Prediksi penyakit ginjal kronis dengan metode pengurangan fitur Symmetrical Uncertainty

Muhamad Kurniawan[1]

1 **Program Magister Teknik Informatika Program Pascasarjana Universitas Amikom Yogyakarta**
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta 55283
muhamad.17@students.amikom.ac.id

Abstrak

Data mining berhubungan dengan pencarian data untuk menemukan pola atau pengetahuan dari data keseluruhan. Data mining dapat digunakan untuk memprediksi suatu keadaan, seperti apakah seseorang terkena penyakit ginjal kronis atau tidak. Dalam penelitian ini metode pengurangan fitur symmetrical uncertainty dengan algoritma klasifikasi Gradient Boosting, Random Forest, Support Vector Machine, dan Naïve Bayes digunakan untuk memprediksi penyakit ginjal kronis. Jumlah atribut yang diklasifikasi adalah 24, 12, 6, 5, dan 4 atribut. Peningkatan nilai akurasi didapatkan pada pengurangan atribut dari 24 ke 12 dengan algoritma Naïve Bayes. Selain itu, diperoleh Support Vector Machine memiliki akurasi terbaik pada semua jumlah atribut, diikuti Gradient Boosting, Random Forest, dan Naïve Bayes. Pada klasifikasi 5 atribut, terlihat algoritma Support Vector Machine dan Gradient Boosting masih memiliki akurasi 1. Kelima atribut tersebut antara lain: hemoglobin, packed cell volume, serum creatinine, albumin, dan specifity gravity. Pengurangan atribut dapat meningkatkan akurasi dan dapat memudahkan proses prediksi karena jumlah atribut lebih sedikit.

Kata Kunci gradient boosting, SVM, pengurangan dimensi, sistem pendukung keputusan

1 Pendahuluan

Seiring dengan cepatnya perkembangan teknologi disemua bidang baik dari sektor pendidikan, pemerintahan, pertanian, dan khususnya kesehatan, teknologi dapat memberikan informasi yang cepat dan akurat baik untuk tim kesehatan, dokter, bahkan untuk pasien sendiri agar lebih mudah mengontrol kondisi kesehatan mereka. Berbagai rumah sakit menghasilkan data pasien dengan jumlah yang besar tiap tahunnya. Dengan teknologi, data tersebut dapat diolah untuk memperoleh pengetahuan baru yang bermanfaat di bidang kesehatan. Penyakit ginjal kronis (PGK) merupakan penyakit yang belakangan ini mencuat sebagai persoalan kesehatan selain penyakit jantung. Pada tahun 1990, penyakit ginjal berada pada peringkat 27 sebagai penyebab kematian, kemudian pada tahun 2010 meningkat menjadi peringkat ke-18 [1]. Di Indonesia sendiri, penyakit ginjal kronis berada pada peringkat ke-2 dalam pembiayaan Badan Penyelenggara Jaminan Sosial (BPJS) terbesar setelah penyakit jantung [2].

PGK adalah kondisi saat fungsi ginjal menurun secara bertahap karena kerusakan ginjal. Secara medis, penyakit ginjal kronis didefinisikan sebagai penurunan laju penyaringan atau filtrasi ginjal selama 3 bulan atau lebih. Pada kondisi penyakit ginjal kronis, cairan dan elektrolit, serta limbah dapat menumpuk dalam tubuh. Gejala dapat terasa lebih jelas saat fungsi ginjal sudah semakin menurun. Pada tahap akhir PGK, kondisi penderita dapat berbahaya jika tidak ditangani dengan terapi pengganti ginjal, salah satunya cuci darah. Data mining berhubungan dengan pencarian data untuk menemukan pola atau pengetahuan dari data keseluruhan. Pengklasifikasian suatu keadaan, seperti apakah seseorang terkena

Jurnal Open Access
Yayasan Lentera Dua Indonesia

penyakit ginjal kronis atau tidak dapat dilakukan dengan data mining. Pada proses klasifikasi tidak terlepas dengan adanya suatu *error* atau kesalahan, oleh karena itu, banyak ilmuwan yang melakukan penelitian untuk meningkatkan tingkat akurasi dari proses ini.

