IMPLEMENTASI *SUPPORT VECTOR MACHINE (SVM)* UNTUK KLASIFIKASI KANKER PAYUDARA BERBASIS *WEBSITE*

Nur Everi Rahmawati[1], Fathulloh[2], Khurotul Aeni[3]
[1,2,3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Peradaban.
nureveri11@gmail.com[1], fathul.peradaban@gmail.com[2], khaeni988@gmail.com[3]
Jl. Raya Pagoejengan Km 3 Paguyangan, Brebes, Jawa Tengah, 52276

Abstract

Breast cancer is one of the leading causes of death in women worldwide. Early detection of breast cancer is essential to increase the chances of recovery and reduce the risk of complications. In this study, a website-based breast cancer prediction system was developed using the Support Vector Machine (SVM) algorithm as a classification method. The data used came from the kaggle site dataset containing statistical information on cell characteristics.

The data processing process included preprocessing stages, including data transformation, feature selection, k-fold cross-validation, and model training using the SVM algorithm. This study used the obtained dataset, resulting in the best accuracy for the best model using k-fold cross-validation with K=10, at 0.9821%. For the aggregation using k-fold cross-validation, the accuracy was 0.9421%, the precision was 0.9357%, the recall was 0.9102%, and the f1-score was 0.921%. The test results showed that the SVM algorithm was able to classify breast cancer types (benign or malignant) with a high level of accuracy and provided fast and accurate predictions. The resulting model is then integrated into a website system using the Flask framework, so that it can be used by users interactively. This website is expected to help the general public in carrying out early detection of breast cancer practically.

Keyword:

Breast Cancer, Support Vector Machine, Python, Machine Learning

Abstrak

Kanker payudara merupakan salah satu penyebab utama kematian pada wanita di seluruh dunia. Deteksi dini terhadap kanker payudara sangat penting untuk meningkatkan peluang kesembuhan dan menekan risiko komplikasi. Dalam penelitian ini, dikembangkan sebuah sistem prediksi kanker payudara berbasis website dengan menggunakan algoritma *Support Vector Machine (SVM)* sebagai metode klasifikasi. Data yang digunakan berasal dari *dataset* situs *kaggle* yang memuat informasi statistik karakteristik sel.

Proses pengolahan data meliputi tahap *preprocessing* yaitu transformasi data, seleksi fitur, *k-fold cross validation*, serta pelatihan model menggunakan algoritma *SVM*. Dalam penelitian ini menggunakan *dataset* yang diperoleh menghasilkan akurasi terbaik pada *best model k-fold cross validation* dengan *K*=10 yaitu 0.9821%. dan untuk agregasi *k-fold cross validation* yaitu akurasi 0.9421%, presisi 0,9357%, *recall* 0,9102 % dan *f1-score* 0,921%. Hasil pengujian menunjukkan bahwa algoritma *SVM* mampu mengklasifikasikan jenis kanker payudara (jinak atau ganas) dengan tingkat akurasi yang tinggi, serta memberikan hasil prediksi secara cepat dan akurat. Model yang dihasilkan kemudian diintegrasikan ke dalam sistem *website* menggunakan *framework Flask*, sehingga dapat digunakan oleh pengguna secara interaktif, *website* ini diharapkan dapat

Kata Kunci:

Kanker Payudara, *Support Vector Machine, Python, Machine Learning*

membantu masyarakat umum dalam melakukan deteksi dini kanker payudara secara praktis.

I. Pendahuluan

Kanker payudara, atau yang juga disebut *carcinoma mammae*, masuk dalam kategori kanker yang paling sering terjadi pada kalangan wanita di tingkat dunia. Penyakit ini muncul karena sel-sel di jaringan payudara mengalami pertumbuhan yang tidak normal dan memiliki kemungkinan untuk menyebar ke organ tubuh lainnya. Berdasarkan keterangan dari *World Health Organization (WHO)*, Kanker payudara menjadi penyumbang signifikan angka kematian akibat kanker pada kaum wanita, dengan menyumbang lebih dari 1 dari 10 kasus baru setiap tahunnya. Secara global, penyakit ini menempati urutan kedua dalam daftar penyebab kematian karena kanker pada wanita. Penyakit ini berkembang secara bertahap tanpa menunjukkan gejala yang jelas, sehingga mayoritas kasus baru terdiagnosis saat dilakukan pemeriksaan rutin[1]. Risiko terjadinya kanker payudara dapat muncul akibat kombinasi faktor keturunan, hormon, pola hidup, dan paparan lingkungan. Pencegahan dan deteksi dini merupakan langkah penting untuk mengurangi tingkat kematian akibat penyakit ini. Di Indonesia, kanker payudara menempati jajaran atas sebagai jenis kanker dengan jenis kanker dengan prevalensi tertinggi dibandingkan kanker lain. Berdasarkan data *Global Cancer Observatory* (*GLOBOCAN*) tahun 2022, terdapat lebih dari 66.271 kasus baru kanker payudara di Indonesia dengan angka kematian mencapai 22.598 kasus[2]. Situasi ini memperlihatkan bahwa kanker payudara menjadi masalah kesehatan yang utama di Indonesia. Keterlambatan dalam mendiagnosis kanker payudara dan kurangnya pemahaman masyarakat tentang pentingnya deteksi dini maupun pengobatan yang benar menjadi faktor penyumbang tingginya angka kematian.

Data mining merupakan teknik pengolahan data yang bertujuan menemukan pola serta mengekstraksi informasi penting dari sekumpulan data berukuran besar. Metode dalam data mining yang umum digunakan adalah klasifikasi, yaitu teknik yang digunakan dalam mengelompokkan data sesuai dengan karakteristik tertentu. *Data mining* merujuk pada kegiatan menggali informasi serta pola penting dari kumpulan data dalam jumlah besa. Proses *data mining* meliputi pengumpulan data, ekstraksi data, analisa data, dan statistik data[3].

