Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
03 Agustus 2023/ Pages: 188-193

Comparison of K-Nearest Neighbor (KNN) and Naive Bayes Algorithms in the Classification of Parkinson's Disease

Komparasi Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penyakit Parkinson

Windy Aprilita Z[1*], Rivaldo Akbar[2], Renaldi Cahyadi Prayogi[3] , Rahmaddeni[4]

[1,2,3,4]Teknik Informatika, STMIK Amik Riau, Pekanbaru

E-Mail: [1]2010031802089@sar.ac.id, [2]2010031802100@sar.ac.id,
[3]2010031802032@sar.ac.id, [4]rahmaddeni@sar.ac.id

Corresponding Author: Windy Aprilita Z

Abstract

Parkinson's disease is one of several motor system disorders that are brought on by the death of dopamine-producing brain cells. Common symptoms include sluggishness during activities and hand tremors or shaking. This disease has a wide range of symptoms that have a significant impact on the quality of life of those who suffer from it as well as the loved ones of those who suffer from it. The diagnosis process is relatively sluggish due to the high cost of a nerve test and the lack of a blood test to detect Parkinson's disease. This study will compare the Classification Method by classifying Parkinson's disease data using the K-Nearest Neighbor (KNN) and Nave Bayes algorithms. The data used are from the kaggle Parkinson's Disease Dataset. With split data of 70:30 and 80:20, respectively, the trial compared the K-Nearest Neighbor (KNN) and Naive Bayes algorithms with 756 data. The K-Nearest Neighbor (KNN) algorithm and a 70:30 data split result in 96% accuracy for the best trial results. The K-Nearest Neighbor (KNN) algorithm is a good choice for data classification, as can be deduced from these findings.

Keyword: Classification, K-Nearest Neighbor, Naive Baye, Parkinson's Disease

Abstrak

Penyakit Parkinson adalah salah satu dari beberapa gangguan sistem motorik yang disebabkan oleh kematian sel otak penghasil dopamin. Gejala umum termasuk kelesuan saat beraktivitas dan tangan gemetar atau gemetar. Penyakit ini memiliki berbagai macam gejala yang berdampak signifikan pada kualitas hidup penderitanya serta orang yang dicintai dari penderitanya. Proses diagnosis relatif lamban karena mahalnya biaya tes saraf dan kurangnya tes darah untuk mendeteksi penyakit Parkinson. Penelitian ini akan membandingkan Metode Klasifikasi dengan mengklasifikasikan data penyakit Parkinson menggunakan algoritma K-Nearest Neighbor (KNN) dan Nave Bayes. Data yang digunakan berasal dari Kaggle Parkinson's Disease Dataset. Dengan split data masing-masing 70:30 dan 80:20, uji coba tersebut membandingkan algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dengan 756 data. Algoritme K-Nearest Neighbor (KNN) dan pembagian data 70:30 menghasilkan akurasi 96% untuk hasil uji coba terbaik. Algoritma K-Nearest Neighbor (KNN) adalah pilihan yang baik untuk klasifikasi data, seperti yang dapat disimpulkan dari temuan ini.

Kata kunci: Klasifikasi, K-Nearest Neighbor, Naive Bayes, Penyakit Parkinson

1. PENDAHULUAN

Penyakit Parkinson adalah suatu kondisi di mana sel-sel saraf di bagian otak tengah yang mengontrol gerakan secara perlahan menyerang. Rasa lemas atau kaku pada tubuh, serta tremor atau getaran halus di satu tangan, merupakan gejala yang dapat dikenali dari penyakit ini. Selain itu, penyakit ini mempengaruhi

subtantia nigra, daerah kecil di otak tengah yang mengirimkan pesan ke berbagai saraf tulang belakang yang mengontrol otot-otot tubuh [1].

Jumlah penderita parkinson terus meningkat, hingga empat kali lipat dari tahun-tahun sebelumnya, menurut Organisasi Kesehatan Dunia. Satu juta orang meninggal karena parkinson setiap tahun. Parkinson dapat disembuhkan jika penderita mengetahui kondisinya dan meminum obatnya sejak dini, sebelum menjadi lebih parah. Konsekuensinya, masyarakat perlu mendapat informasi tentang penyebab penyakit parkinson.

