



BEST FIRST FEATURE SELECTION DAN RADIAL BASIS FUNCTION UNTUK KLASIFIKASI PENYAKIT DIABETES

Nursuci Putri Husain

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Islam Makassar,
Jl. Perintis Kemerdekaan km.9 No. 29 Makassar, Indonesia 90245
Email: nursuciputrihusain.dty@uim-makassar.ac.id

ABSTRAK

Diabetes merupakan salah satu penyebab kematian terbesar di dunia. Penyakit diabetes dapat menyebabkan komplikasi seperti stroke, gagal ginjal, kebutaan yang bisa menyebabkan kecacatan. Sehingga diperlukan model prediksi untuk mengklasifikasikan seseorang mengidap penyakit diabetes atau tidak secara otomatis. Pada penelitian ini diajukan metode kombinasi *Best First Feature Selection* (BF Feature Selection) dan *Radial Basis Function* (RBF) untuk mengklasifikasikan penyakit diabetes. BF Feature Selection sebagai metode untuk mereduksi fitur-fitur pada dataset diabetes. BF Feature Selection mendapatkan 5 fitur yang paling relevan. Selanjutnya, dilakukan klasifikasi terhadap fitur-fitur tersebut menggunakan metode RBF. Uji coba dilakukan dengan menggunakan dataset diabetes dari UCI repository. Pengujian dilakukan dengan membandingkan hasil klasifikasi metode yang diajukan dengan metode yang berbeda pada penelitian sebelumnya. Berdasarkan hasil uji coba, metode yang diajukan mendapatkan nilai akurasi yang tinggi untuk 70-30% training testing partisi yaitu 82,17%. Dari hasil evaluasi tersebut dapat disimpulkan bahwa metode yang diajukan dapat digunakan sebagai metode klasifikasi penyakit diabetes secara otomatis.

Kata kunci: Diabetes, Klasifikasi, Reduksi Fitur, RBF

ABSTRACT

Diabetes is one of the biggest causes of death in the world. Diabetes can cause complications such as stroke, kidney failure, a blindness which can lead to disability. So a prediction model is needed to classify someone as having diabetes or not automatically. In this study, a hybrid method of *Best First Feature Selection* (BF Feature Selection) and *Radial Basis Function* (RBF) is proposed to classify diabetes. BF Feature Selection as a method for reducing features in the diabetes dataset. BF Feature Selection gets the 5 most relevant features. Furthermore, the classification of these features is carried out using the RBF method. Experiment are conducted on diabetes dataset from UCI repository. Experiment was carry out by comparing the result of the proposed method with different methods in previous studies. Based on the trial results, the proposed method has high accuracy for 70-30% training testing partition 82.17%. From the results of the evaluation, it can be concluded that the proposed method can be used as a method of automatic diabetes classification.

Keywords: Diabetes, Classification, Reduction feature, RBF





PENDAHULUAN

Diabetes merupakan salah satu penyakit tidak menular yang paling umum di dunia. Diabetes adalah penyakit yang terjadi ketika gula darah terlalu tinggi. Diabetes adalah penyebab kematian keempat di sebagian besar negara maju. Gejala awal penyakit diabetes yaitu sering buang air kecil, meningkatnya rasa haus, dan lapar. Berdasarkan etiologi, diabetes diklasifikasikan menjadi 3 jenis: diabetes tipe 1, diabetes tipe 2, dan diabetes gestasional (Kharroubi, 2015).

Diabetes tipe 1 terjadi karena adanya kerusakan sel autoimun pada pankreas sehingga menyebabkan hilangnya produksi insulin. Sedangkan, diabetes tipe 2 disebabkan karena adanya resistensi insulin yaitu hilangnya kemampuan tubuh seseorang dalam merespon hormon insulin. Sehingga, hormon insulin tidak dapat membantu penyerapan glukosa dan menyebabkan penumpukan gula dalam darah. Penyakit diabetes dapat menimbulkan komplikasi seperti stroke, gagal ginjal, penyakit arteri coroner, amputasi, dan kebutaan yang bisa mengakibatkan kecacatan. Karena efek diabetes dapat menyebabkan komplikasi bahkan kematian, dibutuhkan model prediksi untuk mengklasifikasikan seseorang mengidap penyakit diabetes untuk mengetahui seseorang mengidap penyakit diabetes atau tidak secara cepat.