Penelitian memprediksi penyakit ginjal kronis, telah banyak dilakukan dengan membandingkan beberapa metode klasfikasi untuk memperoleh metode dengan akurasi terbaik [3; 4; 5; 6; 7; 8; 9] namun masih sedikit yang memperhitungkan metode *pre-processing* pada penelitiannya. Metode pengurangan dimensi merupakan salah satu *pre-processing* yang mana pada beberapa penelitian disebutkan terbukti mampu meningkatkan hasil akurasi [10; 11].

Pada penelitian ini digunakan metode pengurangan fitur, yaitu *Symmetrical Uncertainty*. Hasil dari pengurangan dimensi kemudian diklasifikasi dengan metode *Gradient Boosting, Random Forest*, *Support Vector Machine* (SVM) dan *Naïve Bayes*. Simulasi model pada penelitian ini menggunakan bahasa pemrograman R.

2 Penelitian terkait

N. Tangri, dkk melakukan penelitian dengan membangun dan melakukan validasi model prediksi dari 2 cohort pasien dengan penyakit ginjal kronis stadium 3 sampai 5. Model dibangun menggunakan *Cox proportional hazard regression* dan dievaluasi dengan C-statistics [6]. Model yang dibangun dapat memprediksi dengan cukup akurat penyakit ginjal kronis pada stadium 3 sampai 5.

Pada penelitiannya, Di Noia, dkk. menyajikan sebuah perangkat lunak yang dapat mengklasifikasi status kesehatan pasien berpotensi terkena penyakit ginjal stadium akhir [3]. Perangkat lunak yang digunakan menggunakan algoritma jaringan syaraf buatan yang di latih dengan data yang dikumpulkan selama 38 tahun di Universitas Bari. Perangkat lunak tersedia dalam betuk aplikasi web atau aplikasi android.

Kemudian S. Vijayarani, dkk. melakukan penelitian prediksi penyakit ginjal kronis dengan algoritma klasifikasi *Naïve Bayes* dan SVM [4]. Penelitian ini fokus pada membandingkan akurasi dan waktu eksekusi pada kedua algoritma tersebut . Hasil penelitian tersebut didapat bahwa algoritma SVM lebih baik dari pada *Naïve Bayes*.

Selain itu, L. Jena, dkk. melakukan penelitian dengan memprediksi penyakit kronis ginjal menggunakan aplikasi WEKA [7]. Algoritma yang digunakan adala*h Naïve Bayes, Multilayer Perceptron*, SVM, J48, *Conjunctive Rule*, dan *Decision Table*. Berbeda dengan penelitian yang dilakukan oleh S. Vijayarani, pada penelitian ini algoritma *Naïve Bayes* memiliki akurasi yang lebih baik dibandingkan SVM. Kemudian akurasi terbaik diperoleh dengan menggunakan algoritma *Multilayer Perceptron*. Pada penelitian lainnya, Asif Salekin dkk. menggunakan algoritma klasifikasi K-NN, *Random Forest*, dan *Neural Network* untuk memprediksi penyakit ginjal kronis [8]. Selain itu, untuk mengurangi *overfitting* dan menentukan atribut prediktif, mereka melakukan pengurangan fitur dengan metode *wrapper* dan regularisasi LASSO. Diperoleh bahwa algoritma *Random Forest* dengan klasifikasi atribut sebanyak 12 memiliki akurasi tertinggi, yaitu 0.998 dengan *F1-measure*.

