Implementasi Metode *K-Nearest Neighbor* untuk Diagnosis Kanker Kolorektal dengan *Biomarker Micro-RNA*

Muhammad Sofi Yuniarto[*1)], Eko Adi Sarwoko[*2)]

[**]Jurusan Ilmu Komputer/ Informatika, Fakultas Sains dan Matematika,
Universitas Diponegoro
[1)]msyuniarto@gmail.com, [2)]eko.adi.sarwoko@gmail.com

Abstrak

Kanker Kolorektal merupakan keganasan yang berasal dari jaringan usus besar, terdiri dari kolon dan/ atau rektum. Salah satu metode skrining kanker kolorektal yaitu micro-RNA. Micro-RNA merupakan nukleotida yang berukuran pendek (sekitar 18-25 basa nukleotida) yang berperan dalam berbagai proses intraseluler dengan mengatur ekspresi gen. Micro-RNA yang mempengaruhi kanker kolorektal adalah miR-21, miR-31, miR-135b, miR-183, miR-222, miR-145, dan miR-195. Metode K-Nearest Neighbor digunakan untuk melakukan klasifikasi data micro-RNA. Dataset yang digunakan berjumlah 600 data, terdiri dari 300 data normal, dan 300 data kanker kolorektal. Dataset terdiri dari 7 ekspresi micro-RNA beserta label datanya, kanker kolorektal atau normal. Pembagian dataset menjadi data latih dan data uji menggunakan metode K-Fold Cross Validation. Berdasarkan pengujian yang dilakukan dengan menggunakan 10-Fold Cross Validation, metode K-Nearest Neighbor menghasilkan accuracy terbaik pada K=3 dengan accuracy 94,17%, specificity 94,43%, dan sensitivity 94,41%.

Kata kunci : *Kanker Kolorektal, Micro-RNA, K-Nearest Neighbor, K-Fold Cross Validation*

Abstract

Colorectal cancer is malignancy that originates from the large intestinal tissue, consisting of the colon and/ or rectum. One of the colorectal cancer screening methods is micro-RNA. Micro-RNA is a short-sized nucleotide (about 18-25 nucleotide bases) that plays a role in various intracellular processes by regulating gene expression. Micro-RNA that affects colorectal cancer is miR-21, miR-31, miR-135b, miR-183, miR-222, miR-145, and miR-195. K-Nearest Neighbor is used to classify micro-RNA data. The dataset used is 600 data, consisting of 300 normal data, and 300 colorectal data. The dataset consists of 7 micro-RNA expressions and labelled data, colorectal or normal. Split of dataset into training data and test data using K-Fold Cross Validation method. Based on the tests performed using 10-Fold Cross Validation, K-Nearest Neighbor method produce the best accuracy at K=3 with an accuracy of 94.17%, specificity 94.43%, and sensitivity 94.41%.

Keywords : *Colorectal Cancer, Micro-RNA, K-Nearest Neighbor, K-Fold Cross Validation*

1 PENDAHULUAN

Kanker kolorektal (KKR) adalah keganasan yang berasal dari jaringan usus besar, terdiri dari kolon (bagian terpanjang dari usus besar) dan/atau rektum (bagian kecil terakhir dari usus besar sebelum anus) (Kemenkes RI, 2017). Kanker kolorektal terjadi akibat perkembangan sel yang tidak terkendali pada jaringan usus besar dan merupakan jenis kanker terbanyak ketiga (10,2%) serta penyebab kematian terbanyak kedua (9,2%) dari jumlah seluruh penderita kanker di dunia (WHO, 2018). Di Indonesia, kanker kolorektal menempati posisi kedua terbanyak pada pria. Sedangkan pada wanita, kanker kolorektal menempati urutan ketiga (Yayasan Kanker Indonesia, 2018).

Beberapa penelitian telah menunjukkan bahwa disregulasi ekspresi *micro-RNA* berperan penting dalam proses onkogenesis kolorektal. *Micro-RNA* berperan pada semua aspek perkembangan kolorektal, baik saat inisiasi, perkembangan, sampai ke progresivitas dan metastasis. Bukti penelitian menunjukkan peran *micro-RNA* pada setiap stadium kolorektal. Selain itu *micro-RNA* juga dilaporkan dapat digunakan sebagai marker untuk diagnosis, prognosis, maupun terapi target pada kanker kolorektal (Anwar et al., 2017).

Pada 2015, Lan et al. berhasil mengklasifikasi kanker dengan menggunakan *micro-RNA* (Lan et al., 2015). Pada penelitian lain yang dilakukan oleh Rosenfeld et al. menemukan bahwa *micro-RNA* efektif sebagai *biomarker* untuk melacak jaringan asal kanker dengan mendapatkan nilai *accuracy* di atas 90% (Rosenfeld, 2008). Sedangkan pada 2006, Bishop meneliti data *fluorescence* yang merupakan hasil pengukuran ekspresi *micro-RNA* dapat digunakan untuk pengenalan pola (Bishop, 2006).

K-Nearest Neighbor merupakan salah satu metode yang digunakan untuk melakukan klasifikasi terhadap objek baru berdasarkan sejumlah k tetangga terdekat. Algoritma *K-Nearest Neighbor* relatif sederhana dan mudah dipahami (Primartha, 2018).

Beberapa penelitian menunjukkan bahwa metode *K-Nearest Neighbor* dapat digunakan untuk diagnosis suatu penyakit dan memiliki nilai akurasi yang baik. Pada penelitian dengan menerapkan metode *K-Nearest Neighbor* untuk diagnosis penyakit sapi potong berhasil mendapatkan nilai akurasi sebesar 97,56% (Tyas et al., 2015). Pada penelitian lain, Medjahed menggunakan metode *K-Nearest Neighbor* untuk mendiagnosis *Breast Cancer* dan mendapat hasil akurasi sebesar 98,7% (Medjahed et al., 2013).

Berdasarkan beberapa uraian yang telah dijelaskan, peneliti menerapkan metode *K-Nearest Neighbor* untuk mendiagnosis kanker kolorektal dengan label output terkena kanker kolorektal atau normal menggunakan *biomarker micro-RNA*.

Tujuan dari penelitian adalah untuk mengetahui kinerja metode *K-Nearest Neighbor* dan implementasi metode *K-Nearest Neighbor* pada sebuah aplikasi untuk diagnosis kanker kolorektal dengan *biomarker micro-RNA*.