Dalam upaya menemukan pengobatan penyakit Parkinson yang memanfaatkan teknologi mutakhir, seperti yang saat ini tersedia di bidang ilmu biomedis, berbagai penelitian telah dilakukan. Namun, pasien Parkinson tidak dapat disembuhkan secara khusus. Saat ini pengobatan dan terapi digunakan untuk mengatasi gejala penyakit Parkinson agar tidak berkembang pesat dan tidak menghambat aktivitas sehari-hari. Oleh karena itu, ahli saraf sangat merekomendasikan diagnosis dini penyakit Parkinson sebagai sarana pencegahan [2]. Agar penelitian menjadi lebih efisien, maka diperlukan kerjasama dari berbagai disiplin ilmu.

K-nearest neighbor (k-nn) dan naive Bayes classifier adalah dua pendekatan machine learning yang menjadi bahan perbandingan studi ini. Proses klasifikasi atau prediksi berbeda untuk masing-masing metode ini. Mirip dengan bagaimana jaringan saraf tiruan backpropagation digunakan untuk memprediksi curah hujan dalam metode pembelajaran mesin lainnya [3].

Dalam karya sebelumnya Januarsyah, Zuhairi, & Malik (2019), pendekatan Naive Bayes dan Random Forest dikontraskan. Metode Naive Bayes dengan akurasi 49,06 persen, metode Random Forest dengan akurasi 74,28%, C4.5 dengan akurasi 57,53%, metode Bayesian Network dengan akurasi 48,07% , dan Decision Stump yang memiliki akurasi 49,95% semuanya lebih baik dari metode Random Forest [4]. Menurut penelitian Chandel, Kunwar, Sabitha, Choudhury, & Mukherjee tahun 2017, "Studi perbandingan deteksi penyakit tiroid dengan menggunakan teknik klasifikasi Knearest Neighbor dan Naive Bayes" akurasi metode K-Nearest Neighbor (KNN) adalah 93,44 persen, sedangkan akurasi metode Naive Bayes adalah 22,56% [5].

Masalah penelitian ini adalah membandingkan akurasi algoritma K-Nearest Neighbor dan Naive Bayes untuk klasifikasi penyakit Parkinson menggunakan 756 data yang dibagi menjadi 70-30 dan 80-20 dari kaggle, berdasarkan latar belakang masalah sebelumnya.

2. METODOLOGI PENELITIAN

Dalam penelitian ini, subbab akan menjelaskan metode penelitian yang terdiri dari empat langkah berikut:

2.1 Pendekatan Penelitian

Studi ini menggunakan metodologi kualitatif, mengumpulkan data dalam bentuk catatan lapangan atau laporan resmi lainnya daripada nilai numerik. Hasil penelitian kualitatif ini cenderung lebih ke arah generalisasi [6] karena metode kualitatif melibatkan penggabungan data dari berbagai sumber.

2.2 Pengumpulan Data

Saat mengumpulkan data, hal pertama yang perlu dilakukan adalah mencari informasi di literatur tentang subjek yang terkait dengan judul makalah [7]. Penelitian memerlukan pengumpulan data, dan teknik pengumpulan data meliputi:

1. Data Publik Pendekatan ini bertujuan untuk memperoleh data publik, khususnya dataset Kaggle Parkinson's Disease.
2. Studi Pustaka Langkah pertama adalah melakukan studi literatur dengan banyak membaca literatur dan buku-buku sebelumnya. Anda dapat memperdalam penelitian Anda dan menemukan masalah terbaru dengan membaca banyak karya sebelumnya, mengubahnya menjadi penelitian saat ini. Ini akan berfungsi sebagai referensi untuk penelitian masa depan dengan temuan terbaru [8]. Dalam melakukan penelitian, penelusuran dilakukan melalui dokumen-dokumen terkait baik di media cetak maupun elektronik.
3. Penelitian Pada bagian keempat penelitian, penting untuk melakukan penelitian guna menemukan solusi dari masalah yang diangkat. Dengan melakukan penelitian akan diketahui masalah dan pemecahannya, sehingga menjadi komponen penting dalam penelitian [9].

