

PERBANDINGAN METODE KNN DAN MKNN UNTUK DETEKSI DINI DIABETES MELLITUS

Rizki Mutiara Sari¹, Sutikno^{2*}

^{1,2} Departemen Ilmu Komputer, FSM, Universitas Diponegoro
sutikno@lecturer.undip.ac.id

ABSTRAK

Diabetes Mellitus (DM) adalah kondisi yang berpotensi fatal. Faktor keturunan, kurangnya kesadaran masyarakat akan pola hidup sehat, dan keterlambatan mengenali gejala awal berkontribusi terhadap penyakit DM. Untuk menyiasatinya, deteksi dini DM dibangun menjadi tindakan pencegahan. Terdapat beberapa metode untuk melakukan deteksi DM ini, diantaranya yaitu *K-Nearest Neighbor* (KNN) dan *Modified K-Nearest Neighbor* (MKNN). Keduanya menghasilkan akurasi yang cukup tinggi, namun belum diketahui perbandingan performansi keduanya untuk deteksi DM. Untuk itu pada penelitian ini membandingkan KNN dan MKNN dengan dataset terdiri dari 11 parameter dan 2 kelas diagnosis. Hasil pengujian menunjukkan bahwa metode KNN menghasilkan akurasi lebih tinggi dari MKNN. Metode KNN menghasilkan akurasi terbaik sebesar 0,973333 dan metode MKNN menghasilkan akurasi terbaik sebesar 0,893333.

Keyword : *Diabetes Mellitus, metode klasifikasi, Modified K-Nearest Neighbor, K-Nearest Neighbor.*

1. PENDAHULUAN

Diabetes Mellitus merupakan penyakit tidak menular yang prevalensi dan angka kematiannya terus meningkat [1]. Indonesia menempati urutan ketujuh, menurut data *International Diabetes Federation* [2]. Kurangnya kesadaran akan gejala awal seperti sering buang air kecil, sering haus, dan selalu merasa lapar menjadi penyebab tingginya prevalensi diabetes mellitus. Selain itu, sering muncul tanda-tanda seperti pandangan kabur, kesemutan di tangan atau kaki, gatal (*pruritus*), luka akibat gangguan insulin, luka yang sulit sembuh karena kadar gula darah tinggi, dan berat badan yang tidak ideal [3].

Penderita penyakit ini sering kali tidak menyadari gejala-gejala awal penyakit *Diabetes Mellitus*, sehingga menyebabkan peningkatan jumlah penderita. Selain itu penanganan yang terlambat juga mengakibatkan penyakit ini berkembang menjadi penyakit serius.

Berdasarkan kondisi ini, dibutuhkan sebuah deteksi dini apakah seseorang suspek atau tidak suspek *Diabetes Mellitus*, dengan memasukkan data berupa gejala dan faktor risiko *Diabetes Mellitus* yang dialami. Penelitian tentang masalah ini telah dilakukan diantaranya yaitu dengan menggunakan metode backpropagation neural network (BNN) [4], C4.5 dan Naive bayes [5], KNN [6][7], dan MKNN [8]. Metode KNN dan MKNN menghasilkan akurasi yang cukup baik, namun belum diketahui perbandingan keduanya untuk deteksi *Diabetes Mellitus* pada dataset yang sama. Metode MKNN merupakan pengembangan dari metode KNN [9]. Untuk itu pada penelitian ini membandingkan metode KNN dan MKNN pada dataset yang sama dengan variasi variabel jumlah data latih dan uji, jumlah tetangga, dan nilai *smoothing regulator* yang diterapkan pada deteksi diabetes mellitus.