2 LANDASAN TEORI

2.1 KANKER KOLOREKTAL

Pada kebanyakan pasien kanker kolorektal, perkembangan dari mukosa kolon normal menjadi kanker invasif membutuhkan beberapa perubahan

molekular. Interval waktu yang diperkirakan dari transformasi dari mukosa normal melalui polip adenomatosa menjadi karsinoma invasif adalah 5-10 tahun (Vogelstein, 1988). Interval waktu yang panjang ini dapat digunakan untuk deteksi dini dan bahkan pencegahan kanker kolorektal seperti yang ditunjukkan oleh National Studi Polip. Penelitan tersebut menunjukkan bahwa kanker kolorektal sebagian besar muncul dari adenoma. Selain itu, penelitian tersebut juga menunjukkan bahwa skrining berbasis kolonoskopi dapat mengurangi kejadian dan diagnosis stadium kanker kolorektal dibandingkan dengan populasi yang tidak diskrining (Winawer, 1993). Skrining untuk kanker kolorektal di antaranya *Fecal Occult Blood Testing* (FOBT), prosedur endoskopis, skrining Fecal DNA dan RNA, uji serum darah, dan *micro-RNA* (Mazeh et al., 2013).

2.2 *MICRO-RNA*

Micro-RNA (*miR*) adalah molekul nukleotida yang berukuran pendek (sekitar 18-25 basa nukleotida) berperan dalam berbagai proses intraseluler dengan mengatur ekspresi gen. *Micro-RNA* diketahui berperan penting dalam proses perkembangan embrio, diferensiasi sel, poliferasi sel, apoptosis, perubahan struktur kromosom, resistensi terhadap virus, dan onkogenesis. Komponen *micro-RNA* yang berperan dalam embriogenesis biasanya juga berperan penting dalam proses inisiasi dan progresi terjadinya tumor. Pada kanker, *micro-RNA* dapat digunakan untuk membedakan jaringan kanker dengan yang sehat dan dapat membedakan berbagai subtipe kanker sehingga kemungkinan dapat memberikan manfaat dalam diagnosis, menentukan prognosis, dan memperkirakan responsivitas terhadap terapi (Calin & Croce, 2006; Iorio & Croce, 2012).

2.2.1 PENGAMBILAN SAMPEL DAN METODE EKSTRAKSI *MICRO-RNA*

Micro-RNA dengan kualitas tinggi dapat diekstrasi dari berbagai sumber seperti *sel line*, jaringan segar, parafin blok, plasma, serum, urin, dan cairan tubuh yang lain. *Micro-RNA* dapat dianalisis dari sampel yang diekstrasi dengan metode tradisional yang digunakan untuk mendapatkan RNA total, misalnya menggunakan Trizol atau Qiazol. Berbagai produk yang tersedia secara komersial khusus untuk ekstrasi *micro-RNA* pada umumnya menggunakan garam yang bersifat chaotropik seperti guanidinium thiocyanate (seperti Trizol) yang diikuti dengan ekstrasi menggunakan kolom silica (Accerbi et al., 2013; Liu & Xu, 2011).

2.2.2 DISREGULASI *MICRO-RNA* PADA KANKER KOLOREKTAL

Beberapa penelitian telah menunjukkan adanya perubahan pola ekspresi *micro-RNA* pada kanker kolorektal. Pada kanker kolorektal, sebagian besar *micro-RNA* mengalami peningkatan ekspresi (Luo et al., 2011). Hal ini menunjukkan bahwa *micro-RNA* pada kanker kolorektal lebih berperan sebagai onkogenik, berbeda dengan jenis kanker lainnya yang menunjukkan adanya penurunan ekspresi beberapa *micro-RNA* (Luo et al., 2011; Michael et al., 2003).

Salah satu *micro-RNA* yang paling banyak diteliti tentang kaitannya terhadap kanker kolorektal adalah miR-21 yang menunjukkan adanya peningkatan ekspresi. MiR-21 berperan dalam proses inisiasi, progresi, maupun metastasis kanker kolorektal (Calin & Croce, 2006; Luo et al., 2011). Beberapa *micro-RNA* lainnya yang mengalami peningkatan pada kanker kolorektal adalah miR-31, miR-135b, miR-183, miR-222. Sementara itu, miR-145, dan

miR-195 mengalami penurunan ekspresi pada jaringan kanker kolorektal (Mazeh et al., 2013).

2.3 NORMALISASI DATA

Metode yang digunakan untuk melakukan normalisasi data pada penelitian adalah Normalisasi Min-max. Metode normalisasi min-max merupakan metode normalisasi dengan melakukan transformasi linear terhadap data asli. Rumus penghitungan normalisasi min-max dapat dilihat pada persamaan 1.

$$X_n = \frac{X_0 - X_{min}}{X_{max} - X_{min}} \quad \text{.....} \quad (1)$$

Keterangan :

X_n : nilai baru untuk variabel X

X_0 : nilai lama untuk variabel X

X_{min}: nilai minimum dari suatu fitur

X_{max}: nilai maksimum dari suatu fitur

Keuntungan dari metode ini adalah Keseimbangan nilai perbandingan antardata saat sebelum dan sesudah proses normalisasi, serta tidak ada bias yang dihasilkan oleh metode ini. Sedangkan kekurangan dari metode ini adalah ketika ada data baru, metode ini memungkinkan terjebak "*out of the bound*" error (Mustaffa & Yusof, 2011).

2.4 METODE *K-NEAREST NEIGHBOR*

Algoritma *K-Nearest Neighbor* merupakan algoritma yang digunakan untuk melakukan klasifikasi terhadap objek baru berdasarkan sejumlah K tetangga terdekatnya. Algoritma KNN dapat digolongkan sebagai *supervised learning*, *lazy learning algorithm*, dan *instance-based learning* atau *memory-based learning* (Primartha, 2018).

Algoritma KNN bergantung pada kedekatan antara data latih dengan data uji. Untuk menghitung jarak antara data uji dengan data latih dapat dihitung dengan rumus jarak *Euclidean*, seperti pada persamaan 2.