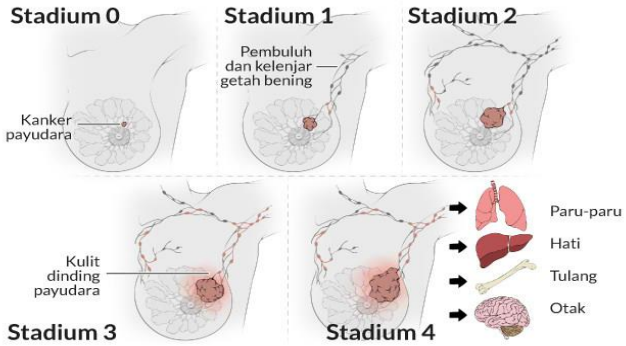
Support Vector Machine (SVM) algoritma *machine learning* efektif untuk regresi data dan klasifikasi data, dengan menyelesaikan permasalahan melalui persamaan *Lagrangian* menggunakan *quadratic programming*[4]. *SVM* bekerja dengan mencari *hyperplane* optimal yang dapat memisahkan dua kelas data dengan margin terluas. Pada bidang medis, *SVM* banyak dimanfaatkan lantaran kehandalannya dalam menangani data yang berdimensi tinggi serta ketahanannya terhadap masalah *overfitting*. Dengan bantuan kernel trick, *SVM* juga efektif dalam mengatasi klasifikasi *non-linear* dan efisien pada dataset dengan jumlah fitur lebih banyak dibandingkan sampel. *SVM* telah terbukti sebagai teknik pengenalan pola yang kuat dengan kinerja baik dalam berbagai tugas klasifikasi biner dan regresi[5]. Dengan penelitian ini, diharapkan kedepannya dapat dikembangkan sistem berbasis *website* yang menggunakan algoritma *SVM* untuk membantu dalam deteksi kanker payudara.

II. Landasan Teori

Kanker payudara merupakan pertumbuhan sel abnormal yang bersifat ganas dan berasal dari jaringan payudara, baik pada epitel duktus maupun lobulus. Kondisi ini muncul ketika sel kehilangan mekanisme kontrol alaminya, sehingga mengalami proliferasi tanpa kendali. Penyakit ini menjadi salah satu kanker yang lebih sering terdeteksi pada wanita, dengan prevalensi 1 dari 10 kasus baru setiap tahun. Kanker payudara adalah penyakit yang dominan lebih banyak menyerang perempuan, meski terkadang pria bisa memiliki

probabilitas terkena penyakit ini dengan prevalensi 1 di antara 1000[1]. Kanker payudara adalah jenis kanker yang memiliki kasus terbanyak di Indonesia terdapat 66.271 kasus kanker payudara baru dengan angka kematian yang mencapai 22.598 kasus pada tahun 2022[2], oleh sebab itu penyakit kanker payudara sudah sangat mengkhawatirkan, perlu edukasi lebih lanjut agar semua orang mengetahui tata cara penanganan dini jika terdiagnosis mengidap kanker payudara terutama pada kaum wanita. Untuk menurunkan resiko terkena kanker payudara, diharapkan wanita menjaga gaya hidup yang sehat dan rutin *check up* payudara di rumah sakit atau fasilitas medis terdekat untuk mendeteksi dini melakukan penangan pertama jika diketahui terdapat gejala-gejala awal terkena kanker payudara.

Gejala yang muncul meliputi benjolan atau pengerasan pada payudara atau ketiak, pembengkakan dan nyeri pada payudara, perubahan kulit seperti munculnya cekungan atau tekstur kulit jeruk (*peau d'orange*), puting tertarik ke dalam, serta cairan yang tidak biasa atau bisa jadi mengeluarkan darah bersumber dari puting. Jika tidak ditangani dapat menyebar ke organ lain (*metastasis*), yang dapat menurunkan kualitas hidup dan meningkatkan risiko kematian[6]. Faktor-faktor risiko yang berperan dalam peningkatan kasus kanker payudara antara lain jenis kelamin perempuan, umur di atas 50 tahun, riwayat anggota keluarga dengan kelainan genetik seperti perubahan genetik pada gen *TP53 (p53), BRCA2, ATM, BRCA1,* atau serta adanya penyakit payudara sebelumnya seperti *DCIS* di payudara yang sama sebelumnya, *LCIS,* atau tingginya intensitas terhadap mamografi. Risiko juga meningkat pada individu dengan riwayat menstruasi awal (<12 tahun) atau bisa jadi saat *menopause* lambat diatas 55 tahun, kondisi dengan reproduksi tertentu (tidak pernah melahirkan atau menyusui), penggunaan hormon, paparan radiasi pada dinding dada, konsumsi alkohol, obesitas, dan faktor lingkungan lainnya.



Gambar 1. Kanker Payudara

Data mining adalah proses mengekstraksi informasi dari suatu kumpulan data dengan memanfaatkan algoritma atau metode, teknik statistika, metode pembelajaran mesin, serta sistem manajemen basis data [7]. Data mining merupakan proses analisis data yang bertujuan untuk memperoleh informasi penting yang dapat dimanfaatkan sesuai kebutuhan tertentu. Adapun fungsionalitas *data mining* ada klasifikasi, *clustering,* regresi, asosiasi, dan deteksi anomaly. Berikut adalah *step-by-step* proses *data mining* pembersihan data, integrasi data, seleksi data, transformasi data, proses *mining,* evaluasi pola, penyajian pengetahuan dan perangkuman.

Seleksi fitur merupakan proses pemilihan *subset* fitur yang relevan dan informatif dari suatu *dataset*, dengan tujuan meningkatkan performa model *machine learning*. Salah satu metode seleksi fitur yang populer adalah *Mutual Information* (*MI*), *Mutual Information* (*MI*) mengukur seberapa besar informasi yang diketahui dari satu variabel jika kita tahu nilai variabel lainnya. seleksi fitur berfungsi untuk mengukur hubungan *non-linier* antara fitur *(X)*

dan label/target *(y)*, Semakin besar nilai *MI* suatu fitur terhadap target, semakin informatif fitur tersebut.