2. TINJAUAN PUSTAKA

2.1. Penyakit Diabetes Mellitus

Diabetes Mellitus (DM) ditandai dengan tingginya kadar gula darah dan gangguan metabolisme karbohidrat, lipid, dan protein. Diabetes jarang memanifestasikan dirinya dengan gejala. Namun, ada beberapa tanda peringatan bahwa diabetes mungkin ada. Pasien diabetes sering mengalami gejala khas berikut [3]:

- a. *Poliuria* adalah kondisi seseorang yang buang air kecil lebih dari delapan kali per hari. Orang dewasa dan anak-anak rata-rata tiga liter urin per hari. Sebaliknya, orang dewasa normal hanya mengeluarkan sekitar 1-2 liter urin.
- b. *Polydipsia* atau sering haus adalah upaya tubuh untuk menghindari dehidrasi (kekurangan cairan). Karena tubuh mengeluarkan banyak air melalui urine, secara otomatis tubuh akan merasa haus untuk menggantikan cairan yang hilang.
- c. *Poliphagia* (banyak makan/selalu merasa lapar).
- d. Penglihatan kabur.
- e. Kesemutan pada tangan atau kaki.
- f. Timbul luka yang disebabkan oleh gangguan insulin.
- g. Berat badan menurun.
- h. Kadar gula darah ≥ 126 mg/dL pada keadaan puasa, atau ≥ 200 mg/dL pada keadaan 2 jam setelah makan.
- i. Luka diabetes sulit kering dan sulit untuk disembuhkan akibat tingginya kadar gula darah.

Selain gejala tersebut, terdapat faktor-faktor yang dapat mempengaruhi *Diabetes Mellitus* [10]:

- a. Untuk penderita muda dan tua, keturunan mungkin memainkan peran lebih besar pada mereka yang berusia di bawah 40 tahun. Lebih dari separuh pasien dewasa memiliki diabetes dalam keluarganya, menunjukkan bahwa kondisi

tersebut cenderung diturunkan daripada diwariskan.

- b. Usia diatas 45 tahun karena banyak terjadi perubahan pada usia tersebut, terutama pada pankreas, organ yang membuat insulin. Ini mengurangi produksi dan fungsi insulin di pankreas seiring bertambahnya usia.

2.2. K-Nearest Neighbor

Metode *K-Nearest Neighbor* (KNN) digunakan untuk mengklasifikasikan objek baru dibandingkan dengan objek yang sudah ada [11]. Jarak antara data ini dengan data lainnya dapat ditentukan dengan menggunakan metode KNN. Nilai jarak digunakan untuk menentukan seberapa mirip atau dekat data uji dan data latih. Jumlah data yang paling dekat dengan data uji adalah nilai k dari metode KNN. Kelas dari data uji adalah kelas dari tetangga terdekat jika k adalah 1. Dua tetangga terdekat akan dipilih dari data pelatihan jika k adalah 2. Jika ada dua kelas berbeda pada k tetangga, maka kelas dari data yang diuji akan dipilih berdasarkan data terbanyak (*voting mayoritas*) [11].

Langkah- pada metode KNN yaitu sebagai berikut [12]:

- a. Menentukan nilai k
- b. Menghitung nilai jarak antara data latih dengan data uji
- c. Mengurutkan data latih berdasarkan jarak terkecil dengan data uji
- d. Memilih nilai yang sering muncul dari tetangga terdekat

2.3. Perhitungan Jarak antar Data

Dalam matematika, jarak adalah sebuah aplikasi yang merumuskan panjang antara dua titik. Jarak akan memungkinkan untuk mengelompokkan individu yang serupa dan memisahkan yang tidak menyerupai [13]. Jarak kedua titik dihitung dengan persamaan (1).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (x_{ir} - x_{jr})^2} \quad (1)$$

dimana :

- x_{ir} : data uji ke- i pada variabel ke-1
 x_{jr} : data latih ke- i pada variabel ke-1
 $d(x_i, x_j)$: jarak
 n : dimensi data

2.4. Modified K-Nearest Neighbor

Keabsahan perhitungan semua data pada data pelatihan merupakan konsep sentral dari metode MKNN. Selain itu, validasi nilai digunakan sebagai faktor pertambahan saat menghitung bobot suara pada semua data pelatihan. Metode MKNN menggunakan bobot kelas tertinggi pada k data latih yang telah divalidasi menggunakan jarak terpendek untuk mengklasifikasikan data uji [9].

