ILKOM Jurnal Ilmiah
Vol. 12 No. 2, Agustus 2020, pp.130-135
E-ISSN 2548-7779

Deteksi Diabetik Retinopati menggunakan Regresi Logistik

Raras Tyasnurita[a,1,*] dan Adhi Yoga Muris Pamungkas[a,2]

[a] *Departemen Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember Surabaya Indonesia 60111*
[1] *raras@is.its.ac.id;* [2] *yogamuris@gmail.com*
**corresponding author*

INFORMASI ARTIKEL

Diterima : 19 Juni 2020
Diulas : 18 Juli 2020
Direvisi : 22 Juli 2020
Diterbitkan : 27 Agustus 2020

Kata Kunci:
Diabetik retinopati
Regresi logistik
Klasifikasi
Pembelajaran mesin

ABSTRAK

Diabetik retinopati adalah penyakit yang diakibatkan oleh komplikasi dari diabetes melitus yang dapat menyebabkan kerusakan pada retina mata. Kondisi ini akan mengganggu penglihatan pada penderita. Mendeteksi penyakit ini sangat penting dilakukan untuk mencegah kebutaan total pada penderita diabetes melitus. Salah satu metode deteksi adalah dengan menggunakan pembelajaran mesin. Penelitian ini menggunakan data ekstraksi fitur dari citra mata yang terdapat tanda-tanda diabetik retinopati atau tidak. Fokus penelitian ini adalah klasifikasi data penderita diabetik retinopati. Klasifikasi dilakukan dengan menggunakan algoritma regresi logistik. Terdapat empat macam kondisi pelatihan data pada model pembelajaran mesin yaitu dengan menggunakan parameter bawaan, standarisasi atribut, pemilihan atribut, dan pengaturan parameter. Model menggunakan nilai kontrol regularisasi (C) sebesar 11,288, nilai iterasi sebanyak 200, dan penalti regularisasi (l1). Hasil dari eksperimen menunjukkan bahwa model tersebut dengan kondisi atribut lengkap menghasilkan akurasi sebesar 80,17% pada data validasi.

Keywords:
Retinopathy diabetic
Logistic regression
Classification
Machine learning

ABSTRACT

Retinopathy diabetic is a disease caused by diabetes mellitus complications that can cause damage to the retina of the eye. It has a direct impact on the disruption of the vision of the patient. Detecting this disease is very important to prevent total blindness on diabetes mellitus patients. One method to do the detection is by using machine learning. This research uses feature extraction data from an image that contains signs of retinopathy diabetic or not. In this research, we focus on retinopathy diabetic classification. We applied logistic regression algorithm for classification. There is four training condition in a machine learning model: using the default parameter, feature standardization, feature selection, and hyper-parameter tuning. The model used a regularization control (C) value of 11.288, iterations 200, and a regularization penalty (l1). The experimental results show that this proposed model with full features produced 80,17% accuracy in data validation.

I. Pendahuluan

Diabetes melitus merupakan penyakit metabolik dimana penderita diabetes tidak bisa memproduksi insulin dalam jumlah cukup atau tubuh tidak mampu menggunakan insulin secara efektif sehingga terjadi kelebihan gula di dalam darah. Kondisi ini baru dirasakan setelah terjadi komplikasi lanjut pada organ tubuh. Pada penderita diabetes melitus, dapat terjadi komplikasi pada tingkat sel dan anatomik. Komplikasi kronik yang disebabkan oleh diabetes melitus dapat terjadi pada pembuluh darah besar (*makrovaskuler*) dan pembuluh darah kecil (*mikrovaskuler*). Komplikasi *makrovaskuler* yang umum terjadi adalah pembekuan darah pada otak, penyakit jantung koroner, dan stroke. Sedangkan komplikasi *mikrovaskuler* yang umum terjadi adalah hiperglikemia dan retinopati (kebutaan) [1].

Retinopati diabetik merupakan kerusakan retina yang diakibatkan oleh komplikasi dari diabetes melitus [2]. Menurut data dari WHO pada tahun 2015, Indonesia menempati posisi ke tujuh di dunia untuk prevalensi penderita diabetes tertinggi di dunia dengan estimasi penderita diabetes sebesar 10 juta orang dari jumlah penderita diabetes di dunia sebesar 415 juta jiwa. Berdasarkan penelitian yang dilakukan oleh

Soewondo dkk (2010) [3], sebanyak 42% dari 1785 penderita diabetes melitus di Indonesia mengalami komplikasi retinopati diabetes.