2.3 Analisis Data

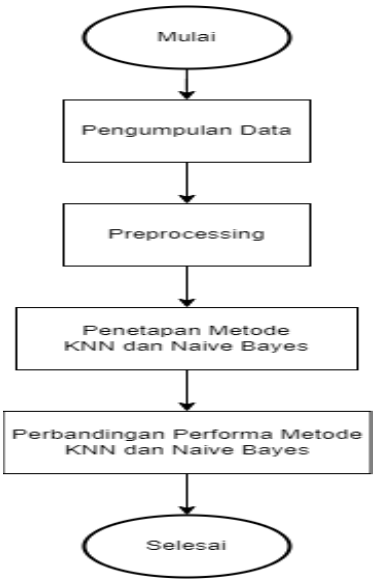
Metode penambangan data tidak diragukan lagi terhubung ke pengumpulan data lapangan, tetapi juga sumber dan jenis data [10]. Pada penelitian ini yang dituju adalah kelas atribut yang diperoleh dari 756 dataset yang berbeda dengan 755 atribut yang berbeda.

Tabel 1. Data Parkinson's Disease

	id	gender	PPE	DFA	RPDE	numPulses	numPeriodsPulses	meanPeriodPulses	stdDevPeriodPulses	locPctJitter	...	tqwt_kurtosisValue_dec_28	tqwt_kurtosisValue_dec_29
0	0	1	0.85247	0.71826	0.57227	240	239	0.008064	0.000087	0.00218	...	1.5620	2.6445
1	0	1	0.76686	0.69481	0.53966	234	233	0.008258	0.000073	0.00195	...	1.5589	3.6107
2	0	1	0.85083	0.67604	0.58982	232	231	0.008340	0.000060	0.00176	...	1.5643	2.3308
3	1	0	0.41121	0.79672	0.59257	178	177	0.010858	0.000183	0.00419	...	3.7805	3.5664
4	1	0	0.32790	0.79782	0.53028	236	235	0.008162	0.002669	0.00535	...	6.1727	5.8416
...	...	...	...	...	...	...	...	...	...	...	...	...	...
751	250	0	0.80903	0.56355	0.28385	417	416	0.004627	0.000052	0.00064	...	3.0706	3.0190
752	250	0	0.16084	0.56499	0.59194	415	413	0.004550	0.000220	0.00143	...	1.9704	1.7451
753	251	0	0.88389	0.72335	0.46815	381	380	0.005069	0.000103	0.00076	...	51.5607	44.4641
754	251	0	0.83782	0.74890	0.49823	340	339	0.005679	0.000055	0.00092	...	19.1607	12.8312
755	251	0	0.81304	0.76471	0.46374	340	339	0.005676	0.000037	0.00078	...	62.9927	21.8152

2.4 Alur Penelitian

Dari proses awal hingga kesimpulan penelitian, alur penelitian peneliti mencakup diagram yang komprehensif dari alur atau langkah-langkah terperinci. Gambar 1 menggambarkan perkembangan penelitian ke dalam sistem klasifikasi penyakit Parkinson. di bawah.



Gambar 1. Tahapan Metodologi Penelitian

2.5 Klasifikasi

Dalam data mining, klasifikasi adalah teknik mempelajari data untuk memprediksi nilai dari sekumpulan atribut. Aturan adalah seperangkat aturan yang dihasilkan oleh algoritma klasifikasi yang dapat digunakan sebagai indikator untuk memprediksi kelas data yang ingin diprediksi [11]. Dataset dijelaskan oleh klasifikasi itu sendiri, di mana setiap tipe data adalah nominal atau biner. Dalam hal klasifikasi, data yang diawasi akan dibagi menjadi dua bagian: data latih dan data uji. Algoritma klasifikasi akan digunakan untuk menganalisis data pelatihan. Berbagai algoritma klasifikasi, antara lain algoritma KNN, algoritma Naive Bayes, algoritma C4.5, dan algoritma C5 [12].