3 Metodologi

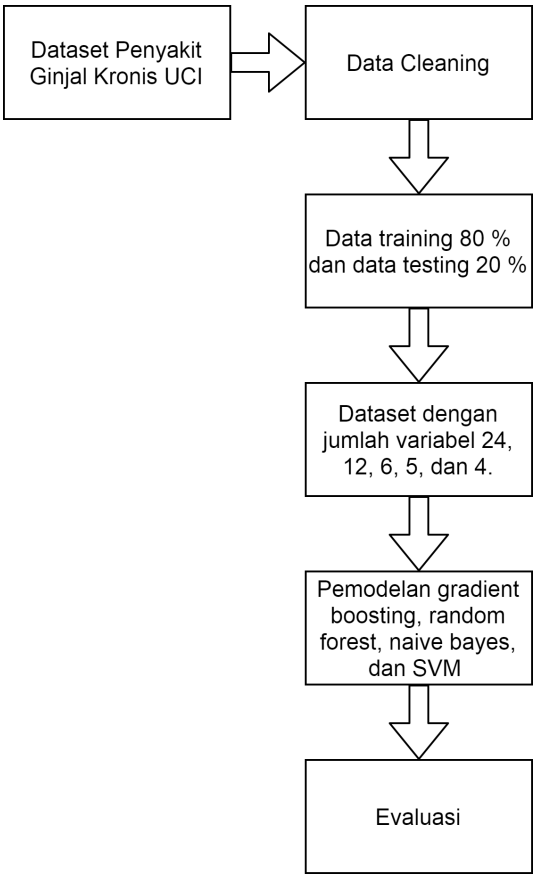
3.1 Dataset percobaan

Dataset penyakit ginjal yang digunakan diambil dari UCI *Machine Learning Repository* tersebut berasal dari rumah sakit Appolo, India yang terdiri dari 25 atribut dan 400 baris, seperti yang ditunjukan dalam tabel 1.

Tabel 1 Tabel dataset

Atribut	Nilai / Satuan Atribut	Keterangan
Umur	Tahun	
Tekanan darah	mm/Hg	
Specific gravity	(1.005,1.010,1.015,1.020,1.025)	
Albumin	(0,1,2,3,4,5)	
Sugar	(0,1,2,3,4,5)	
Red blood cells	(normal,abnormal)	
Pus cell	(normal,abnormal)	
Pus cell clumps	(present,notpresent)	
Bacteria	(present,notpresent)	
Blood glucose random	mgs/dl	
Blood urea	mgs/dl	Kadar urea yang tinggi dalam darah dapat menunjukkan adanya masalah pada ginjal [12].
Serum creatinine	mgs/dl	
Sodium	mEq/L	
Potassium	mEq/L	
Hemoglobin	Gms	Studi memperlihatkan bahwa orang yang terkena penyakit ginjal kronis memiliki hemoglobin yang rendah [13].
Packed cell volume		
White blood cell count	cells/cmm	
Red blood cell count	millions/cmm	Studi memperlihatkan bahwa orang yang terkena penyakit ginjal kronis memiliki red blood cell count yang rendah [13].
Hypertension	(yes,no)	
Diabetes mellitus	(yes,no)	
Coronary artery disease	(yes,no)	
Appetite	(good,poor)	
Pedal edema	(yes,no)	
Anemia	(yes,no)	Beberapa studi memperlihatkan orang yang terkena penyakit ginjal kronis juga mengalami anemia [13].
Class	(ckd,notckd)	

Untuk mengisi nilai *missing value*, digunakan metode *multiple imputation and chained equation*. Dataset yang telah diisi missing value nya kemudian dipisahkan antara data training dan data testing. Data training sebesar 80% dari dataset dengan perlakuan 10 *fold cross validation*. Kemudian dihitung nilai atribut penting nya dengan metode *symmetrical uncertainty*. Setelah itu diambil 12, 6, 5, dan 4 variabel yang paling berpengaruh. Setelah diperoleh beberapa set variabel tersebut kemudian dilakukan klasifikasi dengan algoritma *Naïve Bayes, SVM, Random Forest*, dan *Gradient Boosting*. Evaluasi dilakukan dengan melihat nilai akurasi, sensitivitas, dan spesifitas nya. Alur penelitian yang dilakukan seperti tertampil pada Gambar 1.