Pembelajaran mesin *(machine learning)* dapat didefinisikan sebagai proses dimana sistem dapat meningkatkan performa dari pengalaman. Terdapat tiga jenis *machine learning* yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Salah satu *task* dalam *supervised learning* adalah klasifikasi, dimana ini menjadi fokus permasalahan dari penelitian ini. Pada penelitian ini dibuat model *machine learning* untuk melakukan klasifikasi terhadap penyakit diabetik retinopati. Penelitian terbaru menunjukkan bahwa dari beberapa algoritma pembelajaran mesin, pengklasifikasi regresi logistik menghasilkan akurasi yang paling tinggi (77%) dibanding dengan *Random Forest* (68%), *Decision Tree* (55%), *Adaboost* (67%), dan *K-Nearest Neighbour* (65%) [4]. Selain itu, penelitian sebelumnya menunjukkan bahwa regresi logistik lebih superior dibanding algoritma pembelajaran mesin yang lain dalam hal pemodelan prediksi klinis [5]. Penelitian ini bertujuan untuk melakukan klasifikasi mengenai kejadian diabetik retinopati. Prediksi dilakukan dengan menggunakan metode regresi logistik.

II. Metode

Machine Learning (Pembelajaran Mesin) memiliki peran penting dalam mengenali pola pada data seperti data kesehatan di masa serba digital ini. Herbert Simon mendefinisikan belajar sebagai proses dimana sistem dapat meningkatkan performa dari pengalaman [6]. Tom Mitchell mendefinisikan *machine learning* sebagai studi dari algoritma yang dapat meningkatkan performa *P* pada aktivitas *T* dengan pengalaman *E* [7]. *Machine learning* terdiri dari tiga jenis yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [8]. *Supervised learning* adalah sistem belajar dari data yang memiliki label, contoh dari *supervised learning* adalah klasifikasi. *Unsupervised learning* merupakan sistem belajar dari data yang tidak memiliki label, contohnya merupakan permasalahan klusterisasi. Sedangkan *reinforcement learning* merupakan sistem yang diberikan pada serangkaian aksi dengan *reward*, contoh dari *reinforcement learning* adalah permasalahan penilaian skor kredit pada bank.

Penelitian yang dilakukan penulis berfokus pada permasalahan klasifikasi dimana sistem akan melakukan klasifikasi apakah seseorang memiliki risiko diabetik retinopati atau tidak. Alur klasifikasi pada penelitian ini ditunjukkan pada Gambar 1 dimana terdiri dari 5 tahap [9] yaitu:

A. Pengumpulan data

Langkah pertama yang dilakukan adalah mengumpulkan data. *Dataset* yang digunakan diambil dari *UCI Machine Learning* [10]. *Dataset* yang diambil adalah data mengenai kejadian retinopati diabetik. Pada *dataset* terdapat ciri *(feature)* yang diekstrak dari *dataset* berupa gambar dari *Messidor* untuk memprediksi apakah pada gambar terdapat gejala retinopati diabetik atau tidak.

B. Pemrosesan data

Setelah data dikumpulkan, langkah selanjutnya adalah melakukan pengolahan terhadap data yang sudah dikumpulkan. *Dataset* memiliki jumlah baris sebanyak 1151 baris. Pengolahan data ini meliputi menghilangkan *missing values* yang ada dalam data. Pada *dataset* terdapat 20 atribut dengan rincian 19 atribut ciri dan 1 atribut label. Pada *dataset* dilakukan pembagian data menjadi data pelatihan sebanyak 90% dari keseluruhan data dan data validasi sebesar 10% dari keseluruhan data. Data pelatihan yang dihasilkan kemudian dilakukan pembagian data menjadi 70% sebagai data pelatihan dan 30% sebagai data pengujian.

C. Pelatihan model

Pelatihan model dilakukan dengan menggunakan tiga skenario. Skenario pertama yang digunakan adalah membuat *base model* dengan menggunakan keseluruhan variabel dari data. Skenario kedua adalah melakukan pemodelan dengan melakukan standarisasi pada data. Skenario ketiga adalah melakukan pemodelan dengan melakukan *feature selection* pada ciri (*feature)* dari data.