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

Keterangan:

x_{ir} : data uji ke-i pada variabel ke-r

x_{jr} : data latih ke-j pada variabel ke-r

$d(x_i,x_j)$: jarak data uji i dengan data latih j

n : dimensi data variabel bebas

2.5 *K-FOLD CROSS VALIDATION*

K-Fold Cross Validation merupakan salah satu metode yang digunakan untuk mengevaluasi kinerja *classifier*. Metode *K-Fold Cross Validation* dilakukan dengan membagi *dataset* secara acak menjadi k himpunan bagian (subset). Metode *K-Fold Cross Validation* berguna untuk memvalidasi keakuratan sebuah prediksi atau klasifikasi terhadap suatu data yang belum muncul dalam *dataset* (Kohavi, 1995).

2.6 *EVALUASI SISTEM*

Evaluasi sistem yang digunakan pada penelitian ini adalah *Confusion Matrix*. *Confusion Matrix* merupakan metode yang digunakan untuk menilai sebuah metode klasifikasi dalam mengenali tuple dari kelas target yang berbeda. *Confusion Matrix* melakukan penghitungan berdasarkan *predicted class* dan *actual class* seperti pada Tabel 1. Berdasarkan penelitian (Zhang et al., 2008), evaluasi sistem pada bidang kesehatan digunakan tiga keluaran yaitu *accuracy*, *sensitivity*, dan *specificity*.

Tabel 1 *Confusion Matrix*

Actual class	*Predicted class*	
	C1	C2
C1	TP	FN
C2	FP	TN

Rumus penghitungan *accuracy*, *sensitivity*, dan *specificity* dapat dilihat pada persamaan 3, 4, dan 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (2)}$$

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{..... (3)}$$

$$Specificity = \frac{TN}{TN+FP} \quad \text{..... (4)}$$

Keterangan:
TP : data positif diprediksi positif
TN : data negatif diprediksi negatif
FP : data negatif diprediksi positif
FN : data positif diprediksi negative

2.7 KURVA *RECEIVER OPERATING CHARACTERISTIC* (ROC)

Kurva ROC merupakan salah satu teknik untuk memvisualisasikan akurasi model dan membandingkan perbedaan antarmodel klasifikasi. Kurva ROC menggambarkan grafik dua dimensi dengan *false prositive rate* sebagai garis horizontal dan *true positive rate* sebagai garis vertikal untuk mengukur perbedaan performasi metode. Model klasifikasi yang lebih baik adalah yang mempunyai kurva ROC lebih besar (Vercellis, 2009).

False Prositive Rate (FPR) dihitung menggunakan persamaan 6.

$$FPR = \frac{FP}{FP+TN} \quad \text{..... (5)}$$

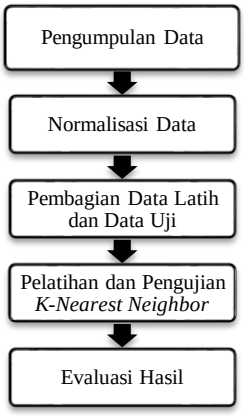
True Positive Rate (TPR) dihitung menggunakan persamaan 7.

$$TPR = \frac{TP}{TP+FN} \quad \text{..... (6)}$$

3 METODOLOGI PENELITIAN

3.1 GARIS BESAR PENYELESAIAN MASALAH

Garis besar penyelesaian masalah dapat dilihat pada Gambar 1.



Gambar 1 Diagram Garis Besar Penyelesaian Masalah

3.2.1 PENGUMPULAN DATA

Dataset yang digunakan merupakan data *micro-RNA* untuk deteksi kanker. Data *micro-RNA* didapatkan dari *National Cancer Institute Genomic Data Commons* dan dapat diakses pada http://portal.gdc.cancer.gov.

Data yang diambil merupakan data dengan label *colorectal* dan normal berjumlah 600 data. Data dengan label *colorectal* (kanker kolorektal) sebanyak 300 data, dan data dengan label normal sebanyak 300 data. Sehingga proporsi data yang digunakan adalah 1:1.

Berdasarkan penelitian (Calin & Croce, 2006; Luo et al., 2011; Mazeh et al., 2013), *micro-RNA* yang berkaitan dengan kanker kolorektal adalah hsa-miR-21, hsa-miR-31, hsa-miR-135b, hsa-miR-183, hsa-miR-222, hsa-miR-145, dan hsa-miR-195. Keterangan ekspresi dan label dapat dilihat pada Tabel 2.

Tabel 2 Keterangan ekspresi

No.	Ekspresi	Tipe
1.	hsa-miR-21	Numerik
2.	hsa-miR-31	Numerik
3.	hsa-miR-135b	Numerik
4.	hsa-miR-183	Numerik
5.	hsa-miR-222	Numerik
6.	hsa-miR-145	Numerik
7.	hsa-miR-195	Numerik
8.	label	String

3.2.2 NORMALISASI DATA

Normalisasi data dilakukan agar data memiliki batas nilai yang sama sehingga

penghitungan jarak *Euclidean* pada algoritma *K-Nearest Neighbor* lebih akurat. Proses normalisasi data menggunakan metode normalisasi Min-Max. Nilai minimum dan nilai maksimum pada *dataset* dapat dilihat pada Tabel 3.

Tabel 3 Nilai Minimum dan Maksimum *Dataset*

Ekspresi	Nilai Minimum	Nilai Maksimum
hsa-miR-21	1247	4887585
hsa-miR-31	0	3424
hsa-miR-135b	0	4389
hsa-miR-183	151	268179
hsa-miR-222	13	2589
hsa-miR-145	160	331708
hsa-miR-195	0	2509

Masing-masing atribut dilakukan penghitungan normalisasi pada data ke-1 hingga data ke-600 berdasarkan rumus penghitungan normalisasi min-max sesuai dengan persamaan 1.

Contoh penghitungan normalisasi pada data ke-1.

$$\text{hsa-miR-21}_1 = \frac{154256-1247}{4887585-1247} = 0{,}03131$$

$$\text{hsa-miR-31}_1 = \frac{0-0}{3424-0} = 0$$

$$\text{hsa-miR-135b}_1 = \frac{0-0}{4389-0} = 0$$

$$\text{hsa-miR-183}_1 = \frac{951-151}{268179-151} = 0{,}002985$$

$$\text{hsa-miR-222}_1 = \frac{21-13}{2589-13} = 0{,}003106$$

$$\text{hsa-miR-145}_1 = \frac{6042-160}{331708-160} = 0{,}017741$$

$$\text{hsa-miR-195}_1 = \frac{450-0}{2509-0} = 0{,}179354$$

Hasil penghitungan normalisasi data ke-1 dapat dilihat pada Tabel 4.