Gambar 1 Alur penelitian

3.2 Metode yang digunakan

Salah satu metode pengurangan fitur adalah *Symmetrical Uncertainty* (SU) yang melihat seberapa berpengaruh suatu variabel terhadap kelas label. Rumus *Symmetrical Uncertainty* tertampil dalam formula 1.

$$SU(X,Y) = \frac{2xMI(X,Y)}{H(X)+H(Y)} \tag{1}$$

Dengan SU adalah nilai attribute importance, MI adalah *Mutual Information* dan H adalah *Entrophy*. Semakin tinggi nilai SU suatu variabel, maka semakin berpengaruh variabel tersebut terhadap kelas label dan sebaliknya [14]. Metode *Naïve Bayes* merupakan algorithma pembelajaran mesin yang menggunakan teorema Bayes yang banyak digunakan untuk mengatasi masalah klasifikasi. Rumus *Naïve Bayes* adalah tertampil dalam formula 2.

$$P(A|B) = \frac{P(B|A)(PA)}{P(B)} \tag{2}$$

Dengan $P(A|B)$ adalah probabilitas event A terjadi terhadap event B, $P(A)$ adalah probabilitas of event A terjadi, $P(B)$ adalah kemungkinan event B terjadi, dan $P(B|A)$ adalah keungkinan event B terjadi terhadap event A. SVM merupakan salah satu algoritma

yang bisa digunakan untuk mengatasi masalah klasifikasi dengan memisahkan sejumlah data menggunakan *hyperplane*. SVM bertanggung jawab dalam memaksimalkan nilai margin– jarak antara *hyperplane* terhadap titik – titik terdekat. Rumus *hyperplane* dapat ditulis seperti dalam formula 3.

$$\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + ... + \beta_n \times x_n \tag{3}$$

Random Forest merupakan salah satu metode lain untuk melakukan peningkatan akurasi dalam klasifikasi. Metode ini berbasis metode *decision tree* yang berbentuk seperti pohon dengan sebuah *root* node yang digunakan untuk mengumpulkan data. Berbeda dengan *tree* pada umumnya yang membagi setiap *node* berdasarkan pembagian terbaik dalam setiap variabel. Sedangkan *Random Forest* setiap *node* dibagi berdasarkan sekelompok prediktor yang terbaik di antara *node*yang dipilih secara acak [15]. Random Forest menggunakan metode bagging dalam melakukan training model – model nya.

Seperti *Random Forest*, *Gradient Boosting* juga merupakan salah satu metode *decision tree*. Perbedaannya adalah pada proses training modelnya menggunakan metode *boosting*. Metode *boosting* melakukan training model secara sekuensial. Tiap model belajar dari kesalahan pada model sebelumnya. Suatu studi memperlihatkan bahwa *Gradient Boosting* sangat baik dalam menyelesaikan masalah klasifikasi dengan variabel berjumlah sedikit sedangkan pada variabel dengan jumlah banyak Random Forest memiliki performa yang lebih baik [16].

3.3 Diskusi dan hasil

Tabel 2 memperlihatkan nilai attribute importance pada dataset penyakit ginjal UCI yang dihitung dengan metode *Symmetrical Uncertainty*. Nilai *Attribute Importance* memperlihatkan seberapa berpengaruh suatu variabel terhadap kelas label. Semakin tinggi nilai *attribute importance* suatu variabel, semakin berpengaruh variabel tersebut pada kelas label. Nilai attribute importance yang rendah memperlihatkan bahwa variabel tersebut kurang berhubungan dengan kelas label.

Terlihat bahwa 6 variabel yang paling berpengaruh terhadap label penyakit ginjal adalah hemoglobin, packed cell volume, serum creatinine, albumin, specifity gravity, dan red blood cell count. Hal ini sedikit berbeda dengan penentuan nilai atribut penting dengan metode LASSO dimana variabel yang paling berpengaruh adalah *specifity gravity, albumin, diabetes melitus, hypertension, hemoglobin*, dan serum creatinine [8]. Dari hasil klasifikasi dataset dengan 24 atribut diperoleh bahwa *Gradient Boosting* dan SVM memiliki akurasi terbaik, diikuti dengan *Random Forest* dan *Naïve Bayes* seperti tertampil dalam tabel 3.

