdasarkan hasil pengujian dan perbandingan dari kedua model yang diusulkan, algoritma K-NN dengan optimasi fitur menggunakan metode *Forward Selection* memiliki nilai akurasi lebih baik dibandingkan dengan algoritma K-NN tanpa seleksi fitur.

***Kata kunci*— *Thoracic Surgery, Forward Selection, K-Nearest Neighbor*.**

I. PENDAHULUAN

Kanker paru merupakan jenis penyakit yang memerlukan penanganan serta tindakan yang cepat dan terarah. Diagnosis penyakit ini memerlukan kompetensi dan dukungan sarana yang tidak sederhana, serta memerlukan pendekatan multidisiplin pada bidang ilmu kedokteran. Penanganan penyakit ini membutuhkan kolaborasi antara ahli paru, ahli radiologi diagnostik, ahli patologi anatomi, ahli radiologi terapi, ahli bedah toraks, ahli rehabilitasi medik, dan lain sebagainya. Merokok merupakan penyebab utama kanker paru-paru dan semakin meningkatnya aktivitas merokok, berdampak pada semakin besarnya resiko menderita kanker paru-paru [1].

Permasalahan utama dalam bedah toraks adalah pemilihan pasien yang tepat untuk operasi dengan resiko dan manfaat bagi pasien baik jangka pendek (seperti komplikasi pasca operasi, termasuk 30 hari) dan juga dalam jangkan panjang (seperti kelangsungan hidup 1 tahun atau 5 tahun setelah operasi [2], sehingga diperlukan penelitian lebih lanjut untuk memprediksi jangka kelangsungan hidup dan matinya pasien setelah menjalani operasi bedah toraks. Operasi bedah toraks memiliki resiko, salah satunya tindakan kardiotoraks pada pasien sangat beresiko terhadap gangguan saraf. Stroke merupakan komplikasi utama dari tindakan kardiotoraks tersebut.

Penelitian yang dilakukan (Tabel 1), mengacu pada penelitian prediksi bedah toraks yang sudah dilakukan, serta penelitian penggunaan *K-Nearest Neighbor* dan seleksi fitur forward selection yang digunakan dalam pemodelan data mining. Penelitian pertama yaitu prediksi hidup dan mati pasien setelah 1 tahun melakukan bedah toraks [3] menggunakan *Artificial Neural Network* dengan hasil akurasi 90%. Pada penelitian selanjutnya, algoritma *Naïve Bayes*, J48 *Decision Tree*, PART (*Partial Decision Tree*), OneR (*One Rule*), *Random Forest Tree*, *Decision Stump* digunakan dengan hasil terbaik pada algoritma *Random Forest Tree* sebesar 95.65% [4] dan penelitian [5] yang menggunakan algoritma *Multilayer Peceptron* (MLP), J48 dan *Naïve Bayes* dengan hasil terbaik sebesar 82.3% pada algoritma MLP. Penelitian penggunaan algoritma KNN (*K-Nearest Neighbor*) pernah dilakukan untuk sejumlah dataset yang diantaranya pada penelitian [6] menggunakan dataset RSS (*Really Simple Sindication*) dengan hasil akurasi baik, penelitian [7] yang menunjukkan performa yang baik sebesar 71.66% menggunakan dataset *Car Evaluation*. Performa KNN dapat dioptimasi menggunakan seleksi fitur *Forward Selection* pada penelitian [8] pada dataset *Financial Distress* dengan akurasi meningkat sebesar 15%.

Untuk menguji performa optimasi seleksi fitur pada algoritma KNN, pada penelitian ini digunakan dataset publik thoracic surgery.