D. Evaluasi model

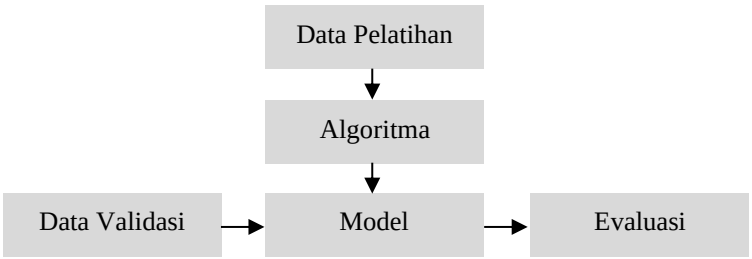
Setelah dilakukan pelatihan model, selanjutnya adalah mengevaluasi performa dari model tersebut. Evaluasi model ini meliputi tingkat akurasi model, presisi, *recall*, sensitivitas, dan *F1 score* [11]. Selain itu metode evaluasi menggunakan *confusion matrix*.

E. Peningkatan performa

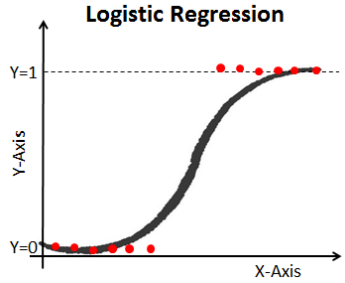
Dari evaluasi model akan diketahui performa dari model. Terdapat beberapa teknik untuk meningkatkan performa dari model, diantaranya adalah *hyper-parameter tuning* [12]. Pada tahap peningkatan performa dilakukan penyetelan parameter atau *hyper-parameter tuning* pada metode logistik regresi. Parameter yang dilakukan penyetelan adalah parameter C (variabel kontrol yang mempertahankan modifikasi kekuatan regularisasi), iterasi maksimum (*max_iter*), dan penalti.

Regresi Logistik adalah metode statistik yang merupakan bagian dari analisis regresi yang biasanya digunakan untuk memprediksi kelas biner [13]. Variabel target bersifat dikotomi atau biner. Dikotomi berarti hanya terdiri atas dua kelas atau nilai, yang biasanya dituliskan dengan angka 1 (ya, sukses, dll) atau 0 (tidak, gagal, dll). Regresi logistik juga merupakan salah satu algoritma klasifikasi yang sering dipakai di dunia kesehatan [14]. Fungsi logistik disebut juga dengan Fungsi Sigmoid karena memberikan kurva berbentuk 'S'

yang dapat mengambil bilangan *real-value* dan memetakannya menjadi nilai antara 0 dan 1 [8]. Kurva regresi logistik ditunjukkan pada Gambar 2. Jika kurva menuju *infinity* positif, nilai y (*output*) akan diprediksi menjadi 1, dan jika kurva pergi ke *infinity* negatif, nilai y (*output*) akan diprediksi menjadi 0. Jika *output* dari fungsi sigmoid lebih dari 0,5, kita dapat mengklasifikasikan hasilnya sebagai 1 atau YA, dan jika kurang dari 0,5, kita dapat mengklasifikasikannya sebagai 0 atau TIDAK.



Gambar 1. Bagan alur klasifikasi pada *machine learning*



Gambar 2. Kurva Regresi logistik [15]

III. Hasil dan Pembahasan

Pada bagian ini diuraikan hasil dari model *machine learning* yang telah dibuat. Model dibuat dengan menggunakan bahasa Python melalui *jupyter notebook*.