Tabel 4 memperlihatkan hasil klasifikasi dengan 12 atribut dengan attribute importance tertinggi. Diperoleh bahwa *Gradient Boosting* dan SVM memiliki nilai akurasi tertinggi diikuti dengan *Random Forest* dan *Naïve Bayes*. *Gradient Boosting* memang memiliki reputasi yang sangat baik dalam menyelesaikan masalah klasifikasi, bahkan banyak penelitian yang memperlihatkan bahwa *Gradient Boosting* menghasilkan akurasi yang lebih baik dibandingkan *Random Forest* [16; 17; 18]. Studi yang dilakkan oleh Rich Caruna, dkk. memperlihatkan bahwa hal ini hanya berlaku pada variabel berjumlah sedikit. Pada variabel berjumlah banyak (> 4000), metode *Random Forest* lebih baik dibandingkan dengan *Gradient Boosting* [16]. Menurut Rich Caruna, dkk, hal ini dapat disebabkan pada variabel berjumlah banyak, metode *Gradient Boosting* mudah mengalami *overfitting*.

Selain itu, terlihat bahwa nilai akurasi *Naïve Bayes* meningkat saat jumlah variabel dikurangi dari 24 ke 12 atribut. Hal ini dapat disebabkan pada klasifikasi 24 atribut terdapat variabel dengan attribute importance rendah, sedangkan pada klasifikasi 12 atribut, *attribute*

Tabel 2 Nilai attribute importance dari dataset

Atribut	Attribute Importance
Hemoglobin	0.571227
Packed cell volume	0.547463
Serum creatinine	0.534347
Albumin	0.532017
Specific gravity	0.496064
Red blood cell count	0.362348
Hypertension	0.354957
Diabetes mellitus	0.301888
Red blood cells	0.290324
Blood urea	0.286323
Pus cell	0.247991
Blood glucose random	0.247143
Sodium	0.217898
Sugar	0.20888
Blood pressure	0.197831
Appetite	0.191279
Potassium	0.180148
Pedal edema	0.178323
Anemia	0.144399
White blood cell count	0.140188
Age	0.107375
Coronary artery disease	0.088814
Pus cell clumps	0.086357
Bacteria	0.061297

Tabel 3 Hasil klasifikasi 24 atribut

	Accuracy	Sensitivity	Specificity
Gradient Boost	1	1	1
Random Forest	0.9625	0.98	0.93
Naïve Bayes	0.95	0.96	0.93
SVM	1	1	1

*importanc*e rendah dihilangkan. Nilai *attribute importance* yang rendah dapat mengacaukan akurasi pada proses klasifikasi.

Tabel 4 Hasil klasifikasi 12 atribut

	Accuracy	Sensitivity	Specificity
Gradient Boost	1	1	1
Random Forest	0.9625	0.98	0.93
Naïve Bayes	0.9625	0.98	0.93
SVM	1	1	1

Pada klasifikasi dengan 6 atribut, nilai akurasi Naïve Bayes menurun seperti terlihat pada tabel 5. Sedangkan untuk algoritma lainnya tidak ada perubahan nilai akurasi, bahkan

pada pada pengurangan atribut dari 6 ke 5 juga tidak terdapat pengurangan nilai akurasi pada semua algoritma. Pada tabel **??** terlihat bahwa dengan hanya 5 atribut masih dapat diperoleh akurasi 100% dengan metode Gradient Boosting dan SVM. Kelima atribut tersebut antara lain *hemoglobin, packed cell volume, serum creatinine, albumin,* dan *specifity gravity.*

Tabel 5 Hasil klasifikasi 6 atribut

	Accuracy	Sensitivity	Specificity
Gradient Boost	1	1	1
Random Forest	0.9625	0.98	0.93
Naïve Bayes	0.95	0.96	0.93
SVM	1	1	1

Tidak berubahnya nilai akurasi ketika atribut ke–6 (*red blood cell count*) dihilangkan dapat disebabkan karena atribut *red blood cell count* memiliki korelasi yang cukup tinggi dengan atribut selain kelas label, yaitu: atribut ke-1 (*hemoglobin*) dan ke-2 (*packed cell volume*). Bahkan korelasi dengan kedua atribut tersebut lebih besar dibandingkan korelasi dengan label kelas. Perubahan terjadi pada semua metode jika atribut dikurangi menjadi 4 seperti tertampil pada tabel 7.