Banyak sekali model prediksi yang telah dikembangkan oleh peneliti sebelumnya seperti *decision tree*, *regresi*, *naïve bayes*, *support vector machine*, dan *random forest* (Ogedengbe & Egbunu, 2020). Algoritma klasifikasi *Decision Tree*, *Support Vector Machine* dan *Naïve Bayes* juga diusulkan (Tapak et al., 2013). Penelitian tersebut menggunakan *dataset Diabetes Pima Indians* dari *UCI repository*. Hasil yang didapatkan menunjukkan bahwa performa *Naïve Bayes* mendapatkan nilai akurasi terbaik yaitu 76,30% dibanding metode lain.

Metode yang mengombinasikan tiga algoritma yaitu *Decision Tree*, *Weighted K-Nearest Neighbor* dan *Logistic Regression* juga diusulkan (Wu et al., 2018). Metode yang diusulkan mengadopsi teknik *vote* pada setiap pengklasifikasi untuk mendapatkan hasil yang optimal. Metode yang diusulkan menghasilkan akurasi tertinggi yaitu 80,60%. Kemudian, diusulkan metode *Classifier Subset Evaluator* (CSE) sebagai teknik reduksi fitur untuk memilih atribut diabetes yang

paling relevan dan metode klasifikasi yang digunakan yaitu *Decision Tree* (Ogedengbe & Egbunu, 2020). Metode CSE-DT menghasilkan akurasi 81,64%. Dalam melakukan klasifikasi penyakit diabetes diperlukan teknik reduksi fitur yang lebih baik karena menghasilkan akurasi klasifikasi yang rendah jika beberapa fitur yang tidak relevan disertakan untuk proses klasifikasi.

Penelitian ini menggunakan metode kombinasi antara *Best First Feature Selection* (BF *Feature Selection*) dan *Radial Basis Function* (RBF) untuk mengklasifikasikan seorang mengidap penyakit diabetes atau tidak. BF *Feature Selection* sebagai metode reduksi fitur untuk menyeleksi fitur-fitur yang relevan pada *dataset* diabetes yang digunakan. Sedangkan, RBF sebagai metode klasifikasi setelah fitur direduksi. Kombinasi metode ini menghasilkan akurasi yang lebih tinggi.