Langkah-langkah metode MKNN yaitu sebagai berikut [9]:

- a. Menghitung jarak antar data latih
- b. Mengurutkan jarak antar data latih berdasarkan jarak terkecil
- c. Menghitung nilai validasi setiap data latih
- d. Menghitung jarak antara data uji dengan data latih
- e. Menghitung bobot setiap data latih
- f. Menentukan kelas dari data uji berdasarkan bobot terbesar

2.5. Validitas Data

Semua data pada data *training* terlebih dahulu harus melalui proses validasi pada metode MKNN dengan menggunakan persamaan (2) [9].

$$validity(x) = \frac{1}{H} \sum_{i=1}^H S(lbl(x), lbl(N_i(x))) \quad (2)$$

Keterangan :

- H : jumlah tetangga
 $lbl(x)$: label data latih
 $N_i(x)$: label tetangga terdekat

Persamaan (3) merupakan persamaan dari fungsi S :

$$S(a, b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \quad (3)$$

dengan a adalah kelas a pada *dataset* dan b adalah kelas selain a pada *dataset*.

2.6. Pembobotan

Metode MKNN mencakup tahapan yang disebut dengan pembobotan data latih. Alih-alih memilih kelas mayoritas di k terdekat, variasi KNN ini membeli kelas dari objek data baru dengan cara memilih berdasarkan bobot data pelatihan. Perhitungan akan digunakan untuk menetapkan bobot pada setiap data pelatihan. Bobot setiap tetangga dilakukan dengan persamaan (4) [9].

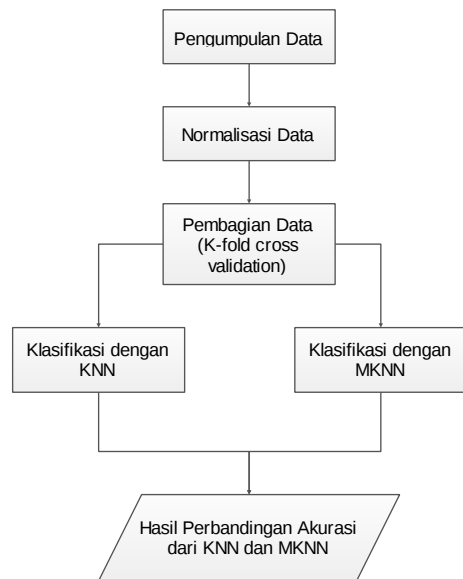
$$W(i) = validity(i) \times \frac{1}{d_e + \alpha} \quad (4)$$

Keterangan :

- $W(i)$: *Weight voting* data latih ke- i
 $validity(i)$: nilai validitas data latih ke- i
 d_e : jarak *Euclidean* antara data latih dengan data uji
 α : konstanta nilai *smoothing regulator* (pemulusan)

3. METODE PENELITIAN

Secara garis penelitian ini dibagi menjadi 5 proses yaitu pengumpulan data, normalisasi data, pembagian data, klasifikasi dengan KNN, dan klasifikasi dengan MKNN, seperti pada Gambar 1.



Gambar 1. Langkah-langkah penelitian

3.1. Pengumpulan Data

Dataset diambil dari data rekam medis pasien pada Poli Penyakit Dalam RSUD. Dr. H. Soewondo Kendal pada periode 25 Januari 2016 sampai 25 Februari 2016 sebanyak 150 data dengan 12 variabel yaitu usia, keturunan, poliuria, polidipsia, poliphagia, kgd, turun berat badan, kesemutan, lelah dan kantuk, rabun mata, luka, dan infeksi [4]. Variabel yang digunakan berjumlah 11 variabel dengan menghilangkan variabel infeksi. Reduksi variabel infeksi dikarenakan gejala infeksi sudah termasuk dalam gejala luka yang sulit sembuh [3]. Selain itu, percobaan dengan menggunakan 12 variabel sudah pernah dilakukan sebelumnya, dari percobaan tersebut dihasilkan akurasi yang lebih rendah dibandingkan dengan menggunakan 11 variabel.