Tabel 4 Hasil Penghitungan Normalisasi

Ekspresi	Data ke-1	Normalisasi
hsa-miR-21	154256	0,31314
hsa-miR-31	0	0
hsa-miR-135b	0	0
hsa-miR-183	951	0,002985
hsa-miR-222	21	0,003106
hsa-miR-145	6042	0,017741
hsa-miR-195	450	0,179354

3.2.3 Pembagian Data Latih dan Data Uji

Tahapan ini digunakan untuk membagi *dataset* menjadi data latih dan data uji. Data latih digunakan dalam proses pelatihan, sementara data uji digunakan untuk memvalidasi model dengan pengujian keakuratan. Pada penelitian ini, metode *K-Fold Cross Validation* digunakan untuk membagi *dataset* menjadi data latih dan data uji.

Pada penelitian ini, nilai K yang digunakan adalah 10, sehingga disebut 10-*Fold Cross Validation*. Pada penelitian ini *dataset* yang digunakan sebanyak 600 data. Sehingga terdapat 10 eksperimen, masing-masing partisi terdapat 60 data. Pada eksperimen ke-1, partisi ke-1 (data ke-1 s.d. data ke-60) digunakan untuk data uji dan sisanya menjadi data latih. Pada eksperimen ke-2, partisi ke-2 digunakan untuk data uji dan sisanya menjadi data latih. Begitu juga untuk eksperimen ke-3 hingga eksperimen ke-10.

3.2.4 Proses *K-Nearest Neighbor*

Dataset yang sudah dibagi menjadi data latih dan data uji kemudian dilakukan pengujian metode *K-Nearest Neighbor* untuk mendapatkan label dari data uji. Contoh data uji ditunjukkan Tabel 5. Sedangkan sampel data latih ditunjukkan Tabel 6.

Tabel 5 Data Uji

Ekspresi	Nilai
hsa-miR-21	2296775
hsa-miR-31	12
hsa-miR-135b	316
hsa-miR-183	46159
hsa-miR-222	368
hsa-miR-145	9881
hsa-miR-195	203

Tabel 6 Sampel Data Latih

Ekspresi	Data ke-1	Data ke-2	Data ke-3
hsa-miR-21	154256	67483	141363
hsa-miR-31	0	0	0
hsa-miR-135b	0	0	0
hsa-miR-183	951	1845	1348

hsa-miR-222	21	38	37
hsa-miR-145	6042	2503	2781
hsa-miR-195	450	286	508

Selanjutnya, data uji pada Tabel 5. dan sampel data latih pada Tabel 6. dilakukan proses normalisasi data dengan menggunakan metode normalisasi min-max sesuai dengan persamaan 1. Sehingga didapat hasil normalisasi data uji yang ditunjukkan pada Tabel 7. dan hasil normalisasi sampel data latih yang ditunjukkan pada Tabel 8.

Tabel 7 Hasil Normalisasi Data Uji

Ekspresi	Nilai
hsa-miR-21	0,469785
hsa-miR-31	0,003505
hsa-miR-135b	0,071998
hsa-miR-183	0,171654
hsa-miR-222	0,137811
hsa-miR-145	0,02932
hsa-miR-195	0,080909

Tabel 8 Hasil Normalisasi Sampel Data Latih

Ekspresi	Data ke-1	Data ke-2	Data ke-3
hsa-miR-21	0,31314	0,013555	0,028675
hsa-miR-31	0	0	0
hsa-miR-135b	0	0	0
hsa-miR-183	0,002985	0,00632	0,004466
hsa-miR-222	0,003106	0,009705	0,009317
hsa-miR-145	0,017741	0,007067	0,007905
hsa-miR-195	0,179354	0,11399	0,202471

Setelah dilakukan proses normalisasi, data uji dihitung jaraknya dengan setiap data latih dengan menggunakan rumus jarak *Euclidean* sesuai dengan persamaan 2.

Contoh penghitungan jarak antara data uji dengan data latih ke-1 berdasarkan persamaan 2.

$$d(x_1,x_1) = \sqrt{\begin{matrix}(0{,}46978-0{,}3131)^2+(0{,}00350-0)^2+\\(0{,}0720-0)^2+(0{,}17165-0{,}00298)^2+\\(0{,}1378-0{,}003)^2+(0{,}029-0{,}0178)^2+\\(0{,}080909-0{,}179354)^2\end{matrix}}$$

$d(x_1,x_1) = 0{,}19307872$

Dengan menggunakan cara yang sama seperti contoh, dilakukan penghitungan jarak antara data uji dengan data latih lainnya. Setelah didapatkan hasil penghitungan jarak antara data uji dengan data latih, hasil penghitungan jarak tersebut diurutkan secara *ascending* (dari yang terkecil ke terbesar). Sehingga didapat hasil seperti pada Tabel 9.

Tabel 9 Hasil penghitungan jarak antara data uji dengan data latih

Urutan	Index	Jarak	*Class*
1	392	0,033087172	*colorectal*
2	507	0,036786086	normal
3	354	0,042034561	*colorectal*
4	324	0,042379338	normal
5	59	0,042461581	*colorectal*
6	286	0,043682078	*colorectal*
7	390	0,04568988	normal

Dengan menggunakan K=3, maka diambil 3 data teratas sehingga menghasilkan 2 kelas *colaretal* dan 1 kelas normal. Berdasarkan kelas mayoritas dengan nilai K=3, data uji menghasilkan nilai kelas="*colorectal*", yang berarti data uji terdiagnosis kanker kolorektal.

3.2.5 EVALUASI

Evaluasi meliputi penghitungan *accuracy*, *specificity*, dan *sensitivity*. Pada penelitian ini evaluasi dilakukan dengan menggunakan 10-*Fold Cross Validation* dengan *dataset* sebanyak 600 data dengan pengujian tetangga terdekat K=3 sampai dengan K=11. Contoh hasil evaluasi dengan nilai tetangga terdekat K=3 pada *fold* ke-3 seperti ditunjukkan pada Tabel 10.