Seleksi fitur dihitung dengan rumus[8]:

$$I(U,C) = \sum_{et\in(1,0)}\sum_{ec\in(1,0)} P(U=et, C=ec) log_2 \frac{P(U=et,C=ec)}{P(U=et)\,P(C=ec)} \quad \text{……..…..…(2.1)}$$

SVM merupakan suatu metode *machine learning* terawasi yang umum diterapkan dalam proses regresi maupun klasifikasi. Prinsip kerjanya adalah proses menyeleksi *hyperplane* terbaik yang mampu membagi suatu data ke dalam dua kelas yang berbeda pada ruang berdimensi tinggi. *Hyperplane* tersebut ditentukan dengan memaksimalkan jarak (margin) antar data dari masing-masing kelas, sehingga kemampuan model dapat mempunyai fitur generalisasi yang lebih baik pada suatu data yang belum pernah diproses sebelumnya. *SVM* diperkenalkan oleh *Vladimir N. Vapnik* dan rekan-rekannya pada tahun 1995 dalam makalah berjudul "*Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing*"[7].

Berikut adalah langkah-langkah dalam menghitung klasifikasi menggunakan *Support Vector Machine (SVM)* beserta rumus-rumus yang digunakan[3]:

1. Persiapan Data

Misalkan kita memiliki *dataset* dengan dua kelas :

$D = \{(\chi_1,\gamma_1),(\chi_2,\gamma_2),\dots(\chi_n,\gamma_n)\}$.....*(2.3)*

Dengan :..

- $\boldsymbol{\chi_i}$ *adalah vektor fitur dari sampel ke-I*
- $\boldsymbol{\gamma_i}$ *adalah label kelas dengan nilai {+1,−1}*

2. Menentukan *Hyperplane*

Hyperplane dalam ruang fitur dapat dinyatakan sebagai :

$\mathcal{W}\cdot x + b = 0$.....(2.4)

Dengan *:*

- $\boldsymbol{w}$ *adalah vektor bobot*
- $\boldsymbol{x}$ *adalah vektor fitur input*
- $\boldsymbol{b}$ *adalah bias*

3. Optimasi Margin Maksimum

SVM adalah mencari margin maksimum, Merupakan jarak antara *hyperplane* dengan titik-titik yang paling dekat dari kedua kelompok data. Jarak margin dihitung dengan :

$Margin = \frac{2}{||w||}$.....(2.5)

SVM memaksimalkan margin dengan menyelesaikan masalah optimasi : $min\frac{1}{2}||w||^2$.....(2.6)

dengan kendala :

$y_i = (\mathcal{W}\cdot x + b) \geq 1, \forall i$.....(2.7)

4. Kasus Data Tidak Sepenuhnya Terpisah (*Soft Margin SVM*)

Jika data tidak dapat dipisahkan secara sempurna, digunakan variabel *slack* ξi *sehingga:*

$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$.....(2.8)

dan fungsi objektif menjadi :

$min\ \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\xi_i$.....(2.9)

dengan *C* merupakan variabel yang mengontrol trade-off antara margin yang lebih besar dan jumlah kesalahan dalam suatu klasifikasi.

5. Penggunaan Fungsi Kernel

Algoritma *SVM* didalamnya menggunakan Fungsi *kernel* adalah untuk mentransformasikan data ke ruang fitur yang memiliki dimensi tinggi :

$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$.....(2.10)

Dalam penelitian ini menggunakan fungsi *Kernel Linear* :

Linear: $K(x_i, x_j) = x_i \cdot x_j$.....(2.11)

6. Klasifikasi Data Baru

Setelah mendapatkan bobot w dan bias b, data baru x' diklasifikasikan berdasarkan:

$f(x') = sgn\ (w \cdot x' + b)$ .....(2.15)

Jika $f(x') > 0$, maka kelas +1, jika $f(x') < 0$, maka kelas -1.

K-Fold Cross Validation adalah metode validasi dengan pembagian data kedalam beberapa *fold k-subset*, dilanjutkan dengan mengulanginya sebanyak *k* kali guna proses pembelajaran dan pengujian [7]. *K-fold* ialah teknik yang diterapkan guna mengurangi bias dan variasi pada suatu proses pemodelan data. Pada setiap putaran, satu *fold* data dijadikan untuk data uji, sedangkan bagian lain berfungsi dijadikan data latih. Keuntungan teknik ini adalah setiap data akan mendapatkan kesempatan memerankan sebagai data uji setidaknya sekali, sekaligus memerankan data latih pada iterasi yang lain. Proses *K-Fold Cross Validation* dilakukan melalui beberapa tahapan utama antara lain pembagian data, proses iterasi, tahap pelatihan dan pengujian, evaluasi performa dan agregasi hasil.

Tabel 1. Ilustrasi *K-Fold Cross Validation*

	Dataset dibagi menjadi 10 bagian(*fold*) secara acak(*random*)										Akurasi
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	100%
Iterasi 1											A1
Iterasi 2											A2
Iterasi 3											A3
Iterasi 4											A4
Iterasi 5											A5
Iterasi 6											A6
Iterasi 7											A7
Iterasi 8											A8
Iterasi 9											A9
Iterasi 10											A10

Keterangan:

Data Uji Data Latih

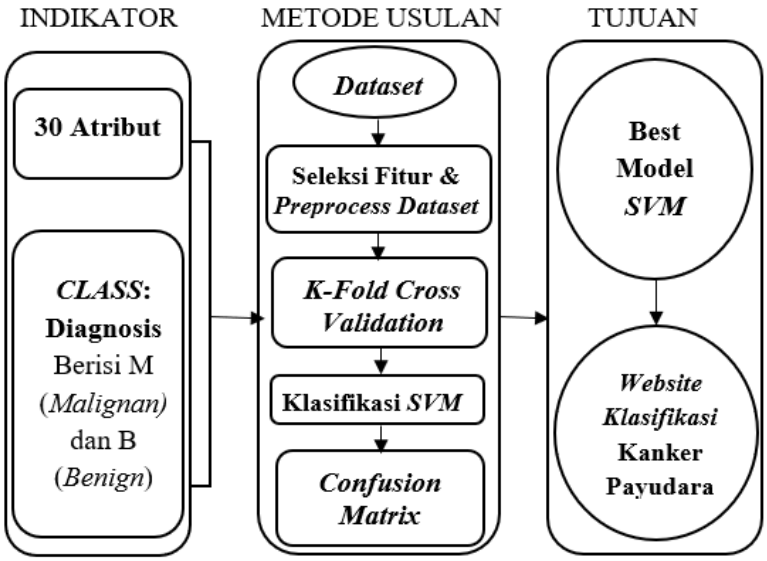
Confusion matrix merupakan alat evaluasi yang diterapkan guna menunjukkan kinerja suatu model melalui empat parameter utama, yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN)*. Ke-4 metrik ini akan menjadi dasar ketika perumusan untuk menghitung *performance metrics* seperti *metric accuracy*, *metric precision, metric recall*, *metric f1-score*[9]. *Confusion matrix* untuk menghitungan metrik evaluasi dengan menggunakan rumus sebagai berikut:

Tabel 2. *Confusion Matrix*

Classification		*Predicted*	
		Class Yes	*Class No*
Actual	*Class Yes*	*TP*	*FN*
	Class No	*FP*	*TN*

Python adalah bahasa pemrograman dengan tingkat tinggi atau yang hamper mendekati bahasa manusia dan bersifat intepretatif, berorientasi pada objek, dan *support* berbagai paradigma pemrograman, termasuk imperatif, fungsionalitas, serta prosedural. Dikembangkan oleh *Guido van Rossum* pada tahun 1980-an dan dipublish resmi di tahun 1991 [10]. *Python* adalah bahasa pemrograman skrip yang mudah dipelajari, tetapi tetap kuat dan fleksibel, sehingga banyak digunakan dalam pengembangan aplikasi. Bahasa ini mendukung paradigma pemrograman berorientasi objek, imperatif, dan fungsional, serta menyediakan berbagai pustaka yang berguna untuk berbagai kebutuhan pemrograman. *Flask* salah satu *framework web di Python* yang merupakan *microframework*. Berfungsi sebagai kerangka kerja yang sudah dikembangkan untuk memudahkan pembuatan situs *web*. Dengan *Flask*, pengembang dapat membangun situs web yang ringan, terstruktur, serta mudah untuk dikelola dan diperbarui[11]. Flask digunakan dalam penelitian ini karena memiliki keunggulan sebagai *microframework* yang ringan, cepat, dan praktis diimplementasikan.

Kerangka pemikiran penelitian ini digambarkan dibawah ini:

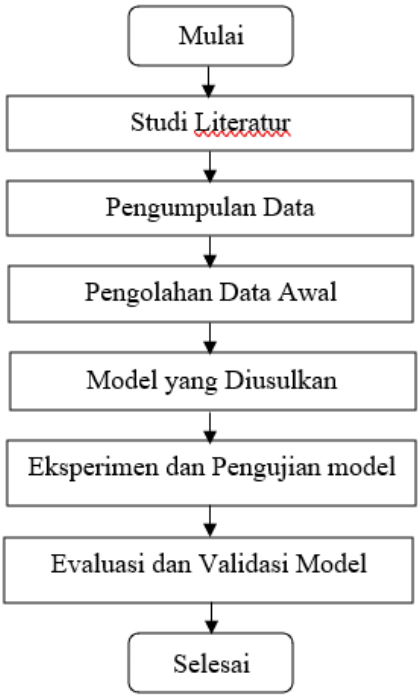


Gambar 2. Kerangka Berfikir

Indikator penelitian ini berisi parameter yang terdiri dari 30 atribut yang mencerminkan karakteristik sel dari sampel jaringan payudara yaitu *Concavity mean, area worst, concave points worst, concavity se, perimeter worst, symmetry worst, fractal dimensions se, concave points mean, smoothness se, compactness mean, radius worst, radius mean, area mean, compactness worst, texture mean, symmetry mean, perimeter mean, texture se, parimeter se, smoothnes mean, smoothness worst, concave points se, compactness se, radius se, concavity worst, smootness worst, fractal dimension mean.* Dengan kelas diagnosis yang terdiri dari dua kategori, yaitu *M* (*Malignan*) yang menunjukkan sel kanker ganas dan *B* (*Benign*) yang menunjukkan sel kanker jinak. Indikator ini menjadi dasar dalam membangun model klasifikasi untuk mendeteksi kanker payudara berdasarkan karakteristik sel yang tersedia. Metode Usulan dalam penelitian ini melibatkan beberapa langkah utama untuk membangun model klasifikasi yang optimal. Proses dimulai dengan penggunaan *dataset* yang kemudian melalui tahap seleksi fitur dan *preprocessing* untuk memastikan data dalam kondisi yang sesuai untuk pelatihan model. Selanjutnya, validasi dengan menerapkan metode *K-Fold Cross Validation* guna menilai kehandalan model melalui pembagian *dataset* menjelma menjadi beberapa *subset* guna proses pelatihan dan prose pengujian. Model kemudian dilatih menggunakan algoritma *Support Vector Machine (SVM)* dalam klasifikasi kanker payudara. Setelah proses pelatihan selesai, model dievaluasi menggunakan *Confusion Matrix* untuk mengukur metrik evaluasi seperti metrik akurasi, metrik presisi, metrik *recall*, dan metrik *F1-Score*. Tujuan penelitian ini mendapatkan model klasifikasi terbaik dengan menggunakan algoritma *SVM* untuk mendeteksi kanker payudara dengan akurasi maksimal. Selain itu, penelitian ini juga bertujuan mengembangkan *website* klasifikasi kanker payudara yang memungkinkan pengguna untuk memasukkan data karakteristik sel dan mendapatkan hasil prediksi secara otomatis berdasarkan model yang telah dikembangkan.

III. Metodologi Penelitian

Penelitian ini mencakup beberapa tahapan, dapat dilihat sebagai contoh pada Gambar 3.1.