TABEL I
PENELITIAN TERKAIT

Peneliti	Dataset	Metode	Hasil
Esteva, Hugo Núñez, Tomás G. Rodríguez, Ricardo O (2007)	Thoracic Surgery	Artificial Neural Network	Akurasi 90%
Sindhu, V Prabha, S A Sathya Veni, S Hemalatha, M (2014)	Thoracic Surgery	Naïve Bayes, J48 Decision Tree, PART (Partial Decision Tree), OneR (One Rule), Random Forest Tree, Decision Stump	Naïve Bayes 82.34%, J48 Decision Tree 85.10%, PART (Partial Decision Tree) 91.91%, OneR (One Rule) 85.31%, Random Forest Tree 95.65%, Decision Stump 85.10%
Danjuma, Kwetishe Joro (2015)	Thoracic Suergery	Multilayer Peceptron (MLP), J48 dan Naïve Bayes	Multilayer Peceptron (MLP) 82.3%, J48 81.8% dan Naïve Bayes 74.4%
Adeniyi, D.A. Wei, Z. Yongquan, Y (2016)	RSS (Really Simple Sindication)	KNN (K-Nearest Neighbor)	Akurasi baik
Samanthula, Barath K Elmehdwi, Yousef Jiang, Wei (2016)	Car Evaluation	KNN (K-Nearest Neighbor)	71.66%
Fallahpour, Saeid Lakvan, Eisa Norouzian Zadeh, Mohammad Hendijani (2017)	Financial Distress	Sequential Floating Forward Feature Selection (SFFS), Support Vector Machine (SVM), Artificial Bee Colony (ABC), Genetic Algoritthm (GA), Sequential Forward Selection (SFS), Principal Component Analysis (PCA), Information Gain (IG)	Akurasi meningkat dengan menggunakan Forward Selection sebesar 15%

K-Nearest Neighbor termasuk kedalam sepuluh algoritma yang paling banyak digunakan untuk penelitian data mining [9]. Sedangkan Feature Selection atau seleksi fitur dapat mengurangi dimensional pada data untuk meningkatkan kinerja dari mesin pembelajaran. Subset selection merupakan metode *feature selection* (seleksi fitur), yang menghilangkan dimensi yang tidak penting dengan menggunakan *error function* yang dapat menyelesaikan permasalahan pada regresi dan klasifikasi. Seleksi fitur adalah bagian penting untuk meningkatkan kinerja model dari algoritma klasifikasi [10]. Sedangkan optimasi waktu pada seleksi fitur dapat diselesaikan dengan menggunakan metode heuristic [11]. Dengan menggunakan seleksi fitur, diharapkan dapat meningkatkan akurasi algoritma KNN [12] [13] pada dataset *thoracic surgery*.

II. METODELOGI PENELITIAN

A. Dataset

Dataset yang digunakan pada penilitian ini adalah dataset *thoracic surgery*. Dataset yang digunakan dalam penelitian ini merupakan dataset pasien yang melakukan bedah toraks. Dataset *thoracic surgery* memiliki 470 record dan 16 atribut serta 1 kelas yang berisi *true* dan *false*. dan dapat diakses melalui UCI *Repository*. Atribut dan deskripsi dari dataset *thoracic surgery* ditunjukkan pada Tabel 2.

TABEL II
DESKRIPSI ATRIBUT *THORACIC SURGERY*

No.	Fitur atau Atribut	Deskripsi
1.	DGN	Diagnosis - kombinasi spesifik kode ICD-10 untuk tumor primer dan sekunder serta beberapa tumor ganda jika ada (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)
2.	PRE4	Kapasitas vital paksa - FVC (numerik)
3.	PRE5	Volume yang telah dihembuskan pada akhir detik pertama masa paksa berakhir - FEV1 (numerik)
4.	PRE6	Status kinerja - Skala Zubrod (PRZ2, PRZ1, PRZ0)
5.	PRE7	Sakit sebelum operasi (T, F)
6.	PRE8	Haemoptysis sebelum operasi (T, F)
7.	PRE9	Dyspnoea sebelum operasi (T, F)
8.	PRE10	Batuk sebelum operasi (T, F)
9.	PRE11	Kelemahan sebelum operasi (T, F)
10.	PRE14	T dalam TNM klinis - ukuran tumor asli, dari OC11 (terkecil) sampai OC14 (terbesar) (OC11, OC14, OC12, OC13)
11.	PRE17	DM tipe 2 - diabetes mellitus (T, F)
12.	PRE19	MI sampai 6 bulan (T, F)
13.	PRE25	PAD - penyakit arteri perifer (T, F)
14.	PRE30	Merokok (T, F)
15.	PRE32	Asma (T, F)
16.	AGE	Usia saat operasi (numerik)
17.	Risk1Y	Masa kelangsungan hidup 1 tahun - (T) nilai true jika meninggal (T, F)