Tabel 6 Hasil klasifikasi 5 atribut

	Accuracy	Sensitivity	Specificity
Gradient Boost	1	1	1
Random Forest	0.9625	0.98	0.93
Naïve Bayes	0.95	0.96	0.93
SVM	1	1	1

Tabel 7 Hasil klasifikasi 4 atribut

	Accuracy	Sensitivity	Specificity
Gradient Boost	0.9375	0.9091	1
Random Forest	0.925	0.92	0.93
Naïve Bayes	0.9	0.88	0.93
SVM	0.975	0.96	1

Tabel 8 meperlihatkan nilai korelasi dengan metode pearson antara atribut *red blood cell count* dengan kelima atribut lainnya. Terlihat bahwa nilai korelasi antara red *blood cell count* dengan *hemoglobin* sebesar 0.7698490, dengan *packed cell volume* sebesar 0.7559931, dan dengan kelas label sebesar -0.6304727. Jika ketiga nilai tersebut diabsolutkan, nilai korelasi *hemoglobin* dan *packed cell volume* masih lebih tinggi dibandingkan kelas label. Nilai negatif pada kelas label tidak menunjukkan bahwa kelas label kurang berkorelasi, namun nilai mendekati -1 menunjukkan hubungan berkorelasi namun berkebalikan sedangkan nilai mendekati 1 menunjukan berkorelasi dan searah. Nilai mendekati 0 menunjukkan hubungan kurang berkorelasi.

Tinggi nya nilai korelasi *red blood cell count* dengan 2 atribut selain kelas label menyebabkan atribut *red blood cell count* dapat diwakilkan dengan kedua atribut tersebut sehingga ketika *red blood cell count* dihilangkan nilai akurasi tidak banyak berubah [19]. Melihat

Tabel 8 Nilai korelasi antara red blood cell

Atribut	Nilai Korelasi
Hemoglobin	0.7698490
Packed cell volume	0.7559931
Serum creatinine	-0.3499213
Albumin	-0.4105895
Specific gravity	0.4129566
class	-0.6304727

fungsinya pada tubuh, red blood cell atau sel darah merah memiliki fungsi yang sama dengan *hemoglobin*, sehingga wajar atribut *red blood cell* dapat diwakilkan oleh *hemoglobin*.

Secara keseluruhan, SVM memiliki akurasi terbaik, diikuti dengan *Gradient Boosting, Random Forest*, dan *Naïve Bayes*. Hasil ini berbeda dengan penelitian L. Jena, dkk. yang menggunakan dataset yang sama 7, dimana pada penelitian tersebut SVM memiliki nilai akurasi yang jauh lebih rendah dari pada *Naïve Bayes*. Hal ini dapat disebabkan proses perlakuan missing value nya yang berbeda. Pada penelitian L. Jena, dkk., perlakuan missing value tidak disebutkan, kemungkinan adanya missing value tidak dihiraukan. Sedangkan pada penelitian ini, *missing value* di isi dengan metode *multiple imputation* and *chained equation*.

SVM cukup sensitif terhadap adanya missing value, karena SVM hanya melakukan pemodelan dengan suatu bagian data saja, sedangkan kebanyakan classifier menggunakan keseluruhan data [20].

4 Kesimpulan dan saran

Pada penelitian ini dilakukan prediksi penyakit ginjal dengan metode pemilihan fitur *Symmetrical Uncertainty*. Algoritma klasifikasi yang digunakan adalah *Naïve Bayes*, SVM, *Random Forest*, dan *Gradient Boosting*. Jumlah variabel yang diklasifikasi 24, 12, 6, 5, dan 4. Secara keseluruhan, SVM memiliki akurasi terbaik, diikuti dengan *Gradient Boosting, Random Forest*, dan *Naïve Bayes*. Dari penelitian ini diperoleh bahwa dengan hanya 5 atribut masih dapat diperoleh akurasi 100% dengan metode *Gradient Boosting* dan SVM. Kelima atribut tersebut antara lain *hemoglobin, packed cell volume, serum creatinine, albumin*, dan *specifity gravity*.