2.6 K-Nearest Neighbor

Salah satu metode klasifikasi yang digunakan dalam data mining adalah algoritma K-Nearest Neighbor, yang menggunakan data pembelajaran berlabel untuk mengklasifikasikan sekumpulan data. KNN adalah bagian dari pembelajaran terawasi, yang menggunakan objek terdekat untuk mengklasifikasikan objek baru. Kelas yang paling sering muncul di KNN akan digunakan sebagai kelas klasifikasi untuk hasil dari instance query baru. Artinya, hasil akan dikategorikan berdasarkan kategori terbanyak di KNN. Nilai kedekatan antara kasus lama dan baru ditentukan dengan menggunakan metode KNN itu sendiri. Klasifikasi KNN tidak menggunakan model memori saja [13]. Alur atau notasi metode algoritma K-Nearest Neighbor dapat dilihat dengan cara sebagai berikut:

$$E(x, y) = \sqrt{\sum_i^n = 0(xi - yi)} \quad (1)$$

Algoritma KNN adalah salah satu yang mudah. Algoritma KNN bekerja dengan menggunakan jarak minimum antara data baru dengan K tetangga terdekat yang telah ditentukan sebelumnya. Mayoritas K tetangga terdekat akan digunakan untuk membuat prediksi kelas dari data baru setelah K tetangga terdekat terkumpul.

Berdasarkan tetangga yang digunakan sebagai acuan perhitungan, ada dua jenis KNN:

- 1-NN, di mana data tetangga terdekat dengan label digunakan untuk mengklasifikasikan;
- K-NN, di mana data tetangga terdekat dengan label digunakan untuk mengklasifikasikan;

2.7 Naïve Bayes

Dalam klasifikasi statistik, algoritma Naive Bayes sering digunakan untuk memprediksi kemungkinan menjadi anggota suatu kelas. Ketika diterapkan pada database dengan banyak data, telah dibuktikan bahwa Naive Bayes bekerja dengan baik dalam kecepatan dan akurasi. Ini memiliki kemampuan klasifikasi yang sebanding dengan pohon keputusan dan jaringan saraf untuk klasifikasi naif Bayes [12]. Rumus Teorema Bayes adalah sebagai berikut:

$$\mathrm{P(H|X)} = \frac{P(X|H)\,P(H)}{P(X)} \quad (2)$$

3. ANALISA DAN PEMBAHASAN

3.1 K-Nearest Neighbor (KNN)

Pengujian dilakukan dengan menggunakan bahasa pemrograman Python pada tools Google colaboratory. Hasil pengujian ditampilkan dengan confusion matrix pada gambar 2 untuk split data 70-30 dan gambar 3 untuk split data 80-20 berikut.

	precision	recall	f1-score	support
F	0.88	0.96	0.92	53
T	0.99	0.96	0.97	174
accuracy			0.96	227
macro avg	0.93	0.96	0.95	227
weighted avg	0.96	0.96	0.96	227

Gambar 2. Confusion Matrix KNN (70-30)

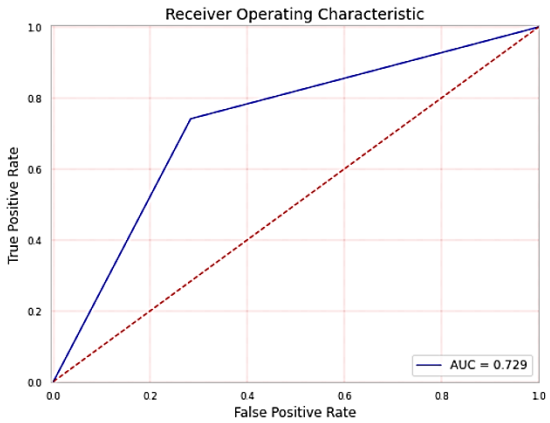
	precision	recall	f1-score	support
F	0.88	0.90	0.89	31
T	0.97	0.97	0.97	121
accuracy			0.95	152
macro avg	0.93	0.94	0.93	152
weighted avg	0.95	0.95	0.95	152