3.2. Normalisasi Data

Normalisasi digunakan untuk mempersempit range data rekam medis pasien. Dalam hal ini gejala dan faktor risiko *Diabetes Mellitus* diubah ke dalam variabel dan kategori dari masing-masing gejala dan faktor risiko tersebut diubah ke bentuk numerik agar dapat diolah sebagai nilai input. Variabel x_1 dan x_6 mendefinisikan usia dan KGD dengan nilai input dihitung dengan menggunakan Persamaan (5).

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A \quad (5)$$

Keterangan :

- v' : data hasil normalisasi
- v : data sebelum dinormalisasi
- new_max_A : batas nilai max baru, dalam hal ini bernilai 1
- new_min_A : batas nilai min baru, dalam hal ini bernilai 0
- $\max_A$: nilai maksimum
- $\min_A$: nilai minimum

Variabel-variabel input dan target lainnya disajikan pada Tabel 1.

Tabel 1. Tabel Pendefinisian Input dan Target

Variabel Gejala	Kategori	Nilai Input
Keturunan (x2)	Ada	1
	Tidak Ada	0
Polyuria (x3)	Sering	1
	Jarang	0,5
	Tidak pernah	0
Polydipsia (x4)	Sering	1
	Jarang	0,5
	Tidak pernah	0
Polyphagia (x5)	Sering	1
	Jarang	0,5
	Tidak pernah	0
Penurunan BB (x7)	Ya	1
	Tidak	0
Kesemutan (x8)	Ya	1
	Tidak	0
Lelah dan kantuk (x9)	Ya	1
	Tidak	0
Rabun mata (x10)	Ya	1
	Tidak	0
Luka (x11)	Ya	1
	Tidak	0
Target (t)	Ya	1
	Tidak	0

3.3. Pembagian Data

Proses pelatihan memanfaatkan data pelatihan (*training set*), sedangkan pengujian akurasi dilakukan dengan data pengujian (*test set*). Teknik *K-Fold Cross Validation* digunakan untuk membagi semua data menjadi data latih dan data uji. Dataset dibagi menjadi k partisi dengan ukuran yang sama, sehingga terdapat k ulangan proses *testing* dan *training*. Data yang tersisa pada iterasi i, Di akan menjadi data latih [14].

3.4. Klasifikasi KNN dan MKNN

Proses klasifikasi KNN dan MKNN dilakukan dengan menggunakan algoritma keduanya. Proses klasifikasi KNN dan MKNN akan menghasilkan akurasi, dan kemudian dilakukan perbandingan. Proses pengujian klasifikasi dilakukan dengan kombinasi 3 variabel yaitu jumlah data latih dan uji, jumlah tetangga, dan *smoothing regulator*.

4. HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan 3 tahap yaitu sebagai berikut:

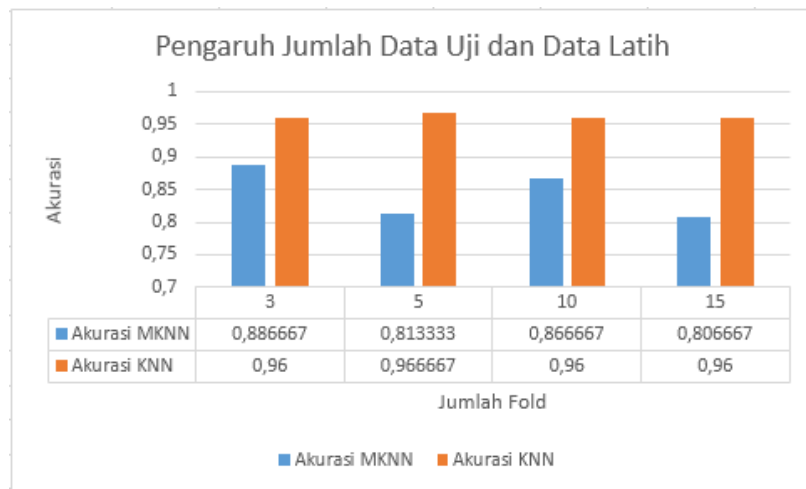
4.1. Hasil Eksperimen 1: variasi jumlah data latih dan uji

