BAB III

METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Tahapan penelitian dijelaskan secara komprehensif melalui *flow chart* yang mengikuti metode penelitian kuantitatif. Diagram alur ini digunakan untuk menyajikan informasi tentang langkah-langkah yang terlibat dalam penelitian ini dengan lebih mudah. Semua tahapan penelitian dipresentasikan dalam Gambar 3.1.

Gambar 3.1 Tahapan Penelitian

3.1.1 Studi Literatur

Studi literatur yang mendukung penelitian, termasuk mengenai ekstraksi citra dengan GLCM dan penggunaan SVM untuk klasifikasi. Sumber literatur yang digunakan berasal dari jurnal dan buku yang dapat dipercaya.

3.1.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data yang diperoleh dari situs https://datasets.simula.no/kvasir/. Dataset Kvasir adalah kumpulan gambar endoskopi saluran cerna. Kumpulan data ini dianotasi dan divalidasi oleh ahli endoskopi bersertifikat. Resolusi gambar dataset Kvasir dengan delapan lapisan berkisar antara 720×576 piksel hingga 1920×1072 piksel. Setiap gambar memiliki sudut kamera, resolusi, kecerahan, zoom, dan titik tengah yang berbeda. Dataset Kvasir memiliki beberapa kelas yang terdiri dari *landmark* anatomi yang menunjukkan beberapa bagian dari usus dan temuan patalogis menunjukkan ketidaknormalan pada usus. Jumlah dataset yang digunakan pada penelitian ini berjumlah 1990 dengan masing-masing kelas yaitu kelas sehat dan kelas kolitis ulseratif memiliki citra berjumlah 995 citra.

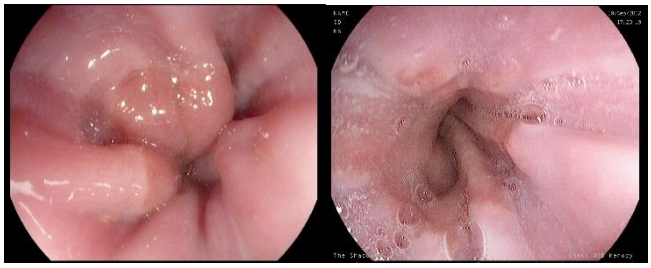
3.1.2.1 Landmark Anatomi Dataset Kvasir

Landmark anatomi adalah karakteristik yang dapat dikenali dalam saluran gastrointernal yang mudah dilihat melaui endoskopi. Karakteristik ini sangat penting sebagai titik referensi untuk menggambarkan lokasi temuan.

a. Z-Line

Z-Line menandakan lokasi transisi antara kerongkongan dan lambung. Ketika dilakukan endoskopi terlihat sebagai batas yang jelas dimana mukosa

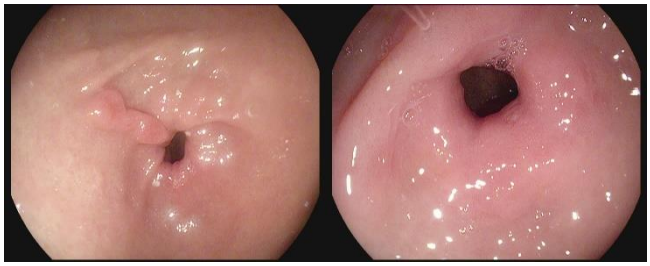
putih di kerongkongan bertemu dengan mukosa lambung merah. Bentuk dari *Z-Line* ditunjukkan pada gambar 3.2. Identifikasi *Z-Line* sangat penting untuk menentukan apakah terjangkit penyakit atau tidak. *Z-line* juga berfungsi sebagai titik referensi saat menggambarkan patologi di kerongkongan.



Gambar 3.2 *Z-Line*

b. Pilorus

Pilorus merupakan daerah sekitar pembukaan dari lambung ke bagian awal dari usus kecil. Bagian ini terdiri dari otot melingkar yang mengatur pergerakan makanan dari perut. Gambar 3.3 menunjukkan gambar hasil endoskopi pilorus normal.



Gambar 3.3 Pilorus

c. Cecum

Cecum atau sekum adalah bagian paling pangkal dari usus besar. Pengenalan dan dokumentasi sekum menjadi penting karena proses endoskopi yang mencapai sekum adalah bukti kolonoskopi lengkap dan terbukti menjadi

indikator kualitas yang valid untuk kolonoskopi. Gambar 3.4 menunjukkan bagaimana bentuk sekum.