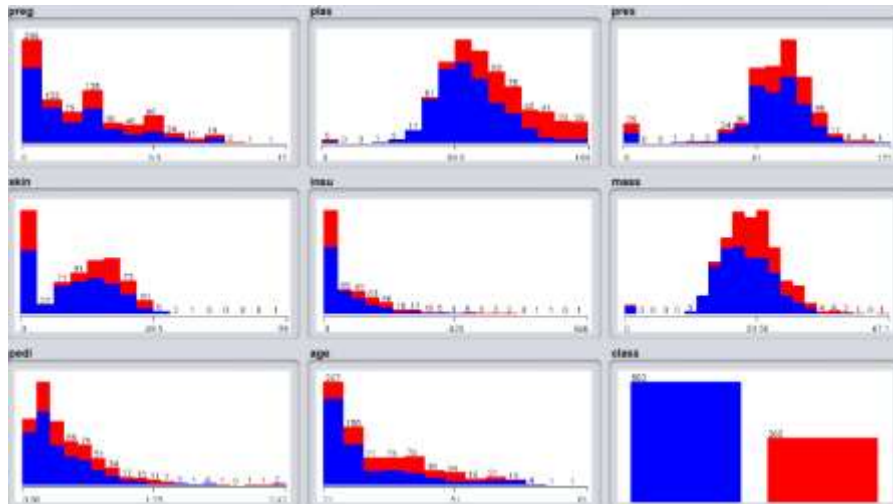
METODE PENELITIAN

Dataset

Dataset diabetes yang digunakan pada penelitian ini adalah *dataset* dari *UCI repository* yang terdiri atas 768 *instance*. Tabel 1 adalah informasi fitur dari setiap *instance*. Hasil tes positif menunjukkan “1” yang berarti pasien mengidap penyakit diabetes dan hasil tes negatif menunjukkan “0” berarti pasien tidak mengidap penyakit diabetes.

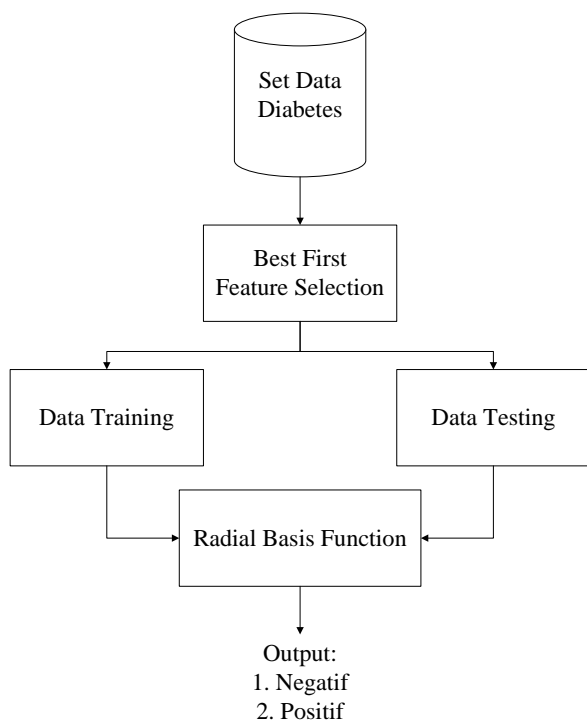
Tabel 1. Fitur pada *dataset* diabetes

No.	Fitur	Singkatan
1.	<i>Body mass index</i>	Bmi
2.	<i>Plasma glucose concentration</i>	Plas
3.	<i>Diastolic blood pressure (mm Hg.)</i>	Pres
4.	<i>Diabetes pedigree function</i>	Pedi
5.	<i>2-Hours serum insulin (mu U/ml)</i>	Insu
6.	<i>Number of times pregnant</i>	Preg
7.	<i>Triceps skinfold thickness (mm)</i>	Skin
8.	<i>Age in years</i>	Age
9.	<i>Class</i>	Class



Gambar 1. Sebaran instance tiap fitur pada dataset

Sebanyak 268 pasien kasus positif diabetes sedangkan 500 kasus pasien negatif diabetes. Dataset diabetes terdiri dari 9 fitur yang merupakan tipe data numerik. Dapat dilihat pada Gambar 1 menunjukkan sebaran instance tiap fitur bervariasi.



Gambar 2. Bagan Alur Metode Penelitian

Best First Feature Selection

Dataset yang memiliki dimensi fitur yang besar dapat dipengaruhi oleh proses klasifikasi. Dimensi fitur dapat direduksi dengan metode reduksi dimensionalitas (Husain et al., 2017). Metode reduksi dimensionalitas dapat dilakukan dengan metode pemilihan fitur. Pemilihan fitur merupakan salah satu metode reduksi dimensi untuk memilih fitur yang memiliki informasi paling relevan dengan data aslinya. Dapat dilihat pada

Gambar 2, penelitian ini menggunakan salah satu metode reduksi fitur yaitu BF Feature Selection.