Variabel yang diuji yaitu jumlah *fold* (k). Nilai *fold* yang diteliti yaitu 3, 5, 10, dan 15. Sementara

kombinasi variabel lainnya yaitu jumlah tetangga (H) = 5, dan *smoothing regulator* (α) = 0,5.

Hasil eksperimen 1 menunjukkan pengaruh jumlah data latih dan data uji yang digunakan

dengan pembagian *dataset* menggunakan *k-fold*. Eksperimen dilakukan dengan mencoba masing-masing nilai *fold*, seperti pada Gambar 2.



Gambar 2. Hasil Eksperimen 1

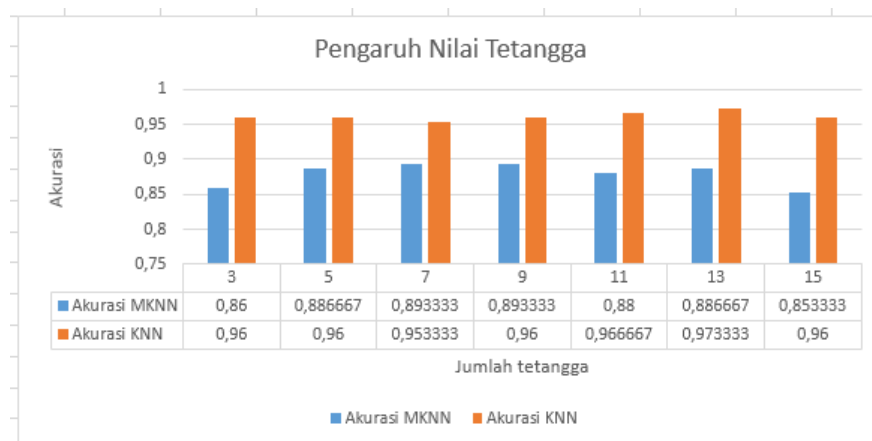
Dari Gambar 2 diketahui bahwa akurasi terbaik untuk metode MKNN sebesar 0,886667 terdapat pada jumlah *fold* = 3 dengan jumlah data uji 50 data dan data latih 100 data. Sedangkan untuk metode KNN akurasi terbaik sebesar 0,966667 pada jumlah *fold* = 5 dengan data uji sejumlah 30 data dan data latih 120 data.

Pada eksperimen 1 hasil menunjukkan bahwa jumlah data latih dan data uji memberi pengaruh terhadap hasil namun tidak besar. *Dataset* yang digunakan memiliki rasio jumlah kelas data pada *dataset* hampir sama pada setiap *fold*. Dapat

diketahui bahwa selisih akurasi antar *fold* pada keempat *fold* tidak terlalu tajam.

4.2. Hasil Eksperimen 2: variasi nilai jumlah tetangga

Hasil eksperimen 2 dilakukan untuk mengetahui pengaruh nilai jumlah tetangga (H). Eksperimen dilakukan dengan mencoba masing-masing nilai jumlah tetangga (H) yang telah disebutkan pada skenario eksperimen 2. Nilai *fold* yang digunakan merupakan nilai *fold* terbaik pada eksperimen 1. Hasil eksperimen 2 ditunjukkan pada Gambar 3.