Tabel 10 *Confusion Matrix* dari K=3 pada *fold* ke- 3

Actual Class	*Predicted Class*	
	Colorectal	Normal
Colorectal	30 (TP)	0 (FN)
Normal	2 (FP)	28 (TN)

Performa dari model klasifikasi diukur dengan menghitung nilai *accuracy*, *specificity*, dan *sensitivity*.

Nilai accuracy dihitung berdasarkan persamaan 3.

$$Accuracy = \frac{30+28}{30+0+0+28} = 96{,}67\%$$

Nilai sensitivity dihitung berdasarkan persamaan 4.

$$Sensitivity = \frac{30}{30+0} = 100\%$$

Nilai specificity dihitung berdasarkan persamaan 5.

$$Specificity = \frac{28}{28+2} = 93{,}33\%$$

3.2.6 DIAGNOSIS

Tahap diagnosis merupakan tahap untuk melakukan klasifikasi terhadap data baru. Setelah data baru dimasukkan ke dalam sistem, data baru akan melalui proses normalisasi terlebih dahulu. Data baru yang sudah dinormalisasi kemudian dilakukan penghitungan jarak *Euclidean* dengan data latih dari model terbaik yang didapatkan pada proses pengujian. Setelah didapatkan hasil penghitungan jarak, kemudian diurutkan hasil penghitungan jarak dimulai dari yang terendah hingga terbesar nilainya. Nilai K pada metode *K-Nearest Neighbor* akan mengambil sejumlah tetangga terdekat K hasil penghitungan jarak yang sudah diurutkan. Hasil dari proses diagnosis akan menunjukkan data tersebut termasuk kelas normal atau kanker kolorektal.

3.2 ANALISIS APLIKASI

3.2.1 DESKRIPSI UMUM APLIKASI

Perangkat lunak yang dibuat dinamakan aplikasi Diagnosis Kanker Kolorektal (DKK). Aplikasi DKK digunakan untuk mendapatkan model klasifikasi terbaik dengan menggunakan metode *K-Nearest Neighbor*. Model terbaik didapat dari proses pengujian beberapa nilai K dengan menggunakan 10-*Fold Cross Validation*. Model yang sudah didapat, digunakan untuk melakukan diagnosis sebuah data baru yang dimasukkan ke dalam aplikasi. Hasil diagnosis data berupa data diagnosis termasuk kelas kanker kolorektal atau normal.

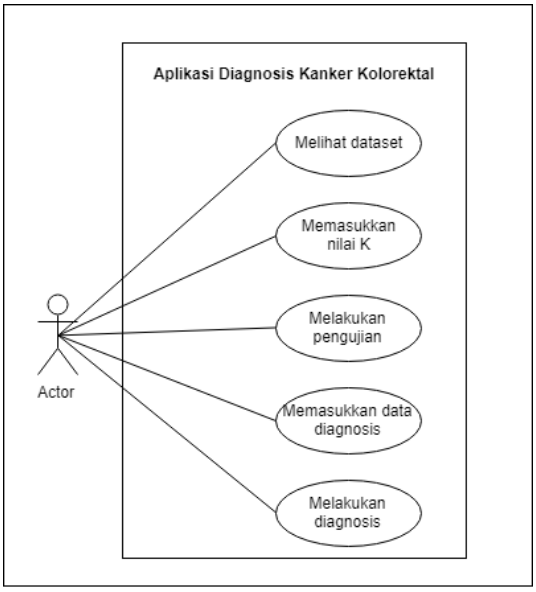
3.2.2 ANALISIS KEBUTUHAN SISTEM

Terdapat lima *use case* dalam pengembangan aplikasi Diagnosis Kanker Kolorektal yang dapat dilihat pada Tabel 11.

Tabel 11 Daftar *Use Case*

Use Case ID	Nama *Use Case*	Deskripsi
UC-DKK-01	Melihat *dataset*	*User* melihat *dataset micro-RNA* yang digunakan pada aplikasi
UC-DKK-02	Memasukkan nilai K	*User* memasukkan nilai K untuk proses pengujian *K-Nearest Neighbor*
UC-DKK-03	Melakukan proses pengujian	*User* melakukan permintaan proses pengujian metode *K-Nearest Neighbor*
UC-DKK-04	Memasukkan data diagnosis	*User* memasukkan data diagnosis
UC-DKK-05	Melakukan proses diagnosis	*User* melakukan permintaan proses diagnosis

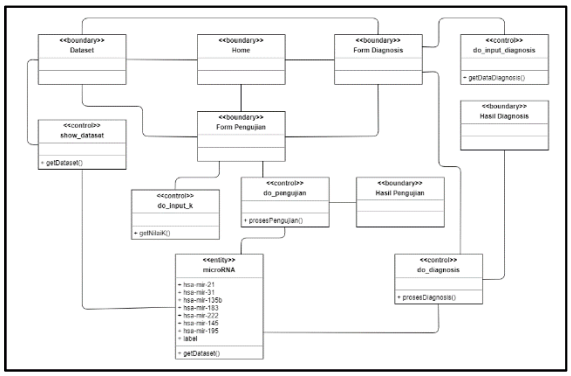
Diagram *use case* menggambarkan apa yang dilakukan sistem dan dengan siapa sistem berinteraksi. Diagram *use case* Aplikasi Diagnosis Kanker Kolorektal ditunjukkan pada Gambar 2.



Gambar 1 Diagram *use case*

Analisis *class diagram* menggambarkan *entity*, *boundary*, dan *control* berdasarkan *use case* yang bersesuaian. Terdapat 6 *boundary* pada aplikasi DKK, yaitu *Home*, *Dataset*, Form Pengujian, Hasil Pengujian, Form Diagnosis, dan Hasil Diagnosis. Terdapat 5 *control* pada aplikasi DKK, yaitu show_dataset, do_input_k, do_pengujian,

do_input_diagnosis, dan do_diagnosis. Terdapat 1 *entity* pada aplikasi DKK, yaitu *microRNA*. Analisis *class diagram* Aplikasi Diagnosis Kanker Kolorektal ditunjukkan pada Gambar 3.



Gambar 2 *Class Diagram*

4 HASIL DAN PEMBAHASAN

Aplikasi Diagnosis Kanker Kolorektal berbasis *web* dibangun menggunakan bahasa pemrograman Python dan menggunakan *web framework* Flask.

4.1 IMPLEMENTASI ANTARMUKA

Implementasi antarmuka Aplikasi Diagnosis Kanker Kolorektal terdapat 6 halaman antarmuka.