BFFS memilih n fitur terbaik untuk memodelkan dataset menggunakan algoritma Greedy. Dimulai dengan membuat N model, masing-masing hanya menggunakan satu dari fitur N dari dataset masukan. Pada iterasi berikutnya dibuat satu set model $N-1$ dengan dua fitur input: satu dipilih di iterasi sebelumnya dan satu lagi dari fitur $N-1$ yang tersisa. Kombinasi fitur yang memberikan performa terbaik yang akan dipilih. Algoritma berhenti ketika mencapai jumlah fitur yang ditentukan sebelumnya oleh pengguna (Hall, 2000).

Radial Basis Function

Radial Basis Function (RBF) dapat digunakan untuk berbagai aplikasi karena pelatihannya lebih cepat daripada multi-layer perceptron. Kecepatan belajar yang lebih cepat ini terjadi karena RBF hanya memiliki tiga layer dan setiap layer dapat ditentukan secara berurutan.

RBF memiliki 3 layer jaringan feed-forward yang terdiri dari satu input layer, satu hidden layer, dan satu output layer (Dash et al., 2016).

1. Input layer terdiri dari node sumber m_0 , dimana m_0 adalah dimensi dari input vector \mathbf{x} .
2. Hidden layer terdiri dari jumlah unit komputasi yang sama dengan ukuran dari sampel pelatihan yaitu N . Setiap unit secara matematis dijelaskan oleh fungsi radial basis:
$$\varphi_j(\mathbf{x}) = \varphi(\|\mathbf{x} - \mathbf{x}_j\|), \quad j = 1, 2, \dots, N \quad (1)$$
Dimana φ_j adalah ukuran lebar fungsi gaussian dengan pusat \mathbf{x}_j .
3. Output Layer, output layer terdiri dari satu unit komputasi, tidak ada batasan pada ukuran output layer. (Haykin, 2009)

Evaluasi Uji Coba

Performa hasil prediksi dari model yang diusulkan didapatkan dengan mencari nilai akurasi, precision, recall, dan f -measure.



$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% , \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% , \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% , \quad (4)$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision+Recall} \times 100\% , \quad (5)$$

Formula akurasi, *precision*, *recall*, dan *f-measure* dapat dilihat pada persamaan (2), (3), (4) dan (5). Pada persamaan tersebut, TP (*True Positive*) adalah jumlah data yang teridentifikasi secara benar, TN (*True Negative*) adalah jumlah data yang tertolak secara benar, FP (*False Positive*) adalah jumlah data yang teridentifikasi secara salah, dan FN (*False Negative*) adalah jumlah data yang tertolak secara salah.

HASIL DAN PEMBAHASAN:

Pada penelitian ini, kami mengusulkan metode reduksi fitur BF Feature Selection untuk memilih fitur yang paling relevan. Setelah dilakukan proses reduksi fitur, dataset Diabetes yang sebelumnya terdiri atas 9 fitur, berkurang menjadi 5 fitur. Lima fitur tersebut adalah plas, mass, pedi, age, dan class. Setelah mendapatkan dataset yang telah direduksi fiturnya, dilakukan klasifikasi menggunakan metode Radial Basis Function (RBF).

Untuk menguji unjuk kerja metode yang diusulkan pada penelitian ini, dilakukan perbandingan terhadap metode yang diusulkan dengan penelitian sebelumnya yang menggunakan metode *Classifier Subset Evaluator* sebagai metode reduksi fitur dan *Decision Tree* sebagai metode klasifikasi (Ogedengbe & Egbunu, 2020). *Classifier Subset Evaluator* dan *Decision Tree* dihibridisasi sebagai *Classifier Subset Evaluator Decision Tree* disingkat CSE_DT. *Dateset* dan partisi *training - testing* yang digunakan sama dengan penelitian yang kami usulkan.

Tabel 2. *Confusion Matrix*

Actual	Tested Negative	Tested Positive
Tested Negative	142	16
Tested Positive	25	47

Tabel 2 merupakan confusion matrix hasil klasifikasi model yang diusulkan, terlihat pada tabel tersebut terdapat 142 pasien yang diprediksi benar tidak mengidap penyakit diabetes, dan 16 pasien yang tidak mengidap penyakit diabetes diprediksi mengidap penyakit diabetes. Kemudian, terdapat 47 pasien benar diprediksi mengidap penyakit diabetes, dan 25 pasien yang mengidap penyakit diabetes diprediksi tidak mengidap penyakit diabetes.