Gambar 33 Hasil Eksperimen 2

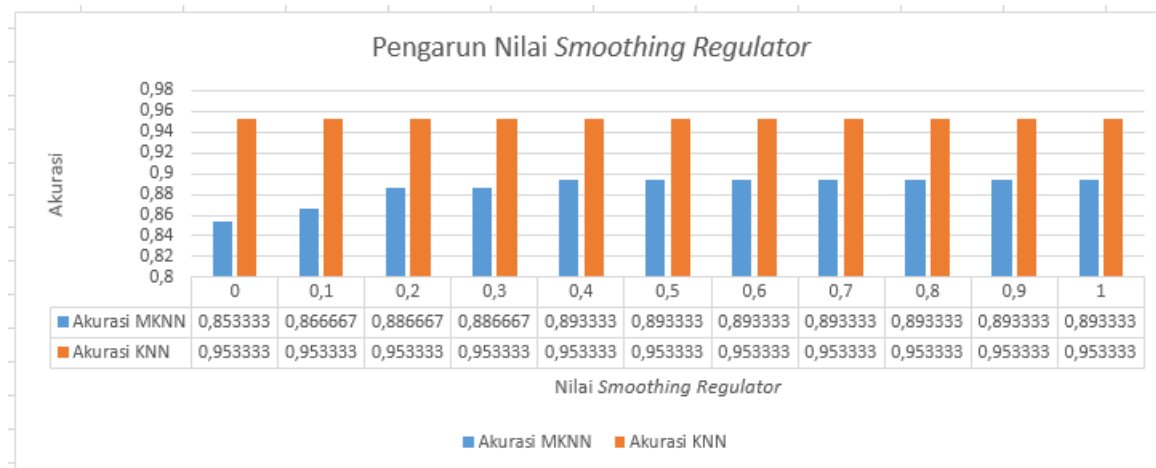
Dari Gambar 4.18 diketahui bahwa akurasi terbaik untuk metode MKNN sebesar 0,893333 terdapat pada jumlah tetangga = 7. Sedangkan untuk metode KNN akurasi terbaik sebesar 0,973333 pada jumlah tetangga = 13.

Hasil eksperimen 2 akurasi metode MKNN bagus ketika jumlah tetangga yang digunakan tidak terlalu banyak. Sedangkan metode KNN, akurasi besar dihasilkan ketika menggunakan jumlah tetangga yang cenderung banyak.

4.3. Hasil Eksperimen 3: variasi nilai *smoothing regulator*

Hasil eksperimen 3 dilakukan untuk mengetahui pengaruh nilai *smoothing regulator* (α). Eksperimen dilakukan dengan mencoba masing-masing nilai *smoothing regulator* (α) yang telah

disebutkan pada skenario eksperimen 3. Nilai *fold* yang digunakan merupakan nilai *fold* terbaik pada eksperimen 1, serta nilai tetangga (H) yang digunakan merupakan nilai terbaik hasil eksperimen 2. Hasil eksperimen 3 ditunjukkan pada Gambar 4.



Gambar 4. Hasil Eksperimen 3

Dari Gambar 4 diketahui bahwa akurasi terbaik untuk metode MKNN sebesar 0,893333 dengan nilai $\alpha = 0,4$ hingga $\alpha = 1$. Sedangkan untuk metode KNN akurasi terbaik sebesar 0,953333.

Pada metode MKNN, nilai *smoothing regulator* (α) yang tidak terlalu besar akan menghasilkan akurasi yang baik. Apabila *smoothing regulator* (α) yang digunakan semakin mendekati 1, dapat mengakibatkan akurasi menurun. Sedangkan pada metode KNN nilai *smoothing regulator* (α) tidak berpengaruh, karena proses perhitungan pada metode ini tidak melibatkan nilai *smoothing regulator* (α).