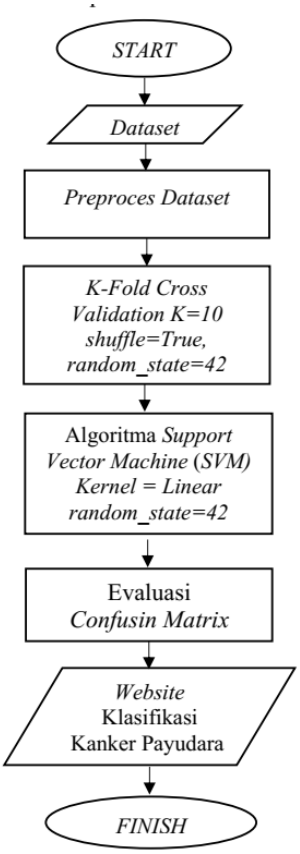


Gambar 3. Tahapan Penelitian

Tahap pertama adalah melakukan kajian literatur terhadap penelitian terkait serta teori-teori relevan dengan topik yang diteliti. Setelah itu, peneliti melakukan tahapan pengumpulan data peneliti melakukan proses pengumpulkan data dengan melakukan riset di berbagai mencari *dataset* kanker payudara diberbagai situs penyedia *dataset*. Tahapan ketiga peneliti melakukan pengolahan data awal terhadap data yang telah terkumpul. Seperti seleksi fitur dan *preprocessing* data awal. Pada tahapan memperoleh data yang akan dipakai di dalam penelitian ini dengan melakukan riset dari berbagai media di internet penyedia *dataset* guna mencari *dataset* kanker payudara. Peneliti dalam riset ini mencari *dataset* kanker payudara dari situs *kaggle* yang memiliki kualitas data yang baik, *open-source*, dan sudah digunakan penelitian-penelitian sebelumnya. *Dataset* yang diharapkan memiliki setiap atribut mewakili karakteristik statistik dari inti sel, termasuk nilai rata-rata, standar deviasi, dan nilai ekstrem dari fitur-fitur seperti radius, tekstur, perimeter, area, kehalusan (*smoothness*), *kompakness*, simetri, dan lain sebagainya. *Dataset* yang dipilih harus memiliki jumlah sampel dan fitur yang cukup untuk melatih model prediksi, sudah melalui proses validasi dan digunakan dalam berbagai studi sebelumnya, Data tersedia dalam format *CSV* yang mudah diolah menggunakan *Python* dan pustaka analisis data seperti *Pandas* dan *Scikit-learn*.

Setelah *dataset* kanker payudara didapatkan, dataset yang diperoleh harus melalui tahap pengolahan awal (*preprocessing*) agar memenuhi syarat untuk dianalisis secara optimal. Tahapan pengolahan data awal meliputi transformasi data dan seleksi fitur. Transformasi data merupakan mengubah format atau nilai pada *dataset* yang didapat guna menyesuaikan format yang didukung oleh pembelajaran mesin algoritma *SVM*. Seleksi fitur berbasis nilai probabilitas terhadap kelas target untuk memilih fitur-fitur dengan relevansi tinggi. Dalam penelitian ini menggunakan metode seleksi fitur *Mutual Information (MI)*, bertujuan untuk mengurangi kompleksitas model dan meningkatkan akurasi prediksi.

Metode yang diusulkan dalam penelitian ini adalah sebagai berikut:



Gambar 4. Metode Yang Diusulkan

Dataset selanjutnya melalui tahap proses seleksi fitur dan *preprocessing dataset,* setelah itu proses validasi *K-Fold Cross Validation K* = 10, *shuffle=True, random_state=42*, setelah menentukan nilai *k* dan parameter lainnya, selanjutnya data latih setiap iterasi di proses menggunakan algoritma *SVM* dengan *Kernel = linear* dan *random_state=42* untuk membuat sebuah model. Setelah terbentuk model dari setiap iterasi selanjutnya validasi menggunakan data uji setiap iterasi untuk mendapatkan akurasi dari setiap model. Setelah mendapatkan akurasi dari setiap model iterasi selanjutnya memilih *best model* dengan akurasi tertinggi dan menghitung rata-rata akurasi dari seluruh model. Setelah mendapatkan *best model* selanjutnya menghitung performa dari *best model* tersebut dengan *confusion matrix* untuk mengetahui metrik akurasi, metrik *recall*, metrik presisi dan metrik *f1-score*. Setelah didapatkan *best model* selanjunya pengembangan *website* klasifikasi kanker payudara menggunakan *framework flask* dari *python* berdasarkan *best model* yang ada.

IV. Hasil dan Pembahasan

Pengumpulan data dalam penelitian ini yaitu dengan melakukan riset dari berbagai media di internet penyedia *dataset* guna mencari *dataset* kanker payudara. Peneliti dalam riset tersebut mendapatkan *dataset* kanker payudara dari situs *kaggle,* (*https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data*), berisi 30 atribut yang mencerminkan karakteristik sel dari sampel jaringan payudara, *Dataset* berisi 569 data pasien penderita kanker ganas maupun pasien dengan kanker jinak. Berikut adalah tabel informasi penjelasan atribut dan kelas dalam *dataset* kanker payudara yang telah didapatkan.

Tabel 3. Informasi Penjelasan Atribut dan Kelas

No	Atribut	Penjelasan
1	*radius_mean*	Rata-rata panjang jari-jari sel tumor.
2	*texture_mean*	Rata-rata variasi intensitas tekstur sel tumor.
3	*perimeter_mean*	Rata-rata keliling sel tumor.
4	*area_mean*	Rata-rata luas sel tumor.
5	*smoothness_mean*	Rata-rata variasi lokal dalam panjang radius.
6	*compactness_mean*	Rata-rata kepadatan sel tumor
7	*concavity_mean*	Rata-rata tingkat cekungan kontur sel tumor.
8	*concave points_mean*	Rata-rata jumlah titik cekungan pada kontur sel tumor.
9	*symmetry_mean*	Rata-rata tingkat simetri sel tumor.
10	*fractal_dimension_mean*	Rata-rata kompleksitas batas sel tumor (dimensi fraktal).
11	*radius_se*	*Standard error* panjang radius sel tumor.
12	*texture_se*	*Standard error* variasi intensitas tekstur sel tumor.
13	*perimeter_se*	*Standard error* keliling sel tumor.
14	*area_se*	*Standard error* luas sel tumor.
15	*smoothness_se*	*Standard error* variasi lokal dalam panjang radius.
16	*compactness_se*	*Standard error* kepadatan sel tumor.
17	*concavity_se*	*Standard error* tingkat cekungan kontur sel tumor.