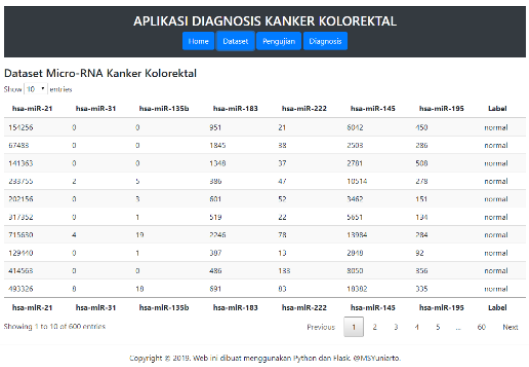
1. Implementasi Antarmuka Halaman Utama

   Halaman Utama merupakan halaman yang pertama kali ditampilkan saat aplikasi Diagnosis Kanker Kolorektal dijalankan. Pada Halaman Utama dan masing-masing halaman terdapat 4 menu yang digunakan untuk menuju halaman yang diinginkan, yaitu menu *Home*, *Dataset*, Pengujian, dan Diagnosis. Implementasi antarmuka Halaman Utama dapat dilihat pada Gambar 4.

Gambar 3 Antarmuka Halaman Utama

2. Implementasi Antarmuka Halaman *Dataset*

   Halaman *Dataset* merupakan halaman yang digunakan untuk melihat *dataset* yang digunakan pada aplikasi Diagnosis Kanker Kolorektal. *Dataset* berupa data *micro-RNA* yang digunakan untuk penelitian. *Dataset* terdiri dari 7 fitur ekspresi *micro-RNA*. Pada Halaman *Dataset* terdapat *dropdown list* yang dapat digunakan untuk menampilkan data sejumlah 10, 25, 50, atau 100 per halaman. Implementasi Halaman *Dataset* dapat dilihat pada Gambar 5.



Gambar 4 Antarmuka Halaman *Dataset*

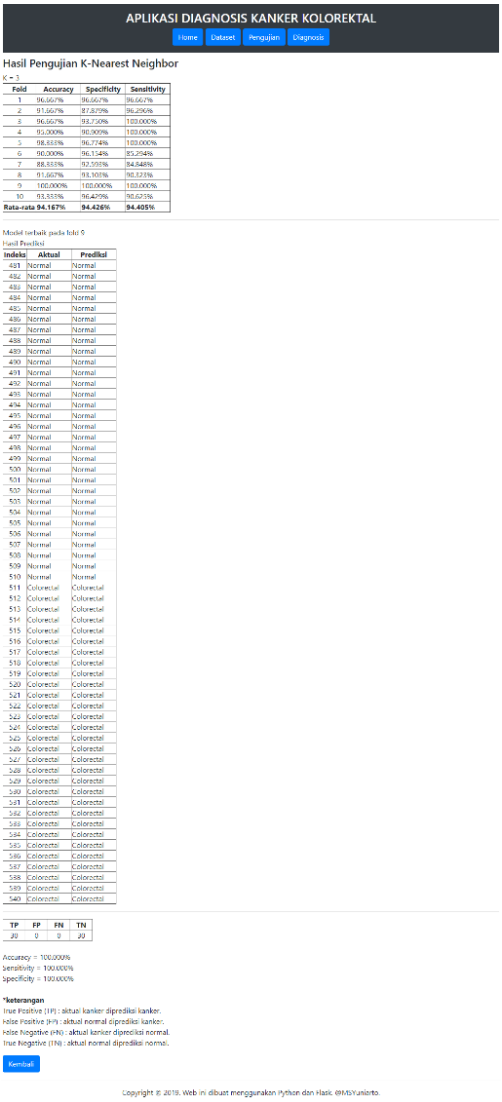
3. Implementasi Antarmuka Halaman Pengujian

   Halaman Pengujian digunakan untuk menguji performa metode *K-Nearest Neighbor* dalam melakukan klasifikasi data *micro-RNA*. Pada halaman pengujian, user diminta untuk memasukkan nilai K. Nilai K yang diperkenankan yaitu berupa angka lebih dari nol. Jika nilai masukan K tidak sesuai maka proses pengujian tidak dapat dilakukan. Berdasarkan 10-*Fold Cross Validation* pada dataset sejumlah 600, maka masukan nilai K maksimal yang dapat digunakan adalah 540. Implementasi halaman pengujian dapat dilihat pada Gambar 6.

Gambar 5. Antarmuka Halaman Pengujian

4. Implementasi Antarmuka Halaman Hasil Pengujian

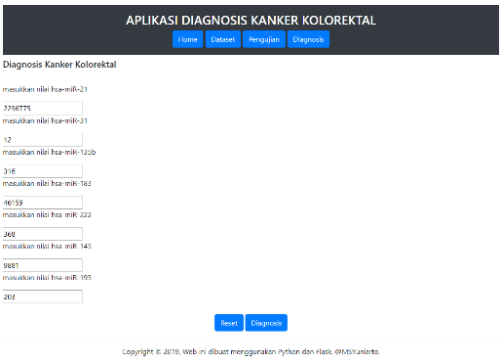
Halaman Hasil Pengujian menampilkan hasil pengujian nilai K dari metode *K-Nearest Neighbor*. Sebagai contoh, dilakukan pengujian nilai K=3. Dengan menggunakan 10-*Fold Cross Validation*, masing-masing *fold* menghasilkan performa berupa *accuracy*, *specificity*, dan *sensitivity*. Hasil dari klasifikasi data uji yang ditampilkan adalah model terbaik yaitu pada *fold* ke 9. Hasil klasifikasi data uji dibandingkan dengan label *actual class* dari data uji, sehingga didapat *Confusion Matrix* (*true positive*, *false negative*, *false positive*, dan *true negative*). *Confusion matrix* digunakan untuk menghitung nilai *accuracy*, *specificity*, dan *sensitivity*. Implementasi Halaman Hasil Pengujian dapat dilihat pada Gambar 7.



Gambar 6 Antarmuka Halaman Hasil Pengujian

5. Implementasi Antarmuka Halaman Diagnosis

   Halaman Diagnosis digunakan untuk melakukan diagnosis data yang dimasukkan oleh pengguna. Terdapat 7 masukan yang diminta aplikasi kepada *user* untuk melakukan diagnosis. Implementasi halaman diagnosis dapat dilihat pada Gambar 8.