Dari ketiga eksperimen rata-rata akurasi MKNN lebih rendah daripada rata-rata akurasi KNN, yaitu akurasi MKNN sebesar 0,893333 sedangkan akurasi KNN sebesar 0,953333. Hal ini disebabkan karena jumlah antara kedua kelas pada *dataset* yang digunakan tidak sama. Jumlah data dengan kelas suspek sebanyak 79 data sedangkan jumlah data dengan kelas tidak suspek sebanyak 71 data. Sedangkan proses pengklasifikasian metode MKNN dengan cara memilih jumlah bobot terbesar dari salah satu kelas. Apabila jumlah data antar kelas data tidak sama, berpengaruh terhadap jumlah bobot akhir yang digunakan sebagai hasil klasifikasi. Berbeda dengan proses pengklasifikasian metode KNN hanya dengan *voting* mayoritas kelas dari tetangga terdekat.

5. KESIMPULAN DAN SARAN

Strategi pengujian KNN dan MKNN untuk deteksi *diabetes mellitus* digunakan dalam penelitian ini. Akurasi metode KNN lebih tinggi daripada metode MKNN. Metode MKNN menghasilkan tingkat akurasi sebesar 0,893333, sedangkan metode

KNN memiliki akurasi tertinggi sebesar 0,973333. Pada metode MKNN, nilai *smoothing regulator* 0,4 sampai dengan 1 menghasilkan akurasi terbaik. Sedangkan pada metode KNN nilai *smoothing regulator* tidak berpengaruh. Algoritma genetika yang secara otomatis menentukan jumlah tetangga yang optimal (H) di MKNN dapat ditambahkan pada penelitian selanjutnya.

DAFTAR PUSTAKA

- [1] S. Soegondo and K. Sukardji, *Hidup secara Mandiri dengan Diabetes Mellitus*. Jakarta: Universitas Indonesia, 2008.
- [2] International Diabetes Federation, "IDF Diabetes ATLAS," 2015.
- [3] Departemen Kesehatan RI, *Pharmaceutical Care Untuk Penyakit Diabetes Mellitus*. Departemen Kesehatan RI, 2005.
- [4] R. Sofiana and Sutikno, "Optimization of Backpropagation for Early Detection of Diabetes Mellitus," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 3232–3237, 2018.
- [5] M. Rifqi, "Aplikasi Data Mining untuk Diagnosis Penyakit Diabetes Menggunakan Algoritma C4.5 dan Naive Bayes Classification," Semarang, 2016.
- [6] F. Yunita, "SISTEM KLASIFIKASI PENYAKIT DIABETES MELLITUS MENGGUNAKAN METODE K-NEAREST NEIGHBOR (K-NN)," *Seloyang Mayang*, vol. 2, no. 1, pp. 223–230, 2016.
- [7] M. F. N. Rilwanu, H. Taufikurachman, and F. Huwaidi, "Penerapan Algoritma K-Nearest Neighbor untuk Mendeteksi Diabetes Berbasis Web Application," *J. Softw. Eng. Inf. Commun. Technol.*, vol. 3, no. 1, pp. 145–152,

- 2022.
- [8] S. I. Fernanda, D. E. Ratnawati, and P. P. Adikara, "Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified KNearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 6, pp. 507–513, 2017.
 - [9] H. Parvin, H. Alizadeh, and B. Minaei-Bidgoli, "MKNN: Modified K-Nearest Neighbor," 2008.
 - [10] S. A. Soelistijo *et al.*, *Konsesus Pengelolaan dan Pencegahan Diabetes Mellitus Tipe 2 di Indonesia*. PB. Perkeni, 2015.
 - [11] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. CV Andi Offset, 2012.
 - [12] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization Diabetic Patients," *Int. J. Eng. Adv. Technol.*, vol. 1, no. 1, pp. 147–151, 2012.
 - [13] S. A. Medjahed, T. Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," *Int. J. Comput. Appl.*, vol. 62, no. 1, pp. 1–5, 2013.
 - [14] J. H. Han, M. Kamber, and J. Pei, *Data Mining: Concept and Techniques*. Elsevier Inc., 2012.