Tabel 3. Nilai Akurasi, *Precision*, *Recall* dan *F-Measure*

Metrics	Metode yang diusulkan	Metode CSE-DT (Ogedengbe & Egbunu, 2020)
<i>Accuracy</i>	82,17%	81,6%
<i>Precision</i>	81,8%	81,7%
<i>Recall</i>	82,2%	81,6%
<i>F-measure</i>	81,8%	80,9%

Pada tabel 3, Untuk mengetahui performa model prediksi yang diusulkan maka pada penelitian ini dicari nilai akurasi, *precision*, *recall*, dan *f-measure* seperti yang ditampilkan pada Table 3. Metode yang diusulkan mendapatkan akurasi lebih tinggi untuk 70-30% *training testing* partisi yaitu 82,17% dibanding metode CSE-DT yaitu 81,6%. Kemudian, nilai *precision*, *recall*, dan *f-measure* yang didapatkan dari metode yang diusulkan adalah 81,8%, 82,2%, dan 81,8%. Metode reduksi fitur BF *Feature Selection* berhasil mereduksi fitur-fitur yang relevan, sehingga hasil klasifikasi menggunakan metode RBF mendapatkan hasil yang lebih tinggi dibanding metode CSE-DT.

KESIMPULAN:

Berdasarkan hasil uji coba, metode yang diusulkan mendapatkan akurasi yang tinggi untuk 70-30% *training testing* partisi yaitu 82,17%. Hasil ini lebih baik dibanding metode lainnya yaitu CSE-DT, hal ini disebabkan karena metode reduksi fitur yang digunakan *Best First Feature Selection* (BF *Feature Selection*) berhasil mendapatkan fitur-fitur yang paling relevan sebelum melalui tahap klasifikasi menggunakan metode *Radial Basis Function* (RBF)

Dari hasil evaluasi tersebut, disimpulkan bahwa metode yang diusulkan dapat digunakan sebagai proses klasifikasi sinyal EEG secara otomatis. Penelitian terhadap automasi reduksi fitur untuk *dataset* diabetes dapat menjadi *future work* pada penelitian ini.

DAFTAR PUSTAKA:

- Dash, C. S. K., Behera, A. K., Dehuri, S., & Cho, S. B. (2016). Radial basis function neural networks: A topical state-of-the-art survey. *Open Computer Science*, 6(1), 33–63. <https://doi.org/10.1515/comp-2016-0005>
- Hall, M. A. (2000). Feature Selection for Discrete and Numeric Class Machine Learning 1 Introduction. *Machine Learning Proc Seventeenth International Conference on Machine Learning*, 1–16.
- Husain, N. P., Arisa, N. N., Rahayu, P. N., Arifin, A. Z., & Herumurti, D. (2017). *Least Squares Support Vector Machines Parameter Optimization Based. 1*, 43–49.
- Kharroubi, A. T. (2015). Diabetes mellitus: The epidemic of the century. *World Journal of Diabetes*, 6(6), 850. <https://doi.org/10.4239/wjd.v6.i6.850>
- Ogedengbe, M. T., & Egbunu, C. O. (2020). CSE-DT



Features selection technique for Diabetes classification. *Applications of Modelling and Simulation*, 4(March), 101–109.

- Tapak, L., Mahjub, H., Hamidi, O., & Poorolajal, J. (2013). Real-data comparison of data mining methods in prediction of diabetes in Iran. *Healthcare Informatics Research*, 19(3), 177–185. <https://doi.org/10.4258/hir.2013.19.3.177>
- Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10(August 2017), 100–107. <https://doi.org/10.1016/j.imu.2017.12.006>