B. Model Penelitian

Pada penelitian ini menggunakan model yang dapat dilihat pada Gambar 1. Dataset bedah toraks dibagi menjadi 2 bagian yaitu data training dan data testing secara otomatis menggunakan *cross validation*. Data training digunakan untuk membentuk model algoritma, sedangkan data testing digunakan untuk menguji kinerja dan performa algoritma. Selanjutnya data tersebut diseleksi menggunakan metode *feature selection* untuk memilih fitur yang paling relevan. Setelah diperoleh fitur atau atribut yang paling relevan berdasarkan seleksi fitur, selanjutnya fitur dan *record* di proses menggunakan metode *machine learning k-nearest neighbor* untuk memunculkan nilai dari kinerja model yang diuji.

1) Forward Selection: Dalam *Forward Selection*, dimulai dengan tidak ada variabel dan menambahkannya satu persatu, pada setiap langkah menambahkan satu kesalahan yang paling kecil, kemudian penambahan selanjutnya tidak ada lagi kesalahan [11].

Prosedur *Forward Selection* [14]:

- Forward Selection dimulai pada saat tidak ada variabel pada model.
- Untuk variabel pertama yang masuk ke model, pilih prediktor yang paling berkorelasi tinggi dengan target. (tanpa kehilangan keumuman, menunjukkan variabel ini) Jika hasilnya model tidak signifikan, berhenti dan laporkan bahwa tidak ada variabel prediktor penting; jika tidak lanjutkan ke langkah 2.
- Untuk setiap variabel yang tersisa, hitunglah F-statistik berurutan untuk variabel tersebut mengingat variabel sudah ada dalam model. Misalnya, dalam operan pertama ini melalui algoritma, ini F-statistik berurutan seperti F , F , dan F . Pada lintasan kedua melalui algoritma, ini mungkin F dan F . Pilih variabel dengan sekuensial terbesar F-statistik.

Untuk variabel yang dipilih pada langkah 2, uji untuk signifikansi sekuensial F-statistik. Jika model yang dihasilkan tidak signifikan, hentikan, dan laporkan arus model tanpa menambahkan variabel dari langkah 2. Jika tidak, tambahkan variabel dari langkah 2 ke dalam model dan kembali ke langkah 2.

2) K-Nearest Neighbor (K-NN): K-NN termasuk dalam 10 algoritma yang paling populer dalam data mining [15]. Algoritma *Nearest Neighbor* atau biasa dikenal dengan *K-Nearest Neighbor* (K-NN), adalah algoritma klasifikasi berdasarkan pada kedekatan lokasi atau jarak satu data dengan data yang lain.

Dasar Algoritma K-NN [16]

Input : D , set objek pelatihan, objek uji, $\boldsymbol{z}$, yang merupakan vektor dari nilai atribut dan L, himpunan kelas yang digunakan untuk melabeli objek