A. Proses pelatihan model

Dataset terdiri dari 1151 baris data dan 20 atribut. Dari 1151 baris data tersebut dilakukan pembagian data dengan porsi sebanyak 90% untuk data pelatihan dan 10% untuk data validasi. Dari 90% data pelatihan kemudian dilakukan pembagian data sebagai data pelatihan dengan proporsi 70% dan data pengujian dengan proporsi sebesar 30%. Model dibuat dengan menggunakan algoritma regresi logistik. Pada skenario pemodelan 1 (Model 1) dilakukan pemodelan dengan menggunakan parameter penalti l1 dan nilai *C* sebesar 0.1. Parameter penalti (regularisasi l1) adalah sama dengan nilai absolut dari besarnya koefisien. Terdapat pula parameter penalti lain yaitu regularisasi l2 dimana nilainya sama dengan sama dengan kuadrat dari koefisien. Dari hasil pemodelan dengan menggunakan nilai parameter tersebut didapatkan hasil akurasi pada data pelatihan sebesar 75,96% dan akurasi pada data pengujian sebesar 75,24%. Selanjutnya dilakukan tes terhadap data validasi sebesar 10% dari total baris data. Tabel 1 menunjukkan hasil tes terhadap data validasi dimana didapatkan hasil akurasi sebesar 78,45%. *Confusion matrix* dari model 1 dapat dilihat pada Tabel 2 dimana detail analisis statistik dirangkum pada Tabel 3.

Tabel 1. Performa Model 1

Model	Akurasi Data Pelatihan	Akurasi Data Pengujian	Akurasi Data Validasi
Model 1	75,96%	75,24%	78,45%

Tabel 2. *Confusion Matrix* Model 1 pada Data Validasi

	0	1
0	50	11
1	14	41

Tabel 3. *Precision, Recall, f1-score,* dan *Support* pada Data Validasi pada Model 1

	Precision	*Recall*	*F1-score*	*Support*
0 (Non RD)	0.78	0.82	0.80	61
1 (RD)	0.79	0.75	0.77	55
Avg / total	0.78	0.78	0.78	116

B. Standarisasi data

Skenario pemodelan kedua adalah dengan melakukan standarisasi pada data. Standarisasi dilakukan dengan menggunakan *class "StandardScaler"* dari *library scikit learn "sklearn.preprocessing"*. Standarisasi dilakukan pada atribut ciri dan tidak dilakukan pada label. Hasil yang didapatkan dari model 2 ini justru mengalami penurunan yang ditunjukkan pada Tabel 4. Akurasi dari data pelatihan menjadi 68,37%, akurasi pada data pengujian menjadi 67,85%, dan akurasi dari data validasi menjadi 66,38%. *Confusion matrix* dan analisis statistik secara berturut-turut ditunjukkan pada Tabel 5 dan 6.

Tabel 4. Performa Model 2 *(standard scaler)*

Model	Akurasi Data Pelatihan	Akurasi Data Pengujian	Akurasi Data Validasi
Model 2	68,37%	67,85%	66,38%