Penelitian kedepannya dapat dengan membandingkan antara metode *symmetrical uncertainty* dengan metode pemilihan fitur lainnya. Selain itu dapat dilakukan dengan dataset lainnya yang memiliki jumlah baris dan variabel yang lebih banyak, serta berasal dari negara Indonesia.

Pustaka

1 *The Global Burden of Disease: Generating Evidence, Guiding Policy.*, Institute for Health Metrics and Evaluation, Seattle, WA: IHME, 2013.

2 *Situasi Penyakit Ginjal Kronis*, Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia, Jl. HR Rasuna Said Blok X5 Kav. 4-9 Lantai 6 Blok C Jakarta Selatan, 2017.

3 T. Di Noia, V. C. Ostuni, F. Pesce, G. Binetti, D. Naso, F. P. Schena, and E. Di Sciascio,

"An end stage kidney disease predictor based on an artificial neural networks ensemble," *Expert systems with applications*, vol. 40, no. 11, pp. 4438–4445, 2013.

4 S. Vijayarani, S. Dhayanand *et al.*, "Data mining classification algorithms for kidney disease prediction," *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 4, pp. 13–25, 2015.

5 I. Pasadana, D. Hartama, M. Zarlis, A. Sianipar, A. Munandar, S. Baeha, and A. Alam, "Chronic kidney disease prediction by using different decision tree techniques," in *Journal of Physics: Conference Series*, vol. 1255, no. 1. IOP Publishing, 2019, p. 012024.

6 N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey, "A predictive model for progression of chronic kidney disease to kidney failure," *Jama*, vol. 305, no. 15, pp. 1553–1559, 2011.

7 L. Jena and N. K. Kamila, "Distributed data mining classification algorithms for prediction of chronic-kidney-disease," *International Journal of Emerging Research in Management &Technology*, vol. 4, no. 11, pp. 110–118, 2015.

8 A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2016, pp. 262–270.

9 Tabassum, Mamatha Bai, and J. Majumdar, "Analysis and prediction of chronic kidney disease using data mining techniques," 2017. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.26856.72965

10 M. Nasution, O. Sitompul, and M. Ramli, "Pca based feature reduction to improve the accuracy of decision tree c4. 5 classification," in *Journal of Physics: Conference Series*, vol. 978, no. 1. IOP Publishing, 2018, p. 012058.

11 H. Xie, J. Li, Q. Zhang, and Y. Wang, "Comparison among dimensionality reduction techniques based on random projection for cancer classification," *Computational biology and chemistry*, vol. 65, pp. 165–172, 2016.

12 P. Madan, O. P. Kalra, S. Agarwal, and O. P. Tandon, "Cognitive impairment in chronic kidney disease," *Nephrology Dialysis Transplantation*, vol. 22, no. 2, pp. 440–444, 2007.

13 O. Latiweshob, H. Elwerfaly, D. Sheriff *et al.*, "Haematological changes in predialyzed and hemodialyzed chronic kidney disease patients in libya," *IOSR J of Dental and Med Sciences*, vol. 16, pp. 106–12, 2017.

14 A. Saikhu, A. Z. Arifin, and C. Fatichah, "Correlation and symmetrical uncertainty-based feature selection for multivariate time series classification," *International Journal of Intelligent Engineering and Systems*, vol. 12, pp. 129–137, 06 2019.

15 A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

16 R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 96–103.

17 I. Babajide Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, vol. 21, no. 8, p. 983, 2016.

18 J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," in *BMC proceedings*, vol. 5, no. S3. Springer, 2011, p. S11.

19 R. Misir, M. Mitra, and R. K. Samanta, "A reduced set of features for chronic kidney disease prediction," *Journal of pathology informatics*, vol. 8, 2017.

20 T. G. Stewart, D. Zeng, and M. C. Wu, "Constructing support vector machines with missing data," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 10, no. 4, p. e1430, 2018.