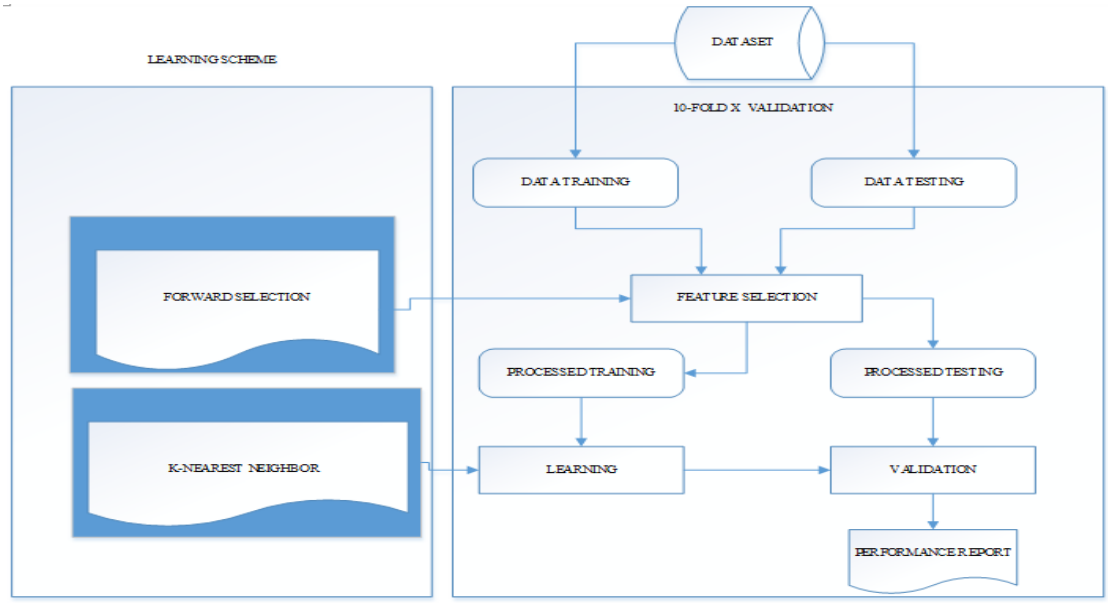
Output : $c_z \in L$, kelas z

foreach objek $\boldsymbol{y} \in D$ **do**

| hitung $d(\boldsymbol{z}, \boldsymbol{y})$, jarak antara $\boldsymbol{z}$ dan $\boldsymbol{y}$;

end

Pilih $N \subseteq D$ himpunan (tetangga) dari objek pelatihan terdekat untuk z;



Gambar. 1 Model penelitian

$c_z = \frac{argmax}{v \in L} \sum_{y \in N} I(v = class(c_y))$;

dimana $I(.)$ adalah fungsi indikator untuk mengambalikan nilai 1 jika argumen benar dan 0 untuk sebaliknya.

Pembelajaran instan berbasis KNN merupakan pembelajaran sederhana dan dapat bekerja dengan baik [17].

III. HASIL DAN PEMBAHASAN

Perhitungan algoritma K-NN menggunakan data sampel yang diambil secara acak pada dataset *thoracic surgery* sebanyak 5, dapat dilihat pada Tabel 3.

TABEL III
CONTOH DATA SAMPEL

No	PRE4	PRE5	AGE	Risk1Yr
1	2	3	15	False
2	3	1	20	False
3	1	3	15	True
4	3	1	24	True
5	2	2	30	True
6	1	3	25	?

Langkah pertama dalam menghitung data menggunakan algoritma K-NN:

1. Menentukan nilai parameter K, dimana nilai K=3
2. Menghitung jarak *Ecludian Distance*
   a. $d1,d6 = \sqrt{(d61-d11)^2 + (d62-d12)^2 + (d63-d13)^2}$
   b. $d1,d6 = \sqrt{(1-2)^2 + (3-3)^2 + (25-15)^2} = 10.049$
   c. $d2,d6 = \sqrt{(1-3)^2 + (3-1)^2 + (25-20)^2} = 5.385$
   d. $d3,d6 = \sqrt{(1-1)^2 + (3-3)^2 + (25-15)^2} = 10$
   e. $d4,d6 = \sqrt{(1-3)^2 + (3-1)^2 + (25-24)^2} = 2.828$
   f. $d5,d6 = \sqrt{(1-2)^2 + (3-2)^2 + (25-30)^2} = 5.196$
3. Menentukan 3 nilai terbaik berdasarkan nilai K, dimana urutan pertama adalah d1, d2 dan d5.
4. Label d1 dan d2 adalah False, sedangkan d5 adalah True. Pada algoritma K-NN ditentukan mayoritas terbanyak sehingga d6 berisi label *False*.

Pada penelitian ini, eksperimen yang dilakukan untuk pengujian model adalah menggunakan aplikasi Rapidminer. Rapidminer merupakan aplikasi pemodelan *data mining* yang paling banyak digunakan dengan hasil yang baik menurut poling KDNuggets [18]. Pengujian dilakukan dua kali, dimana pengujian pertama yaitu K-NN tanpa seleksi fitur *forward selection*, selanjutnya pengujian kedua menggunakan seleksi fitur dengan metode *forward selection*. Tahap selanjutnya adalah mencari nilai terbaik dari kinerja dan performa berdasarkan nilai akurasi dari hasil pengujian kedua model.