No	Atribut	Penjelasan
18	*concave points_se*	*Standard error* jumlah titik cekungan pada kontur sel tumor.
19	*symmetry_se*	*Standard error* tingkat simetri sel tumor.
20	*fractal_dimension_se*	*Standard error* kompleksitas batas sel tumor.
21	*radius_worst*	Nilai terburuk dari panjang radius sel tumor.
22	*texture_worst*	Nilai terburuk dari variasi intensitas tekstur sel tumor.
23	*perimeter_worst*	Nilai terburuk dari keliling sel tumor.
24	*area_worst*	Nilai terburuk dari luas sel tumor.
25	*smoothness_worst*	Nilai terburuk dari variasi lokal dalam panjang radius.
26	*compactness_worst*	Nilai terburuk dari kepadatan sel tumor.
27	*concavity_worst*	Nilai terburuk dari tingkat cekungan kontur sel tumor.
28	*concave points_worst*	Nilai terburuk dari jumlah titik cekungan pada kontur sel tumor.
29	*symmetry_worst*	Nilai terburuk dari tingkat simetri sel tumor.
30	*fractal_dimension_worst*	Nilai terburuk dari kompleksitas batas sel tumor.
31	*diagnosis*	Kelas M (Malignant/Ganas) atau B (Benign/Jinak).

Tabel 4. *Dataset* Kanker Payudara

No	1.	2.	3.	.	.	.	567.	568.	569.
radius_mean	17.99	20.57	19.69	.	.	.	16.6	20.6	7.76
texture_mean	10.38	17.77	132.9	.	.	.	28.08	29.33	24.54
perimeter_mean	122.8	132.9	130	.	.	.	108.3	140.1	47.92
area_mean	1001	1326	1203	.	.	.	858.1	1265	181
smoothness_mean	0.1184	0.08474	0.1096	.	.	.	0.08455	0.1178	0.05263
compactness_mean	0.2776	0.07864	0.1599	.	.	.	0.1023	0.277	0.04362
concavity_mean	0.3001	0.0869	0.1974	.	.	.	0.09251	0.3514	0
concave points_mean	0.1471	0.07017	0.1279	.	.	.	0.05302	0.152	0
symmetry_mean	0.2419	0.1812	0.2069	.	.	.	0.159	0.2397	0.1587
fractal_dimension_mean	0.07871	0.05667	0.05999	.	.	.	0.05648	0.07016	0.05884
radius_se	1.095	0.5435	0.7456	.	.	.	0.4564	0.726	0.3857
texture_se	0.9053	0.7339	0.7869	.	.	.	1.075	1.595	1.428

No	1.	2.	3.	.	.	.	567.	568.	569.
perimeter_se	8.589	3.398	4.585	.	.	.	3.425	5.772	2.548
area_se	153.4	74.08	94.03	.	.	.	48.55	86.22	19.15
smoothness_se	0.006399	0.005225	0.00615	.	.	.	0.005903	0.006522	0.007189
compactness_se	0.04904	0.01308	0.04006	.	.	.	0.03731	0.06158	0.00466
concavity_se	0.05373	0.0186	0.03832	.	.	.	0.0473	0.07117	0
concave points_se	0.01587	0.0134	0.02058	.	.	.	0.01557	0.01664	0
symmetry_se	0.03003	0.01389	0.0225	.	.	.	0.01318	0.02324	0.02676
fractal_dimension_se	0.006193	0.003532	0.004571	.	.	.	0.003892	0.006185	0.002783
radius_worst	25.38	24.99	23.57	.	.	.	18.98	25.74	9.456
texture_worst	17.33	23.41	25.53	.	.	.	34.12	39.42	30.37
perimeter_worst	184.6	158.8	152.5	.	.	.	126.7	184.6	59.16
area_worst	2019	1956	1709	.	.	.	1124	1821	268.6
smoothness_worst	0.1622	0.1238	0.1444	.	.	.	0.1139	0.165	0.08996
compactness_worst	0.6656	0.1866	0.4245	.	.	.	0.3094	0.8681	0.06444
concavity_worst	0.7119	0.2416	0.4504	.	.	.	0.3403	0.9387	0
concave points_worst	0.2654	0.186	0.243	.	.	.	0.1418	0.265	0
symmetry_worst	0.4601	0.275	0.3613	.	.	.	0.2218	0.4087	0.2871
fractal_dimension_worst	0.1189	0.08902	0.08758	.	.	.	0.0782	0.124	0.07039
diagnosis	M	M	M	.	.	.	M	M	B

Tabel 5. Transformasi Data

Diagnosis	
Before	*After*
B (*Benign*)	0
M (Malignan)	1

Tabel *6.* Seleksi Fitur

No	**Atribut**	**Probabilitas > Kelas %**	**No**	**Atribut**	**Probabilitas > Kelas %**
1	*perimeter_worst*	0.4752	16	*concave points_se*	0.1281
2	*area_worst*	0.4643	17	*texture_worst*	0.1242
3	*radius_worst*	0.4537	18	*concavity_se*	0.1183
4	*concave points_mean*	0.4394	19	*smoothness_worst*	0.0989