Gambar 7 Antarmuka Halaman Diagnosis

6. Implementasi Antarmuka Halaman Hasil Diagnosis

   Halaman Hasil Diagnosis menampilkan hasil diagnosis dari data yang telah dimasukkan user pada halaman diagnosis. Hasil diagnosis berupa data termasuk kelas kanker kolorektal atau normal. Implementasi halaman hasil diagnosis dapat dilihat pada Gambar 9.

Gambar 8 Antarmuka Halaman Hasil Diagnosis

4.2 PENGUJIAN

Pengujian yang dilakukan berupa pengujian aplikasi dan pengujian metode *K-Nearest Neighbor*. Pengujian aplikasi Diagnosis Kanker Kolorektal dilakukan dengan menggunakan metode pengujian *black box*. Pengujian dengan metode *black box* bertujuan untuk memastikan bahwa aplikasi telah berjalan baik sesuai dengan

spesifikasi kebutuhan yang telah ditetapkan. Sedangkan, pengujian metode *K-Nearest Neighbor* digunakan untuk mengetahui kinerja metode *K-Nearest Neighbor* dalam melakukan klasifikasi data.

Berdasarkan hasil pengujian dengan menggunakan metode *black box*, pengujian terhadap seluruh butir uji dinyatakan diterima. Sehingga, aplikasi Diagnosis Kanker Kolorektal dapat memenuhi semua spesifikasi kebutuhan yang telah ditetapkan.

Pengujian metode *K-Nearest Neighbor* dilakukan 2 skenario. Skenario 1 digunakan untuk mendapatkan performa terbaik dari pengujian beberapa nilai K. Sedangkan skenario 2 digunakan untuk mendapatkan model terbaik dari nilai K terbaik pada hasil pengujian skenario 1.

4.2.1 SKENARIO 1

Nilai K yang digunakan yaitu bernilai ganjil, dimulai dari 3 hingga 11. Hasil pengujian beberapa nilai K dapat dilihat pada Tabel 12.

Tabel 12 Hasil Pengujian Beberapa Niai K

K	*Accuracy*	*Specificity*	*Sensitivity*
3	**94,17%**	**94,43%**	**94,41%**
5	94,17%	95,28%	93,64%
7	93,17%	94,89%	92,51%
9	92,67%	94,54%	92,07%
11	92,33%	94,50%	91,52%

Berdasarkan hasil pengujian beberapa nilai K yang ditunjukkan pada Tabel 12, nilai *accuracy* tertinggi dihasilkan terdapat pada K=3 dan K=5. Nilai *specificity* pada K=3 lebih rendah daripada pada K=5, artinya pada K=3 terdapat data normal yang diprediksi kanker lebih banyak daripada pada K=5. Sedangkan nilai *sensitivity* pada K=5 lebih rendah daripada pada K=3, artinya pada K=5 terdapat data kanker yang diprediksi normal lebih banyak daripada pada K=3. Sehingga, performa terbaik dihasilkan pada K=3 dengan nilai *accuracy* 94,17%, *specificity* 94,43%, dan *sensitivity* 94,41%.

4.2.2 SKENARIO 2

Skenario 1 didapatkan nilai K terbaik pada K=3. Pada skenario 2 dilakukan pengamatan *fold* terbaik yang dihasilkan pada K=3. Hasil pengujian K=3 dapat dilihat pada Tabel 13.

Tabel 13 Hasil Pengujian K=3

Fold	*Accuracy*	*Specificity*	*Sensitivity*
1	96.667%	96.667%	96.667%
2	91.667%	86.667%	96.667%
3	96.667%	93.333%	100.000%
4	95.000%	90.000%	100.000%
5	98.333%	96.667%	100.000%
6	90.000%	96.667%	83.333%
7	88.333%	93.333%	83.333%
8	91.667%	93.333%	90.000%
9	**100.000%**	**100.000%**	**100.000%**
10	93.333%	96.667%	90.000%

Berdasarkan pengamatan pada Tabel 13, model terbaik didapatkan pada fold 9 dengan nilai *accuracy*, *specificity*, dan *sensitivity* masing-masing sebesar 100%. Pada pengamatan lain, pada pengujian nilai K=3 didapat hasil *confusion matrix* yang ditunjukkan pada Tabel 14.

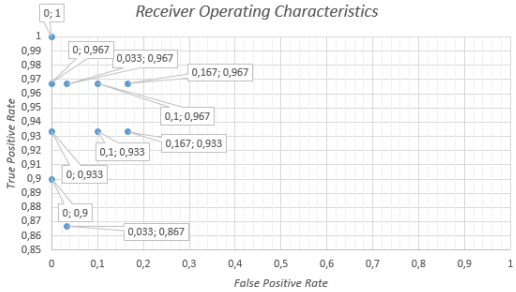
Tabel 14 Hasil *confusion matrix* pengujian K=3

Fold	TP	FN	FP	TN
1	29	1	1	29
2	26	4	1	29
3	28	2	0	30
4	27	3	0	30
5	29	1	0	30
6	29	1	5	25
7	28	2	5	25
8	28	2	3	27
9	30	0	0	30
10	29	1	3	27

Confusion matrix hanya menyajikan informasi dalam bentuk angka. Untuk menampilkan informasi kinerja algoritma klasifikasi dalam bentuk grafik dapat digunakan kurva *Receiver Operating Characteristic* (ROC). Kurva ROC dibuat berdasarkan nilai yang didapatkan pada penghitungan *confusion matrix*, yaitu *False Positive Rate* (FPR) dan *True Positive Rate* (TPR) dengan menggunakan persamaan 6. untuk FPR dan persamaan 7. untuk TPR. Hasil penghitungan FPR dan TPR dapat dilihat pada Tabel 15. Grafik ROC dapat dilihat pada Gambar 10.