Langkah pertama pada penelitian ini, yaitu menguji model K-NN dengan nilai K=5. Penggunaan nilai k dipilih secara acak. Pada penelitian ini dilakukan uji coba pendahuluan pada nilai k=1 dan k=5 terhadap sampel data. Diketahui terdapat peningkatan nilai k=1 dengan nilai k=5, akan tetapi untuk nilai selain k=1 dan k=5 hasil pengujian menunjukkan nilai yang sama. Hasil yang diperoleh dari pengujian model dapat dilihat pada Tabel 4, dimana TF adalah *True False*, TT adalah *True True*, sedangkan PF adalah *Prediction False* dan PT adalah *Predicition True*. *Confusion matrix* dari pengujian model algoritma K-NN tanpa seleksi fitur, dengan hasil akurasi kinerja dari model sebesar 83.40%.

TABEL IV
HASIL K-NN

	TF	TT	class precision
PF	392	70	84.85%
PT	8	0	0.00%
class recall	98.00%	0.00%	

Langkah berikutnya setelah pengujian K-NN tanpa seleksi fitur dilakukan, selanjutnya dilakukan pengujian model K-NN yang dioptimasi oleh seleksi fitur menggunakan metode *forward selection*, dengan nilai K=5. Hasil yang diperoleh dari pengujian model dapat dilihat pada Tabel 5, yang menunjukkan *confusion matrix* dari pengujian model algoritma K-NN yang dioptimasi oleh seleksi fitur menggunakan metode *forward selection*, dengan hasil akurasi kinerja dari model sebesar 85.74%.

TABEL V
HASIL K-NN+FS

	TF	TT	class precision
PF	397	64	86.12%
PT	3	6	66.67%
class recall	99.25%	8.57%	

Setelah dilakukan pengujian serta diperoleh hasil kinerja pada kedua model, selanjutnya dilakukan perbandingan untuk mendapatkan kesimpulan algoritma terbaik diantara K-NN tanpa seleksi fitur dengan K-NN yang dioptimasi oleh seleksi fitur menggunakan metode *forward selection*.

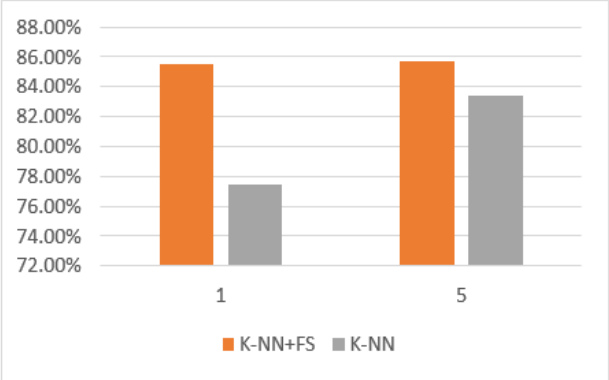
Pada Tabel 6 dan Gambar 2 menunjukkan perbandingan antara model K-NN dan model K-NN yang dilakukan optimasi menggunakan seleksi fitur *forward selection*. Hasil akurasi terbaik dari kedua model, yaitu masing-masing K-NN menghasilkan nilai akurasi sebesar 83.40% dengan nilai K=5. Sedangkan K-NN yang dilakukan optimasi menggunakan seleksi fitur menggunakan metode *forward selection* dengan nilai K=5 menghasilkan nilai akurasi sebesar 85.74%. Berdasarkan perbandingan pada kedua model, disimpulkan bahwa K-NN yang dilakukan optimasi seleksi fitur menggunakan metode *forward selection* merupakan model terbaik yang memiliki nilai akurasi lebih tinggi.

TABEL VI
PERBANDINGAN HASIL K-NN DENGAN K-NN+FS

Nilai K	K-NN	K-NN+FS
1	77.45%	85.53%
5	83.40%	85.74%

Hasil atribut yang relevan dari seleksi fitur menggunakan metode *forward selection* adalah atribut DGN, PRE11 dan PRE7, sehingga dapat diambil kesimpulan bahwa dari 16 atribut hanya 3 atribut yang

paling relevan sehingga dapat meningkatkan kinerja dari algoritma prediksi pasien bedah toraks.