Tabel 5. *Confusion matrix* model 2 pada data validasi

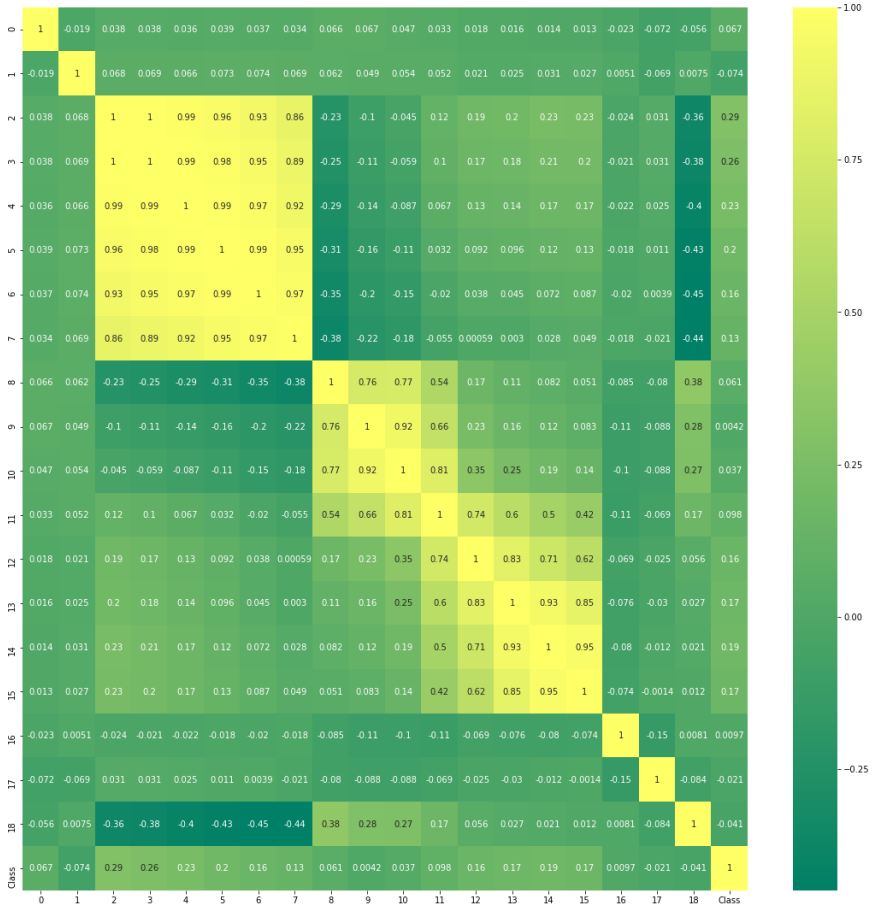
	0	1
0	37	24
1	15	40

Tabel 6. *Precision, recall, f1-score,* dan *Support* pada Data Validasi pada Model 2

	Precision	***Recall***	***F1-score***	***Support***
0 (Non RD)	0.71	0.61	0.65	61
1 (RD)	0.62	0.73	0.67	55
Avg / total	0.67	0.66	0.66	116

C. Feature selection

Pada *feature selection* dilakukan seleksi pada *feature* (ciri) yang memiliki nilai korelasi positif terhadap variabel *class* atau target. Pada Gambar 3. terdapat nilai korelasi antar *feature* (ciri). Dari gambar tersebut terlihat bahwa terdapat 3 *feature* yang memiliki nilai negatif yaitu atribut 1, 17, dan 18. Selanjutnya dilakukan pelatihan model dengan menghilangkan ketiga atribut tersebut. Hasilnya ditunjukkan pada Tabel 7, 8, dan 9 dimana terjadi kenaikan akurasi pada data pengujian dibandingkan dengan skenario pemodelan kedua.



Gambar 3. *Feature correlation*

Tabel 7. Performa Model 3 (*feature selection*)

Model	Akurasi *Train* Data	Akurasi *Test* Data 1	Akurasi *Test* Data 2 (10% dari total)
Model 3	76,52%	74,60%	72,41%

Tabel 8. *Confusion Matrix* Model 3 pada Data Validasi

	0	1
0	46	15
1	17	38

Tabel 9. *Precision, Recall, f1-score,* dan *Support* pada Data Validasi pada Model 3

	Precision	*Recall*	*F1-score*	*Support*
0 (Non RD)	0.73	0.75	0.74	61
1 (RD)	0.72	0.69	0.70	55
Avg / total	0.72	0.72	0.72	116

D. Hyper-parameter tuning

Langkah terakhir yang dilakukan adalah dengan mencari parameter terbaik dari algoritma regresi logistik. Pencarian nilai parameter ini dilakukan pada parameter C, iterasi maksimal (*max_iter*), dan penalti. Nilai yang digunakan dalam pencarian nilai parameter terbaik dari setiap parameter ditunjukkan pada Tabel 10. *Hyper-parameter tuning* dilakukan dengan menggunakan metode *Grid Search*. Pada pencarian parameter terbaik dari model regresi logistik ini menggunakan pembagian data dengan metode *cross validation* dengan jumlah *fold* sebanyak 5 *fold*. Pencarian parameter terbaik dengan menggunakan metode *Grid Search* didapatkan hasil bahwa nilai *C, max_iter,* dan penalti terbaik dari masing-masing model memiliki nilai yang berbeda. Pada Tabel 11. terdapat parameter beserta nilainya dari masing-masing skenario model yang memberikan performa yang terbaik dari keseluruhan kombinasi nilai parameter. Nilai parameter pada masing-masing skenario tersebut menghasilkan model dengan performa yang terangkum pada Tabel 12.