Tabel 15 Hasil Penghitungan FPR dan TPR

Fold	FPR	TPR
1	0.033	0.967
2	0.033	0.867
3	0.000	0.933
4	0.000	0.900
5	0.000	0.967
6	0.167	0.967
7	0.167	0.933
8	0.100	0.933
9	0.000	1.000
10	0.100	0.967



Gambar 9 Grafik ROC Pengujian K=3

Berdasarkan Gambar 10. kurva yang mendekati titik (0,1) berarti kinerjanya bagus, yaitu pada *fold* 9. Sedangkan kurva yang semakin menjauhi (0,1) berarti kinerjanya jelek, yaitu pada titik (0.033, 0.867) atau *fold* 2. Sehingga, model terbaik yang didapatkan dari pengujian K=3 adalah pada *fold* 9 dengan nilai *accuracy* 100%, *specificity* 100%, dan *sensitivity* 100%. Sedangkan berdasarkan grafik ROC, *fold* 2 kinerjanya paling jelek dengan nilai *accuracy* 91,667%, *specificity* 86,667%, dan *sensitivity* 96,667%.

5 KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Kesimpulan yang dapat diambil dari hasil penelitian implementasi metode *K-Nearest Neighbor* untuk diagnosis kanker kolorektal dengan *biomarker micro-RNA* adalah sebagai berikut.

1. Perubahan nilai K pada *K-Nearest Neighbor* mempengaruhi performa metode *K-Nearest Neighbor* yang terdiri dari *accuracy*, *specificity*, dan *sensitivity*.
2. Pengujian nilai K dilakukan dimulai dari 3 sampai dengan 11 yang bernilai ganjil. Hasil pengujian tersebut terdapat hasil rata-rata *accuracy* yang sama, yaitu pada pengujian K=3 dan K=5. Performa K=3 dipilih menjadi model terbaik karena pertimbangan nilai *sensitivity* lebih tinggi dan nilai *specificity* lebih rendah.
3. Hasil performa terbaik dari pengujian beberapa nilai K dengan metode *K-Nearest Neighbor* menghasilkan *accuracy* 94,17%, *specificity* 94,43%, dan *sensitivity* 94,41% pada K=3.
4. Berdasarkan kurva ROC, pada pengujian K=3 didapatkan model dengan performa terbaik pada *fold* 9 dengan nilai *accuracy* 100%, *specificity* 100%, dan *sensitivity* 100%. Sedangkan model dengan performa terjelek pada fold 2 dengan nilai *accuracy* 91,667%, *specificity* 86,667%, dan *sensitivity* 96,667%.

5.2 SARAN

Saran yang dapat diberikan dari penelitian tugas akhir ini untuk penelitian lebih lanjut yaitu pengembangan penelitian tentang penentuan jumlah tetangga (K) optimal pada *K-Nearest Neighbor* dengan otomatis tanpa harus mencoba satu persatu dalam menentukan nilai K. Selain itu, dapat dilakukan pengembangan penelitian tentang metode seleksi atribut (*feature selection*) yang dapat mengetahui bobot masing-masing atribut agar memberikan hasil *accuracy* yang optimal.

DAFTAR PUSTAKA

Accerbi, M., Schmidt, S. A., Paoli, E. De, Park, S., Jeong, D.-H., & Green, P. J. (2013). Methods for Isolation of Total RNA to Recover miRNAs and Other Small RNAs from Diverse Species. *From Plant Genomics to Plant Biotechnology*, *592*, 15–30.

Anwar, S. L., Haryono, S. J., Aryandono, T., & Haryana, S. M. (2017). *Micro-RNA. Biogenesis, Fungsi, dan Perannya dalam Proses karsinogenesis dan Penatalaksanaan Kanker*. Yogyakarta: Gadjah Mada University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.

Calin, G. A., & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nature Reviews Cancer*, *6*(11), 857–866.

Iorio, M. V., & Croce, C. M. (2012). microRNA involvement in human cancer. *Carcinogenesis*, *33*(6), 1126–1133.

Kemenkes RI. (2017). Pedoman Nasional Pelayanan Kedokteran Kolorektal.

Kohavi, R. (1995). A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection, *5*.

Lan, H., Lu, H., Wang, X., & Jin, H. (2015). MicroRNAs as potential biomarkers in cancer: Opportunities and challenges. *BioMed Research International*, *2015*.

Liu, A., & Xu, X. (2011). Formalin-fixed Paraffin-embedded Tissue. *Encyclopedia of Cancer*, 1446–1446.

Luo, X., Burwinkel, B., Tao, S., & Brenner, H. (2011). MicroRNA signatures: Novel biomarker for colorectal cancer? *Cancer Epidemiology Biomarkers and Prevention*, *20*(7), 1272–1286.

Mazeh, H., Mizrahi, I., Ilyayev, N., Halle, D., Brücher, B. L. D. M., Bilchik, A., Protic, M., Daumer, M., Stojadinovic, A., Avital, I., & Nissan, A. (2013). The diagnostic and prognostic role of microRNA in colorectal cancer-a comprehensive review. *Journal of Cancer*, *4*(3), 281–295.

Medjahed, S. A., Ait Saadi, T., & Benyettou, A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, *62*(1), 1–5.

Michael, M. Z., O' Connor, S. M., van Holst Pellekaan, N. G., Young, G. P., & James, R. J. (2003). Reduced accumulation of specific microRNAs in colorectal neoplasia. *Molecular Cancer Research : MCR*, *1*(12), 882–891.

Mustaffa, Z., & Yusof, Y. (2011). A Comparison of Normalization Techniques in Predicting Dengue Outbreak. *International Conference on Business and Economics Research*, *1*, 345–349.

Primartha, R. (2018). *Belajar Machine Learning*. Bandung: Informatika.

Rosenfeld, N. et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnology*, *26*, 462.

Tyas, R. D. O., Soebroto, A. A., & Furqon, M. T. (2015). Pengembangan Sistem Pakar Diagnosa Penyakit Sapi Potong dengan Metode Fuzzy K-Nearest Neighbour. *Journal of Environmental Engineering & Sustainable Technology*, *02*(01), 58–66.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. *Business Intelligence: Data Mining and Optimization for Decision Making*.

Vogelstein, B. et al. (1988). Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, *309*(23), 1426–1434.

WHO. (2018). Global Cancer Observatory: Cancer Today., *876*, 49–50.

Winawer, Z. et al. (1993). Prevention of colorectal cancer by colonscopy polipectomy. *New England Journal of Medicine*.

Yayasan Kanker Indonesia. (2018). Kanker Kolorektal World Cancer Day, (April).

Zhang, S., Zhang, J., Zhu, X., Qin, Y., & Zhang, C. (2008). Missing Value Imputation Based on Data Clustering. *Transactions on Computational Science I*, (60625204), 128–138.