Gambar. 2 Perbandingan model K-NN dengan K-NN+FS

IV. KESIMPULAN

Kanker paru dapat diobati dengan bedah toraks, akan tetapi yang menjadi masalah adalah usia hidup pasien pasca operasi, sehingga diperlukan pemilihan pasien yang tepat berdasarkan resiko dan manfaat dari operasi tersebut terhadap pasien. Dataset bedah toraks memiliki banyak atribut atau fitur sehingga dibutuhkan algoritma untuk seleksi fitur dengan tujuan dapat meningkatkan kinerja dari model penelitian. Hasil penelitian yang dilakukan bahwa model yang menggunakan K-NN tanpa seleksi fitur menghasilkan nilai akurasi terbaik sebesar 83.40%. Sedangkan model yang menggunakan K-NN dan *Forward Selection* (K-NN+FS) menghasilkan nilai akurasi terbaik sebesar 85.74%. Berdasarkan perbandingan dari kedua model yang diuji, dapat disimpulkan bahwa K-NN yang dioptimasi oleh seleksi fitur menggunakan metode *forward selection* memiliki nilai akurasi lebih baik dibandingkan dengan model K-NN tanpa seleksi fitur.

Selain itu, pada penelitian ini juga dapat disimpulkan bahwa pemilihan atribut atau fitur dapat meningkatkan kinerja dari model penelitian. Untuk penelitian selanjutnya dapat digunakan model atau algoritma seleksi fitur yang lain sehingga dapat meningkatkan kinerja atau akurasi dari model penelitian.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada RISTEKDIKTI atas dukungan dana penelitian melalui hibah penelitian dosen pemula tahun 2019, sehingga penelitian ini dapat terlaksana. Peneliti juga mengucapkan terima kasih dan penghargaan kepada Lubicz, Pawelczyk, Rzechonek, dan Kolodziej atas dataset *thoracic surgery* yang tersedia di UCI *Repository* sehingga dataset tersebut dapat digunakan dalam penelitian ini.

REFERENSI

[1] A. Zulkifli, "Kanker Paru," *Buku Ajar iImu Penyakit Dalam*, pp. 2254–2261, 2011.
[2] M. Koklu, H. Kahramanli, and N. Allahverdi, "Applications of Rule Based Classification," 2013, no. November, pp. 1991–1998.
[3] H. Esteva, T. G. Núñez, and R. O. Rodríguez, "Neural Networks and Artificial Intelligence in Thoracic Surgery," *Thorac. Surg. Clin.*, vol. 17, no. 3, pp. 359–367, 2007.
[4] V. Sindhu, S. A. S. Prabha, S. Veni, and M. Hemalatha, "Thoracic surgery analysis using data mining techniques," vol. 5, no. April, pp. 578–586, 2014.
[5] K. J. Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients," *J. Comput. Sci. Issues*, 2015.
[6] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016.
[7] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "K-nearest neighbor classification over semantically secure encrypted relational data," *Int. J. Control Theory Appl.*, vol. 9, no. 27, pp. 437–443, 2016.
[8] S. Fallahpour, E. N. Lakvan, and M. H. Zadeh, "Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem," *J. Retail. Consum. Serv.*, vol. 34, no. March 2016, pp. 159–167, 2017.
[9] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. Boca Raton, London, New York: CRC Press taylor & Francis Group, 2009.
[10] P. Koncz and J. Paralic, "An approach to feature selection for sentiment analysis," *2011 15th IEEE Int. Conf. Intell. Eng. Syst.*, pp. 357–362, 2011.
[11] Alpaydın Ethem, *Introduction to Machine Learning Second Edition*, 2nd ed. London: MIT, 2010.
[12] T. Xu, Q. Peng, and Y. Cheng, "Identifying the semantic orientation of terms using S-HAL for sentiment analysis," *Knowledge-Based Syst.*, vol. 35, pp. 279–289, 2012.
[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.
[14] D. T. Larose, *Data Mining Methods and Models*. 2006.
[15] X. Wu *et al.*, *Top 10 algorithms in data mining*. 2008.
[16] C. Catal and B. Diri, "A systematic review of software fault prediction studies," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7346–7354, 2009.
[17] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining*. 2011.
[18] I. Mierswa, "RapidMiner Voted Most Used Analytics Software in KDNuggets Poll." .