Gambar 3. Confusion Matrix KNN (70-30)

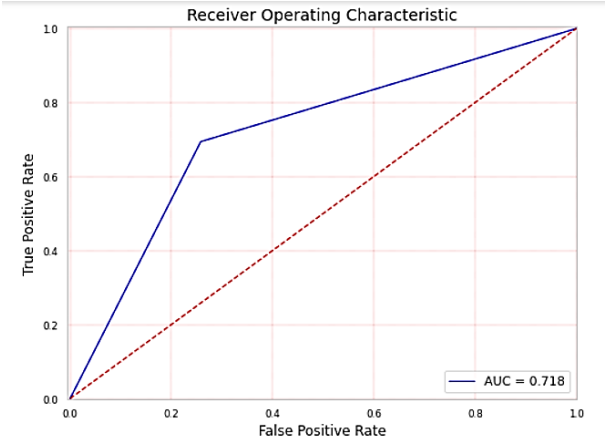
3.2 Naïve Bayes

Naive Bayes adalah teknik klasifikasi probabilistik yang sangat mudah. Dengan menjumlahkan frekuensi dan kombinasi nilai dalam kumpulan data yang diberikan, metode ini menghasilkan sekumpulan probabilitas. Semua atribut pada setiap kategori diasumsikan independen (non-dependent) dengan metode Naive Bayes [14].

Pada alat kolaboratif Google, bahasa pemrograman Python digunakan untuk pengujian. Gambar 4 dan 5 menggambarkan hasil pengujian di samping grafik ROC untuk pembagian data 70-30 dan 80-20.



Gambar 4. Grafik ROC Naïve Bayes (70-30)



Gambar 5. Grafik ROC KNN (80-20)

3.3 Pengujian Sistem

Model Confusion Matrix akan digunakan untuk menguji akurasi prediksi yang dibuat dari hasil yang telah diklasifikasikan menggunakan algoritma perbandingan K-Nearest Neighbor (KNN) dan Naive Bayes dalam klasifikasi diagnosis penyakit Parkinson. Pengklasifikasian jumlah data uji yang benar dan jumlah data uji yang salah ditunjukkan dalam tabel yang disebut matriks konfusi [15]. Data uji split untuk 70-30 dan 80-20, yang menghasilkan hasil yang ditunjukkan pada Tabel 2, akan digunakan.

Tabel 2. Hasil Komparasi Algoritma

Split Data		Akurasi	
Training Set	Test Set	KNN	Naïve Bayes
70	30	96%	74%
80	20	95%	70%

Berdasarkan temuan tersebut, algoritma K-Nearest Neighbor (KNN) memiliki akurasi tertinggi, 96%, dengan data split 70:30.

4. KESIMPULAN

Untuk mengklasifikasikan penyakit Parkinson, penelitian ini membandingkan dan membedakan dua algoritma, K-nearest neighbor (ANN) dan Naive Bayes. Selain itu, penelitian ini menggunakan dua pembagian data. Pada data split 70-30 terlihat bahwa K-Nearest Neighbor (KNN) memiliki skor akurasi yang lebih tinggi dibandingkan Naive Bayes. dimana nilai akurasi tertinggi dari algoritma KNN (K-nearest neighbor) adalah 96 persen. Walaupun akurasi algoritma Naive Bayes adalah 74%, namun dengan menggunakan algoritma K-Nearest Neighbor (KNN) dan Naive Bayes memberikan akurasi yang baik untuk klasifikasi penyakit Parkinson menurut hasil pengujian yang meliputi fungsi preprocessing dan klasifikasi. . Penelitian ini dapat memberikan informasi yang berguna tentang penyakit Parkinson dan algoritma yang mengungguli algoritma K-Nearest Neighbor (KNN) dan Naive Bayes serta dapat berfungsi sebagai model dan pengembangan ilmiah untuk penelitian masa depan.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada semua pihak yang telah membantu sehingga penelitian ini dapat selesai tepat waktu dan benar. Diharapkan penelitian ini bermanfaat untuk penelitian selanjutnya, yang memungkinkan untuk dimasukkan ke dalam aplikasi algoritma lain dengan menggabungkan atau mengkontraskan kedua pendekatan algoritmik tersebut. Ini akan menghasilkan hasil yang lebih beragam dan, tentu saja, informasi yang sangat berharga.