No	Atribut	Probabilitas > Kelas %	No	Atribut	Probabilitas > Kelas %
5	*concave points_worst*	0.4377	20	*symmetry_worst*	0.0979
6	*perimeter_mean*	0.4029	21	*texture_mean*	0.0974
7	*concavity_mean*	0.3747	22	*smoothness_mean*	0.0809
8	*radius_mean*	0.367	23	*compactness_se*	0.0767
9	*area_mean*	0.3607	24	*fractal_dimension_worst*	0.0684
10	*area_se*	0.3407	25	*symmetry_mean*	0.0593
11	*concavity_worst*	0.3149	26	*fractal_dimension_se*	0.0387
12	*perimeter_se*	0.2761	27	*smoothness_se*	0.0162
13	*radius_se*	0.2487	28	*symmetry_se*	0.0157
14	*compactness_worst*	0.2241	29	*fractal_dimension_mean*	0.0069
15	*compactness_mean*	0.2117	30	*texture_se*	0.0

Tabel 7. Fitur Yang Terpilih

No	Atribut	Probabilitas > Kelas %	No	Atribut	Probabilitas > Kelas %
1	*perimeter_worst*	0.4752	9	*area_mean*	0.3607
2	*area_worst*	0.4643	10	*area_se*	0.3407
3	*radius_worst*	0.4537	11	*concavity_worst*	0.3149
4	*concave points_mean*	0.4394	12	*perimeter_se*	0.2761
5	*concave points_worst*	0.4377	13	*radius_se*	0.2487
6	*perimeter_mean*	0.4029	14	*compactness_worst*	0.2241
7	*concavity_mean*	0.3747	15	*compactness_mean*	0.2117
8	*radius_mean*	0.367			

Tabel 8. *Fix Dataset* Kanker Payudara

No	1.	2.	3.	.	.	.	567.	568.	569.
perimeter_worst	184.6	158.8	152.5	.	.	.	126.7	184.6	59.16
area_worst	2019	1956	1709	.	.	.	1124	1821	268.6
radius_worst	25.38	24.99	23.57	.	.	.	18.98	25.74	9.456
concave points_mean	0.1471	0.07017	0.1279	.	.	.	0.05302	0.152	0
concave points_worst	0.2654	0.186	0.243	.	.	.	0.1418	0.265	0

No	1.	2.	3.	.	.	.	567.	568.	569.
perimeter_mean	122.8	132.9	130	.	.	.	108.3	140.1	47.92
concavity_mean	0.3001	0.0869	0.1974	.	.	.	0.09251	0.3514	0
radius_mean	17.99	20.57	19.69	.	.	.	16.6	20.6	7.76
area_mean	1001	1326	1203	.	.	.	858.1	1265	181
area_se	153.4	74.08	94.03	.	.	.	48.55	86.22	19.15
concavity_worst	0.7119	0.2416	0.4504	.	.	.	0.3403	0.9387	0
perimeter_se	8.589	3.398	4.585	.	.	.	3.425	5.772	2.548
radius_se	1.095	0.5435	0.7456	.	.	.	0.4564	0.726	0.3857
compactness_worst	0.6656	0.1866	0.4245	.	.	.	0.3094	0.8681	0.06444
compactness_mean	0.2776	0.07864	0.1599	.	.	.	0.1023	0.277	0.04362
diagnosis	1	1	1	.	.	.	1	1	0

Klasifikasi Pada tahap ini, *dataset* akan diproses menggunakan bahasa pemrograman *Python* melalui aplikasi *Visual Studio Code*. Data yang digunakan berformat *CSV (Comma Separated Values)* dan akan diproses ke dalam *Python* untuk diolah lebih lanjut. *Dataset* yang didapatkan dapat disajikan pada Tabel 4.1, proses transformasi data dijelaskan pada Tabel 4.3. Hasil data setelah melalui tahap pra-pemrosesan disajikan dalam Tabel 4.6. Langkah selanjutnya adalah memisahkan *dataset* menjadi 10 *fold*, di mana pada iterasi terbagi menjadi satu bagian berfungsi sebagai *test data* dan sembilan bagian lainnya digunakan sebagai *train data*. Model yang menghasilkan akurasi tertinggi akan dipilih sebagai model terbaik (*best model*) dan digunakan untuk proses prediksi kanker payudara. Evaluasi akurasi diterapkan dengan menguji model yang telah diproses menggunakan *test data* pada masing-masing iterasi. Setelah tahap pengujian model, proses dilanjutkan dengan evaluasi mengukur performa masing-masing model menggunakan *confusion matrix*. Validasi dilakukan dengan menghitung nilai *TP, FP, FN* dan *TN* dari setiap model. Nilai-nilai tersebut digunakan untuk mengevaluasi tingkat kinerja model. Berdasarkan hasil *confusion matrix*, diperoleh nilai akurasi yang menjadi dasar dalam menentukan model terbaik yang akan dijadikan model klasifikasi, dengan rincian yang didapat antara lain :

Tabel 9. Hasil *K-Fold Cross Validation*

Fold	*TP*	*FP*	*FN*	*TN*	Akurasi	Presisi	*Recall*	*F1-Score*
1	20	0	2	35	0.9649	1.0	0.9091	0.9524
2	21	3	1	32	0.9298	0.875	0.9545	0.913
3	20	2	1	34	0.9474	0.9091	0.9524	0.9302
4	17	1	4	35	0.9123	0.9444	0.8095	0.8718
5	19	1	2	35	0.9474	0.95	0.9048	0.9268
6	18	2	3	34	0.9123	0.9	0.8571	0.878
7	20	2	1	34	0.9474	0.9091	0.9524	0.9302
8	20	3	1	33	0.9298	0.8696	0.9524	0.9091

Fold	*TP*	*FP*	*FN*	*TN*	Akurasi	Presisi	*Recall*	*F1-Score*
9	18	0	3	36	0.9474	1.0	0.8571	0.9231
10	20	0	1	35	0.9821	1.0	0.9524	0.9756
Agregasi					0.9421	0.9357	0.9102	0.921

Dari hasi *K-Fold Cross Validation* diatas maka didapatkan *best model* adalah itersi ke 10 dengan akurasi 0.9821%. selanjutnya adalah proses agregasi *K-Fold*, yaitu menghitung nilai median metrik *f1-score*, metrik akurasi, metrik *recall* dan metrik presisi dari setiap iterasi.