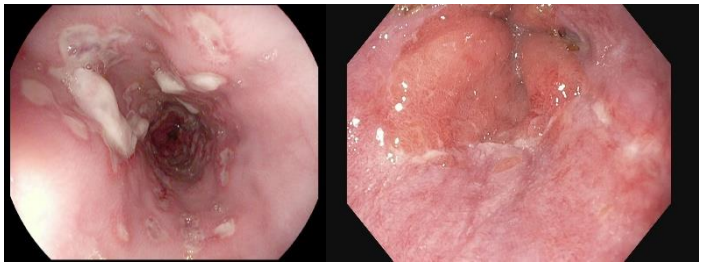
Gambar 3.4 *Cecum*

3.1.2.2 Temuan Patalogis Dataset Kvasir

Temuan patalogis merupakan sebuah ketidaknormalan dalam saluran pencernaan. Pengamatan dalam citra endoskopi dilakukan untuk melihat kerusakan atau perubahan pada mukosa normal. Temuan ini dapat menjadi tanda-tanda adanya suatu penyakit. Deteksi dan klasifikasi patologi menjadi langkah awal dalam pengobatan yang benar dan untuk menindaklanjuti pasien.

a. Esofagitis

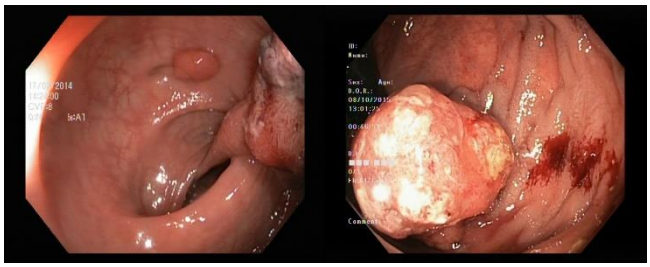
Esofagitis adalah suatu penyakit yang ditandai adanya peradangan kerongkongan pada mukosa esofagus. Gambar 3.5 menunjukkan lidah mukosa merah menonjol di lapisan esofagus putih. Tingkat peradangan dapat ditentukan dengan panjang dari kerusakan pada mukosa dan proporsi lingkar yang terlibat. Peradangan tersebut dapat disebabkan oleh asam lambung mengalir kembali ke kerongkongan. Deteksi diperlukan untuk memulai proses medis untuk meredakan gejala dan mencegah peradangan semakin parah yang dapat menyebabkan komplikasi. Deteksi dengan komputer akan menjadi suatu hal yang baik dalam menilai tingkat peradangan dan pelaporan otomatis.



Gambar 3.5 Esofagitis

b. Polip

Polip adalah gumpalan kecil sel yang terbentuk di bagian usus besar. Polip umumnya tidak berbahaya, tetapi ada beberapa polip yang dapat berpotensi menjadi kanker usus. Deteksi otomatis akan meningkatkan pemeriksaan dan bermanfaat untuk diagnosis, penilaian dan pelaporan. Gambar 3.6 menunjukkan bagaimana usus yang menderita polip.

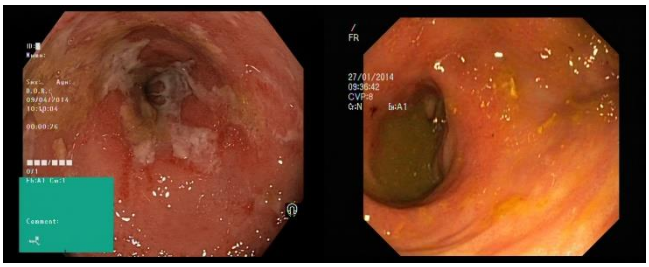


Gambar 3.6 Polip

c. Ulcerative Colitis

Ulcerative Colitis (Kolitis Ulseratif) merupakan penyakit yang disebabkan oleh peradangan mukosa yang mempengaruhi usus besar. Penyakit ini belum ada obatnya dan penderita akan memiliki penyakit ini seuumur hidup. Beberapa terapi dapat dilakukan untuk menekan dampak dari kolitis ulseratif. Sistem penilaian berbasis komputer otomatis akan berkontribusi pada tingkat

keparahan penyakit yang akurat. Gambar 3.7 menunjukkan bagaimana usus yang mengalami kolitis ulseratif.



Gambar 3.7 Ulcerative Colitis

3.1.3 Praproses Data

Praproses data merupakan tahap krusial dalam sebuah penelitian untuk memastikan bahwa data yang akan diolah pada tahap berikutnya memiliki kualitas yang baik. Hal yang dilakukan pada langkah ini adalah mengubah citra yang asalnya *RGB* menjadi citra *Grayscale*. Proses *resizing* atau mengubah ukuran resolusi pada citra dilakukan pada praposes data dengan mengubah resolusinya menjadi 0.5 menjadi lebih kecil.