REFERENSI

[1] I. Swandana, J. Raharjo, and I. Safitri, "Identifikasi Penyakit Parkinson Dengan Metode Discrete Cosine Transform (Dct) Dan Learning Vector Quantization (Lvq) Berdasarkan Vgrf", 2020.
[2] A. R. Onibala, C. D. Mambo, and A. S. R. Masengi, "Peran Vitamin dalam Penanganan Penyakit Parkinson," Jurnal Biomedik (JBM), vol. 13, no. 3, p. 322, Aug. 2021, doi: 10.35790/jbm.13.3.2021.31956.
[3] A. N. , dan M. W. F. Sihananto, "Rainfall Forecasting Using Backpropagation Neural Network," Journal of Information Technology and Computer Science, 2017.
[4] M. Fariz Januarsyah, E. Zuhairi, and R. Firsandaya Malik, Perbandingan Algoritma Random Forest, Decision Stump, Naïve Bayes, Bayesian Network dan Algoritma C4.5 Untuk Prediksi Pola Kartu Poker, vol. 5, no. 1. 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Poker+Hand.
[5] V. K. S. S. T. C. dan S. M. K. Chandel and A. Sudrajat, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," CSI Trans. ICT, 2019.
[6] Z. Abdussamad, Metode Penelitian Kualitatif. 2021.
[7] Herdayati dan Syahrial, "Desain Penelitian Dan Teknik Pengumpulan Data," 2019.
[8] A. Syah Putra et al., "Examine Relationship of Soft Skills, Hard Skills, Innovation and Performance: the Mediation Effect of Organizational Learning," International Journal of Science and Management Studies (IJSMS), 2020, [Online]. Available: www.ijsmsjournal.org
[9] Armaita, E. Barlian, D. Hermon, I. Dewata, and I. Umar, "Policy Model of Community Adaptation using AHP in the Malaria Endemic Region of Lahat Regency - Indonesia," International Journal of Management and Humanities, vol. 4, no. 9, pp. 44–48, May 2020, doi: 10.35940/ijmh.I0855.054920.
[10] A. Rijali, "Analisis Data Kualitatif," 2018.
[11] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making. 2009.
[12] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes," 2018.
[13] W. Puspita Hidayanti, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada 'Lombok Vape On,'" Jurnal Informatika dan Teknologi, vol. 3, no. 2, 2020.
[14] A. Nafalski and A. P. Wibawa, "Machine Translation With Javanese Speech Levels' Classification," Informatics, Control, Measurement in Economy and Environment Protection, vol. 6, no. 1, pp. 21–25, Feb. 2016, doi: 10.5604/20830157.1194260.
[15] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," 2021.
[16] Larose. 2005. Discovering Knowledge in Data.Canada: Wiley-Interscience
[17] Sri Widaningsih (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4.5, Naïve Bayes, Knn, Dan Svm. Jurnal Tekno Insentif | ISSN (p): 1907-4964 | ISSN (e): 2655-089X
[18] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio- Science and Bio -Technology, 5 , 241-266
[19] Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in Data mining: An Overview. International Journal of P2P Network Trends and Technology (IJPTT), 3, 1-5.
[20] Annasaheb, A.B., & Verma, V.K. (2016). Classification Techniques: A Recent Survey. International Journal of Emerging Technologies in Engineering Research (IJETER), 4, 51-54.
[21] Suyatno. (2017) Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika
[22] Gorunescu, F. (2011). Data Mining Concepts, Models and Techniques. Berlin: Springer