Evaluasi dan validasi model dengan menmanfaatkan metode *Confusion Matrix* untuk menilai seberapa baik performa suatu model menghasilkan *TP, FP, FN, TN* kemudian dihitung untuk mengetahui seberapa baik performa (metrik *f1-score*, metrik akurasi, metirk *recall* dan metrik presisi) dari *best* model *SVM* menggunakan *confusion matrix*. *Best model* dari proses *K-Fold Cross Validation* diatas adalah iterasi ke 10:

Tabel 10. *Confusion Matrix Best Model*

	Prediksi	
Aktual	*TP* 20	*FP* 0
	FN 1	*TN* 35

Tabel 11. Perhitungan Performa *Best Model*

Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	*Precision*	$\frac{(TP)}{(TP+FP)}$	*Recall*	$\frac{TP}{(TP+FN)}$	*F1-score*	$\frac{2\,(Presisi\ \times Recall)}{(Presisi+Recall)}$
	$\frac{(20+35)}{(20+35+0+1)}$		$\frac{(20)}{(20+0)}$		$\frac{20}{(20+1)}$		$\frac{2\,(1\ \times 0{,}9524)}{(1+0{,}9524)}$
	0.9821		1		0.9524		0.9756

Dari perhitungan performa di atas *best model SVM* mendapatkan *accuracy* 98,21%, *precision* 100%, *recall* 95,24% dan *f1-score* 0,9756%, ini membuktikan bahwa *best model SVM* yang dibuat menghasilkan performa yang tinggi.

V. Kesimpulan dan Saran

Dari hasil yang diperoleh saat melakukan penelitian klasifikasi kanker payudara menerapkan algoritma *Support Vector Machine (SVM)* yang diimplementasikan di media berbasis *website* menunjukkan akurasi terbaik pada model terbaik dengan *K-Fold Cross Validation K=10* sebesar 98,21%. Sedangkan hasil agregasi dari seluruh iterasi *K-Fold* menunjukkan akurasi 94,21%, presisi 93,57%, *recall* 91,02%, dan *F1-Score* 92,1%. Dari hasil ini, dapat disimpulkan algoritma *SVM* memiliki performa yang tinggi dalam mengklasifikasikan jenis kanker payudara.

Saran yang dapat dijadikan referensi guna yang ingin melakukan penelitian mendatang disampaikan beirkut:

1. Pengembangan sistem dapat diarahkan untuk mendukung multi-klasifikasi, sehingga tidak hanya membedakan antara kanker payudara jinak dan ganas, tetapi juga mampu mengklasifikasikan jenis kanker secara lebih spesifik berdasarkan subtipenya. Contohnya, subtipe ganas (*malignant*) meliputi *HER2-positive Breast Cancer, Triple-negative Breast Cancer*, *Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC),* dan *Inflammator Breast Cancer (IBC)*. Sedangkan subtipe jinak (*benign*) meliputi *Fibroadenoma, Cyst,* dan *Intraductal Papilloma*.
2. Penelitian berikutnya disarankan untuk memanfaatkan teknik klasifikasi berbasis citra, seperti algoritma *Convolutional Neural Network (CNN*) atau metode lain, guna memperoleh model yang lebih optimal dalam klasifikasi kanker payudara.

Daftar Pustaka

[1] A. RIZKA, M. K. Akbar, and N. amelia Putri, "Carcinoma Mammae Sinistra T4bN2M1 Metastasis Pleura," *Jurnal Kedokteran dan Kesehatan Malikussaleh Vol.8 No.1 Mei 2022*, vol. 8, no. 1, pp. 23–31, 2022.

[2] J. Ferlay *et al.*, "Cancer statistics for the year 2020: An overview," *Int J Cancer*, vol. 149, no. 4, pp. 778–789, 2021, doi: 10.1002/ijc.33588.

[3] R. S. Wahono, *Data Mining Data mining*, vol. 2, no. January 2013. 2023.

[4] R. Resmiati and T. Arifin, "SISTEMASI: Jurnal Sistem Informasi Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination," *SISTEMASI*, vol. 10, no. 2, pp. 381–393, 2021, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[5] G. Abdurrahman, "Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid," *JUSTIFY : Jurnal Sistem Informasi Ibrahimy*, vol. 2, no. 1, pp. 74–80, Jul. 2023, doi: 10.35316/justify.v2i1.3370.

[6] Kemenkes RI, "PANDUAN PENATALAKSANAAN KANKER PAYUDARA," *KOMITE PENANGGULANGAN KANKER NASIONAL*, vol. 1, pp. 1–46, 2022, doi: 10.1007/978-1-4419-6076-4_26.

[7] Muttaqin *et al.*, *5-FullBook Pengenalan Data Mining*. 2023.

[8] I. Putu Gede Hendra Suputra, I. Gede Sukadarmika, N. Putra Sastra, I. Gusti Bgs Darmika Putra, and I. Wayan Trisna Wahyudi, "KLASIFIKASI JUDUL BERITA BAHASA INDONESIA MENGGUNAKAN SUPPORT VECTOR MACHINE DAN SELEKSI FITUR MUTUAL INFORMATION," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 22, no. 1, 2025, [Online]. Available: www.kompas.com.

[9] H. I. Islam, M. Khandava Mulyadien, U. Enri, U. Singaperbangsa, and K. Abstract, "Penerapan Algoritma C4.5 dalam Klasifikasi Status Gizi Balita," *Jurnal Ilmiah Wahana Pendidikan*, vol. 8, no. 10, pp. 116–125, 2022.

[10] "What is Python? Executive Summary | Python.org." Accessed: Sep. 07, 2025. [Online]. Available: https://www.python.org/doc/essays/blurb/

[11] D. J. Evan and P. O. N. Saian, "Implementasi Python Framework Flask Pada Modul Transfer Out Toko Di Pt XYZ," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 4, pp. 1121–1131, Nov. 2023, doi: 10.29100/jipi.v8i4.4020.