3.1.3 Ekstraksi Data dengan GLCM

Ekstraksi data dengan GLCM adalah proses mengekstrak suatu informasi dari objek yang terdapat pada citra. Metode yang digunakan pada tahapan ini adalah ekstraksi fitur *Gray Level Co-occurrence Matrix (GLCM)* dengan mencari nilai *energy, contrast, homogeneity, dissimilarity* dan *ASM*.

a. *Energy*

Energy adalah jumlah yang terkait dengan variasi intensitas abu-abu dalam piksel. Persamaan untuk mendapatkan nilai *energy* adalah sebagai sebagai berikut:

$$Energy = \sum_i^m \sum_j^n P(i,j)^2 \quad (3.1)$$

Dimana

P = matriks GLCM normalisasi

i = indeks baris matriks P

j = indeks kolom matriks P

b. *Contrast*

Contrast adalah fitur yang digunakan untuk mengukur kekuatan perbedaan intensitas dalam gambar. Nilai *contrast* akan meningkat jika variasi intensitas pada gambar tinggi dan menurun jika variasinya rendah. Persamaan yang digunakan untuk mengukur *contrast* ditunjukkan oleh persamaan berikut ini:

$$Contrast = \sum_{i=0}^m \sum_{j=0}^n P_{(i,j)}(i-j)^2 \quad (3.2)$$

Dimana

P = matriks GLCM normalisasi

i = indeks baris matriks P

j = indeks kolom matriks P

c. *Homogeneity*

Homogeneity adalah representasi ukuran nilai kesamaan variasi intensitas gambar. Jika semua nilai piksel memiliki nilai yang seragam, maka *homogeneity* memiliki nilai maksimum. Persamaan yang menghitung nilai *homogeneity* adalah sebagai berikut:

$$Homogeneity = \sum_i^m \sum_j^n \frac{1}{1+(i-j)^2} P(i,j) \quad (3.3)$$

Dimana

P = matriks GLCM normalisasi

i = indeks baris matriks P

j = indeks kolom matriks P

d. *Dissimilarity*

Dissimilarity adalah fitur untuk mengukur perbedaan rata-rata tingkat keabuan dalam distribusi citra. Nilai *dissimilarity* dapat diketahui melalui persamaan:

$$Dissimilarity = -\sum_i^m \sum_j^n |i-j| P(i,j) \quad (3.4)$$

Dimana

P = matriks GLCM normalisasi

i = indeks baris matriks P

j = indeks kolom matriks P

e. ASM

Fitur ASM merupakan fitur untuk mewakili keseragaman distribusi tingkat keabuan dalam citra. Nilai ASM dapat dicari dengan menggunakan persamaan:

$$ASM = \sum_i^m \sum_j^n (P(i,j))^2 \quad (3.5)$$

Dimana

P = matriks GLCM normalisasi

i = indeks baris matriks P

j = indeks kolom matriks P

3.1.4 Klasifikasi dengan SVM

Klasifikasi dengan SVM adalah proses memprediksi kelas suatu data baru berdasarkan data pelatihan yang sebelumnya telah diberi label. Proses klasifikasi dengan SVM menggunakan konsep mencari *hyperplane* yang memiliki margin terbesar antara data dari dua kelas yang berbeda yang kemudian, kelas yang memiliki margin terbesar akan memiliki performa yang lebih baik dari kelas yang lain. *Hyperplane* yang telah dicari digunakan untuk membuat model klasifikasi. Penelitian ini menggunakan klasifikasi SVM dengan kernel *Radial Basic Function Gaussian* dengan persamaan:

$$\mathrm{K}\,(x_i, x) = \exp\left(-\gamma\left|\left|x_i - x\right|\right|^2\right), y > 0 \quad (3.6)$$

Penggunaan kernel ini dikarenakan bisa mendapatkan nilai akurasi yang lebih baik dibandingkan dengan menggunakan fungsi kernel yang lainnya (Luthfiana dkk., 2020). Setelah hasil akurasi dari model yang dibuat keluar, maka akan dilakukan proses *hyperparameter tuning* dengan mencari nilai C pada persamaan *soft margin* dan *gamma* pada persamaan 3.6. Nilai tersebut akan dicari yang optimal berdasarkan persebaran data pada ruang dimensi untuk meningkatkan hasil dari akurasi. Persamaan soft margin ditunjukkan pada persamaan 3.7.

$$(w, b, \xi) = \arg min_{w,b,\xi} \frac{1}{2}||w||_2^2 + C\sum_{n=1}^{N} \xi_n \quad (3.7)$$

Nilai C menentukan berapa banyak sampel data yang diizinkan untuk ditempatkan dalam kelas yang berbeda. Jika nilai C rendah, maka probabilitas pencilan akan meningkat, jika nilai C tinggi maka batas keputusan akan ditentukan secara hati-hati. Nilai *gamma* menentukan jarak sampel tunggal memberikan

pengaruh, dengan kata lain parameter *gamma* untuk menyesuaikan kelengkungan batas keputusan.