Tabel 10. Parameter

C	*Max_iter*	Penalti
Log scale -4 sampai 4 dengan step 20	100, 200, 500, 1000	l1, l2

Tabel 11. Nilai Parameter Terbaik pada Masing-Masing Model

Model	C	*Max_iter*	Penalti
Model 1	11,288	200	l1
Model 2	29,763	100	l1
Model 3	11,288	200	l2

Tabel 12. Tabel Performa dari hasil *hyperparameter tuning*

Model	Akurasi Data Pelatihan	Akurasi Data Pengujian	Akurasi Data Validasi
Model 1	76,10%	74,28%	80,17%
Model 2	75,83%	72,03%	75%
Model 3	75,28%	75,56%	74,14%

IV. Kesimpulan

Berdasarkan hasil pemodelan yang dilakukan dengan menggunakan tiga skenario pemodelan didapatkan hasil yang berbeda dari masing-masing skenario pemodelan. Pada skenario pemodelan pertama menghasilkan rata-rata nilai akurasi pada data pelatihan, data pengujian, dan data validasi yang lebih baik dari skenario pemodelan kedua dan ketiga. Nilai akurasi yang didapatkan dari hasil pengujian pada data pelatihan, pengujian, dan validasi pada skenario pemodelan pertama adalah berturut-turut sebesar 75,96%, 75,24%, dan 78,45%. Setelah menerapkan *hyperparameter tuning* menghasilkan model dengan performa lebih baik. Model yang memiliki performa terbaik adalah model 1 dari skenario pemodelan pertama. Nilai akurasi yang didapatkan dari hasil pengujian pada data pelatihan, pengujian, dan validasi pada model ini adalah berturut-turut sebesar 76,10%, 74,28%, dan 80,17%.

Daftar Pustaka

[1] Y. Yuhelma, Y. Hasneli I, and F. Annis N, "Identifikasi dan Analisis Komplikasi Makrovaskuler dan Mikrovaskuler pada Pasien Diabetes Mellitus," *J. Online Mhs.*, vol. 2, no. 1, pp. 569–579, 2015.

[2] R. Casanova, S. Saldana, E. Y. Chew, R. P. Danis, C. M. Greven, and W. T. Ambrosius, "Application of random forests methods to diabetic retinopathy classification analyses," *PLoS One*, vol. 9, no. 6, pp.

1–8, 2014, doi: 10.1371/journal.pone.0098587.

[3] P. Soewondo, S. Soegondo, K. Suastika, A. Pranoto, D. W. Soeatmadji, and A. Tjokroprawiro, "Medical journal of Indonesia," *Med. J. Indones.*, vol. 19, no. 4, pp. 235–44, 2010, [Online]. Available: http://mji.ui.ac.id/journal/index.php/mji/article/view/412/404.

[4] G. T. Reddy *et al.*, "An ensemble based machine learning model for diabetic retinopathy classification," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–6.

[5] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[6] Y. F. Kao and R. Venkatachalam, "Human and Machine Learning," *Comput. Econ.*, no. February, pp. 1–21, 2018, doi: 10.1007/s10614-018-9803-z.

[7] G. F. Cooper *et al.*, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artif. Intell. Med.*, vol. 9, no. 2, pp. 107–138, 1997, doi: 10.1016/S0933-3657(96)00367-3.

[8] T. Sasakawa, J. Hu, and K. Hirasawa, "A brainlike learning system with supervised, unsupervised, and reinforcement learning," *Electr. Eng. Japan*, vol. 162, no. 1, pp. 32–39, 2008.

[9] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn: machine learning without learning the machinery," *GetMobile Mob. Comput. Commun.*, vol. 19, no. 1, pp. 29–33, 2015, doi: 10.1145/2786984.2786995.

[10] B. Antal and A. Hajdu, "Diabetic Retinopathy Debrecen Data Set," *UCI Mach. Learn. Repos.*, 2014.

[11] S. Mohammadian, A. Karsaz, and Y. M. Roshan, "A comparative analysis of classification algorithms in diabetic retinopathy screening," *Proc. Int. Conf. Softw. Eng. Knowl. Eng. SEKE*, vol. 60, no. 10, p. 631, 2017, doi: 10.18293/SEKE2017-207.

[12] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 5, no. 1, pp. 1–15, 2016, doi: 10.1007/s13721-016-0125-6.

[13] G. King and L. Zeng, "Logistic regression in rare events data," *Polit. Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[14] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain," *J. Clin. Epidemiol.*, vol. 54, no. 10, pp. 979–985, 2001, doi: 10.1016/S0895-4356(01)00372-9.

[15] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.