3.1.5 Reduksi Fitur dengan PCA

Reduksi fitur dengan PCA adalah proses untuk mereduksi jumlah fitur pada dataset yang dihasilkan dari proses ekstraksi dengan GLCM yang berjumlah 24 fitur menjadi berjumlah lebih kecil, sehingga dapat membantu dalam visualisasi dari bentuk model yang telah dibuat. Proses reduksi dengan PCA bertujuan untuk menghilangkan faktor yang kurang dominan tanpa mengurangi nilai dari data asli dari variabel acak x. Tahapan dalam PCA adalah sebagai berikut:

a. Menghitung matrik varians kovarian dari data observasi. Varians dihitung untuk menemukan penyebaran data dalam set data untuk menentukan penyimpangan data dalam set data sampel. Matrix kovarians adalah matriks yang nilai kovariansi pada tiap sel nya diperoleh dari sampel. Perhitungan matriks varians dan kovarians dapat menggunakan persamaan 3.8 dan 3.9.

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_{ij} - \mu_j)^2 \quad (3.8)$$

$$Cov\,(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{xj}) \quad (3.9)$$

Dengan μ_x dan μ_y merupakan mean sampel dari variabel x dan y, dimana variabel x_i dan y_i merupakan nilai observasi ke-I dari variabel x dan y. dari data nilai yang ada, maka diperoleh matrik kovarian n x n.

b. Mencari *eigenvalues* dan *eigenvector* dari matrix kovarian yang telah dicari. Nilai *eigen* yang telah dicari kemudian diubah kedalam bentuk *orthogonal varimax* menggunakan persamaan 3.10.

$$Det(A - \lambda I) = 0 \quad (3.10)$$

Dimana:

A = matrik nxn

λ = nilai *eigen*

I = matriks identitas

c. Menentukan nilai proporsi *Principal Component* (%) dengan persamaan 3.11.

$$PC\ (\%) = \frac{NilaiEigen}{VarianceCovarian}\ x100\% \quad (3.11)$$

d. Menghitung *factor loading* berdasarkan *eigenvector* dengan persamaan 3.12.

$$Ax = \lambda x \quad (3.12)$$

Sehingga diperoleh kombinasi linear $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_n$ adalah *eigenvalue* matrik A dan $x1, x2, x3 \ldots xn$ adalah *eigenvector* sesuai *eigenvalue*-nya.

e. Persamaan *eigenvalue* dan *eigenvector* merupakan *Eigen Value Decomposition* (EVD), dengan persamaan 3.13.

$$AX = XD$$

$$A = X\ D\ X^{-1} \quad (3.13)$$

Dimana:

A = matrik nxn yang memiliki n *eigenvalue*

D = *eigenvalue* dari *eigenvector*-nya

X = *eigenvector* dari matrik A

X^{-1} = invers dari *eigenvector* X

3.1.6 Evaluasi dan Pengujian

Model klasifikasi yang digunakan kemudian diuji dengan berbagai pengujian. Pengujian pertama adalah pengujian untuk mencari pengaruh jumlah data latih terhadap akurasi hasil klasifikasi. Evaluasi dilakukan dengan menggunakan *Confusion Matrix*, metode yang umumnya digunakan dalam mengevaluasi model klasifikasi untuk mengukur akurasi, presisi, *recall*, dan *F1-Score*.

Nilai TN, FP, FN, dan TP yang diperoleh dari *Confusion Matrix* tersebut dapat digunakan untuk menghitung nilai Presisi, *Recall*, *F1-Score*, dan Akurasi dengan persamaan sebagai berikut:

a. Presisi

$$\frac{TP}{TP+FP}\ x\ 100\% \tag{3.14}$$

b. Recall

$$\frac{TP}{TP+FN}\ x\ 100\% \tag{3.15}$$

c. F1-Score

$$2\ x\ \frac{Presisi\ x\ Recall}{Presisi+Recall}\ x\ 100\% \tag{3.16}$$

d. Akurasi

$$\frac{TP+TN}{TP+TN+FP+FN} x100\% \tag{3.17}$$

Keterangan:

- TP_i adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi oleh sistem untuk kelas ke-i

- TN_i adalah *True Negative*, yaitu jumlah data bukan positif yang terklasifikasi oleh sistem untuk kelas ke-i
- FN_i adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem untuk kelas ke-i
- FP_i adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem untuk kelas ke-i